Simultaneous identification of m6A and m5C reveals coordinated RNA modification at single-molecule resolution

P. Acera Mateos[1,2,3], A.J. Sethi[1,2,3,&], A. Ravindran[1,2,3,&], M. Guarnacci[1,3,&], A. Srivastava[1,2,3], J. Xu[1], K. Woodward[1,3], Z.W.S. Yuen[1,2,3], W. Hamilton[4], J. Gao[1], L.M. Starrs[1], R. Hayashi[1,3], V. Wickramasinghe[4], T. Preiss[1,3,5], G. Burgio[1,3], N. Dehorter[1,3], N. Shirokikh[1,3*], E. Eyras[1,2,3,6*]

[1] The John Curtin School of Medical Research, Australian National University, Canberra, Australia

[2] EMBL Australia Partner Laboratory Network at the Australian National University, Canberra, Australia

[3] The Shine-Dalgarno Centre for RNA Innovation, Australian National University, Canberra, Australia

[4] Peter MacCallum Cancer Centre, Melbourne, Australia

[5] Victor Chang Cardiac Research Institute, Sydney, Australia.

[6] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

[&] These authors contributed equally

[*] Correspondence to: nikolay.shirokikh@anu.edu.au, eduardo.eyras@anu.edu.au

ABSTRACT

The epitranscriptome embodies many new and largely unexplored functions of RNA. A major roadblock in the epitranscriptomics field is the lack of transcriptome-wide methods to detect more than a single RNA modification type at a time, identify RNA modifications in individual molecules, and estimate modification stoichiometry accurately. We address these issues with CHEUI (CH3 (methylation) Estimation Using Ionic current), a new method that concurrently detects N6-methyladenosine (m6A) and 5-methylcytidine (m5C) in individual RNA molecules from the same sample, as well as differential methylation between any two conditions, using signals from nanopore direct RNA sequencing. CHEUI processes observed and expected signals with convolutional neural networks to achieve high single-molecule accuracy and outperform other methods in detecting m6A and m5C sites and quantifying their stoichiometry. CHEUI's unique capability to identify two modification types in the same sample reveals a non-random co-occurrence of m6A and m5C in mRNA transcripts in cell lines and tissues. CHEUI unlocks an unprecedented potential to study RNA modification configurations and discover new epitranscriptome functions.

INTRODUCTION

The identification of transcriptome-wide maps of two modified ribonucleotides in messenger RNAs (mRNA), 5-methylcytidine (m5C) and N6-methyladenine (m6A) [1–3], over the last decade has sparked a new and expanding area of epitranscriptomics. Techniques based on immunoprecipitation, enzymatic, or chemical reactivity

enrichment methods, coupled with high-throughput sequencing, have uncovered the role of these and other modifications in multiple steps of mRNA metabolism, including translation of mRNA into protein [4,5], mRNA stability [6] and mRNA processing such as alternative splicing of pre-mRNA [7]. Several physiological processes have also been functionally linked with RNA modifications, such as sex determination [8], neurogenesis [9] and learning [10]. Moreover, there is growing evidence that RNA modification pathways are dysregulated in diseases such as cancer [11] and neurological disorders [9]. Most of these studies have focused on changes at global or gene levels, or on the dysregulation of the RNA modification machinery, whereas little is known about how multiple modifications occur in individual mRNA molecules.

A major roadblock preventing rapid progress in research on RNA modifications is the general lack of universal modification detection methods. Although more than 150 naturally occurring RNA modifications have been described [12], only a handful of them can be mapped and quantified across the transcriptome [13,14]. Nanopore direct RNA sequencing (DRS) is the only currently available technology that can determine the primary structure of individual RNA molecules in their native form at a transcriptome-wide level. DRS can capture information about the chemical structure, including naturally occurring covalent modifications in nucleotide residues (nucleotides) [15,16]. Nonetheless, RNA modification detection from DRS signals presents various challenges. The differences between modified and unmodified signals are subtle at single-molecule level and depend on the sequence context. Additionally, due to the variable translocation rate of the molecules through the pores and the potential pore-to-pore variability, different copies of the same molecule present considerable signal variations [17]. These challenges necessitate the application of advanced computational models to interpret the signals and identify their modification status.

Several computational methods have been developed in the past few years to detect RNA modifications in DRS data. These methods can be broadly grouped into two categories. The first one includes methods that rely on comparing DRS signals between two conditions, one corresponding to a sample of interest, often the wild type (WT) sample, and the other with a reduced presence of a specific modification, usually obtained through a knock-out (KO) or knock-down (KD) of a modification 'writer' enzyme or through *in vitro* transcription. This category includes Nanocompore [18], Xpore [19], DRUMMER [20], nanoDOC [21], Yanocomp [22] and Tombo in *sample comparison* mode, all utilizing the collective properties of DRS signals in the two conditions. This category also includes ELIGOS [23] and Epinano [24], which compare base-calling errors between two experiments; and nanoRMS [25], which compares signal features between two samples. The second category of tools can operate in a single condition, i.e., without using a KO/KD or an otherwise control. This category includes MINES [26], Nanom6A [27], and m6Anet [28], all predicting m6A on specific sequence contexts, Tombo in *alternate* mode, which identifies transcriptomic sites with potential m5C modification, and Epinano-RMS, which predicts pseudouridine on high stoichiometry sites [25].

Despite the numerous advances in direct RNA modification detection, some major limitations remain. Approaches comparing two conditions generally require a control sample, which may be difficult to generate. Their modification calling is also indirect, as it relies on changes in the control sample relative to wild type (WT) and these changes may not be related to the modification of interest *per se*. For instance, depletion of m5C leads to a reduction of hm5C [29], hence potentially confounding the results. Regarding the methods that use error patterns, they depend on the specific accuracy of the base caller method, which will vary over time with the base caller version. Moreover, this may not be applicable to all modifications. For instance, it was described that error patterns were not consistent enough to confidently identify m5C methylation [23]. Limitations also exist in methods that work with individual samples. MINES, Nanom6A, and m6Anet only predict m6A modifications in 5′-DR<u>A</u>CH/RR<u>A</u>CH-3′ motifs, and Epinano-RMS only detects pseudouridine in transcriptome sites of high

stoichiometry. Additionally, the ability of current methods from both categories to predict stoichiometry is limited. Some of them cannot predict it, whereas others only estimate the stoichiometry at 5′-DRACH-3′ sites or rely on a control sample devoid of modifications. Importantly, to our knowledge, there are currently no methods that can concurrently predict transcriptome-wide more than one modification type in individual long RNA sequencing reads.

To address the existing limitations, we developed CHEUI (CH3 (methylation) Estimation Using Ionic current), a new computational tool based on a two-stage neural network that provides a series of significant innovations: (1) CHEUI was trained using read signals generated from in-vitro transcripts (IVTs) with specific modification configurations that can be cost-effectively extended to other modifications; (2) CHEUI enables the identification of m6A and m5C from the same sample; (3) CHEUI detects m6A and m5C at a transcriptome-wide level in individual molecules and in any sequence context, without the need for a KO/KD or control sample; (4) CHEUI achieves higher accuracy than other existing methods in predicting m6A and m5C stoichiometry levels while maintaining a lower number of false positives; and (5) CHEUI assesses the differential m6A and m5C deposition per site across the transcriptome between any two conditions. Through a comprehensive set of analyses using data from IVTs, cell lines, and tissues, we uncover a non-random co-occurrence of m6A and m5C in individual mRNA transcripts. CHEUI addresses multiple current limitations in the transcriptome-wide identification of RNA modifications and its broad applicability compared to the previous methods provides a paradigm shift in the transcriptome-wide study of study RNA modifications.

RESULTS

CHEUI enables the detection of m6A and m5C in individual reads and across conditions

For signal pre-processing, CHEUI transforms nanopore read signals into 9-mer groups, composed of five overlapping 5-mers each, and centered at the candidate modification site, adenosine (A) for m6A or cytosine (C) for m5C (**Fig. 1a**) (**Supp. Fig. 1**). Pre-processing includes the derivation of distances between the observed and the expected unmodified signal values from each 5-mer, which become part of the input (**Fig. 1a**) (**Supp. Figs. 2a-2c**). Inclusion of the distance metrics increased accuracy on the validation set by ~10% (**Supp. Fig. 2d**). After preprocessing the signals, CHEUI has two different modules: CHEUI-solo **(Fig. 1b),** which makes predictions in individual samples, and CHEUI-diff **(Fig. 1c),** which tests differential methylation between any two samples. CHEUI-solo predicts methylation at two different levels. It first predicts m6A or m5C at nucleotide resolution on individual read signals (Model 1) and then predicts m6A or m5C at the transcript site level by processing the predicted individual read probabilities with a second model (Model 2) (**Fig. 1b**). Both CHEUI-solo Models 1 and 2 are Convolutional Neural Networks (CNNs) (**Supp. Fig. 3**). CHEUI-diff uses a statistical test to compare the individual read probabilities from CHEUI-solo Model 1 across two conditions, to predict differential m6A or m5C at each transcriptomic site (**Fig. 1c**).

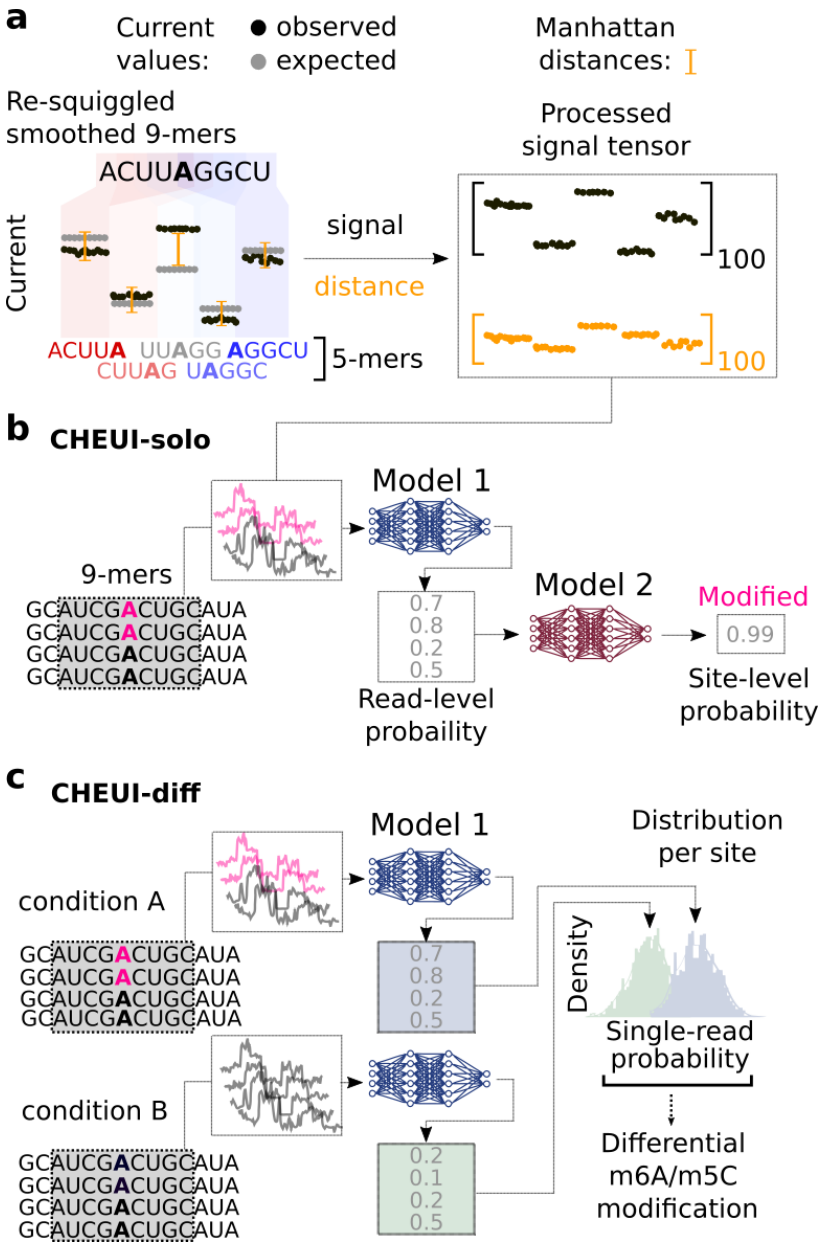

Figure 1. CHEUI modules and signal processing approach. (a) CHEUI processes signals for each 9-mer, i.e., five consecutive 5-mers. The signals for each 5-mer are converted to 20 median values, yielding a vector of length 100. An expected vector of length 100 is calculated for all five 5-mers and a vector of distances between the expected and observed signal values is obtained. These signal and distance vectors are used as inputs for Model 1. **(b)** CHEUI-solo operates in two stages. In the first stage, Model 1 takes the signal and distance vectors corresponding to individual read signals associated to a 9-mer centered at A (indicated as modified in pink, or unmodified in black) or C and predicts the probability of each individual site being modified A (m6A model) or modified C (m5C model). In the second stage, Model 2 takes the distribution of Model 1 probabilities for all the read signals at a given transcriptomic site and predicts the probability of the site being methylated. Then the with stoichiometry is estimated as the proportion of modified reads from Model 1 at that site. **(c)** CHEUI-diff uses the individual read probabilities from Model 1 in two conditions to test for differential m6A or m5C at a specific transcript site.

