Supplementary Material

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

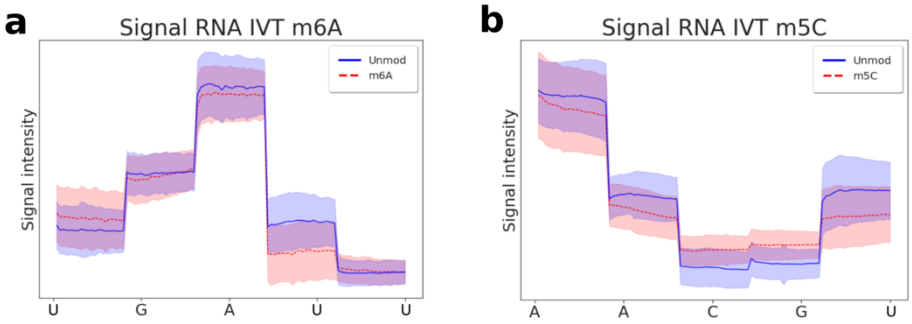
[4] Victor Chang Cardiac Research Institute, Sydney, Australia.

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

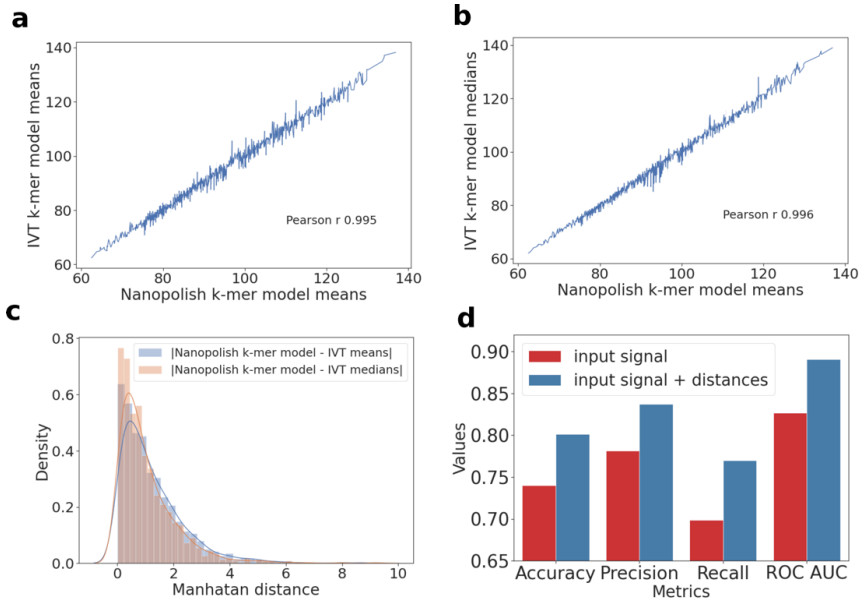
[6] Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain

[&] These authors contributed equally

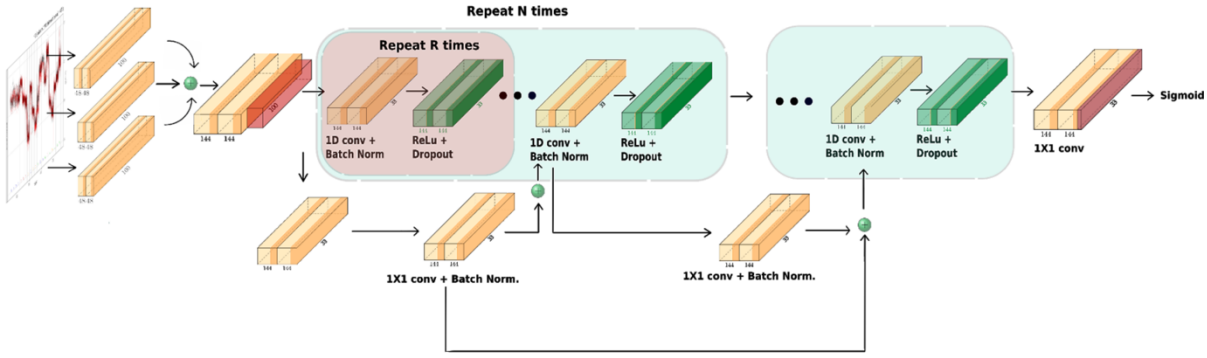
[*] Correspondence to: nikolay.shirokikh@anu.edu.au, eduardo.eyras@anu.edu.au



