SWARM: A Single-Molecule Workflow for High-Precision Profiling of RNA Modifications

Stefan Prodic[1,2,3], Alice Cleynen[4,+], Shafi Mahmud[2,+], Akanksha Srivastava[1,2,3,+], Agin Ravindran[1,2,3], Madhu Kanchi[2], Aditya J. Sethi[1,2,3], Miona Ćorović[5], Ronit Jain[6], Gabriela Santos-Rodriguez[7,8], Grazi Vieira[7,8], Robert J. Weatheritt[7,8], Rippei Hayashi[2], Nicole M. Martinez[6,9,10,11] , Nikolay E. Shirokikh[12], Eduardo Eyras[1,2,3,*]

[1]EMBL Australia Partner Laboratory Network at the Australian National University, Canberra, ACT 2601, Australia [2]The Shine-Dalgarno Centre for RNA Innovation, The John Curtin School of Medical Research, Australian National University, Canberra, ACT 2601, Australia [3]The Centre for Computational Biomedical Sciences, The John Curtin School of Medical Research, Australian National University, Canberra, ACT 2601, Australia [4]France-Australia Mathematical Sciences and Interactions ANU -CNRS International Research Laboratory, Canberra, ACT 2601, Australia [5]Institute of Molecular Biology (IMB), 55128 Mainz, Germany; Theodor Boveri Institute, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany. [6]Department of Chemical and Systems Biology, Stanford University, Stanford, CA 94305, USA [7]School of Biotechnology and Biomolecular Sciences, University of New South Wales, 2052 Sydney, Australia [8]EMBL Australia Partner Laboratory Network at the Garvan Institute of Medical Research, 2010 Sydney, Australia [9]Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA [10]Sarafan ChEM-H Institute, Stanford University, Stanford, CA 94305, USA [11]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA [12]Australian Centre for RNA Therapeutics in Cancer, School of Human Sciences, The University of Western Australia, Perth, WA

* Correspondence to: eduardo.eyras@anu.edu.au
+ These authors contributed equally

Abstract

Nanopore direct RNA sequencing promises to decode the epitranscriptome by detecting multiple modifications on individual RNA molecules, but its potential for biological discovery is hampered by high false-positive rates. We present SWARM, an AI-based framework designed to overcome this fundamental limitation. Its key innovation is a training strategy incorporating non-target modifications to eliminate modification crosstalk, enabling high-precision detection of m6A, pseudouridine (Ψ), and m5C at single-nucleotide and single-molecule resolution from a single sample. Using rigorous in-vitro and cellular RNA benchmarks, SWARM outperforms existing tools and shows strong agreement with orthogonal methods. Applying SWARM across mammalian tissues yielded thousands of novel modification sites with confirmed motifs and localisation patterns. Our high-resolution multi-tissue modification map revealed no evidence of widespread m6A-Ψ interplay, challenging models of a coordinated epitranscriptomic code. We further discovered splicing-shaped pseudouridylation, where TRUB1-mediated modification occurs only after exon-exon ligation stabilises the substrate structure. SWARM provides a robust, universally applicable tool for epitranscriptome discovery.

SWARM is available on GitHub: https://github.com/comprna/SWARM

Introduction

Recent years have seen a remarkable expansion in our understanding of RNA [1–3]. In addition to the four canonical ribonucleotides, which make up the main backbone of RNA, there are over 170 epitranscriptomic modifications that potentially provide an extra dimension of RNA processing and control [1–3]. Multiple research efforts have demonstrated the ability of modifications to modulate the expression and fate of diverse RNA types, including messenger RNA (mRNA), through the interplay with structure, protein-RNA interactions, stability, splicing, translation efficiency, and subcellular localisation [4,5].

The current evidence highlights to two main properties. First, RNA modifications can be dynamic, with modified sites changing across specific conditions, such as developmental stages, cell types, and disease states [1,6]. Second, the position in the mRNA appears to be functionally relevant, with modified sites being evolutionarily conserved and likely triggering positional-dependent outcomes [7,8]. However, the precise mechanisms by which RNA modifications affect mRNA processing and function are not yet fully understood [9,10], largely due to the lack of systematic knowledge and that studies are generally limited to one modification type or one mRNA molecule at a time.

It is thus essential to develop methods that accurately identify multiple modifications at single-molecule and single-nucleotide resolution that are easily applicable across multiple conditions. While several experimental methods exist, they usually detect one single modification at a time, require a non-modified control, and involve complex protocols, including RNA immunoprecipitation [11], chemical conversion [12], or enzymatic treatment [13].

Direct RNA sequencing (DRS) using molecular nanopores offers the potential to identify multiple modifications within an individual RNA molecule at a transcriptome-wide scale[14–18]. However, a fundamental unresolved issue is signal crosstalk, where the presence of one modification alters the nanopore signal of a neighbouring nucleotide, leading to pervasive false positives in multi-modification detection, and confounding biological interpretation. [14–18] is signal crosstalk from neighbouring modifications, which may alter nanopore signals and increase false-positive predictions [16].Furthermore, currently published nanopore-based methods predominantly rely on training with synthetic signals, either IVT-based with unnaturally dense modification contexts [15,16], or synthetic oligos with limited sequence contexts [19]. While there are methods that use signals from cellular mRNA [14,20,21] or rRNA [14], they are affected by a high proportion of low-stoichiometry sites that induce noise during training. Another significant limitation of current methods for nanopore modification detection is a lack of clarity on the optimal parameters to use in single-sample detection, with common use of simple stoichiometry cutoffs. While this may work well for orthogonal experimental methods with negligible false-positive rates, it results in low precision and a high level of noise in nanopore predictions, especially for rare modifications, precluding the identification of biologically sound results.

To address these challenges, we developed SWARM (Single-molecule Workflow for Analysing RNA Modifications), an AI-based computational framework for the transcriptome-wide detection of N6-methyladenosine (m6A), 5-methylcytosine (m5C), and pseudouridine (Ψ) in both nanopore chemistries (SQK-RNA002 and SQK-RNA004) at single-nucleotide resolution in individual molecules from a single sample without the need for controls. We introduce key innovations with respect to current strategies[22] : (1) Incorporation of non-target modifications in training data to minimise signal crosstalk, effectively reducing false positives; (2) diverse training data including cell-line signals from orthogonally validated sites on mRNA, rRNA, and tRNA, enabling robust calling of modified nucleotides and accurate stoichiometry estimates; (3) model selection and evaluation based on both synthetic IVT and cell line data to ensure specificity for the target modification as well as higher precision for low-frequency target modifications ; and (4) systematic benchmarking of modification probability cutoffs using synthetic and cellular RNA, ensuring high-precision and biologically meaningful single-sample analysis.

These innovations have allowed us to obtain two novel biological insights. A comprehensive analysis of multiple tissues across five mammalian species, as well as of perturbation experiments, revealed that, contrary to emerging hypotheses, there is no evidence of widespread coordination between m6A and Ψ. Moreover, through our comprehensive analyses, we discovered a novel splicing-shaped mechanism for TRUB1 candidate sites, whereby Ψ near splice sites only occurs after the right exon-exon pair is expressed, revealing a potential functional link between mRNA maturation and epitranscriptomic modification.

Results

Refined training workflow underpins accurate analysis of multiple RNA modifications

SWARM identifies N6-methyladenosine (m6A), 5-methylcytosine (m5C), and pseudouridine (Ψ), with single-base and single-molecule resolution, testing every A/C/U base at any sequence context on every read, using as input the nanopore signal aligned to a 9-mer sequence centred on the position of interest (Fig. 1a). SWARM read-level models were trained using mixtures of nanopore signals from *in vitro* transcribed (IVT) synthetic RNA templates (Supp. Table 1). To expand the modification contexts training was supplemented with nanopore signals from sites in cellular RNA validated by orthogonal experiments. These cellular sites encompassed mRNA m6A sites validated by GLORI [12], rRNA Ψ sites validated by mass spectrometry and BACS [23,24], and tRNA m5C sites validated by UBS [25] (Fig. 1a). Specifically, positive signals were taken from validated high-stoichiometry sites from direct RNA sequencing in cell types matching the orthogonal experiment samples, while negative signals were obtained from non-modified IVT cellular and synthetic RNA, with genomic coordinates and sequence contexts matching the positive data (see Methods for details).

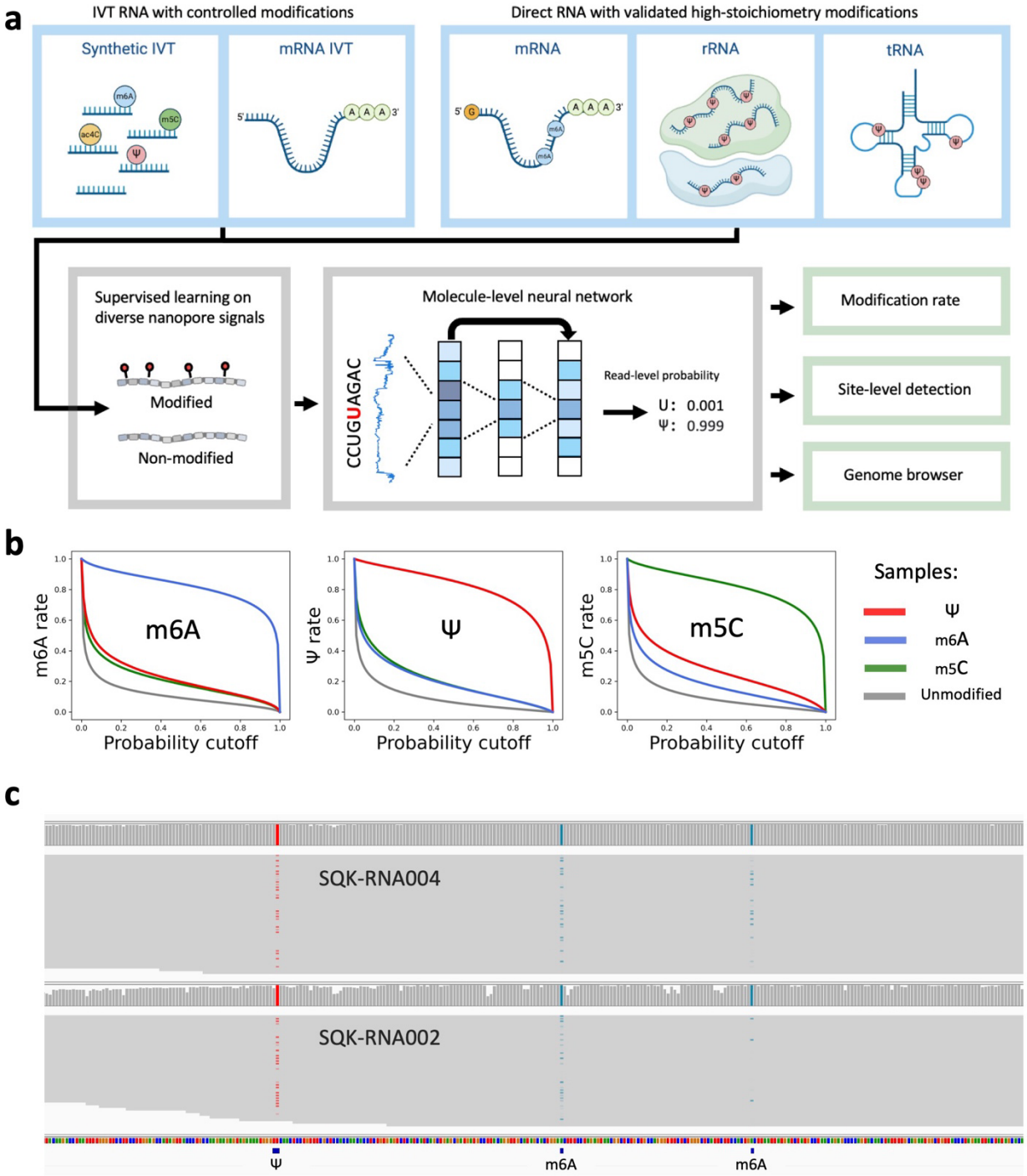

Figure 1. SWARM framework for analysing multiple RNA modifications. (a) SWARM's workflow facilitates the training of single-molecule RNA modification models with synthetic and cell-derived nanopore signals to predict modifications in individual molecules and transcript references at nucleotide resolution. SWARM outputs modSAM, which can be used for visualisation. **(b)** Proportion of correct predictions (y-axis) as a function of the read-level probability cutoff (x-axis) for the SWARM m6A, Ψ, and m5C models, when tested on independent synthetic IVT RNA samples (not used for training) that are unmodified or fully modified for m6A, Ψ, or m5C. **(c)** IGV snapshot of SWARM outputs showing per-read single-base probabilities for a Ψ site (red, position 3567), which was also detected by BID-seq in HeLa [26]. In the same transcript, we show SWARM m6A single-molecule predictions (blue) with sites detected at positions 3654 and 3711, which were also reported by GLORI in HeLa [12]. Validated sites are indicated under the sequence track and were converted from genomic to transcriptomic positions using the Ensembl hg38 GTF annotation.