CHEUI accurately detects m5C and m6A in individual reads and in sequence contexts not seen during training

To evaluate CHEUI's accuracy, we first determined whether CHEUI-solo correctly classifies read signals from k-mer contexts (k=9) used for training but for read signals not previously used, i.e., sensor generalization [30]. For this test, only read signals from 9-mers with a single modified nucleotide were considered, such as those 9-mers where only one A or one C was present, we called this dataset IVT test 1. CHEUI achieved accuracy, precision, and recall values of ~0.8 for m6A and m5C predictions in individual reads (**Fig. 2a,** IVT test 1**)** (**Supp. Figs. 4a and 4b**). Then, to determine CHEUI's ability to classify signals from k-mer contexts not seen during training, i.e., k-mer generalization [30], we used signals from a different IVT sequencing experiment[23], we called this dataset IVT test 2. As before, this test only included signals from 9-mer sites with a single middle A or C. CHEUI achieved accuracy, precision, and recall of ~0.8 for m6A and ~0.75 for m5C (**Fig. 2a,** IVT test 2**)** (**Supp. Figs. 4c and 4d)**.

We next explored whether a double cutoff for the individual read probability would improve the accuracy. In this setting, predictions above a first probability cutoff would be considered methylated, whereas those below a second probability cutoff would be considered non-methylated, with all other read signals between these two cutoff values being discarded. After testing various options, we decided to use probability cutoffs 0.7 and 0.3, which provided an optimal balance between accuracy gain and the number of preserved reads, with an improved area under the receiver operating characteristic curve (ROC AUC) for m6A (from 0.857 to 0.899) and m5C (from 0.827 to 0.877) **(Fig. 2b)**, while retaining ~73% of the reads **(Fig. 2c)**.

To train and test CHEUI-solo Model 2 for predicting the methylation probability at the transcript site level, we *built in* silico-controlled mixtures of reads from the IVT test 1 dataset, with pre-defined proportions of modified and unmodified read signals not included in the training and testing of CHUEI-solo Model 1. CHEUI achieved an AUC of 0.92 for m6A and 0.953 for m5C (**Fig. 2d**). Moreover, at a per-site probability > 0.99, the estimated false positive rate (FPR) on the test data was 0.00074 for m6A and 0.00034 for m5C **(Fig. 2e)**.

CHEUI outperforms other tools at detecting m6A and m5C transcriptomic sites and their stoichiometry levels

We next compared CHEUI-solo with Nanocompore[18], Xpore[19], and Epinano[24] for the ability to detect and quantify RNA modifications. To achieve this, we built positive and negative independent test datasets using mixtures of read signals from IVT test 2 RNAs with known modifications. The positive sites were built with a pre-defined percentage stoichiometry of 20, 40, 60, 80, and 100, using 81 sites for m6A and 84 sites for m5C for each stoichiometry. The negative sites consisted of 512 sites for A and 523 sites for C, using only unmodified IVTs. To build the positive and negative sites, we sampled reads randomly at a variable level of coverage, resulting in a lifelike coverage range of 20 to 149 reads per site. Since Nanocompore, Xpore, and Epinano require a control sample to detect modifications, a second dataset containing only unmodified signals was created for the same sites, randomly sub-sampling independent reads to the same level of coverage. We observed that the number of true positives (TPs) detected by most tools increased with the site stoichiometry (**Fig. 2f**). Notably, CHEUI-solo recovered a higher number of true methylated sites compared to the other tools at all stoichiometry levels for both m6A and m5C. We next estimated the false positive rate (FPR) by predicting with all tools on the built negative sites, using a single sample for CHEUI-solo and two independent negative samples for Xpore, Epinano, and Nanocompore. Xpore and Epinano showed the highest false-positive rate for m6A and m5C. CHEUI-solo had 1 misclassified site for m5C and none for m6A, whereas Nanocompore had no false positives (**Fig. 2g**).

We next evaluated the stoichiometry prediction in a site-wise manner. For this analysis we included nanoRMS [25] and Tombo [31], which estimate stoichiometries at pre-defined sites. Stoichiometries were calculated for the sites that were previously predicted to be modified by each tool. For nanoRMS and Tombo, the predictions for all sites were considered since these tools do not specifically predict whether a site is modified or not. CHEUI-solo outperformed all the other tools, showing a higher correlation for m6A (Pearson r = 0.839) and m5C (Pearson r = 0.839) with the ground truth (**Fig. 2h**). CHEUI-solo was followed by Xpore (r = 0.524) and Nanocompore (r= 0.498) for m6A, and by Xpore (r = 0.556) and NanoRMS (r= 0.46) for m5C. Other tools tested showed low or negative correlations **(Supp. Fig. 5)**.

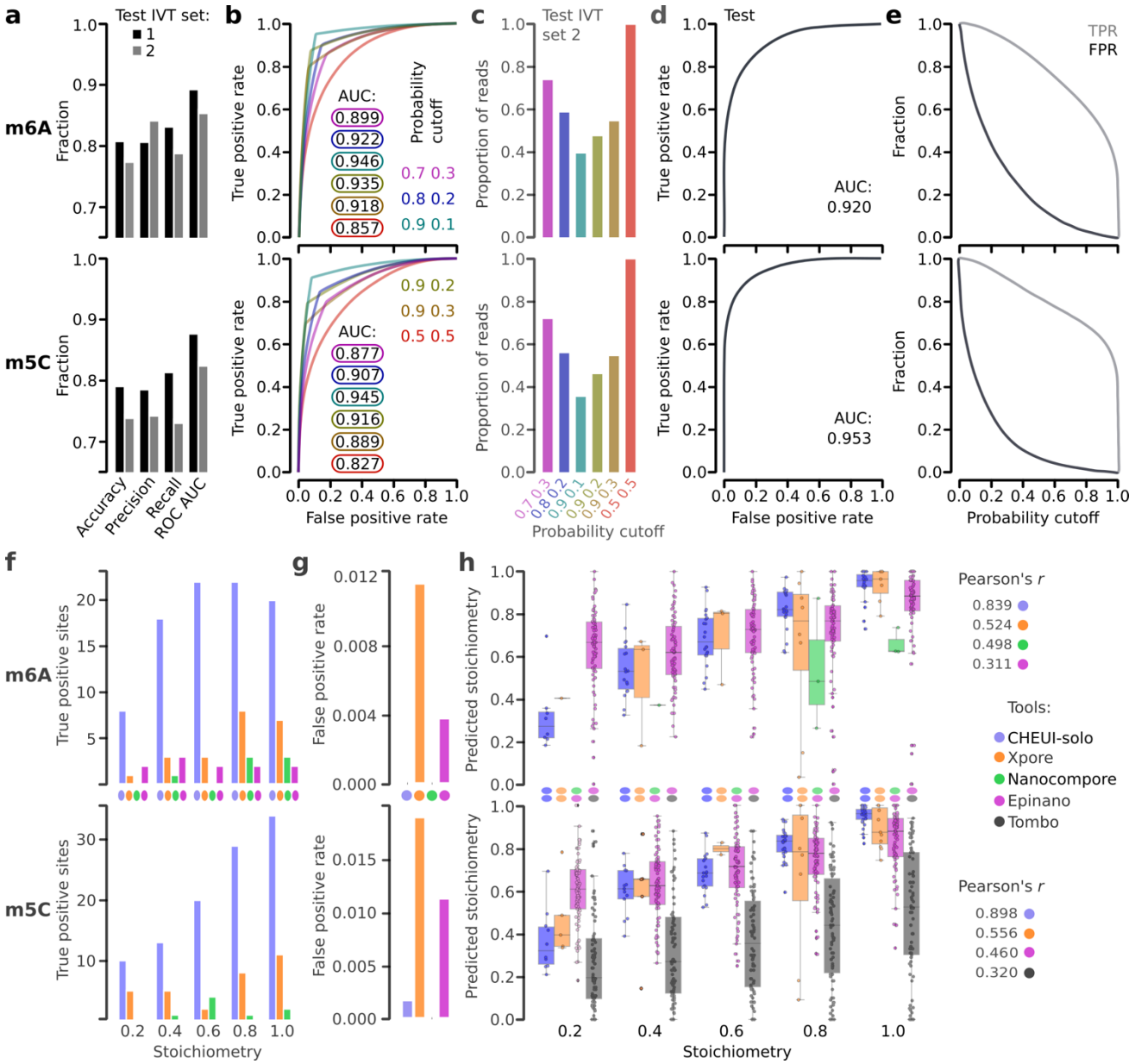

Figure 2. Accuracy metrics at the individual read level and comparison with other RNA modification detection tools. (a) Accuracy, precision, recall, and area (AUC) under the receiver operating curve (ROC) for CHEUI-solo Model 1 for m6A (upper panel) and m5C (lower panel) detections are shown for individual reads containing sequences seen during training (IVT test set 1) and for reads with sequences not seen during training

(IVT test set 2). The metrics (accuracy, precision, recall, AUC) for m6A were (0.835, 0.82, 0.853, 0.91) for IVT test 1 and (0.777, 0.844, 0.791, 0.856) for IVT test 2. For m5C, these metrics were (0.793, 0.788, 0.816, 0.879) for IVT test 1 and (0.741, 0.745,0.733, 0.827) for IVT test 2. **(b)** ROC curves for m6A (upper panel) and m5C (lower panel) for CHEUI-solo Model 1 on the IVT test 2 dataset at different double cutoffs to separate modified and unmodified read signals. The double cutoff is indicated as an X Y pair, where detection probability > X was used to select positives and detection probability < Y was used to select negatives; all other signals being discarded. The ROC curve and double cutoffs are color matched. **(c)** Proportion of reads selected (y-axis) for each double cutoff (x-axis). **(d)** ROC AUC for CHEUI-solo model 2 and the accuracy of predicting m6A (upper panel) or m5C (lower panel) modified transcript sites calculated using independent benchmarking datasets, IVT test 2. **(e)** True positive rate (TPR) and false positive rate (FPR) for CHEUI-solo Model 2 for m6A (upper panel) and m5C (lower panel) as a function of the probability cutoff (x-axis). **(f)** True Positives (TPs) per tool (y-axis) at different stoichiometry levels (x-axis) using and independent benchmarking dataset (IVT test 2), for m6A (upper panel) and m5C (lower panel). **(g)** False Positive Rate (FPR) (y-axis) for each tool (x-axis) returned for 512 m6A negative sites (upper panel) and 523 m5C negative sites (lower panel). Xpore had 14 false-positive site detections (FPR = 0.0273) for m6A and 32 (FPR = 0.0611) for m5C. Epinano detected 2 false-positive sites (FPR = 0.0039) for m6A and 6 (FPR = 0.011) for m5C. CHEUI-solo had 1 false positive site detection for m5C (0.0019 FPR) and none for m6A. Nanocompore had no false positives. **(h)** Correlation between the stoichiometry predicted by each tool (y-axis) and the ground truth stoichiometry using controlled read mixtures (x-axis) for m6A (upper panel) and m5C (lower panel). We included predictions by CHEUI-solo, Xpore, Nano-RMS with the k-nearest neighbors (kNN) algorithm, and Tombo in the alternate mode (only for m5C). The Pearson correlation (r) was calculated between the predicted stoichiometries and the ground truth stoichiometry across all the sites. Correlations for other tools are shown in **Supp. Fig. 5**.

CHEUI accurately identifies m6A modifications in cellular mRNA

We next tested CHEUI's ability to correctly identify m6A in cellular RNA. Using DRS reads from wild-type (WT) HEK293 cells [19] **(Supp. Table S1),** we tested 3,138,914 transcriptomic sites with a coverage of more than 20 reads in all three available replicates. Prior to any significance filtering, these sites showed a high correlation among replicates in the CHEUI predicted stoichiometry and modification probability per site (**Fig. 3a)**. Analyzing the replicates together, we considered as significant those sites with predicted probability > 0.9999, which using an empirical permutation test was estimated to result in an FDR of approximately 0 (see Methods). At this cutoff, CHEUI-solo predicted 10,036 significant m6A transcriptomic sites in 3,905 transcripts, corresponding to 8776 genomic sites **(Supp. Tables S2** and **S3**). Most of the modifications appear on single As, while a minor proportion of AA and AAA sites were predicted as modified (**Supp. Fig. 6a**). 85.12% of the transcriptomic sites (84.63% genomic sites) had the 5′-DRACH-3′ motif, which was higher than the 76.57% identified from m6ACE-seq and miCLIP experiments [32,33]. Interestingly, CHEUI-solo predicted m6A in 1,493 non-DRACH motifs (1,356 genomic sites), with the two most common ones being 5′-GGACG-3′ (203 genomic sites) and 5′-GGATT-3′ (121 genomic sites). These motifs were also the two most common non-DRACH motifs identified by miCLIP2 experiments in the same cell line, occurring at 245 (5′-GGACG-3′) and 96 (5′-GGATT-3′) sites [34].

Next, we considered the DRS data from HEK293 cells with a knockout of the m6A writer METTL3 (METTL3-KO) [19]. Using CHEUI-solo predictions at individual read level, we confirmed a significant decrease in the proportion of m6A nucleotides in METTL3-KO with respect to WT (p-value = 1.3E-254) (**Supp. Fig. 6b**). Using transcriptomic site probability >0.9999 as before, we obtained 4,603 significant m6A transcriptomic sites in METTL3-KO (**Supp. Table S4**), approximately half the number of WT sites, and corroborated an overall decrease in the proportion of modified sites along mRNAs in the KO samples (**Supp. Fig. 6c**).