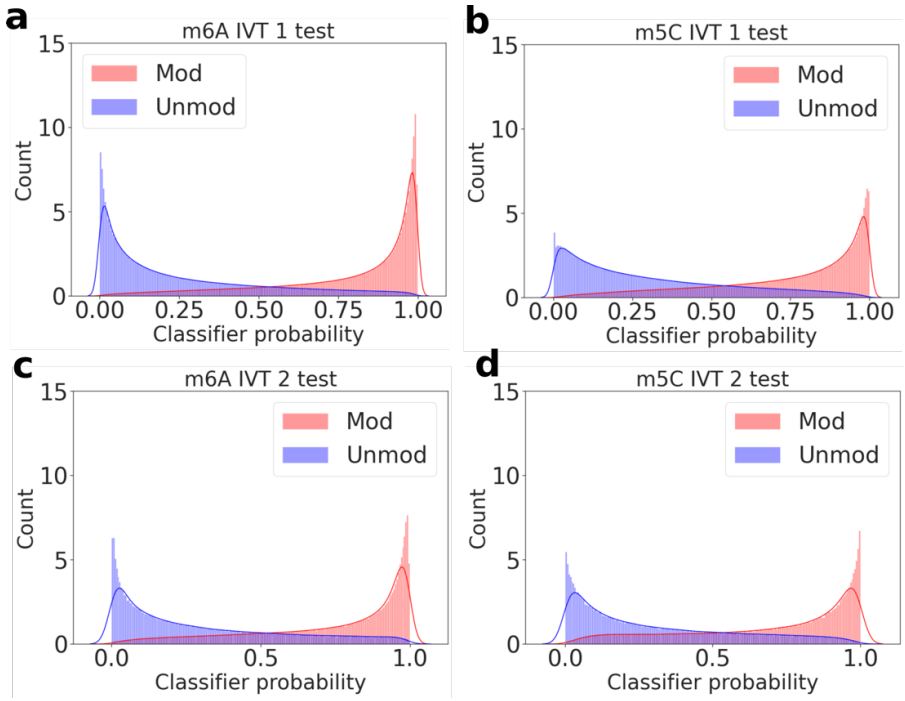
Supplementary Figure 1. **Representation of the input data. (a)** Characteristic profiles of read signals from *in vitro* transcripts (IVTs) with N^6-methyladenosine (m6A) (in red) or adenosine (in blue) located in the middle position. Shown are nanopore current signals from five consecutive nanopore translocation events, i.e., five overlapping 5-mers forming a 9-mer. The signal from each overlapping 5-mer is scaled over the central nucleotide of that 5-mer to twenty values and plotted atop the position of this central nucleotide. Only the central nucleotide of each 5-mer is indicated in the x axis. **(b)** Same as (a), but for read signals from IVTs containing 5-methylcytidine (m5C) (in red) and cytidine (in blue).



