Identification of m6A and m5C RNA modifications at single-molecule resolution from Nanopore sequencing

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

[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] Peter MacCallum Cancer Centre, Melbourne, Australia

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

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

[6] Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain

[&] These authors contributed equally

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

ABSTRACT

The expanding field of epitranscriptomics might rival the epigenome in the diversity of the biological processes impacted. However, the identification of modifications in individual RNA molecules remains challenging. We present CHEUI, a new method that detects N6-methyladenosine (m6A) and 5-methylcytidine (m5C) at single-nucleotide and single-molecule resolution from Nanopore signals. CHEUI predicts methylation in Nanopore reads and transcriptomic sites in a single condition, and differential m6A and m5C methylation between any two conditions. Using extensive benchmarking with Nanopore data derived from synthetic and natural RNA, CHEUI showed higher accuracy than other existing methods in detecting m6A and m5C sites and quantifying the site stoichiometry levels, while maintaining a lower proportion of false positives. CHEUI provides a new capability to detect RNA modifications with high accuracy and resolution that can be cost-effectively expanded to other modifications to unveil the full span of the epitranscriptome in normal and disease conditions.

INTRODUCTION

The identification of transcriptome-wide maps of two modified ribonucleotides in messenger RNAs (mRNA), 5-methylcytidine (m5C) and N6-methyladenine (m6A) (Dominissini et al. 2012; Squires et al. 2012) sparked a new and expanding area of epitranscriptomics that might rival the epigenome in biological importance. Techniques based on immunoprecipitation, enzymatic or chemical reactivity enrichment methods, coupled with high-throughput sequencing, have uncovered the involvement of these and other modifications in multiple steps of mRNA metabolism, including translation (Schumann et al. 2020; Arango et al. 2018), mRNA stability (Gagliardi and Dziembowski 2018) and pre-mRNA alternative splicing (Mendel et al. 2021). Several physiological processes have also been functionally linked with RNA modifications, such as sex determination (Haussmann et al. 2016), neurogenesis (Shafik et al. 2021), and learning (Widagdo et al. 2016). Moreover, there is increasing evidence that RNA modification pathways are dysregulated in cancer (Barbieri and Kouzarides 2020) and neurological disorders (Shafik et al. 2021). Despite these advances, most of the 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 these modifications change in individual mRNA molecules in relation to normal and disease processes.

A major roadblock preventing rapid progress in research on RNA modifications is the general lack of detection methods. Although more than 150 naturally occurring RNA modifications have been described (Boccaletto et al. 2018), only a handful of them can be potentially detected and quantified using transcriptome-wide methods (Anreiter et al. 2021; Linder and Jaffrey 2019). 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. DRS can then capture information about the chemical structure, including naturally occurring covalent modifications in nucleotides (Simpson et al. 2017; Garalde et al. 2018). Nonetheless, RNA modification detection from DRS signals presents various challenges. The differences between individual modified and unmodified signals are usually subtle and depend on the sequence context. Additionally, due to the variable translocation rate of the molecules through the pores and the possible pore-to-pore variability, different copies of the same molecule present considerable signal variations (Rang et al. 2018). These challenges make the application of sophisticated computational models necessary to interpret the signals and modification status.

Several computational methods have been developed in the past few years to detect RNA modifications in DRS data. They can be broadly divided into two categories based on their approaches. The first one includes methods that rely on comparing two conditions, one corresponding to a sample of interest, often the wild type (WT) sample, and the other with a reduced or abolished presence of a specific modification, usually obtained through a knock-out (KO) or knock-down (KD) of a modification 'writer' enzyme. This category includes Nanocompore (Leger et al. 2021), Xpore (Pratanwanich et al. 2021), DRUMMER (Price et al. 2020), nanoDOC (Ueda 2020), Yanocomp (Parker et al. 2021) and Tombo in *sample comparison* mode. These methods compare collective properties of DRS signals in the two conditions. This category also includes ELIGOS (Jenjaroenpun et al. 2021) and Epinano (Liu et al. 2019), which both compare base-calling errors between two experiments; and nanoRMS (Begik et al. 2021), which compares signal features between two samples, to predict stoichiometry on pre-selected sites. The second category of tools can operate on a single condition, i.e., without using a KO/KD or an otherwise control. This category includes MINES (Lorenz et al. 2020), Nanom6A (Gao et al. 2021), and m6Anet (Hendra et al. 2021), all predicting m6A; Tombo in *de novo* and *alternate* modes, which identifies sites with generic or m5C modifications, respectively; and Epinano-RMS, which predicts pseudouridine on high stoichiometry sites (Begik et al. 2021).

Despite the advances provided by current methods to detect RNA modifications in the DRS data, several limitations remain. Methods that compare two conditions generally require a valid KO/KD 'control' sample to identify the modification, which may be difficult or impossible to generate. Furthermore, in these approaches, the identification of the modification is indirect, as it relies on changes in KO/KD relative to WT, which may or may not be directly related to the modification of interest. This makes it impractical to study more than one modification per experiment. A further complication of this approach is that other modifications can be affected by the KO/KD, and the transcriptome could be substantially altered, which may impact the identification of modified sites and transcripts. For instance, depletion of m5C causes depletion of hm5C (Tahiliani et al. 2009), hence potentially confounding the results. Regarding the methods that use error patterns, they depend on the accuracy of the base caller method, which may vary over time or be base caller specific. Moreover, not all modifications produce consistent errors. For instance, it was described that error patterns are not consistent enough to confidently identify m5C methylation (Jenjaroenpun et al. 2021). Limitations also exist in methods that work with individual samples. MINES, Nanom6A, and m6Anet only predict m6A modifications in DRACH/RRACH motifs, Epinano-RMS only detects pseudouridine in transcriptome sites if the proportion of modified molecules is very high, and Tombo does not specify the modification type. These methods also cannot predict a modification in an individual molecule represented by a single read. The ability of current methods from both categories to predict stoichiometry is also limited. MINES does not provide information about stoichiometry, and other methods can only estimate the stoichiometry at 5'-DRACH-3' sites or rely on a control sample depleted of modifications to estimate it. More importantly, there currently are no methods that can predict a modification in a single read in any sequence context.

To address these limitations, we developed CHEUI (**CH**$_3$ (methylation) **E**stimation **U**sing **I**onic current), a new computational framework that provides a series of innovations: (1) CHEUI predicts m6A and m5C at nucleotide resolution and at the level of individual reads in any sequence context; (2) CHEUI also predicts m6A and m5C modifications at the level of transcriptomic sites without the need for a KO/KD or control sample; (3) CHEUI achieves higher accuracy than other existing methods in predicting m6A and m5C stoichiometry levels while maintaining a lower number of false positives; and (4) CHEUI also assesses differential m6A and m5C deposition between any two conditions. Finally, CHEUI was trained with a synthetic RNA dataset that can be cost-effectively generated for other modifications. Thus, CHEUI's ability to detect RNA modifications with high accuracy and resolution can be expanded to virtually any modification to unveil the full span of the epitranscriptome in normal and disease conditions.