To compare CHEUI with other methods, we predicted differential m6A between HEK293 WT and METTL3-KO samples. As expected, CHEUI-diff showed enrichment of significant cases with higher modification stoichiometry in WT (**Fig. 3b**) (**Supp. Table S5**). In comparison with Xpore and Nanocompore, CHEUI-diff detected more sites with higher modification stoichiometry in WT at three different significance thresholds (**Fig. 3c**). CHEUI-diff also predicted a higher proportion of sites with supporting evidence from m6ACE-seq or miCLIP experiments in HEK293 cells [32,33] (**Supp. Fig. 7a**) and containing the 5′-DRACH-3′ motif (**Fig. 3d**), except at the 0.001 significance level, where 0.70 of CHEUI-diff sites and 0.71 of Xpore sites contained the motif. Comparing two METTL3-KO replicates to estimate false positives, CHEUI-diff predicted the lowest number of sites (0, 1, and 3, at the three significance thresholds, respectively) (**Supp. Fig. 7b**). In contrast, Xpore predicted over 2,000 sites at 0.001 significance and over 12,000 sites at 0.05 significance. Only 9.8% of these Xpore sites at significance alpha level=0.05 contained the 5′-DRACH-3′ motif. This was a substantially lower proportion than the 46% found by Xpore in the WT vs METTL3-KO comparison at the same significance level, suggesting that most of the Xpore sites in the comparison of the two METLL3-KO replicates were false positives.

CHEUI accurately identifies m5C modifications in cellular mRNA

We next used CHEUI to discover m5C in cell-derived RNA. To accomplish this, we generated a knock-out (KO) of the RNA methyltransferase NSUN2 (NSUN2-KO), which modifies cytosines in various mRNAs and tRNAs [4,35] using CRISPR-cas9 gene editing technology in HeLa cells. The KO was confirmed by western blot (**Supp. Fig. 8a**) and whole genome sequencing of the WT and KO cells (**Supp. Fig. 8b**). DRS on 3 biological replicates from the WT and NSUN2-KO HeLa cells **(Supp. Table S1)** yielded 2,699,213 transcriptomic sites with a coverage of more than 20 reads in all three replicates for WT and 1,636,369 for NSUN2-KO. Testing these sites with CHEUI-solo (Model 2), prior to any significance filtering, we observed a high correlation in the predicted stoichiometry and modification probability between the replicates (**Fig. 3e**). Analyzing the three replicates together, we considered significant those transcriptomic sites predicted with probability > 0.9999, which we estimated would result in an FDR of approximately 0 using an empirical permutation test (see Methods). We obtained 3,167 significant transcriptomic sites in WT (**Supp. Table S6**) and 1,841 in NSUN2-KO (**Supp. Table S7**). Similar to what we observed for m6A, the prediction of two or more adjacent m5C sites was rare, and most of the predictions were individual m5C sites (**Supp. Fig. 10a**).

We next tested whether CHEUI-solo assigned a high probability for m5C to transcriptomic sites previously detected in HeLa using bisulfite RNA sequencing (bsRNA-seq), using data from three independent studies [4,35,36]. CHEUI-solo probabilities on this union set of 372 sites were significantly higher in WT compared with NSUN2-KO (**Fig 3f**). We further performed a permutation analysis to compare the probability of these sites against the background distribution of probabilities in the same samples (see Methods). Confirming its performance as measured with bsRNA-seq, CHEUI-solo returned higher probability modification values in the WT samples than expected by chance (p-value = 0.001) (**Supp. Fig. 9a**). In contrast, the enrichment of high CHUEI-solo probabilities over the background distribution disappeared in the NSUN2-KO (p-value = 0.025) (**Supp. Fig 9b)**. Furthermore, looking at individual nucleotides with CHEUI-solo Model 1, we observed a reduction in the proportion of m5C over the total cytosine occurrences in NSUN2-KO compared with WT (p-value = 1.2e-35) (**Supp. Fig. 10b**). On the other hand, the profile of significant m5C sites along mRNAs did not change between the WT and NSUN2-KO (**Supp. Fig. 10c**). This is consistent with previous reports showing that a fraction of m5C sites in mRNA are NSUN2-independent [4,35] and potentially regulated by other m5C writers, such as NSUN6 [37,38].

To investigate NSUN2 dependent and independent sites, we used CHEUI-diff to select differentially modified sites between WT and NSUN2-KO (**Supp. Table S8**). This yielded 186 potential NSUN2-dependent unique

genomic sites, 18 of which were previously identified by bsRNA-seq. Furthermore, these 186 sites showed similarity to the previously described sequence motif for NSUN2-dependent sites: 5′-m5CNGGG-3′ [35] (**Fig. 3g**). To identify potential NSUN2-independent sites, we selected sites that were significant according to CHEUI-solo in WT but did not change significantly according to CHEUI-diff and had a stoichiometry difference of less than 0.05. This resulted in 1,250 sites, which showed similarity with the C-rich motif previously described for NSUN2-independent sites [35] (**Fig. 3g**). To further assess the validity of these predictions, we investigated the likelihood of secondary structure formation at the respective sites. Consistent with previous studies [4,35] canonical base-pair probabilities were higher in NSUN2-dependent sites compared to NSUN2-independent sites (**Figs. 3h and 3i**). Also consistent with the previous results [35], the potential base-pairing arrangement suggested a higher occurrence of stem-loops at around 5 nt downstream of the m5C site in NSUN2-dependent sites (**Supp. Fig. 11**). Interestingly, when we used an alternative definition for NSUN2-independent sites to be those that are only significant in HeLa NSUN2-KO (1,841 transcriptomic sites), results were identical in respect of sequence motifs and structural properties **(Supp. Figs. 12a and 12b)**. Further validating CHEUI results, NSUN2-dependent sites identified previously by bsRNA-seq [35] showed significantly higher stoichiometry difference between WT and NSUN2-KO compared with all other m5C sites (**Supp. Fig. 13)**. These results indicate that CHEUI-solo and CHEUI-diff can confidently identify previously discovered m5C sites and discover new ones.

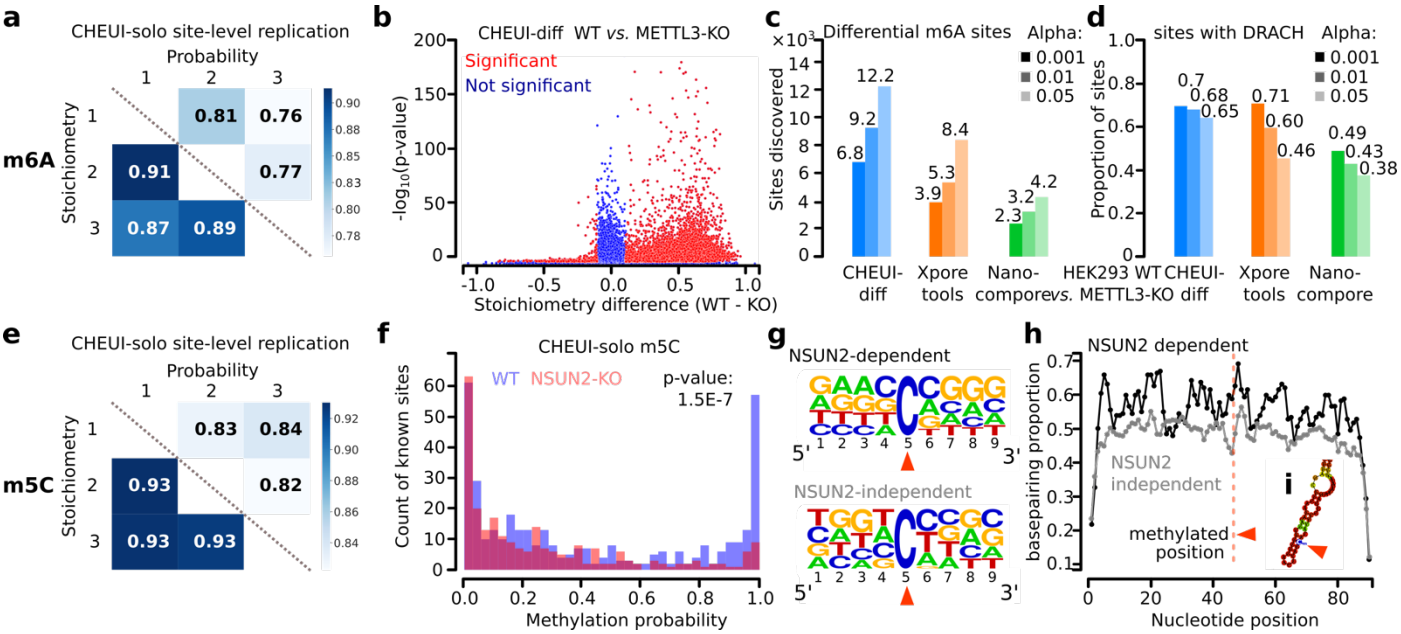

Figure 3. Detection of m6A and m5C in cell lines using CHEUI. (a) Pearson correlation values among HEK293 WT replicates for CHEUI-solo m6A stoichiometry predictions (lower diagonal) and m6A per-site probabilities (upper diagonal) for the 562,628 transcriptomic sites that had a coverage of more than 20 reads in all three replicates. **(b)** Results from CHEUI-diff comparing 3 WT and 3 METTL3-KO replicates. Every dot represents a transcriptomic site, with its significance given as $-\log_{10}$(p-value) (y-axis) and the difference in the stoichiometry between WT and METTL3-KO (x-axis). **(c)** Number of differentially modified m6A sites detected by each tool between HEK293 WT and METTL3-KO using three different levels of significance, Alpha = 0.05, 0.01 and 0.001; i.e., selecting cases with adjusted p-value ≤ Alpha. **(d)** Proportion of differential significant m6A sites containing a DRACH motif for each method at three levels of significance. **(e)** Pearson correlation values among HeLa WT replicates for CHEUI-solo m5C stoichiometry predictions (lower diagonal) and m5C per-site modification probabilities (upper diagonal) for all the 497,439 tested transcriptomic sites with coverage of >20 reads in all three replicates. **(f)** Distribution of CHEUI-solo Model 2 probabilities for HeLa WT and NSUN2-KO sites also

previously identified using bisulfite RNA sequencing. **(g)** Sequence motifs for 32 NSUN2-dependent sites (upper panel) and for the 1,000 most significant NSUN2-independent sites (lower panel) predicted by CHEUI-solo. **(h)** Proportion of base-pairing positions along 90 nucleotides centered at m5C sites predicted by CHEUI-solo. The vertical dashed red line indicates the m5C position. **(i)** Example of RNA secondary structure containing an m5C site in a stem-loop.

Impact of other modifications on the prediction of m6A and m5C

To test if other modifications could impact the accuracy of m6A or m5C in individual reads, we tested CHEUI on read signals from IVTs containing other modifications not used for training, namely, 1-methyladenosine (m1A), hydroxymethylcytidine (hm5C), 5-formylcytidine (f5C), 7-methylguanosine (m7G), pseudouridine (Y), and inosine (I) [23]. All read signals were processed for each 9-mer centered at A or C as before, with the modification either at the same central base (m1A and m6A for A, and m5C, 5fC, and hm5C for C) or in the neighboring bases in the 9-mer (Y, m7G, I, m1A, m6A for C; or Y, m7G, I, m5C, 5fC, hm5C for A) (**Figs. 4a)**. As a general trend, the proportion of signals containing other modifications predicted as positives by CHEUI recapitulated the results for signals without any additional modifications (**Figs. 4b**). This was the case for all modifications, except for predictions by the m6A model in signals containing m1A, a chemical isomer of m6A, which followed a similar trend as m6A (**Fig. 4b, upper panel**).

To investigate whether m1A misclassification was specific to CHEUI, or a phenomenon shared across other methods, we used Xpore and Nanocompore to test the discrimination of m6A and m1A without any *a priori* assumption about the modification type. We used 81 9-mers centered at A and made all possible pairwise comparisons among three sets of reads: one with no modifications, one with all read signals having m1A, and one with all read signals having m6A, with a median coverage of 62 reads per site. When comparing m6A or m1A against unmodified signals, Xpore identified significant differences for 11 and 16 sites, Nanocompore detected 5 and 3 sites, and CHEUI m6A model predicted 19 sites in both cases, consistent with CHEUI's higher recall shown above (**Fig. 4c**). In the comparison of m6A against m1A read signals, Xpore found a significant difference in only two of the sites, whereas Nanocompore found none (**Fig. 4c**). These results suggest that the DRS signals for these two isomers may be indistinguishable with current statistical models and/or pore chemistry (**Supp. Fig 14**). To fully address the m6A and m1A DRS signal similarity, we retrained CHEUI-solo m6A model using m1A signals as negatives and m6A signals as positives. Although this new model achieved accuracy comparable to the original one in the separation of m6A from unmodified signals (**Supp. Fig. 15a**), it showed a trade-off between accurately detecting m6A and correctly separating m6A from m1A (**Supp. Fig. 15b**), further indicating current limitations to separate isomeric RNA modifications using the nanopore signals.

As CHEUI can robustly detect m6A and m5C from the same sample, we further assessed how the presence of one modification may impact the detection of the other at short distances. We analyzed the detection of each modification in individual reads at 9-mers with or without the other modification nearby using reads from the IVT test 2 dataset. CHEUI m5C model showed an increase in the proportion of false positives from 0.08 to a maximum of 0.14 when m6A was at a relative distance of 1-4 nt from C (**Fig. 4d**). In contrast, the proportion of CHEUI m6A model false positives did not increase with a nearby m5C compared to the background level (**Supp. Fig. 16**).