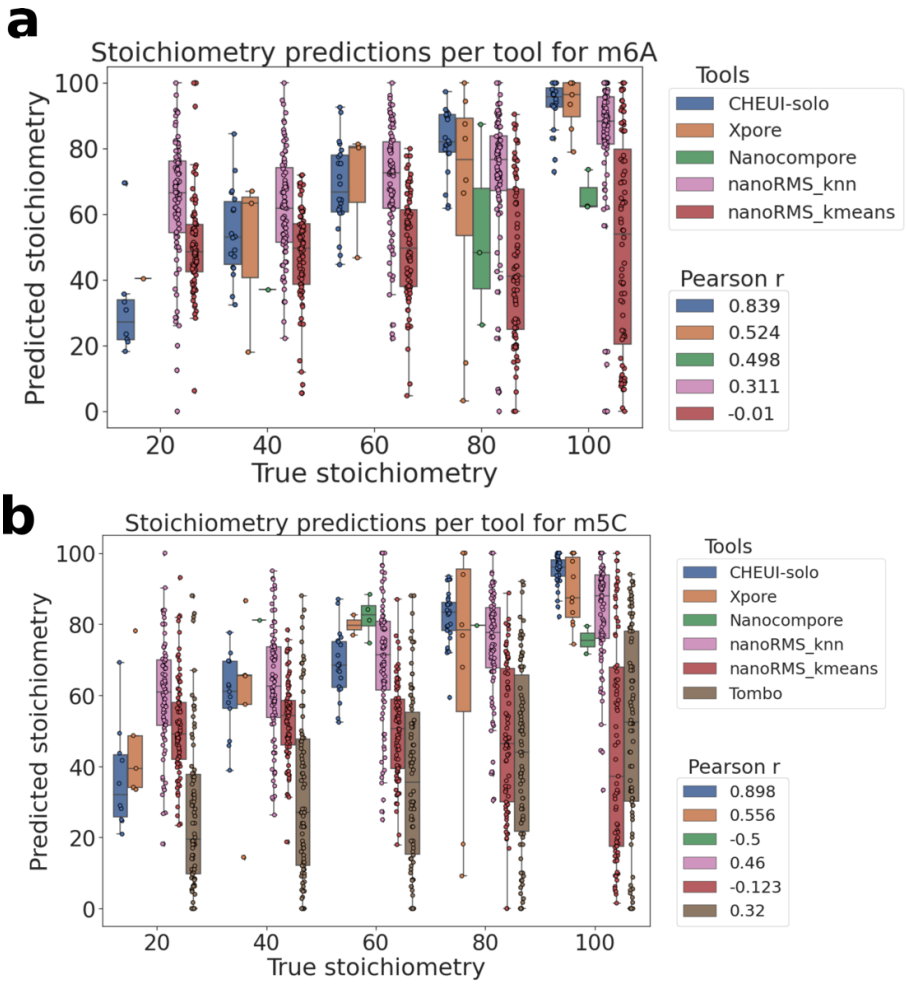
Supplementary Figure 2. Distribution features of nanopore current signals representative of the unmodified RNA nucleotides. (a) Correlation of the values from the Nanopolish k-mer model (x axis) with mean signal values calculated from *in vitro* transcripts (IVTs) for all 5-mers (y axis). **(b)** Correlation of the values from the Nanopolish k-mer model (x axis) with the median signal values calculated from *in vitro* transcripts (IVT) for all 5-mers. **(c)** Distribution of the absolute differences between the values estimated from the IVT signals and those form the Nanopolish k-mer model. **(d)** Comparison of the accuracies of the CHEUI m6A Model 1 with (blue) and without (red) the distance-vector input to the expected values derived from the Nanopolish k-mer model. CHEUI performance is assessed by the accuracy, precision, recall, and area under the receiver operating characteristic (ROC) curve (AUC) from IVT test 1 set (x axis labels).



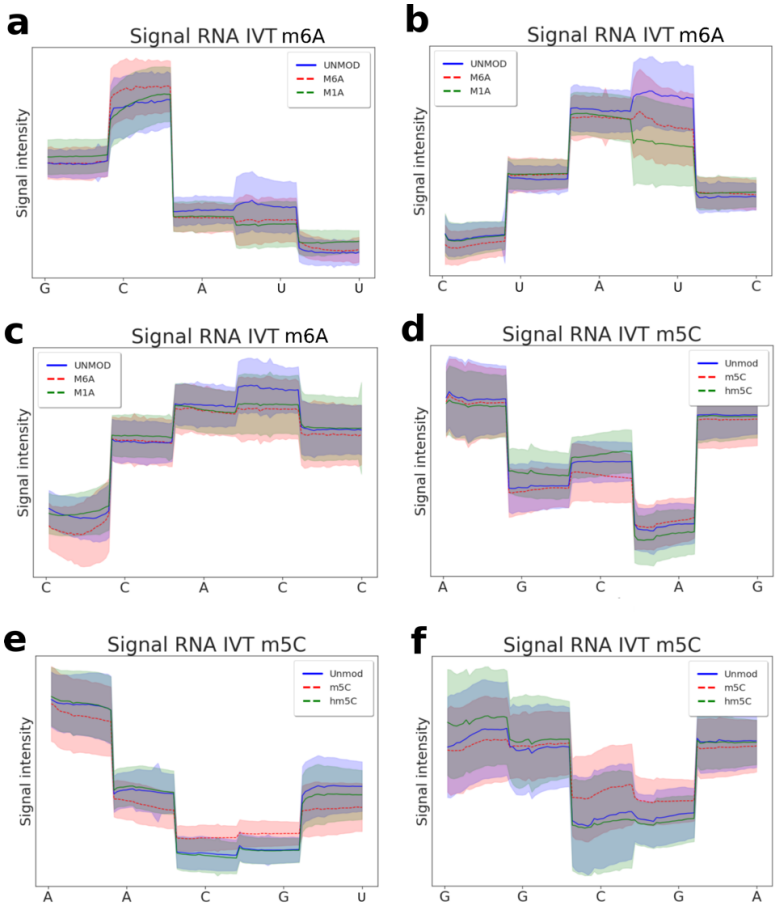
Supplementary Figure 3. Schematic view of the Convolutional Neural Network used in CHEUI-solo. Both models in CHEUI-solo are deep convolutional neural networks (CNNs) defined as shown in the diagram. The models are based on the architecture employed in fast speech recognition algorithms (Li et al. 2019). The main body of the network is made from 4 blocks (green) (N=4 in the diagram), each containing 3 sub-blocks (red) (R=3 in the diagram). Sub-blocks are made of 1D convolutions (1D conv.), followed by batch normalization (Batch Norm), a Rectified Linear Unit (ReLU), and a dropout. The last sub-block of every block is connected directly to all last sub-blocks from previous blocks using a residual connection. The residual connection passes through a 1x1 convolution (1x1 conv.) and batch normalization before being added to the last sub-block convolution.