Given the varying availability of data and the presence of different modifications in cellular RNAs, models for each modification type and for both sequencing chemistries, SQK-RNA002 and SQK-RNA004, were optimised on distinct combinations of IVT and cellular nanopore signals (Supp. Table 2). In addition, the optimal parameters for the neural network architecture and the proportions of each training data source for every target modification and sequencing kit were selected based on two tests: the separation of read-level modification on synthetic IVT controls and the overlap of the predicted mRNA modification sites with orthogonal datasets (Supp. Fig. 1). This training design facilitated high modification specificity, yielding models capable of identifying target modifications among millions of tested sites across transcriptomes.

Each SWARM model accurately separates signals associated with target and non-target modifications or unmodified reads (Fig. 1b). Moreover, SWARM outputs high-resolution modification readouts in the standard modSAM format, enabling visualisation of modified reference sites and individual molecules in genome browsers (Fig. 1c). As an example, we illustrate the SWARM output for the human transcript ENST00000367876.9, with a Ψ site detected by BID-seq [26] and two m6A sites detected in a different study by GLORI [12], which were all identified by SWARM. The IGV snapshot highlights molecule-level detection of Ψ and m6A, enabling high-resolution visualization of epitranscriptomic marks and suggesting relationships between modifications within individual molecules.

SWARM achieves unmatched modification specificity across synthetic controls

Most current methods fail to account for the potential influence of nearby modifications on nanopore signals. [16]This could lead to false positives and affect differential modification tests comparing conditions in which non-target modifications are changing. To address this gap, SWARM incorporates a novel training strategy that includes RNA reads with non-target modifications into the negative training set, thereby aiming to minimise false positives caused by neighbouring modification effects.

To test the specificity of read-level RNA modification prediction models, we calculated the area under the precision-recall curve (AUPRC) for SWARM and various other available tools on independent IVT controls, testing sequence contexts not seen in our training data (Table 1) (Supp. Fig. 2). We used as positives signals containing the target modification and as negatives signals having either no modifications or non-target modifications. Our tests show that the prediction accuracy of other methods decreases when applied to samples with non-target modifications (Table 1). In contrast, SWARM consistently maintains the highest separation of target and non-target modifications for m6A, Ψ, and m5C detection on both sequencing chemistries (Table 1). This high specificity is critical for minimising potential biases and false positives when analysing modification co-occurrence and testing differential modification between conditions.

Table 1. Single-read precision/recall analysis. The table displays the area under the precision-recall curves for the benchmarking of different read-level modification tools for both ONT chemistries (SQK-RNA002 and SQK-RNA004) at single-nucleotide resolution. Benchmarking datasets were built from independent synthetic IVTs and sequence contexts not used for training. Positive cases were built from reads containing the modification indicated in the column "modification". Various negative sets were considered ("negatives"), encompassing reads without modifications ("Unmodified") or reads containing non-target modifications; n.a.= not applicable.

			Negatives for precision-recall AUC			
Target Modification	ONT kit	Tool	Unmodified	Ψ	m5C	m6A
m6A	SQK-RNA002	SWARM	**0.95**	**0.9**	**0.91**	n.a.
		mAFiA	0.92	0.81	0.80	n.a.
		TandemMod	0.90	0.70	0.74	n.a.
		CHEUI	0.86	0.73	0.78	n.a.
		m6anet	0.82	0.78	0.73	n.a.
		m6ABC	0.74	0.73	0.62	n.a.
	SQK-RNA004	SWARM	**0.96**	**0.90**	**0.92**	n.a.
		dorado	0.94	0.88	0.89	n.a.
		m6Anet	0.78	0.75	0.66	n.a.
Ψ	SQK-RNA002	SWARM	**0.97**	n.a.	**0.92**	**0.92**
		TandemMod	0.80	n.a.	0.57	0.69
	SQK-RNA004	SWARM	**0.98**	n.a.	**0.92**	**0.93**
		dorado	0.93	n.a.	0.88	0.90
m5C	SQK-RNA002	SWARM	**0.96**	**0.89**	n.a.	**0.93**
		TandemMod	0.96	0.87	n.a.	0.88
		CHEUI	0.91	0.78	n.a.	0.87
	SQK-RNA004	SWARM	**0.99**	**0.94**	n.a.	**0.96**
		dorado	0.88	0.80	n.a.	0.84

Modification stoichiometry in mRNA and rRNA correlates with orthogonal datasets

To evaluate read-level detection in cellular datasets, we applied SWARM to HeLa and HEK293 sequencing datasets and compared the stoichiometry, i.e. the modification rate of the aggregated reads at each transcriptomic site, with orthogonal experiments performed in the same cell lines. For m6A, SWARM showed a high stoichiometry correlation with GLORI [12] in HeLa mRNA (R>0.9) in both ONT sequencing chemistries, improving over previously published methods (Figs. 2a & 2b) (Supp. Figs. 3a & 3b).

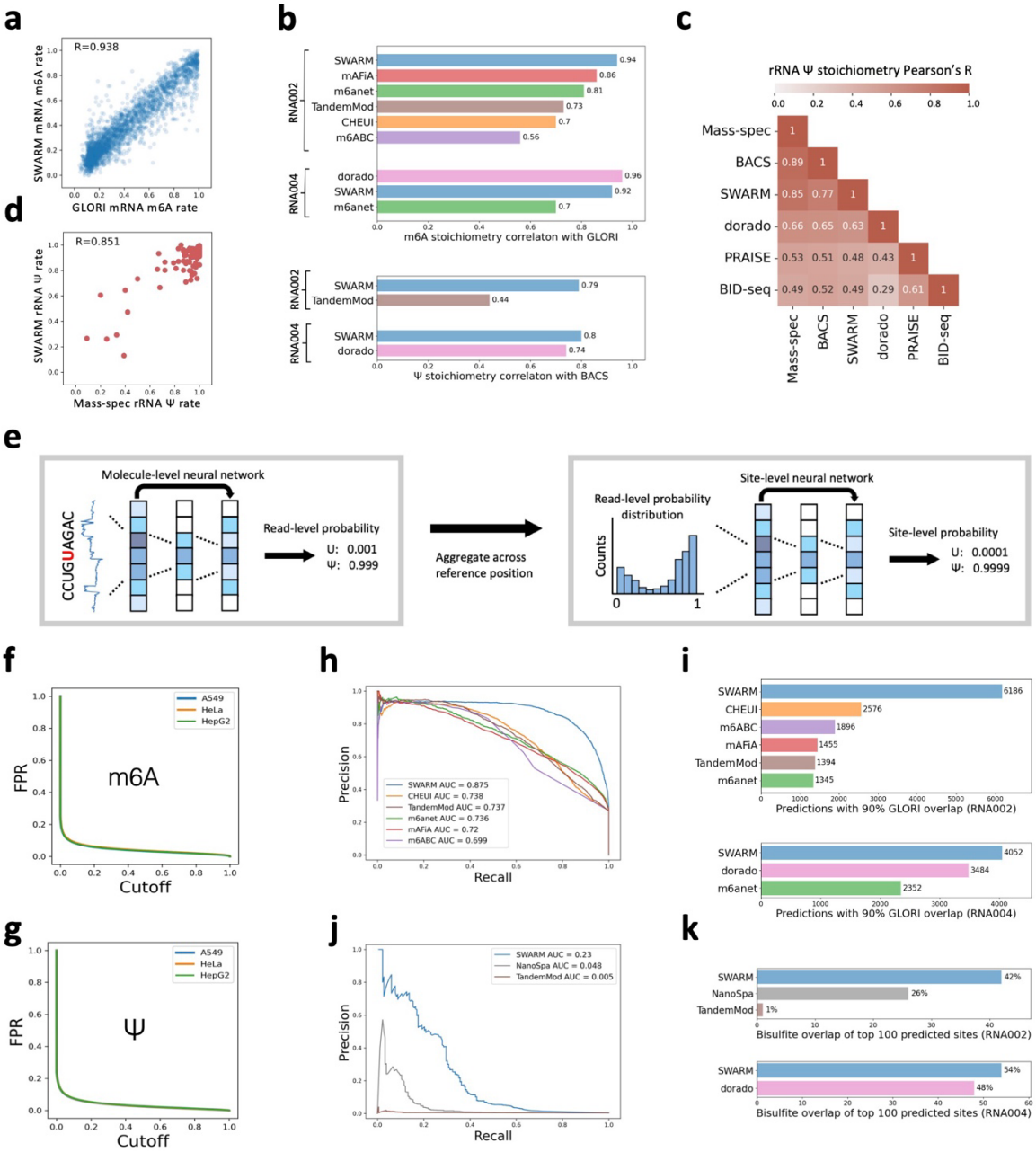

Figure 2. Stoichiometry and site-detection benchmarks on human cell lines. (a) Correlation of stoichiometry from GLORI (x-axis) and SWARM (y-axis) in HeLa SQK-RNA002 sample. **(b)** Pearson's correlation coefficient (y-axis) for m6A (upper panel) and Ψ (lower panel) stoichiometries by each method (x-axis) for SQK-RNA002 and SQK-RNA004 chemistries. The correlations were based on the comparison with GLORI in HeLa for m6A, with BACS in Hek293 for Ψ. **(c)** Matrix of Pearson's correlation coefficients for the pairwise comparison of Ψ stoichiometry reported by mass-spectrometry (MS), nanopore-based (SWARM, dorado), bisulfite-based (BID-seq, PRAISE), and BACS. **(d)** Correlation of Ψ stoichiometry in rRNA from mass spectrometry (x-axis) and SWARM (y-axis) in a HeLa SKQ-RNA004 sample. **(f-g)** SWARM's False positive rate (FPR) (y-axis) as a function of the site-level probability cutoff (x-axis) for the m6A (f) and Ψ (g) models for SQK-RNA002 *in-vitro* transcription experiments in A549, HeLa, and HepG2 cells [31]. **(h)** Precision-recall curve for SWARM and various other methods for the prediction of m6A at METTL3 candidate sites (DRACH) in SQK-RNA002, using as ground truth GLORI sites in HeLa. **(i)** Total number of transcriptomic sites (x-axis) predicted by each method (y-axis) in SQK-RNA002 (upper panel) and SQK-RNA004 (lower panel) at 90% precision, i.e. with 90% of the predictions validated by a GLORI. **(j)** Precision-recall curve for SWARM and other methods for the prediction of Ψ in PUS7 (UNΨAR, N=A,C,G,U) and TRUB1 (GUΨCNA) candidate sequence contexts for SQK-RNA002 DRS in HEK293, using as benchmark the transcriptomic Ψ sites predicted by BID-seq or PRAISE in HEK293. **(k)** From the top 100 sites predicted by each method (y-axis) in HEK293 cells using SQK-RNA002 (upper panel) or SQK-RNA002 (lower

panel), we show proportion of sites (x-axis) that overlap with the bisulfite methods BID-seq and PRAISE in HEK293. Only PUS7 (UNΨAR) and TRUB1 (GUΨCNA) sites were considered.