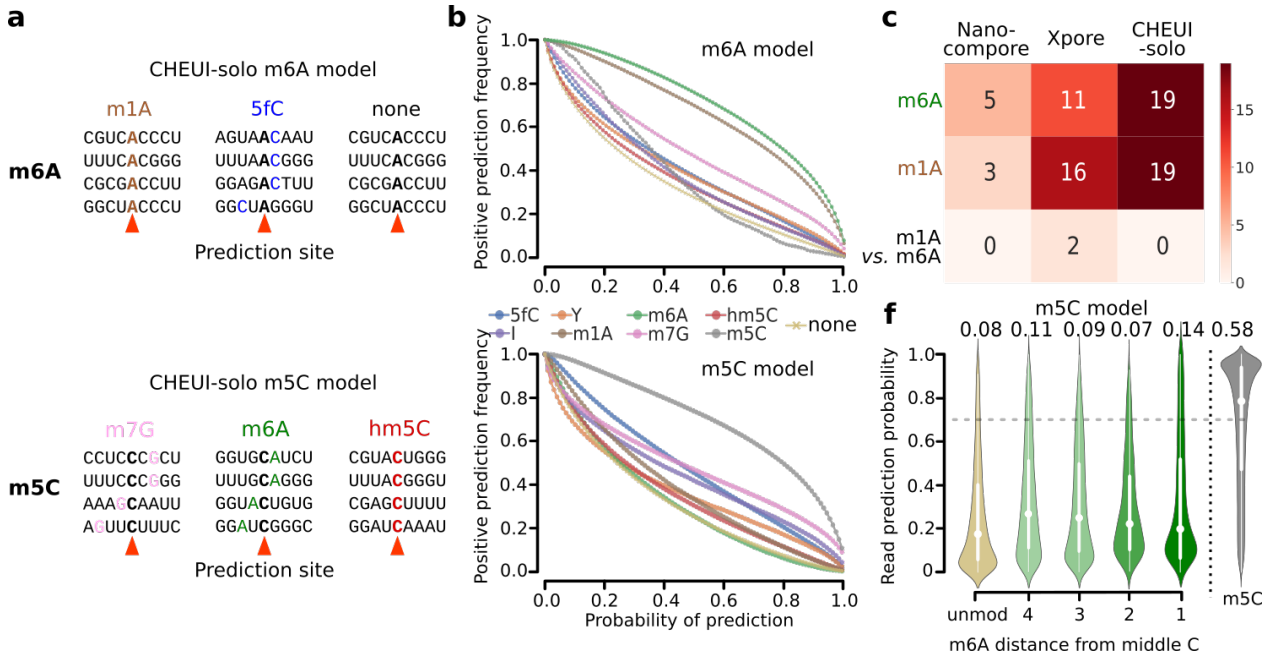

Figure 4. Impact of other RNA modifications on the detection accuracy of m6A and m5. CHEUI-solo's calls were tested in individual reads (Model 1) for m6A (upper panel) and m5C (lower panel) using signals from IVTs containing other modifications. Coverage per site ranged between 20 and 324 reads, with median coverage of 62 reads. **(b)** The number of read signals identified as m6A (upper panel) and m5C (lower panel) modifications by CHEUI-solo Model 1 (y-axis) at different values of the probability cutoff (x-axis). **(c)** The number of significant sites identified by each tool in each of the conditions (y-axis). The 'm6A' and 'm1A' row show the number of sites with 100% stoichiometry predicted as m6A by each method. For Nanocompore and Xpore these were calculated by comparing each sample against the unmodified sample. The 'm6A vs m1A' row shows the number of sites with a significant difference between the two modified samples. For CHEUI, this was calculated as the number of sites that were detected only in one of the samples. **(d)** CHEUI's detection probability of m5C at individual read level (y-axis) using IVT test 2 read signals at 9-mers centered at C with various configurations: 9-mers with no m5C (None), 9-mers with m6A present 1, 2, 3, or 4 nucleotides from the central C, and 9-mers with a modified middle C (m5C). The proportion of read signals identified with probability > 0.7 is indicated above each distribution.

Coordinated m6A and m5C occurrence RNA transcripts

We next exploited CHEUI's unique capability to concurrently identify m6A and m5C to investigate the co-occurrence of modifications in RNA molecules. Using WT HEK293 cell line's data, we calculated whether individual reads covering two predicted modified transcriptomic sites presented the same modification state (i.e., m6A-m5C, m6A-m6A, m5C-m5C) more often or at a similar rate in comparison with random pairs of sites from different transcripts. We observed a modification co-occurrence (proportion of molecules with both sites having the same modification status) was higher than expected by chance for m6A and m5C at distances of more than 5 nucleotides (**Fig. 5a**). At the distance of 5nt, A downstream of C (i.e., CNNNNA) showed a significantly higher co-occurrence compared to the A upstream of C (i.e., ANNNNC) (U-test p-value=0.03), while the latter was close to the random co-occurrence values. At distance 4 or less the co-occurrence was high in both configurations, but our analyses above suggest that co-occurrences at such short distances may result from the impact of an existing modification on the performance of the other modification's model. The co-occurrence of m6A-m6A or m5C-m5C was also higher than expected at short distances (1-4nt) but returned to co-occurrence values close to random

from 5-15 nucleotides (**Figs. 5b** and **5c**). Furthermore, discarding m6A and m5C sites at distances <5 nucleotides from each other, we also observed an enrichment of transcripts harboring both modifications, relative to the total number of m6A and m5C transcriptomic sites, both in HEK293 (**Fig. 5d**) and HeLa (**Supp. Fig. 17**). To examine how CHEUI can resolve m6A and m5C co-occurrences in RNA molecules, we visualized a region of 50 nt from DEAD-Box Helicase 23 transcript ENST00000308025, which encodes DDX23, a protein involved in pre-mRNA splicing and R-loop suppression (**Fig. 5e**). ENST00000308025 presents high-confident m6A and m5C sites (CHEUI-solo model 1 probability >0.7) separated 14 nt apart, with 95% of the individual molecules containing both modifications (**Fig 5e**). Interestingly, this case presents adjacent modifications of the same type, which are predicted to be rare (**Supp. Figs. 6a** and **10a**).

An intriguing question is the possibility of a coordinated m6A and m5C occurrence in a physiological context, where RNA modifications play an important role. We decided to study m6A and m5C during brain development, where m6A has been reported to be relevant [39]. We collected cortex tissue from wild-type mice at three different embryonic stages E12, E15, and E18, and performed DRS of 3′ poly(A)$^+$ RNA (**Supp. Fig. 18**) (**Supp. Table S1**). We tested a total of 1.4M to 2.2M transcriptomic 'A' sites and 1.2M to 2M transcriptomic 'C' sites. Using the probability cutoff of > 0.9999, we obtained 2,876 to 6,040 m6A sites and 1,390 to 2,180 m5C sites (**Supp. Tables S2** and **S9**). The incidence of significant transcriptomic sites identified with CHEUI followed profiles along mRNAs similar to those observed for the cell lines (**Supp. Fig. 19**). We found that in all three conditions, m6A and m5C sites at distances 5nt or more co-occurred in transcripts significantly more often than expected by the random incidence of the two modifications (**Supp. Fig. 20**). The pairs of methylated sites (m6A-m5C, in any order) in each condition showed a wide variation in co-occurrence at the level of individual reads, but the global co-occurrence values were significantly higher than expected by chance at stages E12 and E18 (**Fig. 5f**). Furthermore, co-occurrence values of m6A-m5C sites showed a high correlation among the three embryonic stages, suggesting that the co-occurrence of modifications is transcript-specific and conserved across stages **(Fig. 5g)**. The conservation of the co-occurrence was apparent even for the sites of low stoichiometry across developmental points, which can be exemplified by a 35nt region from the transcript ENSMUST00000014438 (gene Ndufa2), where an m6A and m5C sites were found 13nt apart (**Supp. Fig. 21**). While the modification frequency in these sites was moderate at ~30%, the co-occurrence of modifications for m6A-m5C in molecules were 0.961, 0.957, and 0.913 for E12, E15 and E18, respectively, showing high conservation across conditions (**Supp. Fig. 21**).

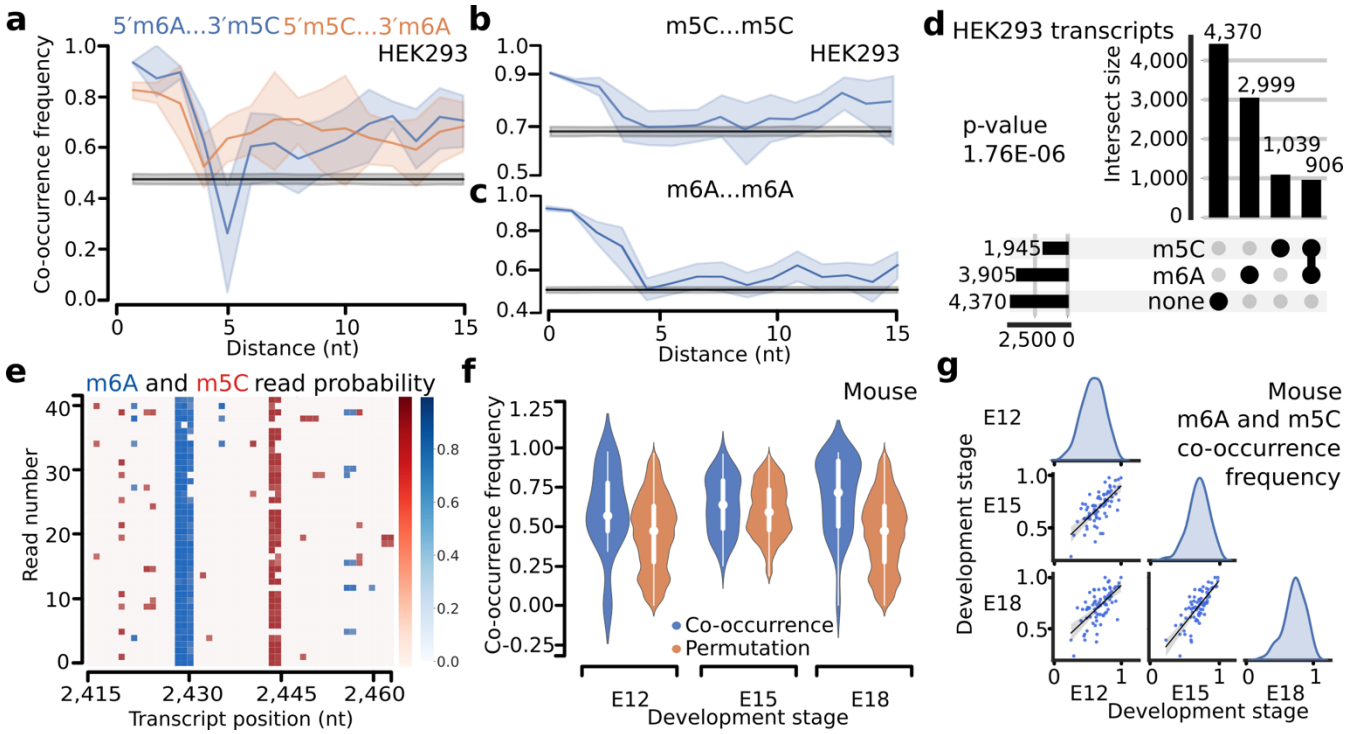

Figure 5. Coordinated occurrence of m6A and m5C in RNA *in vivo*. (a) Co-occurrence (y-axis) of m6A and m5C modifications at the read level at various relative distances (x-axis). Co-occurrence was measured as the proportion of reads with the same modification status. Pairs of sites with A upstream of C are depicted in blue and pairs of sites with A downstream of C are shown in red. The shading indicates the standard deviation across occurrences at each distance. The black line and grey band indicate the mean and standard deviation from the mean of the co-occurrence values for pairs of modifications across different transcripts. Distances are measured as the difference between positions of the two modified nucleotides, e.g., 5′-m6ANNNNm5C-3′ are at the relative distance of 5 nt. **(b)** Same as (a) but for the co-occurrence of m6A-m6A. **(c)** Same as (a) but for the co-occurrence of m5C-m5C. **(d)** Number of human protein-coding transcripts containing m6A and m5C sites, only one of the modifications, or none, in HEK293. Only cases with m6A and m5C at a distance of 5 or larger were considered. The p-value corresponds to a Fisher's test for an increased observed co-occurrence. **(e)** Region of the transcript ENST00000308025 (gene DDX23) showing m6A and m5C modifications occurring in the same RNA reads. Reads are represented along the y-axis and the position along the transcript is represented on the x-axis. The blue scale shows CHEUI's detection probability at the read level for m6A and the red scale for m5C. **(f)** Distribution of m6A-m5C co-occurrence values and co-occurrence permutations (y-axis) from mouse embryonic cortex at three developmental stages E12, E15, and E18 (x-axis). U-tests comparing each distribution of values with its permutations were performed and results are shown on the top of the distributions: ** (<=0.01), *** (<=0.001), and ns (not significant). **(g)** Correlations between the co-occurrence values at the individual-read level for pairs of m6A and m5C sites in a pairwise comparison between mouse frontal cortex developmental stages E12, E15, and E18. Pearson correlation between E12 and E15 was r=0.68 (p-value 7.8E-12), between E12 and E18 was r=0.64 (p-value 4.2E-10), and between E15 and E18 was r=0.78 (p-value 1.3E-16). Density distributions of the co-occurrence values are additionally shown as shaded area plots.