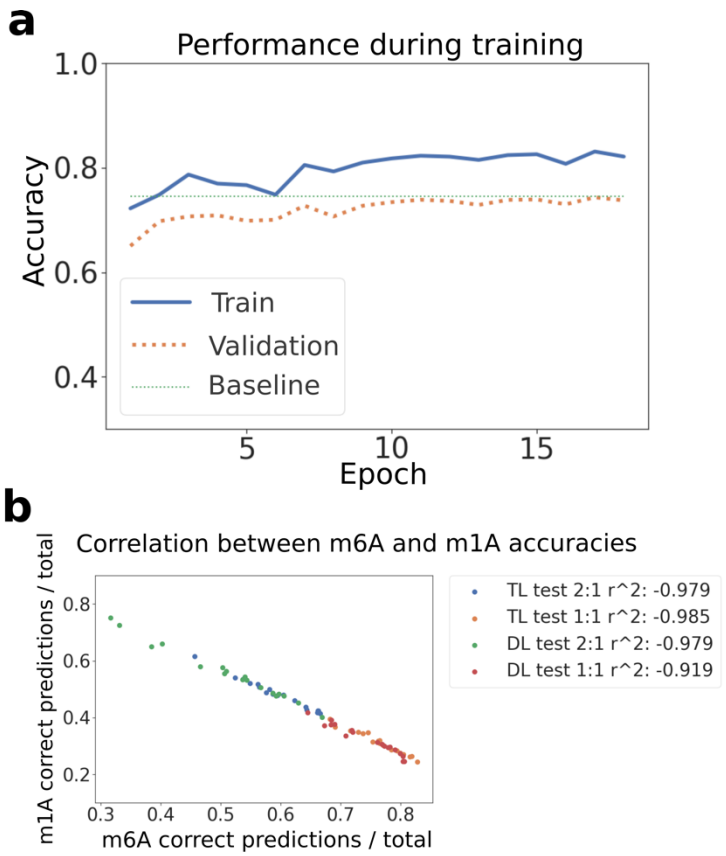
Supplementary Figure 4. Distribution of probabilities assigned by CHEUI for methylation at specific nucleotide positions in individual reads. Shown are distributions returned by CHEUI-solo Model 1 for all the sites and all the reads from the IVT test 1 dataset, for the positive (Mod) and negative (Unmod) cases for m6A **(a)** and m5C **(b)**. **(c, d)** same as (a b), but for the IVT test 2 dataset.



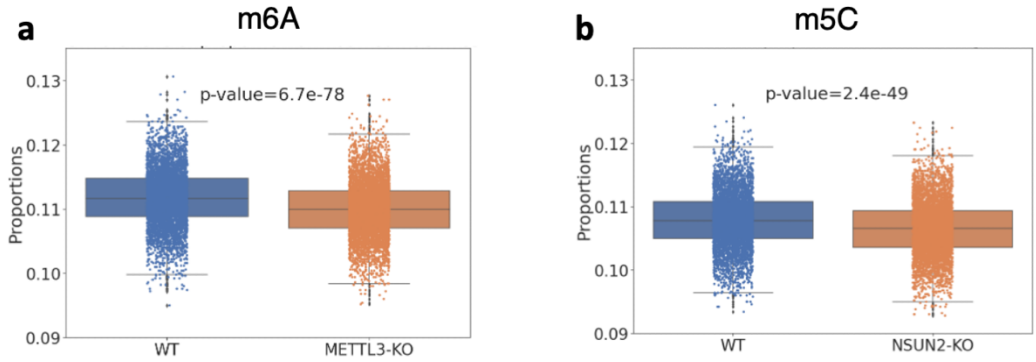
Supplementary Figure 5. Benchmarking of modification stoichiometry identification tools. Stoichiometry values predicted by each tool (y-axis) and the actual stoichiometry values of the controlled read mixtures (x-axis) are shown for m6A **(a)** and for m5C **(b)**. Tool names are given in the legend on the right.