RESULTS

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

CHEUI uses as input the Nanopore read signals corresponding to 9-mers, i.e., five overlapping 5-mers, centered at adenosine (A) for m6A or cytosine (C) for m5C (**Fig. 1a**) (**Supp. Fig. 1**), together with the distances between the observed signal and the expected unmodified signal (**Fig. 1a**) (**Supp. Fig. 2**). Inclusion of the distance metrics increased accuracy by ~10% (**Supp. Fig. 2**). CHEUI has two components: CHEUI-solo **(Fig. 1b),** which makes predictions in individual samples, and CHEUI-diff **(Fig. 1c),** which tests differential methylation between any two conditions. CHEUI-solo predicts methylation at two different levels: it first predicts m6A or m5C at the nucleotide resolution on individual read signals (Model 1) and then predicts m6A or m5C at the transcript site level by processing the individual read probabilities from Model 1 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 predictions derived from CHEUI-solo Model 1 across two conditions to predict differential m6A or m5C at each transcriptomic site (**Fig. 1c**). CHEUI models were trained on *in vitro* transcript (IVT) sequences containing m6A, m5C, and no modifications (see Methods for more details).

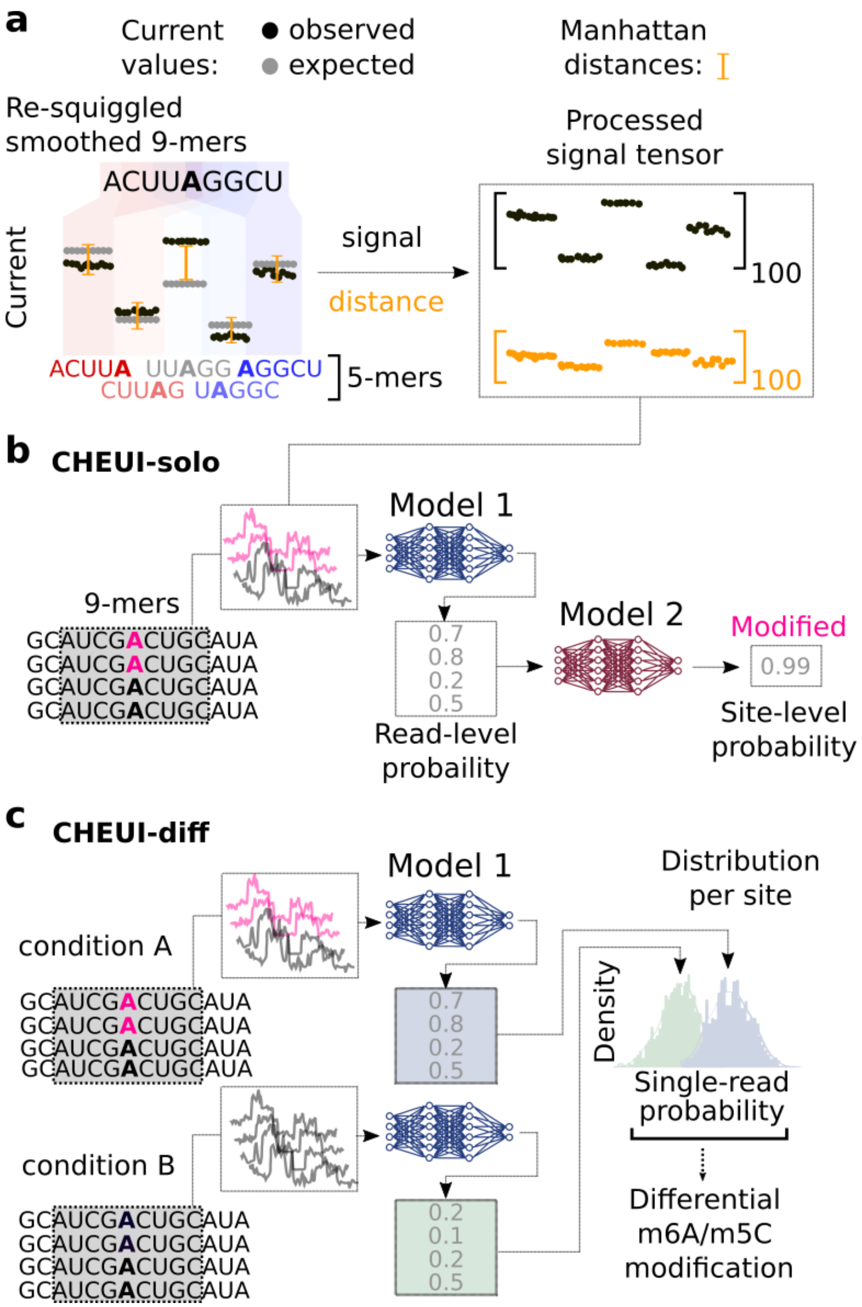

Figure 1. CHEUI modules and signal processing approach. (a) CHEUI processes signals for each 9-mer, i.e., five consecutive 5-mers. The signals for each 5-mer are scaled to 20 values (see Methods), yielding a vector of length 100. An expected vector of length 100 is calculated for all five 5-mers and 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 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 reads 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, with stoichiometry estimated as the proportion of modified reads at that site. **(c)** CHEUI-diff uses the output from Model 1 from two conditions to test for differential m6A or m5C at a specific site. For each site, a non-parametric test is performed to identify significant changes in the distribution of m6A or m5C Model 1 probabilities.

CHEUI accurately detects m5C and m6A modifications 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 (Yao et al. 2021). For this test, only read signals from 9-mers with a single modified nucleotide were considered, i.e., 9-mers where only one A or one C was present. 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 (Yao et al. 2021), we used read signals from a different IVT sequencing experiment (Jenjaroenpun et al. 2021). As before, this test set included signals from 9-mer sites with a single A or a single 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 probability cutoff may improve the accuracy of predictions in individual reads. 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. Using the double probability cutoff of 0.7 and 0.3 provided the optimal balance between accuracy gain and number of preserved reads, as can be gauged by the improved area under the ROC curve from 0.857 to 0.899 for m6A and from 0.827 to 0.877 for m5C **(Fig. 2b),** while retaining ~73% of the reads **(Fig. 2c)**. We thus decided to use (0.7, 0.3) double cutoffs for the rest of the analyses.

To test CHEUI-solo Model 2 for predicting the methylation probability at the transcript site-level, we used controlled mixtures of modified and unmodified signals from reads not included in the training with pre-defined proportions. 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 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 and the modification stoichiometry levels

We next compared CHEUI-solo with Nanocompore (Leger et al. 2021), Xpore (Pratanwanich et al. 2021), and Epinano (Liu et al. 2019) for their ability to detect and quantify RNA modifications using DRS reads. To achieve this, we built positive and negative independent test datasets. The positive dataset consisted of 81 sites for m6A and 84 sites for m5C. We used mixtures with pre-defined stoichiometry of 20, 40, 60, 80, and 100 percent of the IVTs with either modification, for all sites. The negative sites consisted of 512 sites for A and 523 sites for C, and only contained unmodified IVTs. For 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 required a control sample to detect modifications, a second dataset containing only unmodified signals was created for the same sites, randomly sub-sampling to the same level of coverage. We observed that, in general, the number of true positives (TPs) detected by most tools increased with the stoichiometry of the sites (**Fig. 2f**). Notably, CHEUI-solo recovered a higher number of true methylated sites compared to the other tools at all stoichiometry levels. We next estimated the false positives at the site level with the negative sites, using a single sample for CHEUI-solo and two independent samples for Xpore, Epinano, and Nanocompore. Xpore showed the highest false-positive rate for m6A and m5C, followed by Epinano. CHEUI-solo had 1 misclassified site for m5C and none for m6A, whereas Nanocompore had no false positives (**Fig. 2g**).

We next evaluated the prediction of the stoichiometry in a site-wise manner. To this end, we additionally considered nanoRMS (Begik et al. 2021) and Tombo (Stoiber et al. 2017), which can estimate the stoichiometry at a pre-defined site. Epinano was not considered for this test, as it does not provide stoichiometry information. 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 and nanocompore for m6A (Pearson r = 0.524 and 0.498, respectively), and Xpore and NanoRMS for m5C (Pearson r = 0.556 and 0.46, respectively). Tools were not included in this stoichiometry benchmarking if they returned negative correlations **(Supp. Fig. 5)**.