DISCUSSION

With CHEUI we make possible, for the first time, the transcriptome-wide identification of m6A and m5C from the same sample, both in individual molecules as well as in transcriptome reference sites together with their stoichiometry quantification. CHEUI also presents several novelties in design and capabilities with respect to the previous methods that work with nanopore signals. CHEUI abstracts the nanopore signal values into a representation that facilitates the construction of a flexible and generalizable training model agnostic of the sample, pore, detector, and chemistry types. CHEUI identifies modified nucleotides in individual reads and in annotated transcripts, i.e., transcriptomic sites, in a single condition without requiring a KO/KD or control sample, thereby escaping the *sample comparison* paradigm used by most of the other tools. CHEUI also predicts modifications in any sequence context, circumventing the constraints of the contextual methods or those based on indirect evidence. Furthermore, CHEUI detects differential modifications between any two samples.

An in-depth benchmarking across different tools using a ground-truth dataset demonstrated that CHEUI provides a substantial advantage in sensitivity and precision, and accurately calculates the modification stoichiometry. Accuracy assessment of stoichiometry is particularly challenging as it requires complete knowledge of the modification status of reads. To resolve this, we used controlled mixtures of read signals built from fully modified and unmodified *in vitro* transcript sequences selected to reflect variable coverage and realistic stoichiometry values, and without constraining the sequence context or using knowledge from the previous performance. Our analyses suggest that read mixtures provide powerful and effective means to benchmark the accuracy of RNA modification detection methods.

We showed that knocking out a single methylation enzyme in cells to detect modifications may only be effective in certain cases, as modifications can be deposited on mRNA by multiple enzymes, as is the case for m5C. Furthermore, some of the modifications in KO cells could be induced by compensatory effects, since KO cells may adapt and undergo potential compensatory modifications or even genetic selection. In a KD model, where the modification alteration is more transient, cells would have limited adaptation and selection time, and these effects may be mitigated. Still, the possible involvement of multiple enzymes poses limitations on the KD/KO strategy to identify RNA modifications. CHEUI circumvents this challenge and opens new opportunities for unconstrained RNA modification studies.

One of the biggest challenges and principally unresolved questions in testing RNA modification detection tools, and nanopore signal interpretation technologies in general, is the identification of specific modification signatures without complete knowledge of all the modifications present in the sample. There is growing evidence indicating that *in vivo* mRNA harbors multiple modifications in addition to m6A and m5C, but a comprehensive modification catalog of natural mRNA is still lacking. To address this, we analyzed IVT RNAs harboring other RNA modifications not used for training. CHEUI generally separates m6A and m5C from other modifications. Importantly, CHEUI could separate m5C from hm5C, which presents an advantage over bisulfite sequencing that cannot distinguish between these two modifications. We thus prove that the modeling principles implemented in CHEUI offer sufficient generalization power to tackle samples with unknown RNA modification configurations.

Somewhat unexpectedly, CHEUI as well as the other methods tested could not accurately separate the positional isomers m1A and m6A. Visual inspection of the signals for m6A and m1A in the same k-mer contexts showed that they deviate in the same way from the signals corresponding to unmodified nucleotides. In contrast, m5C and hm5C, which have different chemical groups attached to the same position, could be visually distinguished from each other and from the unmodified nucleotides and were separated by CHEUI. This suggests two hypotheses. The first one is that nanopore signals from isomeric modifications may not be distinguishable. This is supported by our analyses and is consistent with the difficulties encountered by other technologies to separate m1A and m6A

[40]. The second hypothesis is that more sophisticated predictive models may separate these modifications. The inclusion of additional features, such as sequence context or evolutionary conservation, could overcome the observed limitation. Nonetheless, the similarity of nanopore signals for m1A and m6A may not have a major impact on the study of mRNAs. Recent evidence indicated that although m1A sites are abundant in tRNAs and rRNAs [41], they are exceedingly rare, possibly absent, in mRNAs [42] and that many of the reported m1A sites in mRNAs could be due to antibody cross-reactivity [43].

CHEUI capacity to predict two modifications concurrently enabled us to measure the co-occurrence of m6A and m5C sites in transcripts and for the first time, identify their entanglement in individual reads. We used systematic analysis of signals to establish at distances of 5 nucleotides or more, the co-occurrence can be reliably identified. At distances closer than 5 nucleotides, there was a measurable mutual signal interference of the modifications, and their prediction remains a general challenge. Although we demonstrated that CHEUI could correctly identify single modified nucleotides with a low false positive rate, there is a residual contribution to false detection from the signal of nearby modified nucleotides. We foresee that this problem could be addressed by incorporating additional predictive features or even with the training of new combinatorial models using training datasets with defined modification co-localizations.

The mechanisms underlying the identified entanglement of modifications in reads and across transcripts remain to be elucidated. Entangled modifications at a single-molecule level may represent the footprint of the 'history' of the RNA molecule, which acquired the modifications by passing through certain processing steps or points of cellular response [44]. Such footprints may contain entangled modifications of various types that are characteristic of a subpopulation of the cell's RNA. Another possibility is that the coordination of modifications is due to the crosstalk between RNA modification enzymes, whereby the binding of RNA by readers or writers for one modification may drive the deposition or removal of the other. A more evolutionary-inspired possibility is the correction of function, whereby a modification is introduced to enhance or compensate for the functional effects of a pre-existing modification.

Finally, CHEUI addresses one of the main challenges associated with the prediction of RNA modifications, the limited availability of suitable training datasets that recapitulate the naturally occurring RNA modifications. Positions of RNA modifications are mostly unknown and sparse; hence specific datasets with abundant observations must be specifically generated to train predictive models. We have shown that *in vitro* transcribed RNAs (IVTs) with modified nucleotides can be exploited to train the identification of specific RNA modifications in individual reads. CHEUI's processing of the signals with convolutional neural networks provides accurate detection of the modifications and generalizes to unseen sequence contexts. IVT datasets with other nucleotide modifications can be straightforwardly produced and are more effective than cellular models with engineered deletions of the modifying enzymes. CHEUI thus provides a convenient and competitive strategy to enable the detection of other RNA modifications, opening new opportunities in epitranscriptomics, synthetic biology, and RNA engineering.

Software availability

CHEUI is freely available from https://github.com/comprna/CHEUI-public under an Academic Public License

Txannotate: https://github.com/comprna/txannotate

Nanocompore: https://github.com/tleonardi/nanocompore

Xpore: https://github.com/GoekeLab/xpore

Epinano: https://github.com/enovoa/EpiNano

Tombo: https://github.com/nanoporetech/tombo

NanoRMS: https://github.com/novoalab/nanoRMS

Keras: https://github.com/keras-team/keras

Tensorflow: https://github.com/tensorflow

Minimap2: https://github.com/lh3/minimap2

Nanopolish: https://github.com/jts/nanopolish

RNAfold: http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi

Data availability

The synthetic sequence templates from [24] were obtained from the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE124309. The nanopore read signals for the in-vitro transcribed (IVT) RNAs obtained from these synthetic sequence templates with m6A, m5C, or no modifications, were obtained from NCBI Sequence Read Archive (SRA) under accessions PRJNA511582 and PRJNA563591. Nanopore data for the synthetic transcripts from [23] was obtained from The Sequence Read Archive (SRA) accession SRP166020. Nanopore data for HEK293 WT and METTL3-KO samples from [19] was obtained from the European Nucleotide Archive (ENA) under accession PRJEB40872. Data from the m6ACE-seq experiments from [33] was obtained from the NCBI Gene Expression Omnibus (GEO) under accession number GSE124509. Nanopore data for HeLa WT and HeLa NSUN2 KO and for the embryonic mouse brain tissues produced in this work have been deposited at NCBI GEO under accession number GSE211762.

METHODS

Nanopore signal preprocessing

All IVTs datasets used with CHEUI were pre-processed using the following steps. First, the FAST5 files from a sample were basecalled using Guppy version 4.0.14 and aligned to the corresponding IVT reference transcriptome using minimap2 [45] with options '-ax map-ont -k 5'. Reads were filtered to select the best match for each read using samtools *-F 2324* [46]. Nanopolish (version 0.13.2) [15] *eventalign* was then used to align signals to the reference using the options '--scale-events --signal-index --samples --print-read-names'. Nanopolish *eventalign* outputs each 5-mer in 3′ to 5′ orientation, whereas the 5-mers (output rows) are given in 5′ to 3′ orientation. To process the signals in the right 5′-3′ orientation, we thus flipped the signals per 5-mer before concatenating the signals from overlapping 5-mers. DRS from mouse (E12, E15, E18) and cell lines (WT and METTL3-KO in HEK293 cells, and WT and NSUN2-KO in HeLa) were basecalled using guppy 5.0. Basecalled sequences were mapped to the reference transcriptome using minimap2 (parameters: '-ax map-ont –k14'). Reads were filtered to select the best match for each read using samtools *-F 2324.* As before, Nanopolish (version 0.13.2) was used to re-squiggle the nanopore signals to the transcript sequences and the signals flipped to the 5′-3′ orientation. All the (per read) signals for every 5 overlapping consecutive 5-mers, together with the read ID and sequence, were then used to create the input for CHEUI-solo Model 1. The genome and annotation references used were GRCh38 and Gencode v38 for the human data, and GRCm39 and Ensembl v104 for the mouse data.

CHEUI-solo Model 1

Model description

CHEUI-solo Model 1 is a convolutional neural network (CNN) modified from the Jasper model [47]. The CNN architecture was implemented using Keras [48] and Tensorflow [49]. CHEUI-solo Model 1 uses as input the signals from each individual read corresponding to 5 consecutive 5-mers, where the middle 5-mer is centered on adenosine (A) for the m6A model or cytosine (C) for the m5C model, i.e., NNNN(A|C)NNNN. To fix the length of the input, the signals associated with every 5-mer were summarized into 20 signal values. If a 5-mer contained more than 20 values, the values were divided into 20 equal subsets, and the median value of each subset was used. If the event had fewer than 20 values, the median was appended to these values until it reached 20 values. As a result, each 9-mer and read signal was mapped to a vector of 100 values.

CHEUI-solo Model 1 also uses as input the distance between the observed and the expected signal for every input. The expected signal is built using the k-mer model from Nanopolish [15], which describes the signal value for each 5-mer in the absence of modifications. For each of the 5 overlapping 5-mers in the observed signals, each expected value was repeated 20 times to obtain a vector of expected values of length 100. Then, a vector of length 100 with the absolute distances between the components of the expected and the observed signals is calculated. The vectors of observed signals and absolute distances are used as input for CHEUI-solo Model 1. Of note, CHEUI-solo Model 1 does not use the actual k-mer (k=9) sequence, only the vector of observed signals and the vector of distances.

Training and testing of CHEUI-solo Model 1

CHEUI-solo Model 1 was trained using read signals generated from *in vitro* transcript (IVT) data [23,24] to produce one model for each modification, m6A or m5C. For the m6A model, the positive training set contained m6A (or m5C) in place of the canonical nucleotide, i.e., replacing every A with m6A (or every C with m5C) [24]. For both models, the negative sets were made from read signals from the same IVTs but with no modifications. In both cases, we constructed non-overlapping datasets for training (IVT train 1), validation (IVT validation 1), and testing (IVT test 1, IVT test 2) (**Supp. Table S10**). IVT train 1 was composed of 9-mers with any number of A's (or C's) in the modified and unmodified sequences. IVT validation 1, used for parameter optimization, was composed of 9-mers containing only one A (or C) at the center of the 9-mer. IVT test 1, which was used to test sensor generalization, was also composed of 9-mers with only one A (or C) at the center. IVT train 1, IVT validation 1, and IVT test 1 datasets we built using publicly available reads [24]. Finally, IVT test 2, used to test k-mer generalization, was built from independent IVT experiments [23]. IVT test 2 was also composed of 9-mers with only one A (or C) at the center of the 9-mer. Importantly, the training and testing was performed on individual read signals.

Binary cross-entropy was used as the objective function, AMSGrad was used as the optimizer, and the Nvidia Tesla V100 was used to accelerate computing. Training was performed for 10 epochs and for every 200,000 read signals the accuracy, precision, recall, and binary cross-entropy loss were calculated on the IVT validation 1 set along with the parameters of the model at that stage. After 10 epochs, there was no improvement on the validation accuracy, so the training was terminated. Accuracy was defined as the proportion of correct cases, i.e. (TN+TP)/(TN+TP+FN+FP); precision was calculated as the proportion of predicted modifications that were correct, i.e. TP/(TP+FP) and recall as the proportion of actual modifications that were correctly predicted, i.e., TP/(TP+FN); where TP = true positive, FP = false positive, TN = true negative, FN = false negative. Binary cross-entropy was defined as

$$H_p(q) = 1/N \cdot \sum_{i=1}^{N} y_i \cdot log2(p(y_i)) + (1 - y_i) \cdot log2(1 - p(y_i)),$$

where: y_i= 1 for a modified base in a specific position of a read and 0 otherwise, and $p(y_i)$is the posterior probability from the Model 1.

CHEUI-solo Model 2

Model description

CHEUI-solo Model 2 is a binary classifier implemented as a CNN like for Model 1. CHEUI-solo Model 2 takes as input the distribution of probabilities generated by Model 1 for all read signals at a given transcriptomic site, i.e., a position in a reference transcript, and predicts the stoichiometry and probability of that site being methylated (m6A or m5C). Model 2 assumes that the distribution of the individual-read probabilities at a given transcriptomic site originates from two classes, one with a subset or all reads having high Model 1 probabilities (modified site), and a second one with low Model 1 probabilities (unmodified site).