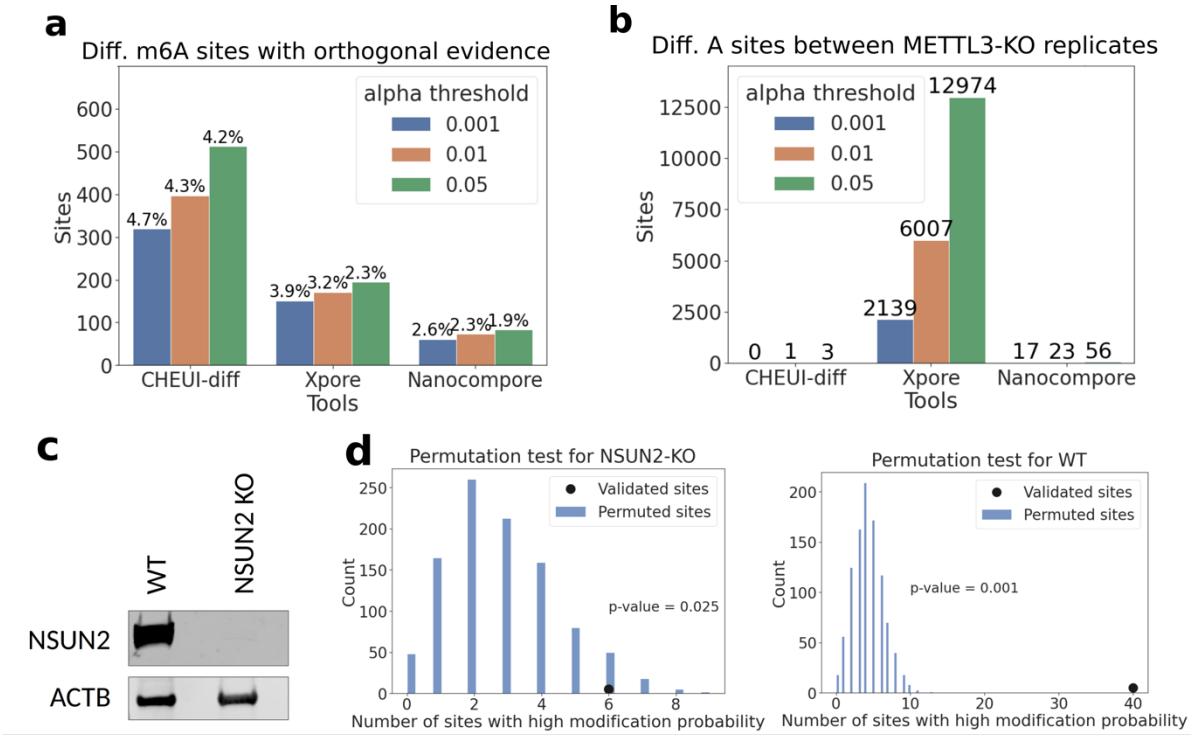
Supplementary Figure 6. Visualization of nanopore signals for A and C modifications. Three different sequence contexts exemplifying typical signal profiles for A (Unmod), m6A, and m1A (**a**, **b**, **c);** and C (Unmod), m5C, and hm5C **(d, e, f)**. Nanopore signals were aggregated as in Supplementary Figure 1.