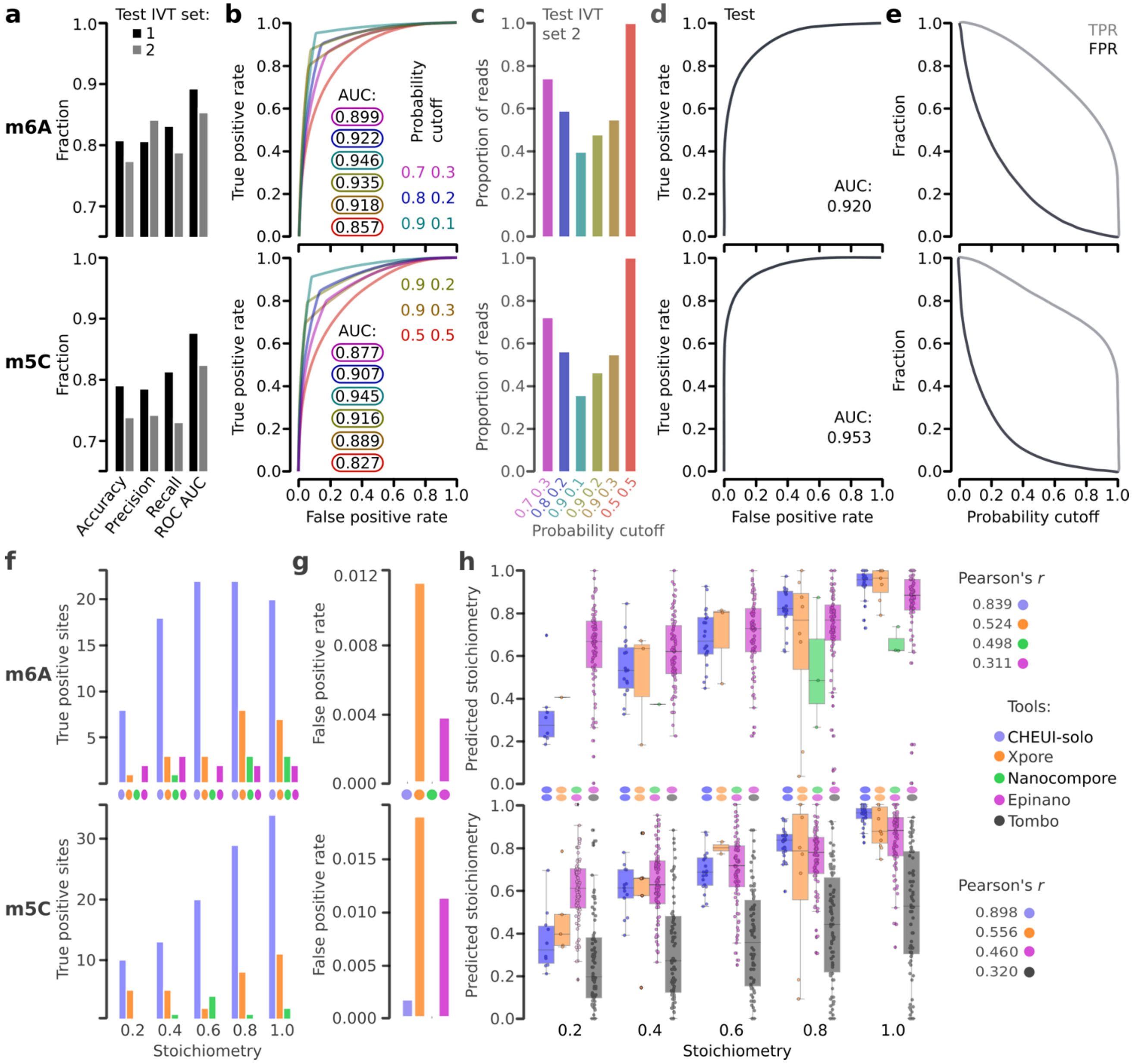

Figure 2. CHEUI-solo's accuracy metrics and comparison with other RNA modification detection tools. (a) Accuracy, precision, recall, and area under the Receiver Operating Characteristic (ROC) curve (AUC) for CHEUI-solo Model 1 for m6A (upper panel) and m5C (lower panel) detections are shown for individual 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 set 2 sets 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. **(c)** Proportion of reads selected (y axis) for each of double cutoff (x axis). **(d)** ROC curves 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. **(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 per tool (y-axis) at different stoichiometry levels (x-axis) using independent benchmarking datasets 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 had 2 false-positive site detections (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.

Impact of other modifications on the prediction of m6A and m5C

To test if other modifications located on the site or at nearby positions could impact the accuracy of predictions made by CHEUI for m6A or m5C in individual reads, we tested CHEUI on read signals from IVTs containing other modifications not used for training CHEUI, namely, 1-methyladenosine (m1A), hydroxymethylcytidine (hm5C), 5-formylcytidine (f5C), 7-methylguanosine (m7G), pseudouridine (Y), and Inosine (I) (Jenjaroenpun et al. 2021). All read signals, regardless of the modification, were processed per the 9-mer centered at an A or C nucleotide, as described for Model 1. Thus, the modifications were either at the same central base (m1A and m6A for A, and m5C, 5fC, and hm5C for C), or in the neighboring bases (Y, m7G, I, m1A, m6A for C; or Y, m7G, I, m5C, 5fC, hm5C for A) (**Figs. 3a and 3b**). As a general trend, the proportion of signals containing other modifications predicted as positives by CHEUI (CHEUI-solo Model 1) recapitulates the results for signals without additional modifications (**Figs. 3c and 3d**). This is the case for all modifications, except for predictions by the m6A model of signals containing m1A, a chemical isomer of m6A, for which the proportion of modified m1A follows a similar trend as m6A (**Fig. 3c**).

To investigate whether m1A misclassification in reads signals was specific to CHEUI, or a phenomenon shared across other methods, we used Xpore and Nanocompore and tested the discrimination of m6A and m1A signals without any *a priori* assumption about the modification type. We used 81 9-mer sites containing a A/m1A/m6A in the middle and made all possible pairwise comparisons among three sets of reads: one with no modifications, one with all reads having m1A, and one with all reads having m6A, with a median coverage of 62 reads per site. When comparing m6A or m1A signals against the unmodified set, Xpore detected 11 m6A sites and 16 m1A sites, whereas Nanocompore detected 5 and 3 sites (**Fig. 3e**). CHEUI-solo detected 19 sites with m1A as m6A and 19 m6A sites (**Fig. 3e**). In the comparison of m6A against m1A reads, Xpore found a significant difference only in two of the sites, whereas Nanocompore found none (**Fig. 3e**). These results showed that the other methods perform similarly to CHEUI when attempting to separate m6A and m1A signals. This indicates that the DRS signals for these two isomers with the current statistical models and pore chemistry might be indistinguishable (**Supp. Fig 6).**

To fully address the m6A and m1A DRS signal similarity, we tested whether CHEUI-solo could be re-trained to separate m6A and m1A. We used m1A and other modifications different from m6A signals as part of the negative set of the training data. This new model achieved accuracy comparable to the original model in the separation of m6A from the unmodified signals **(Supp. Fig. 7a)**. However, it showed a trade-off between accurately detecting m6A and correctly separating m1A from m6A **(Supp. Fig. 7b)**, further suggesting limitations to separate isomers in the Nanopore signals.