Model training and testing

CHEUI-solo Model 2 was trained using controlled mixtures of modified and unmodified reads not used previously for training, validation, or testing of CHEUI-solo Model 1. These controlled mixtures were built to comprise a wide range of values for coverage and stoichiometry, and with a high proportion of low coverage and low stoichiometry sites, to mimic what was previously observed in transcriptomes [4,35,50]. The new read signals were processed as described before and used to make predictions with CHEUI-solo Model 1. The training set for Model 2 consisted of mixtures of modified and unmodified reads from IVTs [24] with their corresponding Model 1 probabilities. To model the low stoichiometry and coverage values, the training sites were built as follows: 1) First, a site was chosen to be modified or unmodified with 50% probability; 2) if unmodified, a coverage was chosen randomly between 0 and 100, using a linear decay, i.e., the higher the coverage, the less likely it was to be selected, and the per-read probabilities were assigned at random from the pool of unmodified signals; 3) if, on the contrary, the site was selected to be modified, the coverage and stoichiometry of the site were chosen using the same linear decay as before, with high coverage and stoichiometry values less likely to be chosen. The linear decay was implemented using the *random.choices* function from the general python distribution using the weights *(10 - coverage) x 0.01 + 0.9* as argument. Weights indicate the relative likelihood of each element on the list to be chosen, with each incremental unit of coverage or stoichiometry corresponding to a decrease in their weight by one unit. Using this procedure, we generated approximately 1.5M synthetic sites per modification with variable coverage and stoichiometry. These sites were randomly split into training and testing in a 9:1 proportion.

Comparison with other tools

Tools selected for comparison

We chose tools available for each specific benchmarking comparison. We used Epinano [24], which implements a linear regression with two samples, one depleted of modifications to detect outliers, i.e., observations with large residuals, to identify modifications. We used *EpiNano-Error,* which combines all types of read errors (mismatches, insertions, and deletions) in the pairwise mode. We also used nanoRMS [25], which does not predict modified sites but uses predictions from another method to calculate the stoichiometry using a sample comparison approach. Specifically, nanoRMS uses the signals processed by Tombo or Nanopolish and implements a supervised *k*-NN method based on the sample labels, or an unsupervised method based on *k*-means with *k*=2, to separate modified and unmodified signals. For nanoRMS, the stoichiometry was calculated from the proportion of reads from the

WT sample in the modified cluster, divided by the total number of WT reads. We also tested Nanocompore [18], which uses the assignment of raw signals to a transcriptome reference with Nanopolish and uses the mean current value and mean dwell time of all the signals per 5-mer and compare the distributions for all read signals aligning on the same site between two conditions. Nanocompore fits a Gaussian mixture model with two components to the data and performs a statistical test to determine whether each cluster is significantly associated with a sample. We also tested Xpore [19], which operates similarly to Nanocompore, using the assignment of raw signals to the transcript reference with Nanopolish and comparing the mean current values between two or more conditions for each transcriptomic site. Xpore uses information from unmodified k-mers as a prior for Gaussian distributions and variational Bayesian inference to infer the mean and variance of each distribution. After fitting the data into clusters, Xpore labels clusters with values closer to the expected unmodified signals as unmodified and then performs a statistical test on the differential modification rates between samples and assigns a p-value per site. We also tested Tombo in *sample comparison* mode, which performs a statistical test comparing the signal values between two conditions; and Tombo in *alternative mode,* which predicts a proportion of m5C modification per transcriptomic (not individual read) site, although it does not provide a score or probability.

IVT controlled mixtures for benchmarking

To create a controlled and independent dataset to benchmark the accuracy in the prediction of stoichiometry and transcript-site modification, we used the reads from [23] not used in the previous tests to generate mock 'WT' and 'KO' samples. The mock 'WT' sample was generated by randomly sampling reads from the modified and unmodified sets to create multiple stoichiometry mixtures with 20, 40, 60, 80, and 100 percent. The mock 'KO' sample was created by randomly sampling reads from the unmodified pool of reads. We ran Epinano, Nanocompore, Xpore, and CHEUI, using default parameters to predict RNA modifications. Epinano, Nanocompore, and Xpore were run using the generated WT and KO mock samples. CHEUI was run using only the generated WT sample as the KO was not necessary. Predicted sites were considered at three levels of significance or alpha values, i.e., predicted sites were considered significant if after correcting for multiple testing, the adjusted p-values were ≤ alpha, where alpha = 0.05, 0.01, 0.001.

Transcript-site predictions, i.e., the methylation state of a position in the reference sequence, in the IVT-based mixtures were classified as positive if they had a probability > 0.99 from CHEUI-solo Model 2, and negative otherwise. Nanocompore, Xpore, Epinano, and CHEUI were run using thresholds recommended by the documentation for each tool. For Xpore (https://github.com/GoekeLab/xpore), sites containing a k-mer (k=9) centered in adenosine, in the evaluation of m6A, or a cytosine, in the evaluation of m5C, that had a predicted p-value lower than 0.05 were considered significant. For Nanocompore, the same selection of k-mers centered in adenosine or cytosines was done, and sites with a p-value lower than 0.05 were selected as positives. For Epinano, we used Guppy version 3.0.3 and *EpiNano-Error* with the combined errors *Epinano_sumErr* method to detect modifications, as recommended in the Epinano documentation. We then used the linear regression model and 'unm' or 'mod' from the 'linear model residuals z score prediction' column to classify sites as unmodified or modified, respectively.

To estimate the false positive rate for Epinano, Nanocompore, and Xpore we evaluated the number of sites each tool predicted as modified when comparing two sets of reads with no modifications. For CHEUI, we used only one of those datasets with no modifications. We evaluated all sites with A or C, regardless of whether they had other As or Cs nearby in the same k-mer (k=0) sequence context. In contrast, to determine the true positive rate and stoichiometry, we only evaluated k-mers (k=9) containing one centered m6A and no additional A's, or one centered m5C and no additional C's to avoid the influence of having 2 or more modified nucleotides affecting the tested site, since the IVTs were built with all nucleotides of one type either modified or not modified.

Stoichiometry benchmarking

Stoichiometries were calculated in the following way. CHEUI-solo calculates the stoichiometry as the proportion of modified reads in the 'WT' sample. For the analyses presented, we used a (CHEUI-solo Model 1) probability higher than 0.7 for modified individual read sites, lower than 0.3 for unmodified ones, and rejecting reads with probability values in the range [0.3, 0.7]. Stoichiometry was only calculated in sites predicted as positive by CHEUI, i.e., with a probability > 0.99 from CHEUI-solo Model 2. For Xpore, we used the values of the column 'mod_rate_WT-rep1', which we interpreted as the modification rate of the mock 'WT' sample. In the case of Nanocompore, we used the column 'cluster_counts' that contains the number of WT and KO reads that belong to the two clusters, one modified and the other unmodified. Stoichiometry was then calculated as the percentage of modified reads in the 'WT' sample, i.e., we divided the number of WT reads in the modified cluster by the total number of WT reads. We also included nanoRMS with *k*-NN and *k*-means for the stoichiometry comparison. In this case, since nanoRMS only predicts the stoichiometry on sites predicted by another method and since Epinano predicted very few sites in our test set, we applied nanoRMS to all tested sites (81 for m6A and 84 for m5C) to obtain a more unbiased assessment. The percentage of modified reads per site was obtained from the nanoRMS output tables, dividing the number of modified reads in the WT by the total number of WT reads. Finally, Tombo assesses every site and gives a fraction of modified reads but does not specify the site as modified or not. As most of the sites had a fraction of modified reads above 0, even for the unmodified sample (75 out of 84 sites), we only considered Tombo for the stoichiometry comparisons.

Testing m6A and m5C accuracy in read signals with other modifications

For this test, we used the Nanopore signals for the IVT transcripts from [23]. Each dataset contained either unmodified signals, or signals for modified nucleotides with m6A, m5C, 1-methyladenosine (m1A), hydroxy-methylcytidine (hm5C), 5-formylcytidine (5fC), 7-methylguanosine (m7G), pseudouridine (Y), and Inosine (I). We considered all 9-mers centered at A or C in the IVT reads containing modifications other than m6A (for A-centered 9-mers) or m5C (for C-centered 9-mers). Thus, the modifications were either at the same central base (m1A and m6A for A, and m5C, 5fC, and hm5C for C) or in neighboring bases (Y, m7G, I, m1A, m6A for C; or Y, m7G, I, m5C, 5fC, hm5C for A). We used CHEUI-solo Model 1 to predict m6A in the middle A or m5C in the middle C for all these read signals to determine the influence of these other modifications on CHEUI's ability to correctly separate A from m6A and C from m5C.

CHEUI-solo for transcriptome-wide analyses

Reads from the three replicates for each condition WT HeLa, NSUN2-KO HeLa, WT HEK293, and METTL3-KO HEK293 were aligned to the Gencode v38 transcriptome (GRCh38) using minimap2 as described above. CHEUI-solo (Model 1 and Model 2) was run on pooled replicates from each condition, except when comparing replicates within the same condition. In each case, CHEUI-solo Model 1 was run on all the reads, whereas CHEUI-solo Model 2 was run only on transcriptomic sites with more than 20 reads coverage. This produced a methylation probability and estimated stoichiometry in all tested transcriptomic sites. To establish a probability cutoff of significance for CHEUI-solo Model 2, we calculated the probability distribution of modified sites expected by chance, without a biological signal. To do so, in each given condition, we shuffled all read signals across all transcriptomic sites, maintaining the same number of transcriptomic sites and the same coverage at each site. We then run CHEUI-solo Model 2 over these sites with the new read signal distributions obtained after shuffling the reads. For each potential probability cutoff, the proportion of candidate transcriptomic sites selected as methylated from the shuffled configuration was then considered as an estimate of the false discovery rate (FDR). We found that a probability cutoff of 0.9999 for CHEUI-solo Model 2 would yield an FDR = 0 for m6A, and an FDR =

0.000384 for m5C. We thus consider modified transcriptomic sites the ones having a model 2 probability equal to or higher than 0.9999 for both modifications.

Comparison with other methods for m6A detection in HEK293 cell lines

Xpore, Nanocompore, and CHEUI-diff were used to call differential RNA modifications on all A sites, using 3 WT and 3 KO replicates for HEK293. CHEUI-diff was run on sites that had >20 reads in both conditions, WT and KO. We used three distinct levels of significance: 0.05, 0.01, and 0.001. For Xpore and CHEUI-diff, FDR correction was performed with Benjamini-Hochberg. Since Nanocompore already provides adjusted p-values, the threshold was applied without FDR correction. To compare the transcriptomic sites predicted as m6A in WT, we selected those sites predicted by each method to have increased stoichiometry in WT. By default, CHEUI-diff does not test sites where the difference in stoichiometry between the two conditions is smaller than 0.1 in absolute value. For Xpore, we used the module *xpore postprocessing* to filter the output. To calculate the potential number of m6A false positives we used each tool to compare two replicates from the same KO condition with the highest number of reads, METTL-KO rep2 and rep3. The KO was used instead of the WT samples to minimize the chances of including variable m6A sites that may occur in WT samples. To compare the Nanopore-based predictions with m6A transcriptomic sites with previous evidence we used the union of m6ACE-seq and miCLIP sites [32,33].

CHEUI on HeLa NSUN2-KO and WT cells

CHEUI-solo (Models 1 and 2) was run pooling together the 3 samples from each condition, WT and NSUN2-KO. Information from previous m5C sites in HeLa was collected from 3 different bisulfite RNA sequencing experiments (bsRNA-seq) [4,35,36] and the union of the three sets was considered for subsequent comparisons. The probabilities from CHEUI-solo Model 2 corresponding to sites with orthogonal evidence were compared between WT and NSUN2-KO using Mann-Whitney U-test.

The permutation analysis to test the enrichment of high probability values in the candidate sites detected by bsRNA-seq was performed in the following way. First, we calculated how many bsRNA-seq candidate sites were tested by CHEUI-solo (total sites) and how many of these were 'high probability sites', defined to have a (Model 2) probability>0.99. Then, we randomly sampled the same number of transcriptomic sites tested with CHEUI-solo Model 2 and counted how many of these were high probability sites. We repeated this procedure 1000 times and calculated an empirical p-value.

Sequence logos were performed using https://weblogo.berkeley.edu/logo.cgi. To study the secondary structure of NSUN2 dependent and independent m5C sites, we used RNAfold 2.4.18 [51] to estimate the base-pair probabilities in the 90 nucleotides around the m5C site (45nt on either side). For each sequence, we calculated the nucleotide positions that had pair-wise interactions with other nucleotides according to RNAfold. Then, we calculated at each position the proportion of nucleotides with interactions with respect to the total number of sequences. These proportions were plotted for WT and NSUN2-KO samples.

CRISPR-Cas9 knockout (KO) of NSUN2

HeLa cell lines and culture

HeLa cells (human cervical cancer) were obtained from ATTC and confirmed *via* short tandem repeat (STR) profiling with CellBank Australia. Cells were grown in DMEM medium (Gibco) supplemented with 10% FBS and 1× antibiotic-antimycotic solution (Sigma) and passaged when 70–90% confluent.

Guide sequence design

Two CRISPR (cr)RNAs were designed, targeting the 5′- (exon 2 crRNA "AGGCUACCCCGAGAUCGUCA") and 3′-proximal (exon 19 crRNA "AAUGAGAGUGCAGCCAGCAC") regions of the gene. Gene sequences from Ensembl (Asia server) were processed *via* CCTop [52] to check for efficacy and predict potential off-target cleavage effects. The two sequences with highest predicted efficacy and minimal off-target effects were selected as crRNA and ordered as Alt-R CRISPR-Cas9 crRNA from Integrated DNA Technologies (IDT).