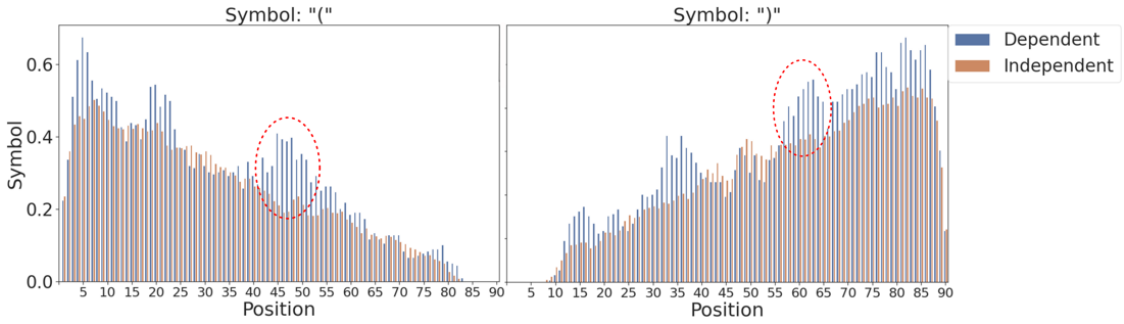
Supplementary Figure 7. Accuracy metrics returned by CHEUI-solo Model 1 upon retraining with other modifications as negatives. **(a)** Accuracy (y axis) of CHEUI Model 1 retrained to predict m6A using other available RNA modifications read signals (Y, f5C, m7G, m1A, I, hm5C, m5C), plus unmodified signals, as negatives, and using the m6A read signals from IVT train 1 as positives. Accuracy was calculated using m6A and unmodified signals during training (train) and on the IVT test 2 dataset (test), compared to the original CHEUI-solo Model 1 for m6A (baseline). **(b)** Association between the correctly classified m1A nucleotides divided by the total number of m1A nucleotides (y axis) and the correctly classified m6A nucleotides divided by the total of m6A nucleotides (x axis) for different versions of CHEUI-solo Model 1, using transfer learning from the original model (TL) or not (DL) and using different label weights, i.e., relative contribution of the negative examples in the loss function (1:1, 2:1).



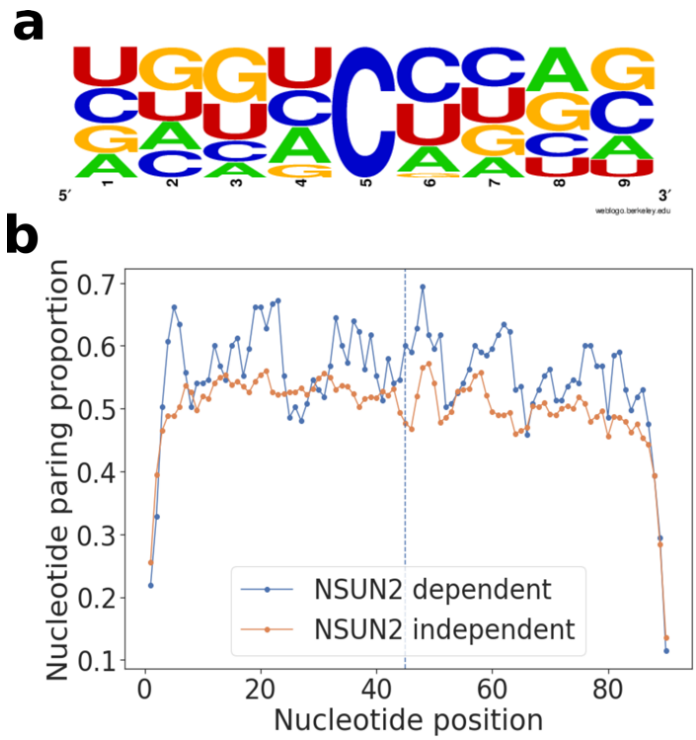
Supplementary Figure 8. Proportion of modified nucleotides observed with CHEUI in poly(A)$^+$ mRNA from HeLa cells. Shown are the proportions of modified A (m6A) over the total of A nucleotides (modified and unmodified) in the WT and METTL3-KO samples **(a)** and the proportions of m5C over all C nucleotides in the WT and NSUN2-KO samples **(b)**. Each dot indicates the proportion of individual nucleotides from individual reads assigned as modified with significance by CHEUI-solo Model 1 (probability > 0,7), from a random pool of 5,000 individual nucleotides at individual reads tested. Each box plot represents 5,000 points obtained from 5,000 random samples of those 5,000 nucleotides. Non-parametric rank sum test was used to test for the statistical significance.