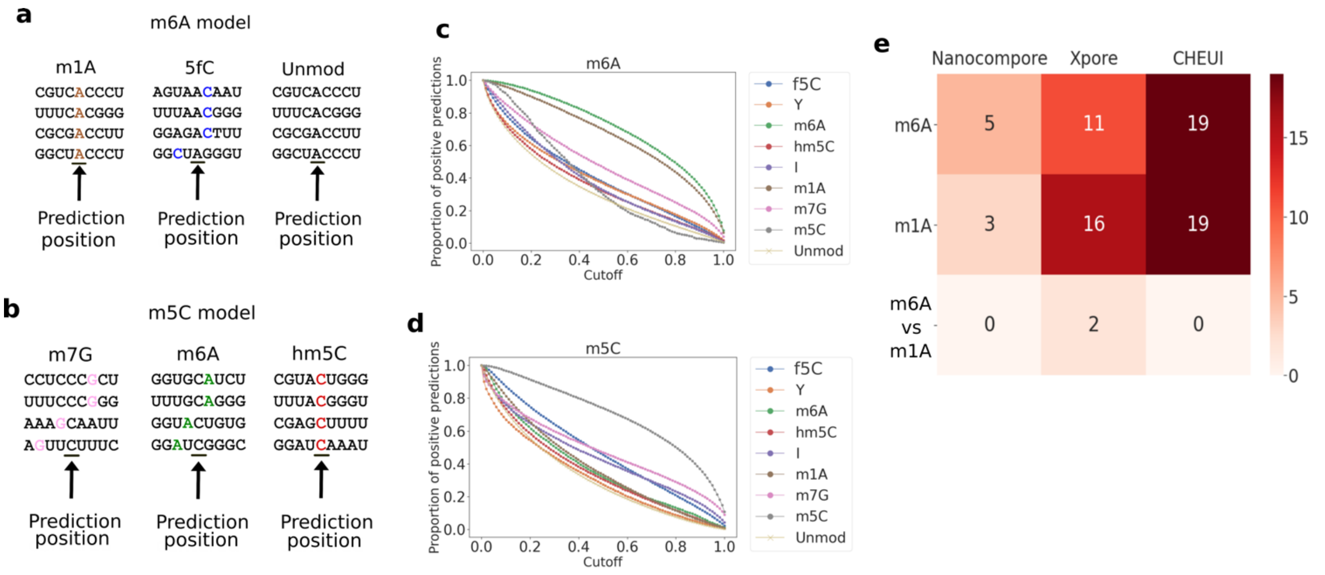

Figure 3. Impact of other RNA modifications on the prediction accuracy of m6A and m5C by CHEUI and the other tools. CHEUI-solo's predictions were tested in individual read signals (Model 1) for m6A **(a)** and m5C **(b)** using read signals from IVTs containing other modifications. Coverage per site ranged between 20 and 324 reads, with median coverage of 62 reads. **(c)** The number of individual read signals predicted as m6A modifications by CHEUI-solo Model 1 (y-axis) at different values of the (Model 1) probability cutoff (x axis). **(d)** Same as (c) but for m5C. (**e)** The number of significant (transcriptomic-like) sites predicted by each tool in each of the conditions (y axis). For Nanocompore and Xpore, the 'm6A' and 'm1A' rows indicate predictions in the comparison of modified sites at 100% stoichiometry against a control sample with no modifications. CHEUI values for the same rows correspond to the number of sites predicted as m6A by CHEU-solo Model 2 in m6A or m1A sites with 100% modification stoichiometry. For Nanocompore and Xpore, the 'm6A vs m1A' row represents the number of significant sites predicted when comparing an m6A sample against an m1A sample, both with 100% stoichiometry. For CHEUI, it indicates the number of sites that were detected by CHEUI-solo Model 2 only in the m6A or m1A samples.

CHEUI accurately identifies m6A modifications in HEK293 cell lines

We next tested CHEUI's ability to correctly identify m6A and m5C in naturally occurring RNAs. First, we used CHEUI-solo to predict m6A using DRS data from WT HEK293 cells **(Supp. Table S1)** (Pratanwanich et al. 2021). We recovered 562,628 transcriptomic sites that had a coverage of more than 20 reads in all three available replicates. Testing these sites with CHEUI-solo resulted in a high correlation among replicates in the predicted stoichiometry and modification probability per transcriptomic site (**Fig. 4a)**. Analyzing the three replicates together, we considered as m6A modified those transcriptomic sites with predicted probability > 0.9999, which we estimated would result in FDR~0 using an empirical permutation test (see Methods). At this cutoff, CHEUI-solo returned 9,857 significant m6A transcriptomic sites **(Supp. Table S2)**. These were found in 3,830 transcripts and corresponded to 8,643 unique genomic positions, as 898 of the sites occurred in more than one transcript in the same gene at the same genomic location (**Supp. Table S3**). 85.12% of the transcriptomic sites (84.63% of the unique genomic sites) corresponded to the 5'-DRACH-3' motif, which was higher than the 76.57% identified from m6ACE-seq and miCLIP experiments (Linder et al. 2015; Koh et al. 2019). Interestingly, CHEUI-solo predicted m6A in 1,493 non-DRACH motifs (1,356 within the unique genomic sites), 25 of them also previously identified by m6ACE-seq and miCLIP (Linder et al. 2015; Koh et al. 2019). The two most common non-DRACH motifs identified by CHEUI-solo were 5'-GGACG-3' (203 unique genomic positions) and 5'-GGATT-3' (121 unique genomic sites). These motifs coincided with the two most common non-DRACH motifs identified by miCLIP2 experiments in the same cell lines and occurred at 245 (5'-GGACG-3') and 96 (5'-GGATT-3') sites (Körtel et al. 2021). Most of the sites with DRACH or non-DRACH motifs occurred in mRNAs.

Next, we considered the DRS data from HEK293 cells with a knockout of the m6A writer METTL3 (METTL3-KO) (Pratanwanich et al. 2021). Using the predictions from CHEUI-solo in individual reads, we could confirm a significant decrease in the proportion of modified A in the METTL3-KO sample with respect to the WT sample (p-value = 6.7e-78) (**Supp. Fig. 8**). Next, we predicted differential m6A sites between HEK293 wild type (WT) cells and HEK293 cells with a knockout of the m6A writer METTL3 (METTL3-KO) (Pratanwanich et al. 2021). CHEUI-diff showed an enrichment of significant cases with higher modification stoichiometry in WT (**Fig. 4b**).

In comparison with Xpore and Nanocompore, CHEUI-diff detected more sites with higher modification stoichiometry in WT at all three different significance thresholds (**Fig. 4c**). For these sites, CHEUI-diff also predicted more sites with supporting evidence from m6ACE-seq or miCLIP experiments in HEK293 (Linder et al. 2015; Koh et al. 2019) (**Supp. Fig. 9a**) and containing the 5'-DRACH3-3' motif (**Fig. 4d**), except at the 0.001 significance level, where 0.70 of CHEUI-diff sites and 0.71 of Xpore sites had the motif. Comparing two METTL3-KO replicates to estimate false positives, CHEUI-diff predicted the lowest number of significant sites (0, 1, and 3, at the three significance thresholds, respectively) (**Supp. Fig. 9b**). In contrast, Xpore predicted over 2,000 sites at 0.001 significance and over 12,000 sites at 0.05 significance. Moreover, only 9.8% of the Xpore sites at 0.05 significance contained the 5'-DRACH-3' motif, a substantially lower proportion compared to the 46% found by Xpore in the WT vs METTL3-KO comparison at the same significance level. This suggested that most of the Xpore sites in the comparison of the two METLL3-KO replicates may be false positives.