Ribonuclear protein preparation

2.5 µM of NSUN2 exon 2 crRNA was combined with equimolar amounts of NSUN2 exon 19 crRNA and annealed with 5 µM Alt-R CRISPR-Cas9 trans-activating CRISPR (tracr)RNA, ATTO 550 (IDT) in 10 µl of 1× IDT Duplex Buffer. The ribonuclear protein (RNP) assembly reaction was then performed by combining 0.575 µM of the annealed crRNA:tracrRNA with 30.5 pmol of IDT Alt-R S.p. Cas9 Nuclease V3 in 2.2 µl Neon Transfection System 'R' resuspension buffer (Invitrogen) for 5 minutes at 37 °C; the resultant mixture was kept at room temperature until transfection.

Transfection

Electroporation was conducted using Neon Transfection System (Invitrogen) and following the manufacturer's protocol, with the following modifications: HeLa cells were resuspended in Neon Transfection System 'R' resuspension buffer (Invitrogen) to a concentration of 2.8×10^7/ml. For each electroporation reaction, 2×10^5 cells prepared as above were incubated with 1× v/v RNP at 37 °C for 5 minutes, before being electroporated at 1005 volts, 35 milliseconds with 2 pulses. Two reactions were seeded per well of a 24-well plate. Cells were recovered in complete medium under standard incubation conditions of 37 °C and 5% v/v CO_2 for 24 to 36 hours.

Single cell sorting

Cells were sorted for singlets and ATTO 550 positivity on a FACSAria II Cell Sorter (BD) hosted at the Flow Cytometry Facility of the John Curtin School of Medical Research, the Australian National University. Although all singlets were positive when compared with negative controls, only cells with high intensity ATTO 550 ($>10^{33}$ RFU) were sorted into 96-well plates for subsequent culturing. Cells were maintained in complete media and expanded to 6-well plates for genomic DNA (gDNA) extraction upon reaching 70% confluency.

Amplicon analysis

The gDNA was extracted by incubating cell pellets with 30 µl of in-house rapid lysis buffer (40 µg Proteinase K, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.1% v/v Tween-20) at 56 °C for 1 hour followed by denaturation at 95 °C for 10 minutes. Amplification of NSUN2 gene was conducted with standard protocols under 35 cycles in Mastercycler Nexus (Eppendorf), using Q5 High-Fidelity DNA Polymerase (New England BioLabs) and 5 µl of extracted gDNA. Amplicons were purified with ExoSAP-IT (Applied Biosystems) and sequenced on an AB 3730xl DNA Analyzer, by the ACRF Biomolecular Resource Facility (BRF) from the John Curtin School of Medical Research, Australian National University, following the manufacturer's protocol (Applied Biosystems 2002). Sequencing data was analyzed manually using SnapGene software (from Insightful Science; available at snapgene.com) to confirm alteration of the target loci.

Protein analysis

Cells were grown in DMEM medium (Gibco) supplemented with 10% FBS and 1× antibiotic-antimycotic solution (Sigma) and passaged when 70-100% confluent. Unmodified wild-type (WT) and NSUN2 KO cells were scraped in 200-500 µl of protein extraction buffer (50 mM Tris pH 7.5 at 25 °C, 5 mM EDTA, 150 mM NaCl, 21.5 mM $MgCl_2$, 10% glycerol, 1% v/v Triton X-100, 1× Complete EDTA-free Protease Inhibitor Cocktail (Sigma)) and incubated for 10 minutes on ice, then incubated for 30 minutes at 4°C on a rotator. The mixture was centrifuged at 13,000 g for 10 minutes at 4 °C. The supernatant was transferred to a clean tube, used, or stored at –80 °C. Total protein concentration was then estimated by taking a Qubit measurement via Protein Assay Kit (Thermo Fisher Scientific) following the manufacturer's instructions. 30 µg of total protein was loaded on NuPage 4-12% w/v Bis-Tris Protein Gels (Invitrogen), and proteins were electrophoretically separated using NuPAGE MES SDS Running Buffer under recommended conditions. Separated proteins were transferred onto PVDF membrane using iBlot 2 Transfer Stacks, PVDF, mini (Thermo Fisher Scientific, cat. no. IB24002), following manufacturers' instructions. The membrane was blocked in Odyssey Blocking Buffer (LI-COR, cat. no. 927-40000) and probed with primary antibodies: anti-NSUN2 (1:1000; Proteintech, cat. no. 20854-1-AP), anti-ACTB (1:1000; SantaCruz, cat. no. sc-4778 AF790). The membranes were then incubated with the anti-rabbit-IR-Dye680 secondary antibody (1:10,000; LI-COR, cat. no. 925-68071) and scanned using the Odyssey CLx Imaging System (LI-COR). The KO's effect was assessed by the specific intensity alteration of the fluorescent signal of the respective band with mobility corresponding to that expected of NSUN2.

Extraction of polyadenylated mRNA from HeLa cells

3 WT and 3 NSUN2-KO 80% confluent 10 cm plates were washed twice in ice-cold PBS and scraped in 500 µl of denaturing lysis and binding buffer (100 mM Tris-HCl pH 7.4, 1 % w/v lithium dodecyl sulfate (LiDS), 0.8 M lithium chloride, 40 mM EDTA and 8 mM DTT; LBB). The cell lysate was thoroughly pipetted with 200 µl tip until the sample viscosity was reduced, and pipetting was seamless. 500 µl of oligo$(dT)_{25}$ magnetic beads (New England BioLabs) suspension was used per replicate. The beads were washed with 1 ml of LBB twice, each time collecting the beads on a magnet and completely removing the supernatant. Upon washing, the oligo$(dT)_{25}$ beads were resuspended in the cell lysate and placed in a rotator set for 20 rpm at 25 °C for 5 minutes, followed by the same rotation at 4 °C for 30 minutes. The suspension was briefly spun down at 12,000 g, separated on a magnet, and the supernatant was discarded. The beads were then resuspended with 1 ml of wash buffer (20 mM Tris-HCl pH 7.4, 0.2 % v/v Titron X-100, 0.4 M lithium chloride, 10 mM EDTA and 8 mM DTT; WB) and washed on a rotator set for 20 rpm at 4 °C for 5 minutes, using 3 rounds of washing. The beads were collected on a magnetic rack and the supernatant was discarded. The wash procedure was repeated three times. The elution was carried out stepwise. Washed bead pellet was first resuspended in 50 µl of the elution buffer (25 mM HEPES-KOH, 0.1 mM EDTA; HE). The first suspension was heated at 60 °C for 5 minutes to facilitate the elution, and the eluate was collected upon placing the bead-sample mixture on a magnetic rack, separating the beads, and recovering the clean supernatant. The resultant pellet was next resuspended in another 50 µl of HE buffer, and the process was repeated. The eluates were then combined and subjected to an additional solid phase reversible mobilization (SPRI) bead purification step and stored frozen.

The eluate from oligo(dT) bead extraction was further purified using AMPure XP SPRI beads (Beckman Coulter Life Sciences) according to the manufacturer's recommendations. Briefly, the eluate samples were supplemented with 1.2x volumes of the SPRI bead suspension in its standard (supplied) binding buffer, and the resultant mixture incubated at room temperature for 5 minutes with periodic mixing. The SPRI beads were brought down by a brief 2,000 g spin and separated from the solution on a magnetic rack. The supernatant was removed, and the beads were resuspended in 1 ml of 80 % v/v ethanol, 20 % v/v deionized water mixture and further washed by tube

flipping. The bead and solution separation procedure were repeated. The ethanol washing process was repeated one more time. Any remaining liquid was brought down by a brief spin and removed using a pipette, and the beads were allowed to air-dry while in the magnetic rack for 2 minutes. The purified RNA was then eluted in deionized water and the RNA content was assessed using absorbance readout *via* Nanodrop and fluorescence-based detection *via* Qubit RNA high sensitivity (HS) assay kit (Thermo Fisher Scientific).

RNA DRS Library Preparation for HeLa samples

The library preparation followed the manufacturer's recommendations. 650-800 ng from HeLa cells, were used for each 2× library preparation within every replicate (all recommended volumes doubled-up) with direct RNA sequencing kit (SQK-RNA002) as supplied by Oxford Nanopore Technology. The modifications were that Superscript IV RNA Polymerase (Thermo Fisher Scientific) was used, RNA Control Standard (RCS) was omitted, and RNasin Plus (Promega) was included at 1 U/ µl in all reaction solutions until the SPRI purification step after the reverse transcription reaction. The final adaptor-ligated sample was eluted in 40 µl.

Whole genome sequencing and analysis of HeLa samples

To confirm the NSUN2 gene knockout and characterize its genomic alteration, NSUN2 KO HeLa cells were sequenced against their WT counterparts. Cells were grown in 10 cm plates to 80% confluence collected using standard trypsin-based detachment and pelleted by centrifugation for 3 minutes at 1,000 × g. Genomic DNA was then extracted using Monarch HMW DNA Extraction Kit for Cells & Blood (New England Biolabs) following manufacturer's instructions. The extracted DNA was quality-checked using Femto Pulse 165 kb kit (Agilent) and subjected to additional size-selection with 20 kbp high pass cut-off using BluePippin Size-Selection System (Sage Science). The DNA input was quantified by Qubit dsDNA broad range assay (Thermo Fisher Scientific), and libraries were prepared using the DNA ligation kit SQK-LSK110 (ONT), as per manufacturer's instructions. Samples were sequenced at the Australian National University's Biomolecular Resource Facility on a PromethION X24 instrument using flowcell FLO-PRO002 for about 1 day each. Flow Cell Wash Kit EXP-WSH004 (ONT) was used to flush the flowcell between loading samples of different types. 2 million or more reads were generated per sample, with N50 of 55-60 kbp. Raw ONT sequencing data from WT and NSUN2-KO were basecalled in real-time with high accuracy (HAC) model and Guppy (v5.1.13), generating nanopore FASTQ reads. Only reads with mean quality >7 (passed reads) were used for downstream analysis. The FASTQ files were then aligned to the Telomere-to-Telomere (T2T) human reference genome (T2T-CHM13 v2.0) [53] using minimap2 (v2.22). The resulting aligned reads were used for visualization in Integrative Genomics Viewer (IGV) (v2.13). With the known crRNA sequences used for NSUN2 gene knockout (AGGCUACCCCGAGAUCGUCA in exon 2; AAUGAGAGUGCAGCCAGCAC in exon 19), manual inspection of the alignment in IGV was carried out to confirm the KO in exons 9 and 12 of the NSUN2 gene by identifying the presence of deletion.

Embryonic mouse brain tissue experiments

Brain tissue extraction

Mice were dissected on embryonic day (E) 12, E15 and E18. All procedures were conducted in accordance with the Australian National University Animal Experimentation Ethics Committee (protocol number A2019/46). Pregnant females were cervically dislocated, and embryos extracted in cold sterile PBS. The frontal area of the cortex, i.e., the pallium, was then dissected with micro-knifes under a Zeiss STEMI 508 stereomicroscope and tissue samples were immediately placed in a 1.5 ml microcentrifuge tube (Eppendorf, DNA) containing 300 µl of denaturing lysis and binding buffer (100 mM Tris-HCl pH 7.4 at 25 °C, 1 % w/v lithium dodecyl sulfate (LDS),

0.8 M lithium chloride, 40 mM EDTA and 8 mM DTT; LBB). Samples were immediately agitated by vigorous pipetting until almost complete tissue dissolution, flash-frozen on dry ice and stored at – 80 °C until sequencing.

Polyadenylated mRNA extraction from brain samples

150 mg of cortex tissue was lysed immediately upon extraction. The tissue/LBB mixture was thoroughly pipetted with 200 µl tip until the sample viscosity was reduced, and pipetting was seamless. 500 µl of oligo$(dT)_{25}$ magnetic beads (New England BioLabs) suspension was used per replicate. The beads were washed with 1 ml of LBB twice, each time collecting the beads on a magnet and completely removing the supernatant. Upon washing, the oligo$(dT)_{25}$ beads were resuspended in the tissue/LBB mixture and placed in a rotator set for 20 rpm at 25 °C for 5 minutes, followed by the same rotation at 4 °C for 30 minutes. The suspension was briefly spun down at 12,000 g, separated on a magnet, and the supernatant was discarded. The beads were then resuspended with 1 ml wash buffer (20 mM Tris-HCl pH 7.4, 0.2 % v/v Titron X-100, 0.4 M lithium chloride, 10 mM EDTA and 8 mM DTT; WB) and washed on a rotator set for 20 rpm at 4 °C for 5 minutes, 3 wash rounds in total. The beads were collected on a magnetic rack and the supernatant was discarded. The wash procedure was repeated three times. The elution was carried out stepwise. Washed bead pellet was first resuspended in 50 µl of the elution buffer (25 mM HEPES-KOH, 0.1 mM EDTA; HE). The first suspension was heated at 60 °C for 5 minutes to facilitate the elution, and the eluate was collected upon placing the bead-sample mixture on a magnetic rack, separating the beads, and recovering the clean supernatant. The resultant pellet was next resuspended in another 50 µl of HE buffer, and the process was repeated. The eluates were then combined and subjected to an additional solid-phase reversible immobilization (SPRI) bead purification step and stored frozen.