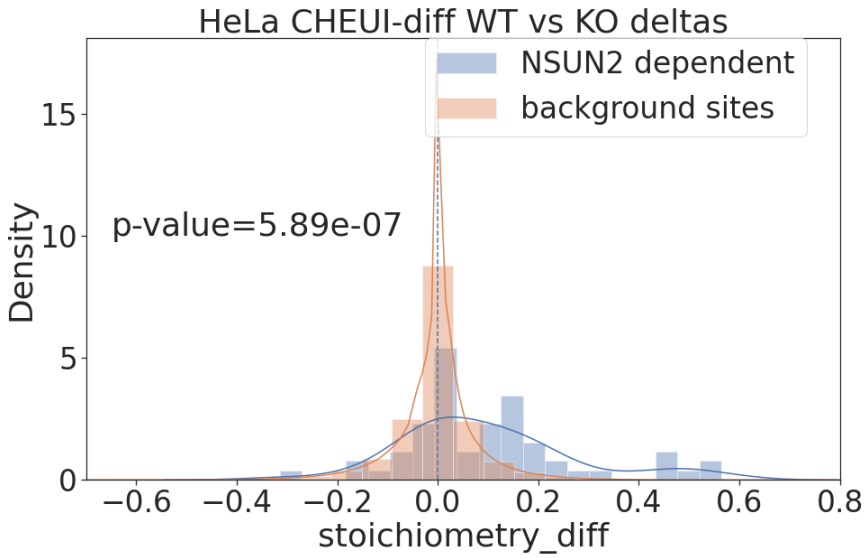
Supplementary Figure 9. Comparative CHEUI analysis of the modification sites in methyltransferase knockout. (a) The number of differentially modified m6A sites between HEK293 WT and METTL3-KO predicted by CHEUI-diff, Xpore and Nanocompore at three levels of significance and overlapping the experimental site set from m6ACE-seq or miCLIP obtained for HEK293. The percentages on the top of the bars correspond to the percentage of sites with the orthogonal evidence with respect to the total number of predicted differential sites. **(b)** Differential m6A significant sites discovered by CHEUI, Xpore and CHEUI comparing two different METTL3-KO biological replicates. **(c)** Western blot with an NSUN2 antibody in WT and KO HeLa cells. Actin B (ACTB) is shown in each condition as a control. **(d)** Bisulfite-based m5C sites are enriched in high CHEUI probabilities. The number of transcriptomic m5C sites predicted by bisulfite sequencing in HeLa cells that had a CHEUI-solo Model 2 probability > 0.99 (black dot) is shown in comparison to the number of cases obtained by chance (blue bars). The cases obtained by chance were calculated by shuffling the CHEUI probabilities across tested sites 10,000 times and calculating again each time how many of the bisulfite-based m5C sites had a probability > 0.99. The left panel shows the result using Nanopore signals from the NSUN2-KO HeLa cells, whereas the right panel shows the results for the Nanopore signals from HeLa WT cells.



Supplementary Figure 10. RNA secondary structures associated with CHEUI m5C predictions. Shown is the proportion of paired bases, indicated with the symbol '(' (left panel) or ')' (right panel) in the predicted secondary structures using 90 nt around (45 nt on either side) the predicted m5C sites on transcriptomic sequences, removing sites that appeared in more than one transcript in the same gene. Position 45 contains the m5C predicted by CHEUI-solo **(a)**. The dashed red circle indicates positions that have an increase of base-paired nucleotides potentially forming a stem-loop on sequences from NSUN2 dependent sites.



Supplementary Figure 11. Sequence motifs and RNA secondary structures for an alternative definition of NSUN2-independent m5c sites. (a) Sequence motifs for NSUN2-independent sites defined as significant sites by CHEUI-solo in the NSUN2-KO sample. **(b)** Proportion of base-paired positions along 90 nt centered at the m5C sites (as in FIgure 4h) predicted by CHEUI-solo. The vertical line indicates the m5C position. NSUN2-dependent sites were defined as sites predicted by CHEUI-diff to have a significant decrease in modification stoichiometry in the NSUN2-KO sample, whereas NSUN2-independent sites were defined in this case as sites significant according to CHEUI-solo for the NSUN2-KO sample.



Supplementary Figure 12. Distribution of the differential stoichiometry in NSUN2-dependent and background sites. Blue bars indicate the stoichiometry difference between WT and NSUN2-KO samples in sites that were reported previously to be NSUN2-dependent by Huang et al. 2019. Orange bars indicate the stoichiometry difference between WT and NSUN2-KO sites that were not in the NSUN2-dependent list. Non-parametric test using rank-sum was performed to compare the two distributions.