CHEUI accurately identifies m5C modifications in HeLa cell lines

We next used CHEUI for predicting m5C in cell-derived RNA. To accomplish this, we used the CRISPR-cas9 system in HeLa cells to engineer a knock-out (KO) of the RNA methyltransferase NSUN2 (NSUN2-KO) (**Supp. Fig. 9c**), which modifies cytosines in various mRNAs and tRNAs (Huang et al. 2019; Schumann et al. 2020). We performed DRS on 3 biological replicates from the WT and NSUN2-KO HeLa cells **(Supp. Table S1)**. We recovered 497,439 transcriptomic sites with a coverage of more than 20 reads in all three replicates for each condition. Testing these sites with CHEUI-solo (Model 2), we observed a high correlation among the replicates in the predicted stoichiometry and modification probability values per transcriptomic site (**Fig. 4e**). Analyzing the three replicates together, and using a transcriptomic site predicted probability > 0.9999, which we estimated would result in FDR~0 using an empirical permutation test (see Methods), we obtained 3,318 significant sites in WT and 1,906 in NSUN2-KO (**Supp. Table S4**). 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 = 2.4e-49) (**Supp. Fig. 8**).

We then tested whether CHEUI-solo assigned a high probability for m5C to transcriptomic sites previously detected in HeLa using bisulfite RNA sequencing (bsRNA-seq). We used data from three independent studies and took the union of these sites as an orthogonally validated set (Schumann et al. 2020; Huang et al. 2019; Yang et al. 2017). CHEUI-solo probabilities on this union set in the WT samples were significantly higher compared to the probabilities assigned to the same sites in the NSUN2-KO samples (**Fig 4f**). 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. 9d).** In contrast, the enrichment of high CHUEI-solo probabilities over the background distribution disappeared in the NSUN2-KO (p-value = 0.025) (**Supp. Fig 9d)**. However, not all sites exhibited low probabilities of modification, which is consistent with the previous proposal that a fraction of the m5C sites is NSUN2-independent in mRNA (Huang et al. 2019; Schumann et al. 2020).

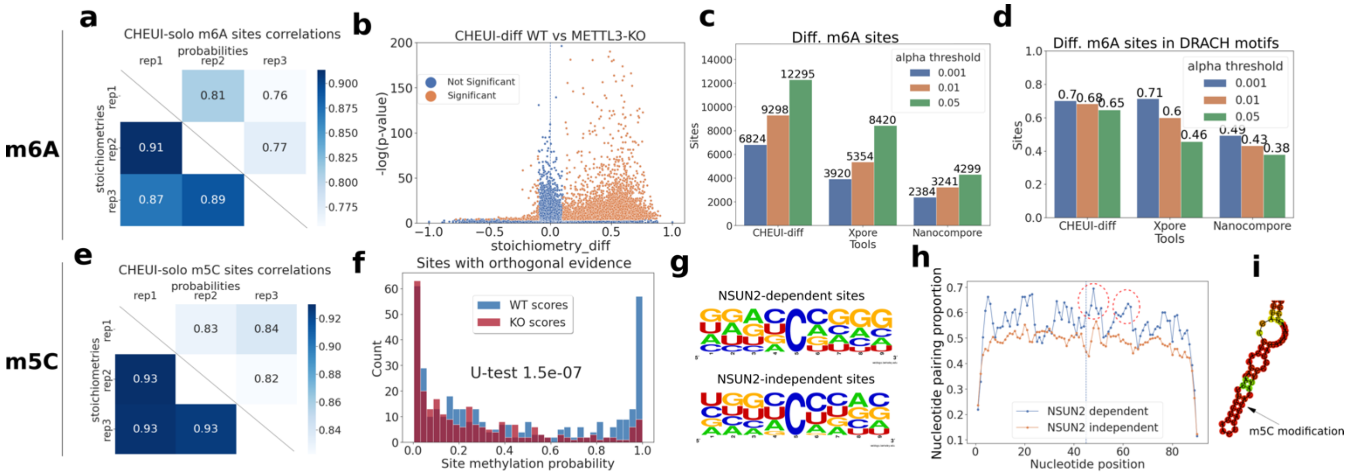

Figure 4. Detection of m6A and m5C in cell lines using CHEUI. (a) Pearson correlation values among HEK293 WT replicates for CHEUI-solo m6A stoichiometry predictions (lower diagonal) and m6A per-site probabilities (upper diagonal) for the 562,628 transcriptomic sites that had a coverage of more than 20 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)** The number of differentially modified m6A sites detected by each tool between HEK293 WT and METTL3-KO using three different alpha levels of significance (0.05, 0.01 and 0.001) on the adjusted p-values. **(d)** The 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) 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 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 line indicates the m5C position. Dashed red circles indicate a higher proportion of base-pairs indicating a stem on the NSUN2 dependent sites. **(i)** Example of RNA secondary structure containing an m5C site at the base of a stem-loop.

To investigate NSUN2-dependent and -independent sites, we used CHEUI-diff to select differentially modified sites between WT and NSUN2-KO. 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' (Huang et al. 2019) (**Fig. 4g**). 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 (Huang et al. 2019) (**Fig. 4g**). To further assess the validity of these CHEUI predictions, we investigated their secondary structure. Consistent with previous studies (Schumann et al. 2020; Huang et al. 2019), we found an enrichment of base-pairing probabilities in NSUN2-dependent sites compared to NSUN2-independent sites (**Figs. 4h and 4i**). Moreover, the potential base-pairings suggested a higher occurrence of stem-loops at around 5 nt downstream of the m5C site in NSUN2-dependent sites (**Supp. Fig. 10**), also consistent with previous results (Huang et al. 2019). Interestingly, when we used an alternative definition for NSUN2-independent sites to be those that are only significant in HeLa NSUN2-KO (1,940 transcriptomic sites), we found the same results in terms of sequence motifs and structural properties **(Supp. Fig. 11)**. Furthermore, we found that most NSUN2-independent sites found by Huang et al. based on bsRNA-seq data had a higher delta stoichiometry between WT vs NSUN2-KO using CHEUI-diff, compared to all the other sites (**Supp. Fig. 12)**. These results indicate that CHEUI-solo and CHEUI-diff can correctly assign high probabilities to previously discovered m5C sites and potentially discover new ones.

DISCUSSION

In this work, we described CHEUI, a new method to accurately identify m6A and m5C, two of the most abundant RNA modifications in mRNA, from Nanopore signals. CHEUI presents several novelties with respect to previous methods describing m6A or m5C in the Nanopore signals. CHEUI performs prediction at nucleotide resolution in any sequence context, and at the level of individual reads as well as at transcriptomic sites. Furthermore, CHEUI can operate on a single condition, without requiring a KO/KD or control sample. CHEUI also accurately assess differential m6A and m5C between any two conditions in any sequence context. CHEUI thus escapes the 'sample comparison' paradigm used by most of the previous tools and the constraints presented by other computational or experimental methods to detect RNA modifications that are limited to certain sequence contexts or that are based on indirect evidence like errors in sequence identification. With CHEUI, we have also demonstrated that synthetic transcripts containing a modification for all copies of a given nucleotide can be used to train predictive models that are able to generalize and detect *in vivo* RNA modifications. This approach provides a simple and cost-effective pathway towards the generation of models for virtually any modification.