Regarding Ψ, our analyses in HEK293 mRNA showed that SWARM had the highest stoichiometry correlation with BACS [24] compared to other available methods for both ONT chemistries (Fig. 2b) (Supp. Figs. 3c & 3d). Furthermore, we measured the Ψ stoichiometry using SWARM and dorado in a HeLa rRNA SQK-RNA004 DRS dataset and compared the predicted values with the stoichiometry reported in HeLa using mass-spectrometry (MS) [23] and bisulfite-based methods [24,26,27] (Figs. 2c & 2d) (Supp. Fig. 4). MS, considered to be closest to the ground truth, displayed the highest correlation with BACS (R=0.89), followed closely by SWARM (R=0.85). SWARM and BACS showed a high correlation (R=0.77), while dorado and bisulfite-based methods displayed considerably lower correlation values (Fig. 2c) (Supp. Fig. 4).

In contrast to m6A and Ψ, currently available nanopore methods showed only a modest correlation of m5C stoichiometry with bisulfite-based techniques [25] in quantifying m5C in HEK293 mRNA (Supp. Figs. 3e & 3f). This discrepancy highlights a key limitation in the field, possibly due to inherent inaccuracies in existing detection methods or the challenges of reliably identifying m5C sites with low stoichiometry, a hallmark feature of this modification [28].

SWARM bolsters precise transcriptome-wide detection of low-frequency modification sites

Except for m6A, RNA modifications tend to occur at fewer than 1% of mRNA nucleotides and often display low stoichiometries [29,30]. This makes the transcriptome-wide detection of modifications particularly challenging for all machine learning methods, as it requires a false positive rate lower than the actual modification frequency to yield meaningful results. To address this challenge, we implemented a site-level neural network designed to estimate the level of confidence about the presence of a modification at a specific reference coordinate site from the distribution of probabilities across individual reads (Fig. 2e). Site-level models were trained for each read-level model, with training data composed of synthetic combinations of IVT signals in assorted proportions of modified and unmodified reads spanning a wide range of stoichiometry and coverage levels, considering sites as modified if their stoichiometry was over 10% (Methods).

To select optimal cutoffs for the *de novo* detection of modified sites, we applied each site-level model on unmodified IVT transcriptomes from HeLa, HEK293, and HepG2 [31] and estimated the false positive rate across all 9-mer sequence contexts. The false positive rate values as a function of the site-level probability cutoffs remained similar across all three cell lines (A549, HeLa, HepG2) (Figs. 2f & 2g) (Supp. Fig. 5), implying that these estimated false positive rates are generalisable to other samples.

To evaluate the precision of the nanopore-detected modified sites, we applied the site-level models on human cell lines and built precision-recall curves by testing against sites reported by orthogonal methods. Our results demonstrate that SWARM models trained on all-context sequences performed better when restricted to biologically relevant sequences, recovering more validated sites with higher precision despite testing fewer sites. Specifically, nanopore-detected m6A sites showed higher overlap with GLORI when restricted to DRACH motifs (Supp. Fig. 6a & 6b). This is relevant as some of the current m6A models for SQK-RNA002 only apply to DRACH motifs. Notably, overlap of non-DRACH sites with GLORI was low for all tools and sequencing kits (Supp. Figs. 7a & 7b). When comparing SWARM with other SQK-RNA002 models restricted to DRm6ACH sites, SWARM achieved a higher agreement in terms of AUPRC with GLORI sites in HeLa (Fig. 2h). SWARM also maintained a high overlap with GLORI sites on SQK-RNA004, with a close AUPRC to dorado at DRm6ACH sites (Supp. Fig. 7c). Importantly, at 90% precision, SWARM recalled more GLORI sites than any other available method for SQK-RNA002 or SQK-RNA004, demonstrating top performance on m6A sites across both chemistries (Fig. 2i).

Similarly to m6A, we observed that SWARM-detected Ψ sites had higher overlap with bisulfite methods [26,27] when restricted to sites with motif for TRUB1 (GUΨCNA) or PUS7 (UNΨAR, N=A,C,G,U) (Supp. Fig. 6c & 6d), the main Ψ writers modifying mRNA [32,33]. SWARM also surpassed other nanopore methods in the overlap of detected Ψ sites with orthogonal methods, having the highest AUPRC with BID-seq and PRAISE sites in HEK293 (Fig. 2j) in TRUB1 and PUS7 sequence contexts as well as in all-context detection (Supp. Figs. 7d & 7f). Furthermore, SWARM exhibited significantly higher agreement than competing approaches across both chemistries when evaluating the top 100 most confident nanopore predictions with TRUB1 and PUS7 sites detected by bisulfite methods (Fig. 2k), demonstrating best-in-class performance for precise Ψ site detection.

In contrast to m6A and Ψ, all available methods for nanopore m5C site detection showed low overlap with bisulfite sites [25] in all sequence contexts (Supp. Figs. 5 & 6). Despite an increase in the overlap when restricted to sites for NSUN6, a major m5C synthase with a well-defined motif (CTm5CCA) [34] (Supp. Figs. 5 & 6), m5C site detection with nanopore remains lagging behind m6A and Ψ detection in terms of orthogonal support [6].

Limited co-occurrence of m6A and Ψ across mammalian tissues

Leveraging SWARM's specificity and ability to detect modified sites with high precision, we used SWARM to map the m6A and Ψ landscapes across six tissues (Cerebellum, Frontal Cortex, Hippocampus, Liver, Testis, Skeletal Muscle) from five mammals (Human, Mouse, Rat, Dog, Cow) [7]. Given the demonstrated higher precision when restricting to well-established motifs, we focused the analysis on sites potentially associated with METTL3 (DRm6ACH) for m6A and with TRUB1 (GUΨCNANNC) and PUS7 (UVΨAR, V=A,C,G) for Ψ d. We detected a total of 376,674 m6A sites (Supp. Table 3) across 79,284 transcripts, and 3,682 Ψ sites (Supp. Table 4) across 3,468 transcripts (Fig. 3a). The consensus sequence of detected sites (Fig. 3b) recapitulated previously established motifs for TRUB1 [33], PUS7 [24] , and

METTL3 [12] beyond restricted nucleotides, providing confidence into low level of noise among detected sites.

Furthermore, 2,563 unique transcripts across all species harboured both Ψ and m6A sites, comprising on average 1.24% of transcripts per condition. We assessed whether the observed m6A-Ψ co-occurrence within transcripts is different from expected by chance. Considering transcripts with both METTL3 and TRUB1 or PUS7 candidate sites tested by SWARM, we compared the observed co-occurrence with the mean and standard deviation of the overlaps obtained after permuting all sites tested 1000 times. Our analysis indicates that the m6A-Ψ co-occurrence within transcripts is mildly lower than expected, with an average 16% reduction across all tissues and species. The resulting z-scores were consistently negative in all samples, with a statistically significant ($p < 0.05$) decrease in modification co-occurrence in 27 out of the 30 samples tested (Fig. 3c) (Supp. Fig. 8a).

No evidence of direct mutual exclusion between m6A and Ψ

Following up on the observed reduction of transcript-level site co-occurrence, we investigated the m6A-Ψ co-occurrence patterns in individual reads. Specifically, we selected m6A-Ψ site pairs based on their closest distance in transcripts where both modifications were detected and compared the observed co-occurrence with the mean and standard deviation obtained after performing 100 permutations of the same reads covering each pair of transcript sites. Only 16 out of the 2,848 site pairs tested showed a statistically significant difference from expected co-occurrence (adjusted z-score test $p < 0.05$) (Fig. 3d). In addition to the high proportion of non-significant sites, the distribution of z-scores computed among these sites did not diverge from normality (Supp. Fig. 8b), with a median very close to zero (-0.04), supporting the notion that m6A and Ψ occur in individual molecules independently of each other. Interestingly, a similar z-score distribution was observed regardless of the distance between m6A and Ψ sites, with molecule-level co-occurrence appearing random even when the m6A and Ψ sites were close to each other (Fig. 3d).

To further assess the potential association between m6A and Ψ, we analysed the possible differences in m6A and Ψ stoichiometries in METTL3-knockout [35] , PUS7-knockdown [36] , and TRUB1-knockdown [14] conditions and corresponding controls. Focusing only on sites previously reported by orthogonal methods [12,24], we observed a dramatic depletion of m6A with the METTL3 knockout (Figs. 3e & 3f) compared to the control, with no detectable significant Ψ differences between the same samples. In turn, PUS7 knockdown (Figs. 3g & 3h) significantly reduced Ψ stoichiometry, with no significant m6A changes across DRACH sites. Similarly, TRUB1 knockdown (Figs. 3i & 3j) induced reduction of Ψ, while only 0.11% of tested sites showed a significant m6A change, illustrating a lack of widespread TRUB1 impact on m6A.