The eluate from oligo(dT) bead extraction was further purified using AMPure XP SPRI beads (Beckman Coulter Life Sciences) according to the manufacturer's recommendations. Briefly, the eluate samples were supplemented with 1.2× volumes of the SPRI bead suspension in its standard (supplied) binding buffer, and the resultant mixture was incubated at room temperature for 5 minutes with periodic mixing. The SPRI beads were brought down by a brief 2,000 g spin down and separated from the solution on a magnetic rack. The supernatant was removed, and the beads were resuspended in 1 ml of 80 % v/v ethanol, 20 % v/v deionized water mixture and further washed by tube flipping. The bead and solution separation procedure were repeated. The ethanol washing process was repeated one more time. Any remaining liquid was brought down by a brief spin and removed using a pipette, and the beads were allowed to air-dry while in the magnetic rack for 2 minutes. The purified RNA was then eluted in 20 µl of deionized water and the RNA content was assessed using absorbance readout *via* Nanodrop and fluorescence-based detection *via* Qubit RNA high sensitivity (HS) assay kit (Thermo Fisher Scientific).

Flow cell priming and library sequencing

Nanopore sequencing was conducted on an Oxford Nanopore MinION Mk1B using R9.4.1 flow cells for ~72 hours in each run. Initially, the flow cell was left at 25 °C for 30 minutes to reach ambient temperature. The flow cell was inserted into the MinION Mk1B and a quality check was performed to ensure that the pore count was above manufacturer warranty level (800 pores). Prior to sample loading, the priming solution (Flush Buffer + Flush Tether) was degassed in a vacuum chamber for 5 minutes. A similar approach was repeated when loading the RNA library. The run set up on the loaded libraries was performed according to Standard running options on the MinKNOW software (Version 4.3.25). The SQK-RNA002 sequencing option was selected, and the bulk file output was switched from OFF to ON to export the output. For real-time assessment of the quality of the run, the output FAST5 files were base called in-line with sequencing using the MinKNOW-provided Guppy software running with 'fast' base calling preset and model.

Expression analysis of the mouse data

Basecalled reads were aligned to the mouse reference genome (GRCm39) using minimap2 v2.1.0 (parameters: 'minimap2 -ax splice -k14 -B3 -O3,10 --junc-bonus 1 --junc-bed'). During alignment, splice junction coordinates were provided to minimap2 in BED format using the 'junc-bed' flag to improve the accuracy of the spliced alignments. Splice junction BED files were generated using minimap2 paftools.js gff2bed function, using the the gene structure reference (Ensembl 2014 mouse GTF). Primary genomic alignments were assigned to genes using Subread featureCounts v2.0.1 in stranded, long-read mode (using parameters --primary -L -T 48 -s 1 --extraAttributes "gene_biotype, gene_name"). DESeq2 v1.26.0 [55] was used to obtain log-normalised gene counts. PCA plots were generated from regularized log transformed gene counts, using DESeq2's plotPCA function.

Lift over of transcriptomic to genomic sites and calculation of metatranscript coordinates

We wrote the txannotate software (https://github.com/comprna/txannotate) to annotate RNA methylation calls in transcriptomic space with metatranscript coordinates and to transpose the coordinates of annotated calls from transcriptomic to genomic space. In short, *txannotate* uses the *genomicFeatures* R-package [54] to parse gene structure from a GTF annotation to map transcriptomic coordinates to genomic coordinates. The input is RNA methylation calls in bed-like format, i.e., tab delimited file where column 1 represents reference sequence, column 2 represents interval start, and column 3 represents interval end. Firstly, the shell script *cheui_to_bed.sh* (contained within txannotate) with default parameters converts CHEUI methylation calls to a bed-like format. Secondly, the txannotate script *annotate.R* with default parameters assigns metatranscript coordinates to the methylation calls using the relevant reference annotation, Ensembl v104 (GRCm39) GTF for mouse and Gencode v38 (GRCh38) for human. Further, *annotate.R* uses the gene structure information to assign sites on protein-coding transcripts to metagene locations (5′UTR, CDS, or 3′UTR) and calculate the distances from a given site to the nearest upstream and downstream splice junctions annotated (if there is any) in the same transcript where the modified site was predicted. Finally, using the txannotate script *lift.R*, and providing the relevant gene-structure reference, the annotated methylation calls are transposed from transcriptomic coordinates (i.e., position on a specific transcript) to genomic coordinates, i.e., position on a specific chromosome.

RNA methylation metatranscript plots

During the conversion of site-level methylation calls from transcriptomic to genomic coordinates using our *txannotate* package, the absolute distance (in nucleotides) and relative metagene position (as a fraction of the overall UTR or CDS length) of each site were calculated with respect to the original isoform to which the underlying reads were uniquely aligned. The relative meta-transcript coordinates were derived as previously described [57], placing the modifications along three equal-sized segments of length L. Position 0 represents the transcript start site (TSS), position L represents the CDS start, position 2L represents the CDS end, and position 3L represents the polyadenylation site (PAS). For our graphical representation, we used L=40. Meta-transcript plots showing the abundance of tested and significant sites, alongside the proportion of significant sites per tested region, were made using ggplot2.

Co-occurrence of modifications in transcripts and reads

To study the co-occurrence of modifications in annotated transcripts, we considered all protein-coding transcripts (mRNAs) with at least two tested sites, i.e., having 20 or more reads at both sites. For the co-occurrence of m6A and m5C, we partitioned all these mRNA transcripts into four sets according to whether they contained two significant m6A and m5C sites, only one of the modifications, or had no significant sites (even though both were

tested). Based on this partition, we performed a Fisher's exact test to determine whether the association of m6A and m5C in transcripts was higher than expected. To study the co-occurrence of modifications in reads, we considered those transcripts with two modified sites at a relative distance from 1 to 15. We then calculated the co-occurrence as the proportion of reads with both modifications, i.e., the number of reads that had both sites predicted as modified (CHEUI-solo Model 1 probability > 0.7) divided by the total number of reads considered. To calculate the expected level of co-occurrence in the same sample, we calculated the co-occurrence for 1000 pairs of modified sites located in different transcripts. For this analysis, we discarded any possible reads and sites on the ribosomal RNAs (rRNAs). It is known that rRNAs are hypermodified in multiple positions. Considering our analysis of the effects of other modifications on the identification of m6A and m5C, we would expect that they would be affected by other modifications.

Supplementary Tables description

Supp. Table S1: Description of the samples used in this study.

Supp. Table S2: Number of m6A and m5C sites predicted in each of the samples from Supp. Table S1.

Supp. Table S3: Significant sites CHEUI-solo predictions m6A in HEK293 WT.

Supp. Table S4: Significant sites CHEUI-solo predictions m6A in HEK293 METTL3-KO

Supp. Table S5: Significant sites CHEUI-diff predictions differential m6A between WT and KO

Supp. Table S6: Significant sites CHEUI-solo predictions m5C in HeLa WT.

Supp. Table S7 - Significant sites CHEUI-solo predictions m5C in HeLa NSUN2-KO.

Supp. Table S8 - Significant sites CHEUI-diff predictions differential m5C between WT and KO

Supp. Table S9 - Significant sites CHEUI-solo predictions of m6A and m5C in three developmental time points from mouse embryonic brain.

Supp. Table S10 – Number of IVT inputs used for training and testing of CHEUI.

Funding

This research was supported by the Australian Research Council (ARC) Discovery Project grants DP210102385 (to TP, RH and EE), DP220101352 (to EE and TP), and DP180100111 (to TP and NS); by the National Health and Medical Research Council (NHMRC) Senior Research Fellowship APP1135928 (to TP) and Investigator Grant GNT1175388 (to NS). 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), the ANU Merit Allocation Scheme (ANUMAS), and Phenomics Australia. 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), and particularly to Tiffany Cripps, for their assistance with Sanger sequencing.

References

1. Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201–6 (2012).

2. Squires, J. E. *et al.* Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* **40**, 5023–5033 (2012).

3. Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635–46 (2012).

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

5. Arango, D. *et al.* Acetylation of Cytidine in mRNA Promotes Translation Efficiency. *Cell* **175**, 1872-1886.e24 (2018).

6. Gagliardi, D. & Dziembowski, A. 5' and 3' modifications controlling RNA degradation: from safeguards to executioners. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **373**, (2018).

7. Mendel, M. *et al.* Splice site m6A methylation prevents binding of U2AF35 to inhibit RNA splicing. *Cell* **184**, 3125-3142.e25 (2021).

8. Haussmann, I. U. *et al.* m6A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. *Nature* **540**, 301–304 (2016).

9. Shafik, A. M. *et al.* N6-methyladenosine dynamics in neurodevelopment and aging, and its potential role in Alzheimer's disease. *Genome Biol.* **22**, 17 (2021).

10. Widagdo, J. *et al.* Experience-Dependent Accumulation of N6-Methyladenosine in the Prefrontal Cortex Is Associated with Memory Processes in Mice. *J. Neurosci.* **36**, 6771–7 (2016).

11. Barbieri, I. & Kouzarides, T. Role of RNA modifications in cancer. *Nat. Rev. Cancer* **20**, 303–322 (2020).

12. Boccaletto, P. *et al.* MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* **46**, D303–D307 (2018).

13. Anreiter, I., Mir, Q., Simpson, J. T., Janga, S. C. & Soller, M. New Twists in Detecting mRNA

Modification Dynamics. *Trends Biotechnol.* **39**, 72–89 (2021).

14. Linder, B. & Jaffrey, S. R. Discovering and Mapping the Modified Nucleotides That Comprise the Epitranscriptome of mRNA. *Cold Spring Harb. Perspect. Biol.* **11**, (2019).

15. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

16. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).

17. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).

18. Leger, A. *et al.* RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.* **12**, 7198 (2021).

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

20. Price, A. M. *et al.* Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. *Nat. Commun.* **11**, 6016 (2020).

21. Ueda, H. nanoDoc: RNA modification detection using Nanopore raw reads with Deep One-Class Classification. *bioRxiv* 2020.09.13.295089 (2020). doi:10.1101/2020.09.13.295089

22. Parker, M. T., Barton, G. J. & Simpson, G. G. Yanocomp: robust prediction of m6A modifications in individual nanopore direct RNA reads. *bioRxiv* 2021.06.15.448494 (2021). doi:10.1101/2021.06.15.448494

23. Jenjaroenpun, P. *et al.* Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res.* **49**, e7 (2021).

24. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.* **10**, 4079 (2019).

25. Begik, O. *et al.* Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.* (2021). doi:10.1038/s41587-021-00915-6

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

27. Gao, Y. *et al.* Quantitative profiling of N6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *Genome Biol.* **22**, 22 (2021).

28. Hendra, C. *et al.* Detection of m6A from direct RNA sequencing using a Multiple Instance Learning framework. *bioRxiv* 2021.09.20.461055 (2021). doi:10.1101/2021.09.20.461055

29. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–5 (2009).

30. Yao, B. *et al.* Nanopore callers for epigenetics from limited supervised data. *bioRxiv* 2021.06.17.448800 (2021). doi:10.1101/2021.06.17.448800

31. Stoiber, M. *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* 94672 (2017). doi:10.1101/094672

32. Linder, B. *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* **12**, 767–72 (2015).

33. Koh, C. W. Q., Goh, Y. T. & Goh, W. S. S. Atlas of quantitative single-base-resolution N6-methyl-adenine methylomes. *Nat. Commun.* **10**, 5636 (2019).

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

35. Huang, T., Chen, W., Liu, J., Gu, N. & Zhang, R. Genome-wide identification of mRNA 5-methylcytosine in mammals. *Nat. Struct. Mol. Biol.* **26**, 380–388 (2019).

36. Yang, X. *et al.* 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m5C reader. *Cell Res.* **27**, 606–625 (2017).

37. Liu, J. *et al.* Sequence- and structure-selective mRNA m5C methylation by NSUN6 in animals. *Natl. Sci. Rev.* **8**, nwaa273 (2021).

38. Selmi, T. *et al.* Sequence- and structure-specific cytosine-5 mRNA methylation by NSUN6. *Nucleic Acids Res.* **49**, 1006–1022 (2021).

39. Livneh, I., Moshitch-Moshkovitz, S., Amariglio, N., Rechavi, G. & Dominissini, D. The m6A epitranscriptome: transcriptome plasticity in brain development and function. *Nat. Rev. Neurosci.* **21**, 36–51 (2020).

40. Wang, H., Todd, D. A. & Chiu, N. H. L. Enhanced differentiation of isomeric RNA modifications by reducing the size of ions in ion mobility mass spectrometric measurements. *J. Anal. Sci. Technol.* **11**, 46 (2020).

41. Khoddami, V. *et al.* Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6784–6789 (2019).

42. Safra, M. *et al.* The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* **551**, 251–255 (2017).

43. Grozhik, A. V *et al.* Antibody cross-reactivity accounts for widespread appearance of m1A in 5'UTRs. *Nat. Commun.* **10**, 5126 (2019).

44. 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).

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

46. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

47. Li, J. *et al.* Jasper: An End-to-End Convolutional Neural Acoustic Model. (2019).

48. Chollet, F. & and others. Keras. Available at: https://github.com/fchollet/keras. (2015).

49. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2016).

50. Garcia-Campos, M. A. *et al.* Deciphering the 'm6A Code' via Antibody-Independent Quantitative Profiling. *Cell* **178**, 731-747.e16 (2019).

51. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

52. Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLoS One* **10**, e0124633 (2015).

53. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

54. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

55. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

56. Begik, O. *et al.* Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. *Genome Biol.* **21**, 97 (2020).

57. Olarerin-George, A. O. & Jaffrey, S. R. MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites. *Bioinformatics* **33**, 1563–1564 (2017).