We presented an in-depth, rigorous benchmarking across different RNA modification detection methods using a ground-truth test dataset. CHEUI has demonstrated substantially improved sensitivity and stoichiometry prediction accuracy over the other tools. The stoichiometry accuracy is challenging to assess as it requires the complete knowledge of the modification status of reads. To resolve this, we used control mixtures of read signals from *in vitro* transcript sequences, where the test sites were selected with variable and realistic coverage and stoichiometry values, and without constraining the sequence context or using any knowledge of previous performance. We argue that these unbiased read mixtures provide a powerful and effective approach to benchmark the general accuracy of RNA modification detection methods that should be used jointly with testing naturally occurring RNAs.

Benchmarking with cell line data showed that CHEUI's performance is consistent across replicates, it controls the false-positive rates and recover methylation sites previously detected by orthogonal techniques. One limitation of the cell line datasets is that known sites must be defined by other techniques that present their own biases and limitations. As we have seen, the use of KO cells may be only effective for some of the enzymes, as modifications such as m5C may be deposited onto mRNA by multiple enzymes. Furthermore, some of the modifications detected in KO cells could have been induced by compensatory effects, given that the KO cells will have the time to adapt and undergo potential compensatory modifications or even genetic selection. These effects may not happen in a KD model, where the changes are more rapid, and cells would have a limited adaptation and selection timeline.

One of the biggest challenges for machine learning RNA modification detection algorithms is the lack of ground truth 'real world' test datasets. A test set should ideally recapitulate the distribution of the future data where the model will be applied. In naturally occurring RNA, we would expect multiple modifications to occur at different rates along the RNA molecules. As a proxy of this, we showed that other RNA modifications do not generally affect CHEUI's accuracy to identify m6A and m5C. CHEUI could separate m5C from hm5C, which presents an advantage over bisulfite sequencing, which has been one of the preferred methods to detect m5C transcriptome-wide but nonetheless does not discriminate between the two modifications.

Unexpectedly, we observed that CHEUI and other methods could not separate the Nanopore signals corresponding to m1A and m6A nucleotides, which are positional isomers. Visual inspection of the signals for m6A and m1A nucleotides in the same k-mer contexts showed that they deviate similarly from the unmodified nucleotides. In contrast, m5C and hm5C, which have different chemical groups attached to the same position, were separated by CHEUI and the signals in the same k-mer context could be visually distinguished from each other and from unmodified nucleotides. This suggests two hypotheses. The first one is that Nanopore signals from isomeric modifications are not distinguishable. This is supported by our analyses and is consistent with the difficulties encountered by other technologies, such as mass spectrometry, to separate m1A and m6A (Wang and Wang 2020). The second hypothesis is that more powerful deep learning models may have to be developed to separate these modifications. The inclusion of additional predictive features could increase the capabilities of the current model and overcome the observed limitation. Nonetheless, the similarity of Nanopore signals for m1A and m6A may not have a major impact in the study of mRNAs. Recent evidence indicated that although m1A sites are abundant in tRNAs and rRNAs (Khoddami et al. 2019), they are exceedingly rare or potentially absent in mRNAs (Safra et al. 2017) and that many of the reported m1A sites in mRNAs could be due to antibody cross-reactivity (Grozhik et al. 2019).

Importantly, 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 must be generated to train predictive models. *In vitro* transcribed (IVT) sequences with and without modified nucleotides are useful to understand the properties of the Nanopore signals in relation to modifications but had not been exploited yet to train predictive models at single read and single nucleotide resolution. Although IVT sequences may not contain all possible sequence contexts and may include configurations that do not occur naturally, we have shown that our training strategy and deep learning model implemented in CHEUI can address these challenges and accomplish high accuracy and low false positive rate. *In vitro* datasets with any nucleotide modification can be systematically produced much more efficiently than specific cellular models with engineered deletions or insertions of modifications or associated enzymes. CHEUI thus provides an effective framework to exploit this type of data.

CHEUI could be expanded to detect other RNA modifications at single read and single nucleotide resolution. Moreover, as IVT datasets can be built with natural as well as non-naturally occurring modifications, CHEUI opens new opportunities in the field of synthetic biology and RNA engineering.

Software availability

CHEUI will soon be available for academic use

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 sequences from (Liu et al. 2019) were obtained from the NCBI Gene Expression Omnibus (GEO) database through accession GSE124309. The Nanopore read signals for these sequences with and without m6A and m5C modifications were obtained from NCBI Sequence Read Archive (SRA) under accession PRJNA511582 and PRJNA563591, respectively. Nanopore data for the synthetic transcripts from (Jenjaroenpun et al. 2021) was obtained from The Sequence Read Archive (SRA) accession SRP166020. Nanopore data for HEK293 WT and METTL3 KO from (Pratanwanich et al. 2021) was obtained from the European Nucleotide Archive (ENA), accession PRJEB40872. Data from the m6ACE-seq experiments from (Koh et al. 2019) was obtained from the NCBI Gene Expression Omnibus (GEO) under accession number GSE124509. Nanopore data for HeLa WT and NSUN2 KO will soon be available at ENA.

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 reference transcriptome using minimap2 (Li 2018) with options '-ax map-ont -k 5'. Reads were filtered to select the best match for each read using samtools *-F 2324* (Danecek et al. 2021). Nanopolish (Simpson et al. 2017) *eventalign* was then used to align signals to the reference using the options '--scale-events --signal-index --samples --print-read-names'.

Nanopolish *eventalign* outputs the signals for each 5-mer in 3' to 5' orientation, whereas the 5-mers are given (output rows) 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. All the (per read) signals for every 5 overlapping consecutive 5-mers, together with the read ID and sequence, were then used to create a file to be used as input for CHEUI-solo Model 1.

CHEUI-solo Model 1

CHEUI-solo Model 1 is a convolutional neural network (CNN) modified from the Jasper model (Li et al. 2019). The CNN architecture was implemented using Keras (Chollet and and others 2015) and Tensorflow (Abadi et al. 2016). 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 reduce the noise of the input signals and fix the length of the input for CHEUI-solo Model 1, the signals associated with every 5-mer were summarized into 20 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 reaching 20 values. As a result, each 9-mer was mapped to a vector of 100 signal values.

CHEUI-solo Model 1 also considers the distance between the observed and the expected signal for every input. The expected signal is built using the k-mer model from Nanopolish (Simpson et al. 2017), which describes the expected signal value for each 5-mer in the absence of modifications (**Supp. Fig. 2**). 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. A vector of length 100 with the absolute distances between the components of the expected and the observed signals was then calculated. The vectors of observed signals and absolute distances were 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 corresponding vector of observed signals and the vector of distances to the expected signals.

CHEUI-solo Model 1 training and testing

We trained and tested CHEUI-solo Model 1 using only *in vitro* transcript (IVT) data to produce one trained-model for each modification, m6A and m5C. For the m6A (or m5C) model, the positive sets contained m6A (or m5C) in place of the canonical nucleotide, i.e., replacing every A with m6A (or every C with m5C) (Liu et al. 2019). For both models, the negative sets were made of the same IVT sequences containing 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 S5**). 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. Finally, IVT test 2, used to test k-mer generalization, was built from independent IVT experiments (Jenjaroenpun et al. 2021). IVT test 2 was also composed of 9-mers with only one A (or C) at the center of the 9-mer. A summary of the number of reads and different 9-mers used for each dataset can be found in Supp. Table S5. 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 in the validation accuracy and the training was finished. 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.

CHEUI-solo Model 2

CHEUI-solo Model 2 implements a CNN binary classifier that takes as input the distribution of probabilities generated by Model 1 for all read signals at the same transcriptomic site and predicts the stoichiometry and probability of that site being methylated (m6A or m5C). Model 2 assumes that the distribution of the individual-read methylation-probabilities at a given transcriptomic site originates from two groups, one with high Model 1 probabilities (modified site), and a second on with low Model 1 probabilities (unmodified site).