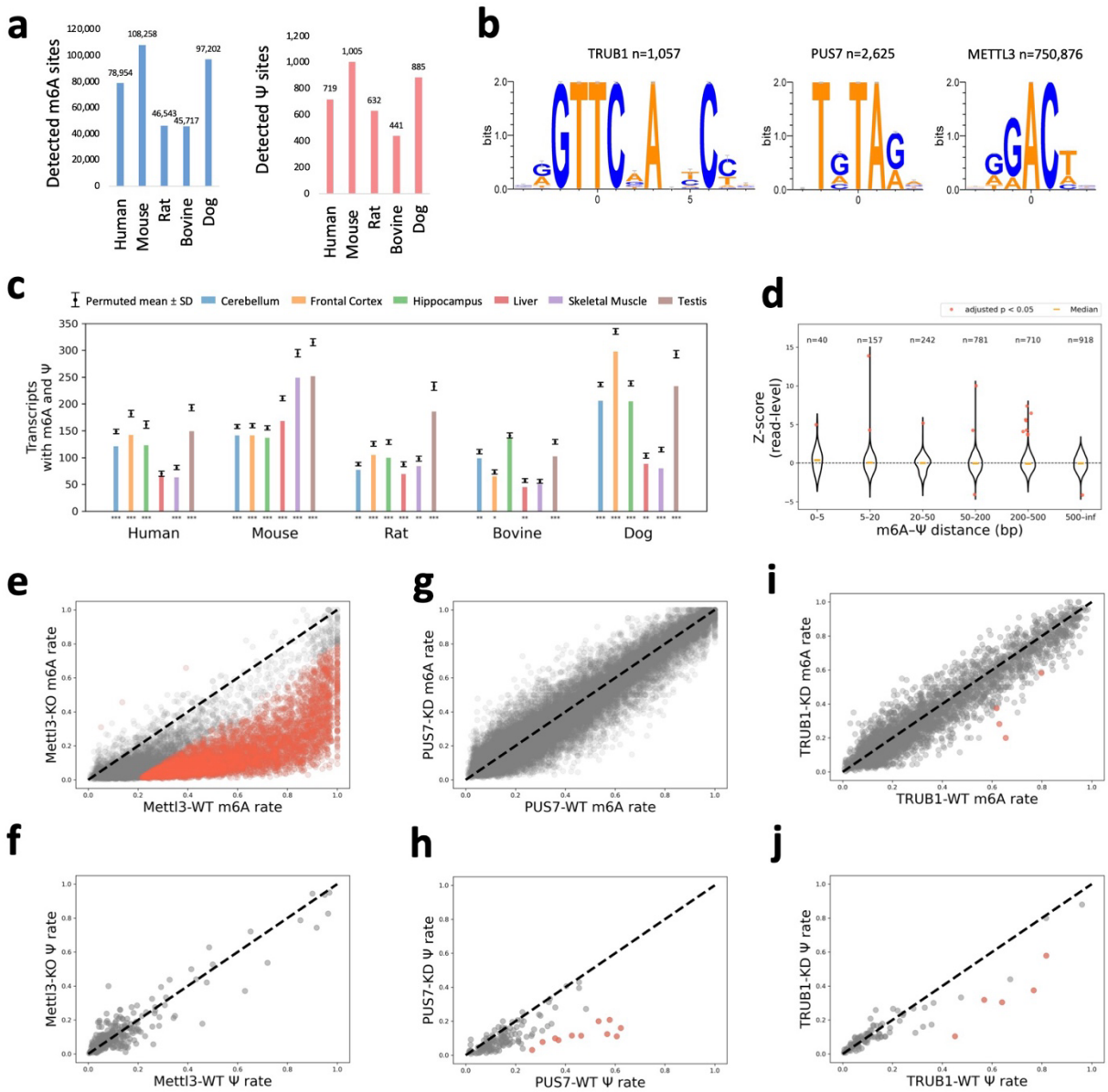

Figure 3. m6A and Ψ co-occurrence in mammalian tissues. (a) Number of m6A sites (left panel) and Ψ sites (right panel) detected in at least one tissue of each species. **(b)** Logo plots of detected sites with TRUB1, PUS7, and METTL3 substrate motifs (left to right). **(c)** Number of transcripts with at least one METTL3 candidate m6A site and one PUS7 or TRUB1 candidate Ψ site for each species and tissue (x-axis). On top of each bar, we show the expected number of cases (mean and standard deviation) from permuted modification states across all the tested METTL3 and TRUB1/PUS7 candidate sites in transcripts. Significant p-values are indicated: * = 0.05; ** = 0.005; *** = 0.001. **(d)** For pairs of m6A and Ψ sites in the same transcript at a distance range (x-axis), we plot the z-score distribution obtained by comparing the observed read-level co-occurrence, calculated as the proportion of reads where both modifications are present, with the expected co-occurrence calculated by randomising the read probabilities. **(e-j)** m6A (e,g,i) and Ψ (f,h,j) stoichiometry in KO/KD (y-axes) backgrounds of: Hek293 Mettl3 KO (e-f), HeLa PUS7 KD (g-h), and Hek293 TRUB1 KD (i-j) against the same cell-line control (x-axis). m6A sites restricted to DRACH sites reported by GLORI (), Ψ sites restricted to ones reported by BACS () with motifs for PUS7 (UNΨAR) and TRUB1 (GUΨ) (f), PUS7 (h), and TRUB1 (j). Sites significant after Benjamini-Hochberg correction (p-value < 0.05) and stoichiometry change > 0.2 are coloured red. Significance was tested using a generalised linear model comparing stoichiometry with observed coverage between the two conditions.

Splicing-shaped pseudouridylation

Our high-precision mapping revealed a distinct spatial organisation for Ψ, confirming its localisation peak near exon-exon junctions [32]. This pattern, observed for Ψ sites with both TRUB1 (Fig. 4a) and PUS7 (Supp. Fig. 9a) motifs, contrasts with the well-documented depletion of m6A marks across exon-exon boundaries [16,37–40] (Supp. Fig. 9b).

We found that Ψ localisation is largely driven by a pre-existing overrepresentation of the respective PUS7 and TRUB1 recognition motifs near splice sites (Supp. Fig. 9a). Consequently, when normalised for this underlying positional bias, the actual Ψ modification rate remains relatively constant across exon-exon boundaries (Supp. Fig. 9c). However, the splice-site proximity of the Ψ sites raises the question of whether the substrate sequence for pseudourydilation may be affected by the splicing boundaries, especially for TRUB1, which in mRNA shows affinity for an extended motif that includes a stem-loop similar to the tRNA TRUB1 substrate [33].

Across the six tissues from five mammals analysed, we detected a total of 791 unique candidate TRUB1 Ψ sites, i.e. Ψ sites predicted by SWARM on GUΨCNANNC sequence contexts. These sites showed an extended motif compared with control sites (tested GUUCNANNC sites but negative for Ψ), suggesting the expected stem-loop structure [33] (Fig. 3b)(Supp. Fig. 10). Confirming this, the 791 TRUB1-motif sites showed sequence complementarity between the expected positions, which was absent in the control unmodified sites (Fig. 4b) (Supp. Fig. 11). From these 791 sites, 42 were located close enough to the exon-exon boundary that the expected structure would inevitably cross it. Specifically, the location of the boundary relative to the Ψ site ranged from 5 nucleotides upstream (i.e., Ψ occurs downstream of the boundary) to 9 nucleotides downstream (i.e., Ψ occurs upstream of the boundary). These 42 sites showed higher average basepairing probability in the stem of the TRUB1 substrate structure when the exon-exon sequence was considered (i.e., spliced mRNA) compared to when the exon and adjacent intron sequences were considered (i.e., pre-mRNA) (Fig. 4c), suggesting that many TRUB1 Ψ sites nearby exon-exon boundaries may only occur after splicing (Fig. 4d). Furthermore, we also detected a significantly lower minimum free energy (MFE) for the TRUB1-substrate structures in the spliced transcripts compared to the pre-mRNA ones (paired t-test p-value = 0.00343) (Fig. 4e). From all the 42 cases detected, 20 showed a more stable structure in the spliced transcript, implying that these TRUB1 sites may only be modified after splicing. For the rest, 15 showed no difference in stability, and 7 showed a more stable structure in the unspliced transcript, suggesting that some TRUB1 substrates may also form before splicing.

To validate the potential link of TRUB1 pseudouridylation with splicing, we used direct RNA sequencing of pre-mRNA from HeLa cells [41]. From the six Ψ sites detected in human tissues at close distance from the splice site (out of the 42 detected across the five mammals), we found readcoverage for three of them in HeLa pre-mRNA. In all three cases, Ψ was detected by SWARM in the spliced reads, but was not observed above detectable levels in the pre-mRNA reads that included the adjacent intron (Figs. 4f

& 4g) (Supp. Fig. 12), confirming the expected modification status based on differences in TRUB1 structure stability. As a control, we measured Ψ stoichiometry in unspliced pre-mRNA reads at sites harbouring TRUB1 motifs fully contained within a single exon. The stoichiometry at those sites was generally lower in pre-mRNA reads than in reads from total RNA sequencing samples (Supp. Fig. 13), raising the question of whether modification depletion in unspliced reads near exon boundaries might be due to timing. To assess this, we analysed pre-mRNA reads from samples treated with Pladienolide B (Pla-B) to inhibit splicing, or with DMSO as controls [41]. Although Ψ levels were generally low in unspliced reads in the control, the modification levels increased in unspliced reads across most within-exon sites following Pla-B treatment (Supp. Fig. 13), supporting that the internal TRUB1 sites had sufficient time to be modified. In contrast, for the one splice-interrupted Ψ site for which we had sufficient coverage, the modification remains absent in unspliced reads even after PlaB treatment (Supp. Fig. 13). This confirms that TRUB1 sites disrupted by splice-site boundaries can be modified only after exon-exon ligation has occurred.