CHEUI-solo Model 2 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 comprised 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 using other techniques (Garcia-Campos et al. 2019; Schumann et al. 2020; Huang et al. 2019). 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 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 of probabilities, 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 probability 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 on 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.

Other tools considered for benchmarking comparisons

We chose tools available for each specific benchmarking comparison. We used Epinano (Liu et al. 2019), 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* that combines all types of read error (mismatches, insertions, and deletions) in the pairwise mode. We also used nanoRMS (Begik et al. 2021), 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. The stoichiometry was estimated 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 (Leger et al. 2021), which uses the assignment of raw signals to a transcriptome reference with

Nanopolish, 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 (Pratanwanich et al. 2021), 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 means and variances of each distribution. After fitting the data into clusters, Xpore assigns as unmodified the cluster with values closer to the expected unmodified signals 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.

Benchmarking of RNA modification detection using IVT controlled mixtures

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 (Jenjaroenpun et al. 2021) 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.

To perform per-site predictions with CHEUI, we classified as positive those sites with 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 centered in an 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 the 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. Moreover, we evaluated all sites with A or C, regardless of whether they had other surrounding As or Cs in their sequence context. On the other hand, to evaluate the true positive rates and the stoichiometry, we only evaluated sites containing one centered modified adenosine or cytosine surrounded by 4 (non-A or non-C) unmodified bases on each side to avoid the influence of having 2 or more modified nucleotides affecting one site.

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 (Jenjaroenpun et al. 2021). 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 mapped to the Gencode human annotation (v38) using minimap2 as described above. CHEUI-solo (Model 1 and Model 2) was run on the 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. 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 all the samples used to test the m6A model, and an FDR = 0.000384 for the samples used to test m5C model. We thus used the cutoff of 0.9999 to select sites from CHEUI-solo at the transcriptomic site level for both models.

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: alpha value = 0.05, 0.01, 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 modified 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 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 (Linder et al. 2015; Koh et al. 2019).

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) (Schumann et al. 2020; Huang et al. 2019; Yang et al. 2017) and the union was taken as the list of orthogonal evidence sites for subsequent comparisons. 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 arbitrarily to have probability>0.99). Then, we randomly sampled the same number of sites from the transcriptome 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 (Lorenz et al. 2011) 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) and 3'-proximal (exon 19 crRNA) regions of the gene (**Supp. Table S6**). Briefly, gene sequences from Ensembl (Asia server) were processed *via* CCTop (Stemmer et al. 2015) 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 mins 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 harvested by scraped from the flasks in 200-500 µL of the Protein Extraction Buffer (50 mM Tris pH7.5 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 manufacturer's instructions. ~30 µg of total protein was loaded on NuPage 4-12% Bis-Tris Protein Gels (Invitrogen) and proteins electrophoretically separated with using NuPAGE™ MES SDS Running Buffer under recommended conditions. Separated proteins were transferred onto PVDF membrane using iBlot™ 2 Transfer Stacks, PVDF, mini (cat. No. IB24002), following manufacturers' instruction. 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 **(Supp. Fig. 9c)**.

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 ºC for 5 minutes, followed by the same rotation at 4 ºC 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 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 ºC for 5 minutes. 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 ºC 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 1.2x 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 was 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. 650ng-800 ng from HeLa cells, were used for each 2x 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/ml in all reaction solutions until the SPRI purification step after the reverse transcription reaction. The final adaptor-ligated sample was eluted in 40 ml.

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 ºC 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.

Liftover of transcriptomic to genomic sites

Using the *genomicFeatures* R-package (Lawrence et al. 2013), we transposed site-level methylation calls from transcriptomic coordinates to genomic coordinates, using a transcriptome annotation (GTF) as a reference for gene structure. We also calculated the distance of a given site from the nearest upstream and downstream splice site (where present) and assigned sites on protein-coding transcripts to metagene locations (5' UTR, CDS, or 3' UTR). Given a genomic site with multiple transcriptomic sites arising from alternative isoforms, the probability and stoichiometry of the transcriptomic site with the highest stoichiometry were assigned to the genomic site for analyses purposes.

Funding

This research was supported by the Australian Research Council (ARC) Discovery Project grants DP210102385 (to TP, RH and EE) and DP180100111 (to TP and NS) and 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) and the ANU Merit Allocation Scheme (ANUMAS). The funding bodies had no role in study design, data collection or data analysis.

Acknowledgements

We are grateful to the personnel from the Biomolecular Resource Facility at JCSMR (ANU), and particularly to Tiffany Cripps, for their assistance with Sanger sequencing.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. http://arxiv.org/abs/1603.04467.

Anreiter I, Mir Q, Simpson JT, Janga SC, Soller M. 2021. New Twists in Detecting mRNA Modification Dynamics. *Trends Biotechnol* **39**: 72–89. http://www.ncbi.nlm.nih.gov/pubmed/32620324.

Arango D, Sturgill D, Alhusaini N, Dillman AA, Sweet TJ, Hanson G, Hosogane M, Sinclair WR, Nanan KK, Mandler MD, et al. 2018. Acetylation of Cytidine in mRNA Promotes Translation Efficiency. *Cell* **175**: 1872-1886.e24. http://www.ncbi.nlm.nih.gov/pubmed/30449621.

Barbieri I, Kouzarides T. 2020. Role of RNA modifications in cancer. *Nat Rev Cancer* **20**: 303–322. http://www.ncbi.nlm.nih.gov/pubmed/32300195.

Begik O, Lucas MC, Pryszcz LP, Ramirez JM, Medina R, Milenkovic I, Cruciani S, Liu H, Vieira HGS, Sas-Chen A, et al. 2021. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol*. http://www.ncbi.nlm.nih.gov/pubmed/33986546.

Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, de Crécy-Lagard V, Ross R, Limbach PA, Kotter A, et al. 2018. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res* **46**: D303–D307. http://www.ncbi.nlm.nih.gov/pubmed/29106616.

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

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giab008/6137722.

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**: 201–6.

Gagliardi D, Dziembowski A. 2018. 5' and 3' modifications controlling RNA degradation: from safeguards to executioners. *Philos Trans R Soc Lond B Biol Sci* **373**. http://www.ncbi.nlm.nih.gov/pubmed/30397097.

Gao Y, Liu X, Wu B, Wang H, Xi F, Kohnen M V, Reddy ASN, Gu L. 2021. 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. http://www.ncbi.nlm.nih.gov/pubmed/33413586.

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.

Garcia-Campos MA, Edelheit S, Toth U, Safra M, Shachar R, Viukov S, Winkler R, Nir R, Lasman L, Brandis A, et al. 2019. Deciphering the 'm6A Code' via Antibody-Independent Quantitative Profiling. *Cell* **178**: 731-747.e16. http://www.ncbi.nlm.nih.gov/pubmed/31257032.

Grozhik A V, Olarerin-George AO, Sindelar M, Li X, Gross SS, Jaffrey SR. 2019. Antibody cross-reactivity accounts for widespread appearance of m1A in 5'UTRs. *Nat Commun* **10**: 5126. http://www.ncbi.nlm.nih.gov/pubmed/31719534.

Haussmann IU, Bodi Z, Sanchez-Moran E, Mongan NP, Archer N, Fray RG, Soller M. 2016. m6A potentiates

Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. *Nature* **540**: 301–304. http://www.ncbi.nlm.nih.gov/pubmed/27919081.

Hendra C, Pratanwanich PN, Wan YK, Goh WSS, Thiery A, Göke J. 2021. Detection of m6A from direct RNA sequencing using a Multiple Instance Learning framework. *bioRxiv* 2021.09.20.461055. http://biorxiv.org/content/early/2021/09/22/2021.09.20.461055.abstract.