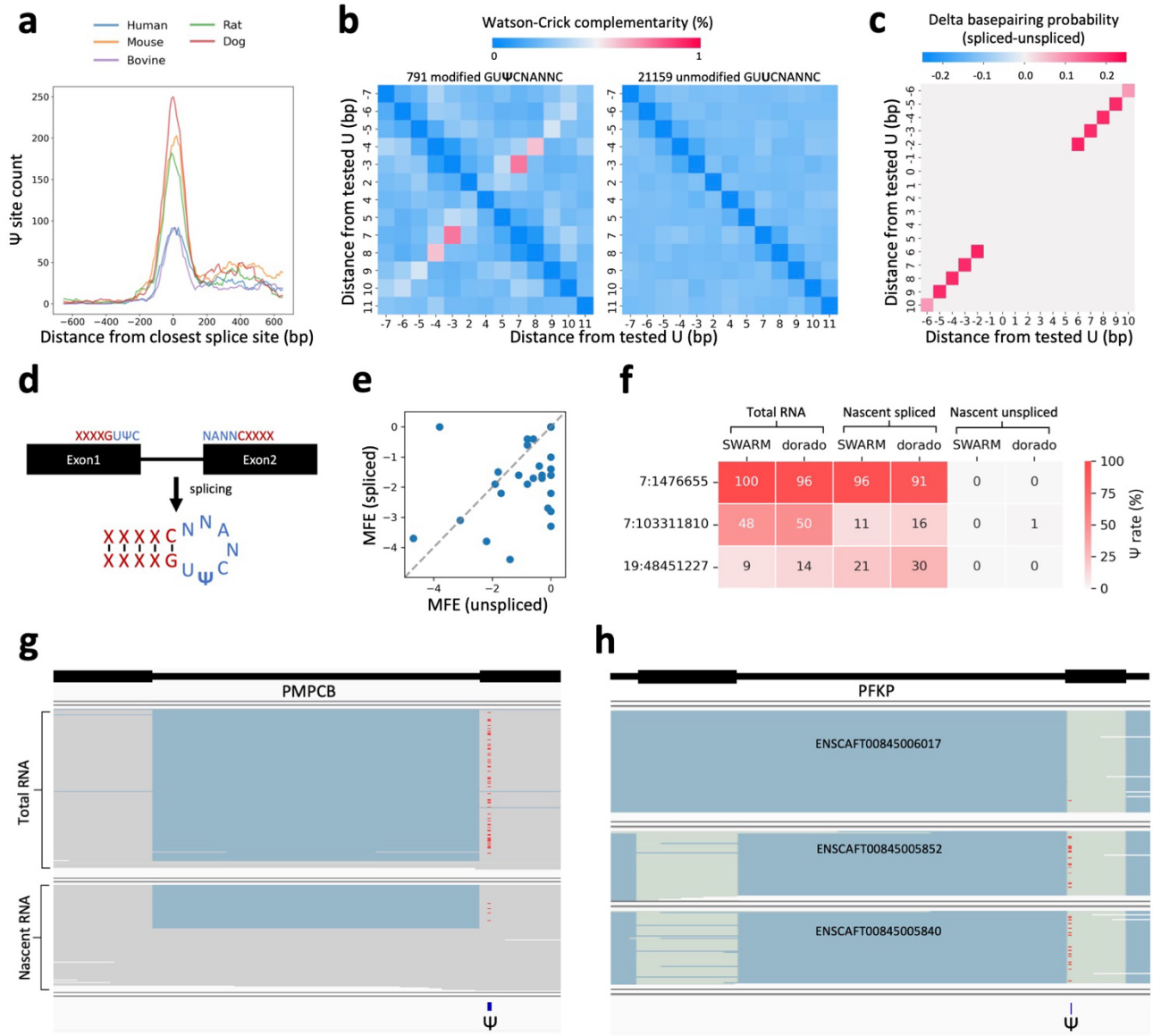

Figure 4. Splicing-shaped TRUB1 pseudouridine sites. (a) Counts of Ψ sites detected in each species in 100bp sliding windows centred on the closest exon-exon junction. **(b)** Base-complementary matrix across 791 Ψ sites (left) and 21,159 tested unmodified sites (right) with TRUB1 consensus motifs in 5 mammals and 6 tissues. Unique genomic sites detected in at least one transcript in at least one tissue were analysed in each species. Complementarity was calculated as in Safra et al. [33], using only Watson-Crick base-pairing. **(c)** Delta of base-pairing probability values for all positions around the 42 TRUB1-candidate Ψ sites identified in mammals with the substrate structure crossing the exon-exon boundaries, comparing the spliced and the pre-mRNA (i.e., including the adjacent intron) sequence. Unlike in (b), non-Watson-Crick base-pairs were also considered. **(d)** Schematic showing TRUB1 substrate structure formation after splicing. **(e)** Scatter plot showing the minimum free energies (MFE) of the candidate TRUB1 substrate structure in the spliced (y-axis) and pre-mRNA (x-axis) for the 42 mammalian sites where the TRUB1 substrate structure motif crossed the exon-exon boundary. **(f)** Ψ stoichiometries in three HeLa sites with the TRUB1 substrate structure interrupted by an exon-exon boundary, measured by SWARM and dorado in three RNA fractions: reads from total RNA sequencing, and spliced or unspliced reads from nascent RNA sequencing. **(g)** IGV snapshot of the human PMPCB gene showing reads from total RNA (upper track) and nascent RNA fractions (lower track) in HeLa. Gray blocks represent positive-strand reference matches, and blue lines represent introns. Below, the position of the Ψ site (7:103311810) is indicated. Read-level ΨSWARM predictions with probability > 0.9 are shown in red. **(h)** IGV snapshot of the dog PFKP gene in testis tissue samples with three isoforms present in each alignment track. Green blocks represent negative-strand reference matches, and blue lines represent introns. Below, the position of the Ψ site (2:32615598) is indicated. Read-level Ψ SWARM predictions with probability P>0.9 are shown in red.

To further validate the phenomenon of splicing-shaped pseudourydilation, we next considered whether it could also happen in the context of alternative splicing. Out of the 42 splicing-shaped candidate TRUB1sites detected across mammalian tissues, 20 were associated with alternative splicing events calculated with SUPPA [42], with only one site captured in multiple isoforms having different TRUB1 motif sequences in the same tissue sample. A Ψ site in the *PFKP* gene was detected in dog testis near an exon skipping event with both alternative isoforms expressed in the same tissue, with Ψ detected in reads including the alternative exon, while Ψ was depleted in reads skipping the exon (Fig. 4h). Confirming the splicing-shaped nature of this site, we confirmed a higher stability of the TRUB1 substrate structure in the reads including the alternative exon compared with the reads skipping the alternative exon (Fig. 4h, Supp. Table 5).

Discussion and Conclusions

SWARM provides a significant methodological advance in epitranscriptomics by establishing a robust, high-resolution workflow for the simultaneous detection of m6A, Ψ, and m5C from individual nanopore RNA sequencing reads. By integrating synthetic RNA training data with signals from orthogonally validated cellular sites, we developed predictive models that achieve high specificity and accuracy, effectively mitigating the confounding influence of neighbouring modifications on nanopore signals. A key innovation in SWARM is its training strategy, which includes non-target modifications to eliminate signal crosstalk, resulting in a drastic reduction in false positives, a significant limitation of existing tools. This capability is a fundamental leap beyond existing tools, as demonstrated by SWARM's superior performance in rigorous benchmarking against independent IVT controls and orthogonal datasets. Moreover, unlike most other detection tools, SWARM is compatible with both SQK-RNA002 and SQK-RNA004 chemistries, ensuring broad applicability to new and legacy datasets. We have illustrated this

capability with our application of SWARM to diverse mammalian transcriptomes, a rich nanopore-based resource generated with the SQK-RNA002 chemistry [7]. The unique SWARM's accuracy for SQK-RNA002 data for m6A and Ψ enabled us to obtain new biological insights, challenging existing paradigms about modification crosstalk, and revealing new principles of pseudouridine modification.

A key finding of our study is the absence of a direct coordination between m6A and Ψ modifications. Despite their independent enrichments in mRNA and potential roles in regulating mRNA metabolism, our analyses of the transcript and read-level co-occurrence and perturbation experiments, consistently indicated no evidence for dramatic positive or negative widespread association. However, the slight but significant underrepresentation of transcripts with both modifications suggests these marks may be subject to selective pressures that favour their mutual exclusivity, perhaps to avoid conflicting regulatory outcomes. While these findings challenge the emerging hypothesis of a widespread coordinated epitranscriptomic code, these results do not discard the possibility that in specific mRNAs, different combinations of modifications might be co-regulated or define more nuanced functional outcomes. Tools like SWARM can help dissect these complex relationships.

Furthermore, our analyses provide compelling evidence that a subset of TRUB1-mediated pseudouridylation events may be shaped by RNA splicing. We identified Ψ sites whose predicted local RNA structures become significantly more stable in the mature, spliced mRNA compared to the pre-mRNA context. This splicing dependence was validated in several cases, where Ψ modification was evident in spliced isoforms but absent or depleted in unspliced transcripts or alternatively spliced variants. However, we also observed cases where the TRUB1-specific structure could be more stable in the pre-mRNA context. This suggests a model in which splicing remodels local RNA structure near splice sites, making it a more suitable substrate for TRUB1 recognition and modification, specifically in pre-mRNA or mature mRNA, thereby functionally linking mRNA maturation and epitranscriptomic marking.

Importantly, our high-precision framework and comprehensive benchmarking exposed a current limitation in the field: nanopore sequencing and bisulfite-based techniques show poor concordance for m5C detection. This likely reflects both the low stoichiometry of m5C in mRNA and the technical limitations present across all existing assays. Rather than a failure of any single technology, this result highlights a critical lack of standardized, reliable ground truth for this modification. Thus, SWARM not only advances detection for m6A and Ψ but also establishes a new benchmark that reveals the urgent need for community-wide efforts to develop validated controls and orthogonal standards for m5C [43].

The observed discrepancy is also occurring across various orthogonal methods, not just nanopore-based methods, in both the number of modified sites and their stoichiometry. This could be related not only to the low stoichiometry of the modifications but also to the low reproducibility of the various techniques. Future improvements in nanopore-based detection could involve more sophisticated strategies for integrating IVT and cellular data to enhance accuracy. Furthermore, the development of comprehensive standardised benchmarking datasets will help improve the detection of various

modifications using nanopore sequencing and other technologies [44], especially rare ones or those that appear only under very specific conditions [6].

Recent studies have highlighted that biologically-sound results using nanopore-based predictions might be limited to m6A [22]. The nature of m6A, being modified mostly by one single complex and having overall high stoichiometry, likely makes detection and biological discovery much easier. In contrast, pseudouridine can be placed by up to 13 different synthases with diverse affinities and generally present low stoichiometry in mRNA . In our study, we demonstrate that a training strategy that improves predictions and yield biologically meaningful results for pseudouridine using nanopore.

The modular design of SWARM is a key strength, providing a framework that can be extended to incorporate additional RNA modifications as high-quality training and testing data for cellular RNA become available. While recent studies follow this principle by including more diverse sequence contexts and additional modifications [15,18], evaluation remains limited to IVT signals, without a comparison with cellular transcriptomes. As we have shown, this is often where models lose precision. Future iterations would benefit from training that include more diverse sequence contexts and additional modification backgrounds. Furthermore, the observed discrepancies among orthogonal methods highlight the need for a collective effort to improve current technologies and develop new ones[43].

In conclusion, SWARM equips the epitranscriptomics community with a powerful, versatile, and highly accurate tool for decoding the complex landscape of RNA modifications. Its ability to profile multiple modifications from the same molecule, across both current and legacy sequencing chemistries, makes it an indispensable resource for mining existing datasets and designing new experiments. As an open-source tool with standardised benchmarks, SWARM opens new avenues for investigating the dynamics, coordination, and functional interplay of the epitranscriptome in health and disease.

Methods

Synthetic RNAs and nanopore sequencing

Reference *in-vitro* transcribed RNAs (IVTs) i1-i4 and m1-m4 [16] were obtained by run-off *in-vitro* transcription from plasmids encoding bacteriophage T7 RNA polymerase promoter followed by a unique artificial sequence covering the entire length of the desired RNAs, with 2.5-5.0 μg/μl in 20 μl plasmids were ordered from GeneScript Biotech. Eight pUC57-mini-X (X = i1-i4, m1-m4) DNA templates were generated by plasmid linearization with different restriction endonucleases as used previously [16]. [16] The i1-i4 sequences contained random 5-mers, whereas m1-m4 sequences were built from four non-overlapping fragments of the mouse canonical pre-rRNA (~13kb long) (Supp. Table 1) [16]. RNA was preheated at 65°C for 2 min before being polyadenylated by E. coli poly(A) polymerase. About 2.5μg of RNA, 6μl of EPAP, 1mM ATP, 2μl of Rnasin, and 2μl of polymerase, were incubated for 30 min at 37°C.

The poly A-tailed RNA was cleaned up and purified with 1ml freshly prepared 80% ethanol and 2x ratio of AMPure XP beads to RNA. The polyadenylated RNA was further used for the library preparation and sequencing using the ONT direct RNA sequencing protocol.

The sequencing library was prepared as per the manufacturer's recommendation: 1000ng of RNA was used for 2X library preparation for IVT construct. The RNA Control Standard (RCS) was eliminated and Superscript IV and Rnasin were used for incubation and reverse transcription. The sample was purified with 1x ratio of AMPure XP beads to RNA and 80% freshly prepared ethanol. The sample went through a second incubation with NEBNext quick ligation buffer and T4 DNA ligase for 15 minutes at room temperature. The RNA sample was further cleaned up with AMPure XP beads and wash buffer (Nanopore). Then we eluted the reverse transcribed and adapted RNA by elution buffer (ELB). The final nanopore sequencing was conducted on a MinION for 48-72 hours. A quality check was performed to ensure each flowcell had more than 800 active pores. The direct RNA sequencing was conducted using both chemistries, SQK-RNA002 and SQK-RNA004. For SQK-RNA004, the full protocol was identical to the SQK-RNA002 except for minor adjustments for the reverse transcriptase reagents and incubation time: ProtoScript II was used for the reverse transcription reaction: 25°C for 5 minutes, then 42°C for 60 minutes, 65° C for 20 min, and bringing the sample temperature to 4° C.

Synthesis of in vitro transcribed cDNA libraries

15 ng of Human Brain, Hippocampus Marathon®-Ready (Catalogue number *639319*) and Mouse 7-day (D7) Embryo Marathon®-Ready (Catalogue number *639407*) were purchased from Clontech (Takara Bio, USA). All procedures relating to the reverse transcription of poly(A)+ RNA for the synthesis of Marathon-Ready cDNA for these samples can be found in Marathon®-Ready instruction manual by the manufacturer. cDNA amplification of the samples was conducted to increase the yield of the initial starting amount of cDNA from 15 ng to at least 1 mg for sufficient RNA production and hence generate very high throughput libraries with as many diverse sequence contexts for SWARM training models. The cDNA amplification procedure to amplify human hippocampus and mouse embryo D7 cDNA samples was adapted from the Marathon® end-to-end PCR procedure.