Huang T, Chen W, Liu J, Gu N, Zhang R. 2019. Genome-wide identification of mRNA 5-methylcytosine in mammals. *Nat Struct Mol Biol* **26**: 380–388. http://www.ncbi.nlm.nih.gov/pubmed/31061524.

Jenjaroenpun P, Wongsurawat T, Wadley TD, Wassenaar TM, Liu J, Dai Q, Wanchai V, Akel NS, Jamshidi-Parsian A, Franco AT, et al. 2021. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res* **49**: e7. https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa620/5876284.

Khoddami V, Yerra A, Mosbruger TL, Fleming AM, Burrows CJ, Cairns BR. 2019. Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc Natl Acad Sci U S A* **116**: 6784–6789. http://www.ncbi.nlm.nih.gov/pubmed/30872485.

Koh CWQ, Goh YT, Goh WSS. 2019. Atlas of quantitative single-base-resolution N6-methyl-adenine methylomes. *Nat Commun* **10**: 5636. http://www.ncbi.nlm.nih.gov/pubmed/31822664.

Körtel N, Rücklé C, Zhou Y, Busch A, Hoch-Kraft P, Sutandy FXR, Haase J, Pradhan M, Musheev M, Ostareck D, et al. 2021. Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning. *Nucleic Acids Res* **49**: e92. http://www.ncbi.nlm.nih.gov/pubmed/34157120.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**: e1003118. http://www.ncbi.nlm.nih.gov/pubmed/23950696.

Leger A, Amaral PP, Pandolfini L, Capitanchik C, Capraro F, Miano V, Migliori V, Toolan-Kerr P, Sideri T, Enright AJ, et al. 2021. RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat Commun* **12**: 7198. http://www.ncbi.nlm.nih.gov/pubmed/34893601.

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

Li J, Lavrukhin V, Ginsburg B, Leary R, Kuchaiev O, Cohen JM, Nguyen H, Gadde RT. 2019. Jasper: An End-to-End Convolutional Neural Acoustic Model. http://arxiv.org/abs/1904.03288.

Linder B, Grozhik A V, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. 2015. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* **12**: 767–72. http://www.ncbi.nlm.nih.gov/pubmed/26121403.

Linder B, Jaffrey SR. 2019. Discovering and Mapping the Modified Nucleotides That Comprise the Epitranscriptome of mRNA. *Cold Spring Harb Perspect Biol* **11**. http://www.ncbi.nlm.nih.gov/pubmed/31160350.

Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS, Smith MA, Novoa EM. 2019. Accurate detection of m6A RNA modifications in native RNA sequences. *Nat Commun* **10**: 4079. http://www.ncbi.nlm.nih.gov/pubmed/31501426.

Lorenz DA, Sathe S, Einstein JM, Yeo GW. 2020. Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**: 19–28. http://www.ncbi.nlm.nih.gov/pubmed/31624092.

Mendel M, Delaney K, Pandey RR, Chen K-M, Wenda JM, Vågbø CB, Steiner FA, Homolka D, Pillai RS. 2021. Splice site m6A methylation prevents binding of U2AF35 to inhibit RNA splicing. *Cell* **184**: 3125-3142.e25. http://www.ncbi.nlm.nih.gov/pubmed/33930289.

Parker MT, Barton GJ, Simpson GG. 2021. Yanocomp: robust prediction of m&lt;sup&gt;6&lt;/sup&gt;A modifications in individual nanopore direct RNA reads. *bioRxiv* 2021.06.15.448494. http://biorxiv.org/content/early/2021/06/16/2021.06.15.448494.abstract.

Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, Poon P, Goh YT, Yap PML, Chooi JY, et al. 2021. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat Biotechnol* 2020.06.18.160010-2020.06.18.160010. http://biorxiv.org/content/early/2020/06/20/2020.06.18.160010.abstract.

Price AM, Hayer KE, McIntyre ABR, Gokhale NS, Abebe JS, Della Fera AN, Mason CE, Horner SM, Wilson AC, Depledge DP, et al. 2020. Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. *Nat Commun* **11**: 6016. http://www.nature.com/articles/s41467-020-19787-6.

Rang FJ, Kloosterman WP, de Ridder J. 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* **19**: 90. http://www.ncbi.nlm.nih.gov/pubmed/30005597.

Safra M, Sas-Chen A, Nir R, Winkler R, Nachshon A, Bar-Yaacov D, Erlacher M, Rossmanith W, Stern-Ginossar N, Schwartz S. 2017. The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* **551**: 251–255. http://www.ncbi.nlm.nih.gov/pubmed/29072297.

Schumann U, Zhang H-N, Sibbritt T, Pan A, Horvath A, Gross S, Clark SJ, Yang L, Preiss T. 2020. Multiple links between 5-methylcytosine content of mRNA and translation. *BMC Biol* **18**: 40.

Shafik AM, Zhang F, Guo Z, Dai Q, Pajdzik K, Li Y, Kang Y, Yao B, Wu H, He C, et al. 2021. N6-methyladenosine dynamics in neurodevelopment and aging, and its potential role in Alzheimer's disease. *Genome Biol* **22**: 17. http://www.ncbi.nlm.nih.gov/pubmed/33402207.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410.

Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, Parker BJ, Suter CM, Preiss T. 2012. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res* **40**: 5023–5033. http://www.ncbi.nlm.nih.gov/pubmed/22344696.

Stemmer M, Thumberger T, Del Sol Keyer M, Wittbrodt J, Mateo JL. 2015. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLoS One* **10**: e0124633. http://www.ncbi.nlm.nih.gov/pubmed/25909470.

Stoiber M, Quick J, Egan R, Eun Lee J, Celniker S, Neely RK, Loman N, Pennacchio LA, Brown J. 2017. &lt;em&gt;De novo&lt;/em&gt; Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* 94672. http://biorxiv.org/content/early/2017/04/10/094672.abstract.

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**: 930–5. http://www.ncbi.nlm.nih.gov/pubmed/19372391.

Ueda H. 2020. nanoDoc: RNA modification detection using Nanopore raw reads with Deep One-Class Classification. *bioRxiv* 2020.09.13.295089.

http://biorxiv.org/content/early/2020/09/13/2020.09.13.295089.abstract.

Wang J, Wang L. 2020. Deep analysis of RNA N6-adenosine methylation (m6A) patterns in human cells. *NAR genomics Bioinforma* **2**: lqaa007. http://www.ncbi.nlm.nih.gov/pubmed/33575554.

Widagdo J, Zhao Q-Y, Kempen M-J, Tan MC, Ratnu VS, Wei W, Leighton L, Spadaro PA, Edson J, Anggono V, et al. 2016. Experience-Dependent Accumulation of N6-Methyladenosine in the Prefrontal Cortex Is Associated with Memory Processes in Mice. *J Neurosci* **36**: 6771–7. http://www.ncbi.nlm.nih.gov/pubmed/27335407.

Yang X, Yang Y, Sun B-F, Chen Y-S, Xu J-W, Lai W-Y, Li A, Wang X, Bhattarai DP, Xiao W, et al. 2017. 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m5C reader. *Cell Res* **27**: 606–625. http://www.ncbi.nlm.nih.gov/pubmed/28418038.

Yao B, Hsu C, Goldner G, Michaeli Y, Ebenstein Y, Listgarten J. 2021. Nanopore callers for epigenetics from limited supervised data. *bioRxiv* 2021.06.17.448800. http://biorxiv.org/content/early/2021/06/17/2021.06.17.448800.abstract.