To set up the amplification, AP1 primer from the Marathon® kit was used as the negative control, as recommended by manufacturer protocol. Since all sequences in the pool of cDNA samples harboured the reverse complement sequence of AP1 adaptor nucleotides at the 5' end, we designed in-house primers AP1-F: *CTAATACGACTCACTATAGGGCTCGAGCG* (T_m = 67.5°C) as the forward primer and AP1-R: *TTCTAGAATTCAGCGGCCGCTTTTTTTT* (T_m = 67.5°C) as the reverse primer, designed around the poly(A) tail of adaptor sequence. These primers were used to amplify all transcripts in the cDNA sample. The primers were designed and ordered from Integrated DNA Technologies (IDT), and 100 mM of master stock solutions of the forward (AP1-F) and reverse primers (AP1_R) as well as the original AP1 primer were made in deionised mQ water. To make the PCR master mix for each cDNA sample, 5 mM of AP1-F and AP1-R were mixed with 1×SuperFi II buffer and Platinum SuperFi II Polymerase (Thermo Fisher

Scientific), supplemented with dNTPs added to a final concentration of 200 mM. The total PCR master mix was made in a final volume of 50 ml. A parallel negative control for each cDNA sample was set using the same conditions as described above, except replacing the primers with the kit-provided AP1 primer.

The PCR master mix for each cDNA sample was incubated in a PCR thermocycler with preset amplification conditions. The steps involved denaturation at 98°C for 30 seconds, followed by an annealing step at 60°C for 30 seconds and the final extension step at 72 °C for 4 minutes. This entire protocol is cycled to 25 cycles. The amplification was confirmed by running the samples on a 2% agarose gel to visually confirm for the presence of bands indicating successful amplification of the cDNA samples. Upon visually confirming the presence of cDNA amplification, the samples were further purified by adding 2.5× volumes of AMPure XP SPRI bead suspension (Beckman Coulter) to the cDNA samples. The bead-sample suspension was allowed to mix and bind by incubating the samples at 25 °C in a hula mixer. The beads were pelleted on a magnet, and the supernatant was discarded. The pelleted beads were washed twice with 80% v/v ethanol, with each wash involving the same bead re-suspension and pelleting. The cDNA samples were eluted in 22ml of denoised mQ water. The concentration of the cDNA samples was obtained using a nanodrop spectrophotometer and its quality was further validated with Qubit Broad Range (BR) DNA assay using Qubit spectrophotometer (Thermo Fisher Scientific). The cDNA samples were stored at -80°C prior to conducting the *in vitro* transcription reaction.

***In vitro* transcription reaction of cDNA libraries**

Following amplification, *in vitro* transcription of both cDNA samples was performed using HiScribe T7 High Yield RNA Kit (New England Biolabs) following manufacturer's instructions and recommendations. For each transcription reaction, 2 µg of each cDNA sample was mixed with nucleotide buffer mix (NTP) to yield a final concentration of 10 mM of each nucleotide, 1 unit of RNasin Plus (Promega), 5 mM of dithiothreitol (DTT) and 2 units of T7 RNA polymerase mix with the total reaction volume scaled to 40 µl. The transcription reaction mixture was incubated at 37°C for 2.5 hours, after which the sample was diluted to 100 µl using deionised water. The diluted sample was then mixed with 1× DNase I buffer (Thermo Fisher Scientific) and 2 units of DNase I (Thermo Fisher Scientific) and was incubated for 15 minutes at 37°C. Following the 2.5-hour reaction time, 2.5× volumes of AMPure XP SPRI bead suspension (Beckman Coulter) was added to this mixture, and binding was conducted for 10 minutes at 25°C with periodic mixing. As recommended by the bead supplier, the binding step was followed by two sequential washing steps each with 1 ml of 80% v/v ethanol. The RNA sample was eluted in 30 µl of deionised MQ water. The concentration of the RNA samples was assessed initially with nanodrop spectrophotometer, and its quality was additionally assessed with Qubit Broad Range (BR) assay using Qubit spectrophotometer (Thermo Fisher Scientific). The RNA samples were stored at -80°C prior to *in vitro* polyadenylation.

***In vitro* Polyadenylation of cDNA libraries**

To facilitate sequencing with the standard direct RNA sequencing protocol, which relies on 3' poly(A) priming, *in vitro* polyadenylation of the human hippocampus and mouse embryo D7 RNA samples was performed following the transcription reaction. 25 µg of the total RNA from each sample was diluted in deionised MQ water and made up to a final volume of 94 µl. This sample was heated at 65°C for 5 minutes to remove any potential unwanted RNA secondary structures. 12 µl of *Escherichia coli* Poly(A) Polymerase buffer (New England Biolabs) was then added to the sample, followed by the addition of 8 µl of 1 mM ATP. The resultant solution was also supplemented with 15 units of *Escherichia coli* Poly(A) Polymerase (New England Biolabs) and 3 µl of 40 U/µl RNasin Plus (Promega). The reaction mixture was mixed by pipetting mixing and was incubated at 37°C for 15 minutes. The mixture was then diluted with 200 µl with deionised MQ water, and the sample was mixed with 2.5× AMPure XP bead suspension. A similar RNA extraction process, as described above in the transcription reaction, was repeated. Two independent replicates of each sample were separately subjected to the *in vitro* polyadenylation reaction.

Direct RNA library preparation and sequencing of cDNA libraries

Poly(A)+ RNA samples from human hippocampus and mouse embryo D7 were prepared for direct RNA sequencing (DRS) using the SQK-RNA004 chemistry according to ONT protocol, with minor alterations to obtain a high number of reads for SWARM training. Unless otherwise indicated, all libraries were prepared in double volumes of the SQK-RNA004 protocol. The library preparation was initiated with the first ligation step, where the RNA samples were sequentially mixed with T4 DNA Ligase Buffer and T4 DNA ligase (Thermo Fisher Scientific), 1 unit of RNasin Plus (Promega) and RTA adaptor required for the initial ligation of RT Adaptor. The first ligation reaction mixture was incubated for 15 minutes as recommended by the ONT SQK-RNA004 protocol. Subsequently, reverse transcription (RT) was carried out by supplementing the mixture with 5x Induro RT Reaction Buffer along with 200 units of Induro Reverse Transcriptase (Thermo Fisher Scientific), 10 mM dNTP, 0.1 M DTT and nuclease-free water as per manufacturer's instructions. The Induro RT reaction was conducted at 55°C for 10 minutes followed by 4°C hold in the PCR thermocycler.

Following RT, the resulting sample underwent purification using AgenCourt AMPure XP beads (Beckman Coulter). The steps for purification of the library involved a 5-minute incubation at the Hula mixer, magnetic pelleting, ethanol washing, and resuspension steps as described in the SQK-RNA004 protocol. In the subsequent second ligation step, the purified library was combined with T4 DNA Ligase along T4 DNA Ligase buffer (Thermo Fisher Scientific), denoised mQ water and RNA Ligation Adaptor (RLA) to facilitate ligation of the library with a motor protein adaptor as per manufacturer's instructions. A final AgenCourt AMPure XP beads based cleanup procedure was performed with two rounds of Wash Buffer (WB), as provided by the kit. Finally, the sample was eluted in elution buffer ELB as a part of the SQK-RNA004 protocol. All incubation durations, concentrations, and temperatures were rigorously maintained in accordance with SQK-RNA004 protocol provided by ONT throughout the entire duration

of the library preparation protocol. The resulting RNA library was directly used for loading or preserved at 4 °C prior to loading.

All cDNA libraries including replicates for the human hippocampus and mouse embryo D7 were sequenced on a PromethION device using FLO-PRO004RA flow cells. For the library loading, the flow cell was first inserted under the clip of the PromethION to ensure proper thermal and electrical contact. Quality control of the flow cell was performed to assess the number of pores, and the priming port was opened to remove any air bubbles. A priming mix, comprising of the RNA Flush Tether (RFT) and Flush Cell Flush (FCF) (Activated Flush Buffer), was prepared and loaded into the flow cell as per manufacturer's instructions. Following this, the loading library was prepared by mixing the ELB eluted RNA with Sequencing Buffer (SB) and Library Solution (LiS). Upon mixing, the library was gently loaded into the flow cell via the SpotON sample port in a dropwise manner. After loading both the activated flush buffer and the library into the FLO-PRO004RA flow cell, the SpotON sample port cover was replaced, the priming port was closed and PromethION was secured in place, following the recommended guidelines. The DRS sequencing runs were conducted at room temperature (~25 °C) up to a maximum duration of 72 hours with wash and reload, where the second half of the library was loaded for an additional 24 hours. Live basecalling was disabled during the run and the file format for data acquisition was set to POD5 for downstream analysis.

Basecalling and alignment

Reads from the synthetic and cDNA IVTs and the cell lines were basecalled using guppy 6.4.6 configured with the rna_r9.4.1_70bps_hac model for SQK-RNA002, or with dorado version 0.7.2 with the rna004_130bps_sup@v5.0.0 model for SQK-RNA004. Basecalling was accelerated using BLOW5 files [45] and buttery-eel [46]. Basecalled reads were aligned to references using minimap2 [47] with parameters: -ax map-ont -k 5 for the reads from the synthetic IVTs, and -k 14 for all other samples. Signals from the synthetic IVT reads were mapped to the sequence templates, whereas all other reads (from cell lines, cDNAs, and mammalian tissue) were mapped to the Ensembl annotation (cDNA references, version 109). The reference used for analysing rRNA reads included the human cDNA reference and GenBank accession numbers U13369 (for 28S rRNA nts 7935-12969), X03205 (for 18S rRNA), U13369 (for 5.8S rRNA nts 6623-6779). tRNA reads were mapped to hg38 high-confidence mature tRNA sequences obtained from GtRNAdb [48] (release 22). Event alignment was performed using f5c [49] with parameters: f5c eventalign --rna --signal-index --scale-events --samples --print-read-names --sam.

Read-level features processing

SWARM parses SAM files with aligned nanopore signal events for every 9-mer sequence context (five consecutive 5-mers) across all provided reads. For every base of interest, a 180x7 vector is generated by interpolating scaled signal events for each 5-mer into fixed-length vectors (n=36) and using 7 feature channels including: (1) interpolated signal events, (2-5) four channels holding the distances from the

expected k-mer signals varying the central 9-mer base (A/C/G/T), (6) dwelling times, (7) per-base q-scores. Expected event mean signal values were obtained from https://github.com/nanoporetech/kmer_models (for SQK-RNA002), and from https://raw.githubusercontent.com/hasindu2008/f5c/v1.3/test/rna004-models/rna004.nucleotide.5mer.model (for SQK-RNA004). Signal features are computed using C++ and directly streamed for parallel neural network prediction, circumventing long-term storage of intermediate files.

Read-level training datasets

Bespoke training datasets were curated for each sequencing kit and target modification. Every training dataset contained up to 500 signals per 9-mer from the all-5-mer synthetic constructs (IVT-i) with IVT sample harbouring target modification for positive data, and an equivalent number of negative signals comprised of 80% unmodified IVT signals, and 20% combination of equally represented IVT signals with non-target modifications (Ψ, m6A, m5C, ac4C).

Training dataset for m6A included above mentioned IVT-i signal mixture, supplemented with positive HEK293 mRNA signals (up to 5000 per reference position) from sites with high stoichiometry (>80%) reported by GLORI [12], and the same number of negatives originating from the same reference positions in fully unmodified HeLa mRNA IVT [31] for SQK-RNA002, and HEK293 mRNA IVT [50] + Human Hippocampus IVT (this study) for SQK-RNA004.

Dataset for training Ψ models was composed of IVT-i mixtures supplemented with modified signals (up to 5000 per reference position) from rRNA Ψ sites reported by mass-spec [23] and BID-seq [26] with high stoichiometry (>80%) in an *in-vitro* polyadenylated HeLa DRS sample (this study), and negative counterparts in the same number from equivalent positions from unmodified IVT sample synthesised from the mouse rRNA reference (IVT-m) [16]. In addition to rRNA, the SQK-RNA004 Ψ dataset contained tRNA modified signals (up to 5000 per central 5-mer) from an *in-vitro* polyadenylated HeLa sample (this study) at Ψ sites with high stoichiometry (>80%) reported by BACS [24], and negative signals within equivalent 9-mers from human hippocampus IVT (this study).

The training dataset for m5C for the SQK-RNA002 chemistry only contained IVT-i signal mixtures, while for the SQK-RNA004 chemistry it additionally included signals from an *in-vitro* polyadenylated HeLa sample (this study) (up to 5000 per central 5-mer) at high-stoichiometry sites (>80%) reported by UBS-seq [25], and negative signals within equivalent 9-mers from human hippocampus IVT (this study).

Each dataset was split into training/validation/testing using a 60/20/20 ratio, with stratified sampling ensuring the same proportions of 9-mer sequences in each split. In addition, five sub-classes were defined, which comprised different signal contexts and were used to compute sample weights for

balancing contributions of each context. Namely, IVT-i-multi, containing multiple target modifications in the same 9-mer, IVT-i-single, with only one target modification in the 9-mer, as well as mRNA, rRNA, and tRNA signals, where applicable.

Read-level network architectures and training

Considering differences in signal patterns and training data composition specific to each modification, three distinct neural network architectures were trained for each modification labelled according to the number of parameters (Mini/Mid/Large). Specifically, the *Mini* network was a custom ResNet [51] with 5 residual blocks and 96k parameters, the *Mid* network was the same one used by Mateos et al. [16] (128k parameters), and the *Large* network was a customised MobileNet [52] with variable filter size (3,7,11) and 6.7m parameters. Each neural network was trained using Tensorflow 2.15.0 with an Adam optimiser for 100 epochs with early stopping and cyclic learning rate. The best epoch was selected based on the minimal validation loss. Each model was trained using either no weights (w0) or balanced weights for all four above-mentioned sub-classes (w1).

Site-level model

A site-level model was trained for each read-level model by generating artificial mixtures of modified and unmodified IVT-i signals independent from the training dataset. For generating artificial sites, pools of up to 5000 read-level predictions per reference position were assembled from each of the target modification and unmodified IVT samples. For each reference site, 80 artificial sites were generated with random coverage uniformly sampled between 10-1000 reads, with 50/50 split of positive and negative sites. Positive sites contained a mixture of positive and negative signals with random stoichiometry uniformly sampled between 10% and 100%. Negative sites were generated using primarily unmodified IVT signals at a given reference site, with the additional noise added, including a random sample of 0-4% of predictions from a sample with target modification at the same site, and an identical proportion of random probabilities between 0.2 and 0.8 to represent less confident and noisy read-level predictions. The artificial site mixtures were split 60/20/20 for training/validation/testing.

Input features for all site-level models contained vectors of 100 values representing binned counts of read-level probabilities. All site-level models were trained using the previously described Mini network with the Adam optimiser for up to 100 epochs. Early stopping and a cyclic learning rate schedule were applied, and the best-performing epoch was selected based on the lowest validation loss.

Model selection

Optimal neural network architecture and training sample weights were chosen for each model based on both modification specificity on synthetic IVT controls and overlap of detected sites with orthogonal

methods. Modification accuracy was measured as area under the precision-recall curve (AUPRC) computed on synthetic IVT-i signals from the test split with balanced weights for each non-target modification context. Overlap with orthogonal methods was measured as AUPRC using site-level probabilities evaluated against polyA+ site labels from orthogonal methods: GLORI [12] sites for m6A models (excluding training sites) in HEK293, BACS [24] sites for Ψ models in HeLa, and UBS [25] sites for m5C models in HeLa. Since most models consistently displayed high AUPRC in IVT controls, models with the highest site overlap with orthogonal methods were selected, with the exception being the m6A model for SQK-RNA004, which was selected according to the highest AUPRC in IVTs, given the small differences in orthogonal overlap and the observed variability in performance on IVT controls.

Modification specificity benchmarking

Available nanopore modification models were applied to IVT-m samples (IVT-m1, IVT-m2, IVT-m3, IVT-m4), independent from the training data constructs and generated from Mouse rDNA templates containing Ψ, m6A, m5C, and no modifications. Due to coverage constraints, models for SQK-RNA002 were benchmarked on IVT-m1 and IVT-m2 constructs only, while models for SQK-RNA004 were benchmarked using the IVT-m4 construct. Up to 500 read-level predictions per reference position were sampled, and all models were benchmarked using the signals with target modification as positives, with separate AUPRC computed using negatives with different non-target modifications (Ψ, m6A, m5C) and unmodified IVTs. All reference positions supported by the models were used, with all context detection available for SWARM, CHEUI [16], TandemMod [15], dorado (0.8.2), m6ABasecaller [21], and DRACH-only constraint used for m6anet [20] and mAFiA [19].

Stoichiometry benchmarking

SWARM estimates stoichiometry at each tested reference site as the number of observations where read-level probability is over 0.5 out of all tested signals at a given site. To benchmark m6A stoichiometry correlation with GLORI in mRNA sites in HeLa, models were applied to publicly available HeLa nanopore samples [53] for SQK-RNA002, and to an in-house HeLa sample used for SQK-RNA004. Ψ and m5C mRNA stoichiometries were benchmarked against BACS [24] and UBS [25] measurements using publicly available HEK293 samples [35] for SQK-RNA002, and an in-house HEK293 sample for SQK-RNA004. Ψ stoichiometry in rRNA was benchmarked against mass-spec [23] in a publicly available HeLa *in-vitro* polyadenylated sample [53] for SQK-RNA002, and in-house HeLa *in-vitro* polyadenylated sample for SQK-RNA004. Pearson's R was computed for each stoichiometry comparison, restricted only to reference sites reported by all compared methods. Transcriptomic sites aligned to the cDNA reference were lifted over to the hg38 primary assembly using R2DTool [54], and the stoichiometry was calculated for each genomic coordinate by averaging across covered transcripts and weighting by coverage. Minimum coverage of 20 reads per transcript and default probability cutoffs were used for all nanopore methods.

Site-level overlap benchmarking

Nanopore-detected sites were evaluated against site labels reported by orthogonal methods using the same set of samples as described for the stoichiometry benchmarking. AUPRC was computed for each tool using site-level probabilities when available (SWARM, CHEUI, m6Anet, NanoSpa), or using stoichiometry (TandemMod, dorado 0.8.2, mAFiA, m6ABasecaller). For AUPRC analysis, true modification sites were considered those reported by orthogonal methods, while negative sites were all tested sites that were not orthogonally reported. Benchmarks were restricted to the reference sites that were tested by all compared methods. Transcriptomic sites aligned to the cDNA reference were lifted over to the genome as described above, and maximal site-level probability (or stoichiometry if site-probability was not available from the method) was selected from transcripts covering the same genomic coordinate. A minimum coverage of 20 reads per transcript site was used for all methods.

Site-level cut-offs for *de-novo* modification detection

The false positive rate (FPR) at different site-level probability cut-offs was estimated from predictions on fully unmodified polyA+ transcriptome IVTs. Publicly available IVT samples from HeLa, A549 and HepG2 [31,53] were used for SQK-RNA002, while publicly available HeLa and HEK293 samples [50] were used for SQK-RNA004. Minimum coverage of 20 reads per transcript was used. The site-level probability was set to 0 for all sites with stoichiometry under the detection threshold (10%). FPR was calculated as the number of predicted sites with a probability larger than a given cut-off out of all tested sites in that sample. For each model, the cut-off that resulted in an FPR < 0.1% in HeLa transcriptomes was selected for *de novo* detection of modified sites in further analyses.

Modification co-occurrence

Transcript-level m6A and Ψ co-occurrence was defined as the detection of at least one of each modification type in the same transcript. The expected number of transcripts displaying m6A and Ψ co-occurrence was computed by permuting site labels among tested sites for each tissue sample 1000 times. The mean and standard deviation of the permuted number of transcripts with co-occurrences were used to calculate z-scores as z-score $=(x-\mu)/\sigma$, with x = observed value; μ = sample mean; σ = sample standard deviation. Two-tailed p-values were obtained from the z-scores using $p = 2(1-\Phi(|z|))$, where Φ = the cumulative distribution function of the standard normal distribution.

Read-level co-occurrence was assessed for all m6A–Ψ site pairs detected within the same transcripts, restricted to pairs where each site was the nearest neighbour of the other modification type. For each site pair, only reads where both m6A and Ψ were tested were considered, and positions in individual reads were considered as modified if the predicted probability exceeded 0.5. The observed number of reads with both modifications detected was compared to the null distribution by permuting the tested

read-level modification states within the same reference position. Permuted mean and standard deviation were used to compute z-scores and p-values using the above formula.

Enrichment near splice-sites

m6A and Ψ sites detected across tissue samples were annotated with transcript metadata regarding the closest splice site distance and position with respect to the CDS/UTRs using R2DTool [54] and GTF annotation files for corresponding species (Ensembl release 109). Each transcriptomic position was considered as modified if detected above the described thresholds in at least one tissue of a given species, while background distribution was computed for transcriptomic sites tested in at least one tissue with coverage over 20 and writer motifs (DRACH for m6A, UVUAR for PUS7 Ψ, and GUUCNANNC for TRUB1 Ψ). DRACH m6A, PUS7 Ψ, and TRUB1 Ψ motifs were handled separately, and within each species, the modification site labels were shuffled 1000 times, computing modification site frequency in sliding density windows with window size=100bp, step=10bp, and considering only sites within 1kb from the closest splice junction. Observed modification frequency was compared to the permuted mean and standard deviation in each window to compute z-scores and p-values, adjusted with the Benjamini-Hochberg method.

TRUB1 substrate secondary structure analysis

Sequence complementarity matrices were generated as in Safra et al. [33]. RNA secondary structure of TRUB1 substrate was modeled using ViennaRNA python API version 2.7.0 on sequences 6bp upstream to 10bp downstream of the candidate pseudouridine site, i.e. `NNNNGU`**`U`**`CNANNCNNNN`, using `(((((.....)))))` as a constraint.

Nascent RNA pseudouridine detection at TRUB1-candidate sites

Raw nanopore signals from Sethi et al. [41] were basecalled with dorado 0.7.2 and aligned to the hg38 primary assembly using minimap2 -ax splice. Reads covering the region of each target site were extracted using samtools view and pod5 subset. Target reads were basecalled with dorado 0.8.2 and pseU model, converted to fastq with -T* flag and aligned to the genome using minimap2 -ax splice -y to keep modification tags for dorado analysis. To perform reference-free SWARM modification detection in target reads, dorado 0.7.2 basecalls were aligned using minimap2 -ax map-ont and resquiggled via f5c eventalign using a reference with basecalled read sequences, followed by SWARM read-level modification detection. SWARM modsam predictions were converted to fastq and aligned to the genome with same commands as described for dorado 0.8.2. Aligned mod.bam files were processed to only contain modification readouts at target sites, and modification probabilities over 0.9 were considered as modified for IGV visualisation and stoichiometry quantification. Minimum coverage of 5 reads was used for stoichiometry quantification.

Data availability

Datasets produced for this work will be uploaded to NCBI/SRA. These include: IVT RNA from human and mouse hippocampus cDNA libraries (SQK-RNA004), HeLa WT total and small-size-selected in-vitro polyadenylated (SQK-RNA004), and HEK293 WT (SQK-RNA004). Data from mammalian tissues (SQK-RNA002) [7] was obtained from PRJNA1188790. Data for cell lines (IVT HeLa, A549, HEK293) (SQK-RNA002) and WT HeLa (SQK-RNA002) [31,53] was obtained from https://www.ncbi.nlm.nih.gov/bioproject/PRJNA777450.https://www.ncbi.nlm.nih.gov/bioproject/PRJNA777450. Data for HEK293 WT and METTL3-KO data (SQK-RNA002) [35] was obtained from https://www.ebi.ac.uk/ena/browser/view/PRJEB40872. Data for HEK293 and HeLa IVT (SQK-RNA004) [50] was obtained from ENA (PRJEB80229). HeLa WT mRNA and pre-mRNA datasets (SQK-RNA004) [41] were obtained from . Data for HEK293 WT and METTL3-KO data (SQK-RNA002) [35] was obtained from https://www.ebi.ac.uk/ena/browser/view/PRJEB40872. Data for HEK293 and HeLa IVT (SQK-RNA004) [50] was obtained from ENA (PRJEB80229). HeLa WT mRNA and pre-mRNA datasets (SQK-RNA004) [41] were obtained from https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1256087.

Software Availability

SWARM v1 (https://github.com/comprna/SWARM)
Minimap2 (https://github.com/lh3/minimap2/releases/tag/v2.24)
Dorado 0.8.2 (https://github.com/nanoporetech/dorado/releases/tag/v0.8.2)
TandemMod v1.1.0 (https://github.com/yulab2021/TandemMod)
M6Abasecalller (https://github.com/novoalab/m6ABasecaller)
CHEUI v1.0 (https://github.com/comprna/CHEUI)
F5c v1.5 (https://github.com/hasindu2008/f5c/releases/tag/v1.5)
Battery-eel v0.7.2 (https://github.com/Psy-Fer/buttery-eel/releases)
R2DTool v2.0.1 (https://github.com/comprna/R2Dtool/releases/tag/v2.0.1)

Funding

This research was supported by the Australian Research Council (ARC) Discovery Project grants DP250100070 and DP250103133; by the National Health and Medical Research Council (NHMRC) Ideas Grant 2018833; by a PhD scholarship from the ANU TALO Computational Biology Talent Accelerator, by European Union's Horizon 2020 Research and innovation program Marie Sklodowska-Curie grant (agreement No 890462), by an NHMRC Investigator Grant GNT1175388, and by the ARC Centre of Excellence for the Mathematical Analysis of Cell Systems (MACSYS) (CE230100001). This research was also indirectly supported by the Australian Government's National Collaborative Research

Infrastructure Strategy (NCRIS) through access to computational resources provided by the National Computational Infrastructure (NCI) through the National Computational Merit Allocation Scheme (NCMAS) and the ANU Merit Allocation Scheme (ANUMAS). The funding bodies had no role in study design, data collection, or data analysis.

Acknowledgements

We are grateful to the personnel from the Biomolecular Resource Facility at JCSMR (ANU) for their technical assistance. We are also grateful to the personnel of the Ecogenomics and Bioinformatics Lab, a joint initiative of the Research School of Biology (ANU) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO), and particularly to Niccy Aitken and Ashley Jones, for their continued support and feedback regarding ONT sequencing.

References

1. Delaunay, S., Helm, M. & Frye, M. RNA modifications in physiology and disease: towards clinical applications. *Nat. Rev. Genet.* **25**, 104–122 (2024).
2. Shi, H., Wei, J. & He, C. Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. *Mol. Cell* **74**, 640–650 (2019).
3. Borchardt, E. K., Martinez, N. M. & Gilbert, W. V. Regulation and Function of RNA Pseudouridylation in Human Cells. *Annu. Rev. Genet.* **54**, 309–336 (2020).
4. Lewis, C. J. T., Pan, T. & Kalsotra, A. RNA modifications and structures cooperate to guide RNA-protein interactions. *Nat. Rev. Mol. Cell Biol.* **18**, 202–210 (2017).
5. Boo, S. H. & Kim, Y. K. The emerging role of RNA modifications in the regulation of mRNA stability. *Exp. Mol. Med.* **52**, 400–408 (2020).
6. Liu, J. *et al.* Developmental mRNA m5C landscape and regulatory innovations of massive m5C modification of maternal mRNAs in animals. *Nat. Commun.* **13**, 2484 (2022).
7. Santos-Rodriguez, G. *et al.* The conserved landscape of RNA modifications and transcript diversity across mammalian evolution. *bioRxiv [PREPRINT]* 2024.11.24.624934 (2024). doi:10.1101/2024.11.24.624934
8. Li, Q. *et al.* NSUN2-Mediated m5C Methylation and METTL3/METTL14-Mediated m6A Methylation Cooperatively Enhance p21 Translation. *J. Cell. Biochem.* **118**, 2587–2598 (2017).
9. Meyer, K. D. *et al.* 5' UTR m(6)A Promotes Cap-Independent Translation. *Cell* **163**, 999–1010 (2015).
10. Guca, E. *et al.* N6-methyladenosine in 5' UTR does not promote translation initiation. *Mol. Cell* **84**, 584-595.e6 (2024).

11. Körtel, N. *et al.* Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning. *Nucleic Acids Res.* **49**, e92 (2021).

12. Liu, C. *et al.* Absolute quantification of single-base m6A methylation in the mammalian transcriptome using GLORI. *Nat. Biotechnol.* **41**, 355–366 (2023).

13. Xiao, Y.-L. *et al.* Transcriptome-wide profiling and quantification of N6-methyladenosine by enzyme-assisted adenosine deamination. *Nat. Biotechnol.* **41**, 993–1003 (2023).

14. Huang, S., Wylder, A. C. & Pan, T. Simultaneous nanopore profiling of mRNA m6A and pseudouridine reveals translation coordination. *Nat. Biotechnol.* **42**, 1831–1835 (2024).

15. Wu, Y. *et al.* Transfer learning enables identification of multiple types of RNA modifications using nanopore direct RNA sequencing. *Nat. Commun.* **15**, 4049 (2024).

16. Acera Mateos, P. *et al.* Prediction of m6A and m5C at single-molecule resolution reveals a transcriptome-wide co-occurrence of RNA modifications. *Nat. Commun.* **15**, 3899 (2024).

17. Sendinc, E. *et al.* Mapping multiple RNA modifications simultaneously by proximity barcode sequencing [PREPRINT]. *bioRxiv* 2024.10.09.617509 (2024). doi:10.1101/2024.10.09.617509

18. Diensthuber, G. *et al.* Systematic benchmarking of basecalling models for RNA modification detection with highly-multiplexed nanopore sequencing. (2025).

19. Chan, A., Naarmann-de Vries, I. S., Scheitl, C. P. M., Höbartner, C. & Dieterich, C. Detecting m6A at single-molecular resolution via direct RNA sequencing and realistic training data. *Nat. Commun.* **15**, 3323 (2024).

20. Hendra, C. *et al.* Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat. Methods* **19**, 1590–1598 (2022).

21. Cruciani, S. *et al.* De novo basecalling of RNA modifications at single molecule and nucleotide resolution. *Genome Biol.* **26**, 38 (2025).

22. Luo, T., Xu, M., Wang, M., Chen, F. & Shi, J. Systematic evaluation of computational tools for multitype RNA modification detection using nanopore direct RNA sequencing. *Nat. Methods* (2025). doi:10.1038/s41592-025-02974-y

23. Taoka, M. *et al.* Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res.* **46**, 9289–9298 (2018).

24. Xu, H. *et al.* Absolute quantitative and base-resolution sequencing reveals comprehensive landscape of pseudouridine across the human transcriptome. *Nat. Methods* **21**, 2024–2033 (2024).

25. Dai, Q. *et al.* Ultrafast bisulfite sequencing detection of 5-methylcytosine in DNA and RNA. *Nat. Biotechnol.* **42**, 1559–1570 (2024).

26. Dai, Q. *et al.* Quantitative sequencing using BID-seq uncovers abundant pseudouridines in

mammalian mRNA at base resolution. *Nat. Biotechnol.* **41**, 344–354 (2023).

27. Zhang, M. *et al.* Quantitative profiling of pseudouridylation landscape in the human transcriptome. *Nat. Chem. Biol.* **19**, 1185–1195 (2023).

28. Schumann, U. *et al.* Multiple links between 5-methylcytosine content of mRNA and translation. *BMC Biol.* **18**, 40 (2020).

29. Ge, R. *et al.* m6A-SAC-seq for quantitative whole transcriptome m6A profiling. *Nat. Protoc.* **18**, 626–657 (2023).

30. Li, X. *et al.* Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.* **11**, 592–7 (2015).

31. McCormick, C. A. *et al.* Multicellular, IVT-derived, unmodified human transcriptome for nanopore-direct RNA analysis. *GigaByte (Hong Kong, China)* **2024**, gigabyte129 (2024).

32. Martinez, N. M. *et al.* Pseudouridine synthases modify human pre-mRNA co-transcriptionally and affect pre-mRNA processing. *Mol. Cell* **82**, 645-659.e9 (2022).

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

34. Liu, J. *et al.* Epitranscriptomic subtyping, visualization, and denoising by global motif visualization. *Nat. Commun.* **14**, 5944 (2023).

35. Pratanwanich, P. N. *et al.* Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat. Biotechnol.* **39**, 1394–1402 (2021).

36. Rodell, R. *et al.* Nano-Mod-Amp reveals RNA sequence, structural and cell type specific features of pseudouridylation by PUS7. (2025). doi:10.1101/2025.10.30.685621

37. Luo, Z. *et al.* Exon-intron boundary inhibits m6A deposition, enabling m6A distribution hallmark, longer mRNA half-life and flexible protein coding. *Nat. Commun.* **14**, 4172 (2023).

38. Uzonyi, A. *et al.* Exclusion of m6A from splice-site proximal regions by the exon junction complex dictates m6A topologies and mRNA stability. *Mol. Cell* **83**, 237-251.e7 (2023).

39. He, P. C. *et al.* Exon architecture controls mRNA m6A suppression and gene expression. *Science* **379**, 677–682 (2023).

40. Yang, X., Triboulet, R., Liu, Q., Sendinc, E. & Gregory, R. I. Exon junction complex shapes the m6A epitranscriptome. *Nat. Commun.* **13**, 7904 (2022).

41. Sethi, A. J. *et al.* Single-molecule multimodal timing of in vivo mRNA synthesis. *bioRxiv* 2025.04.27.650906 (2025). doi:10.1101/2025.04.27.650906

42. Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).

43. International Human RNome Project Consortium. Unlocking the regulatory code of RNA: launching the Human RNome Project. *Genome Biol.* **26**, 367 (2025).

44. National Academies of Sciences and Medicine, E. *Charting a Future for Sequencing RNA and Its Modifications: A New Era for Biology and Medicine*. (The National Academies Press, 2024). doi:10.17226/27165

45. Gamaarachchi, H. *et al.* Fast nanopore sequencing data analysis with SLOW5. *Nat. Biotechnol.* **40**, 1026–1029 (2022).

46. Samarakoon, H., Ferguson, J. M., Gamaarachchi, H. & Deveson, I. W. Accelerated nanopore basecalling with SLOW5 data format. *Bioinformatics* **39**, (2023).

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

48. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93-7 (2009).

49. Gamaarachchi, H. *et al.* GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC Bioinformatics* **21**, 343 (2020).

50. Zou, Y. *et al.* A Comparative Evaluation of Computational Models for RNA modification detection using Nanopore sequencing with RNA004 Chemistry. *bioRxiv Prepr. Serv. Biol.* (2025). doi:10.1101/2025.02.03.636352

51. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. (2015).

52. Howard, A. G. *et al.* MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017).

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

54. Sethi, A. J., Acera Mateos, P., Hayashi, R., Shirokikh, N. E. & Eyras, E. R2Dtool: integration and visualization of isoform-resolved RNA features. *Bioinformatics* **40**, 2022.09.23.509222 (2024).