Supplementary Material

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

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

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

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

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

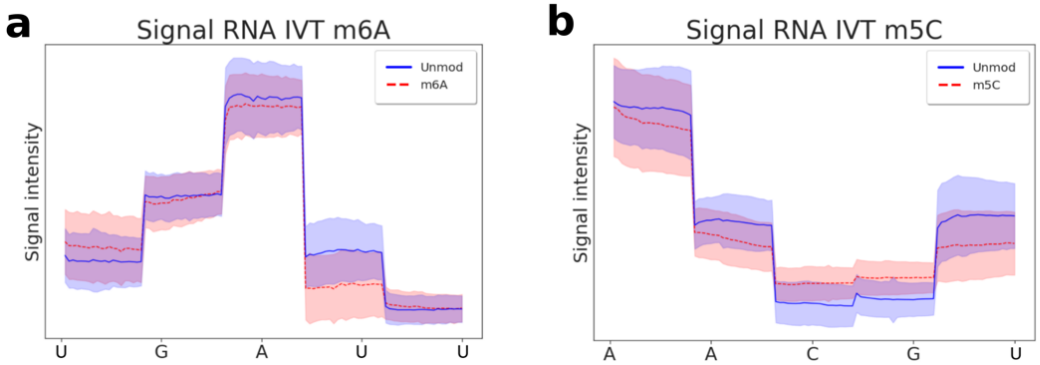
[4] Peter MacCallum Cancer Centre, Melbourne, Australia

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

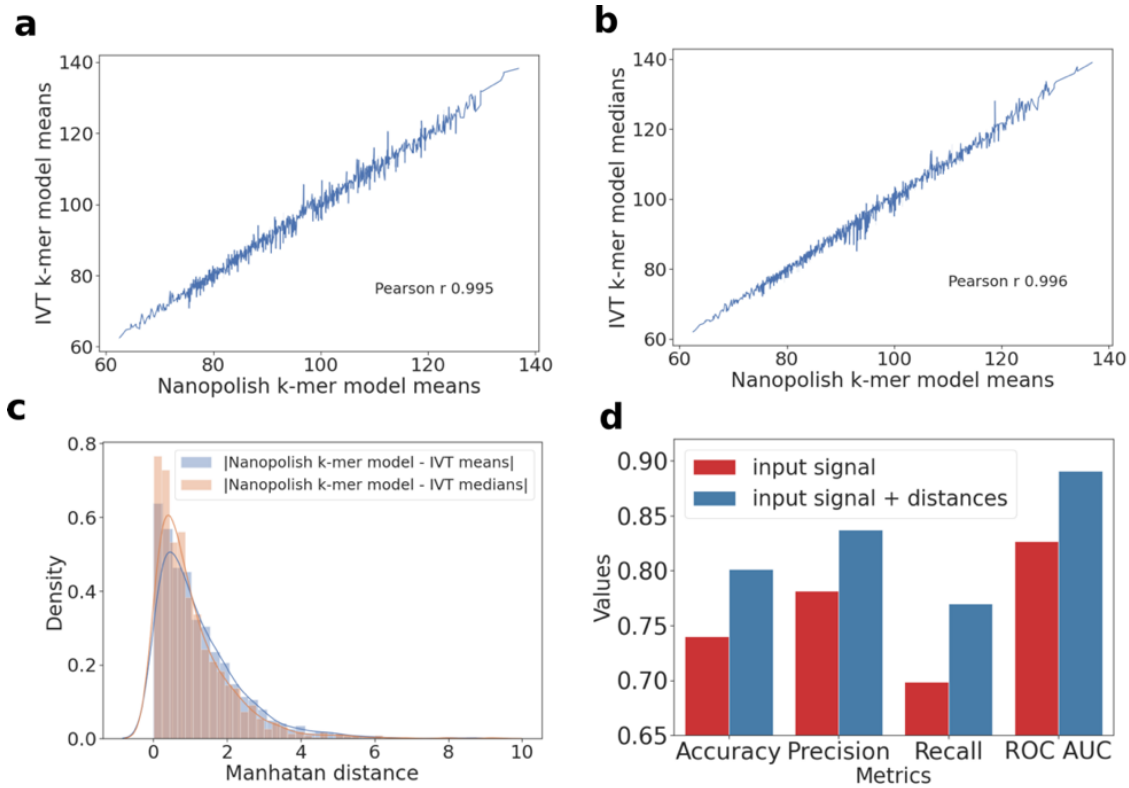
[6] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

[&] These authors contributed equally

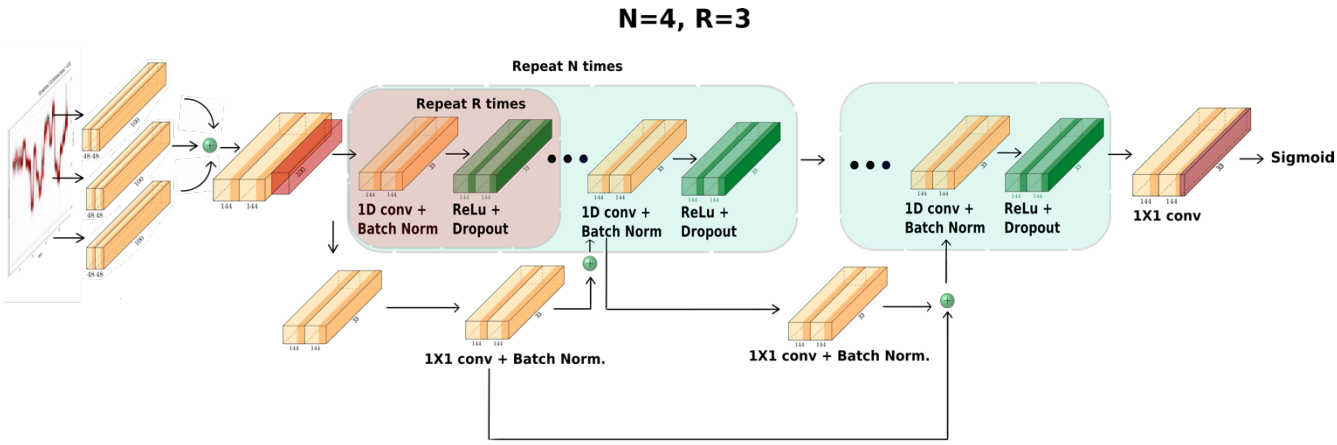
[*] Correspondence to: nikolay.shirokikh@anu.edu.au, eduardo.eyras@anu.edu.au



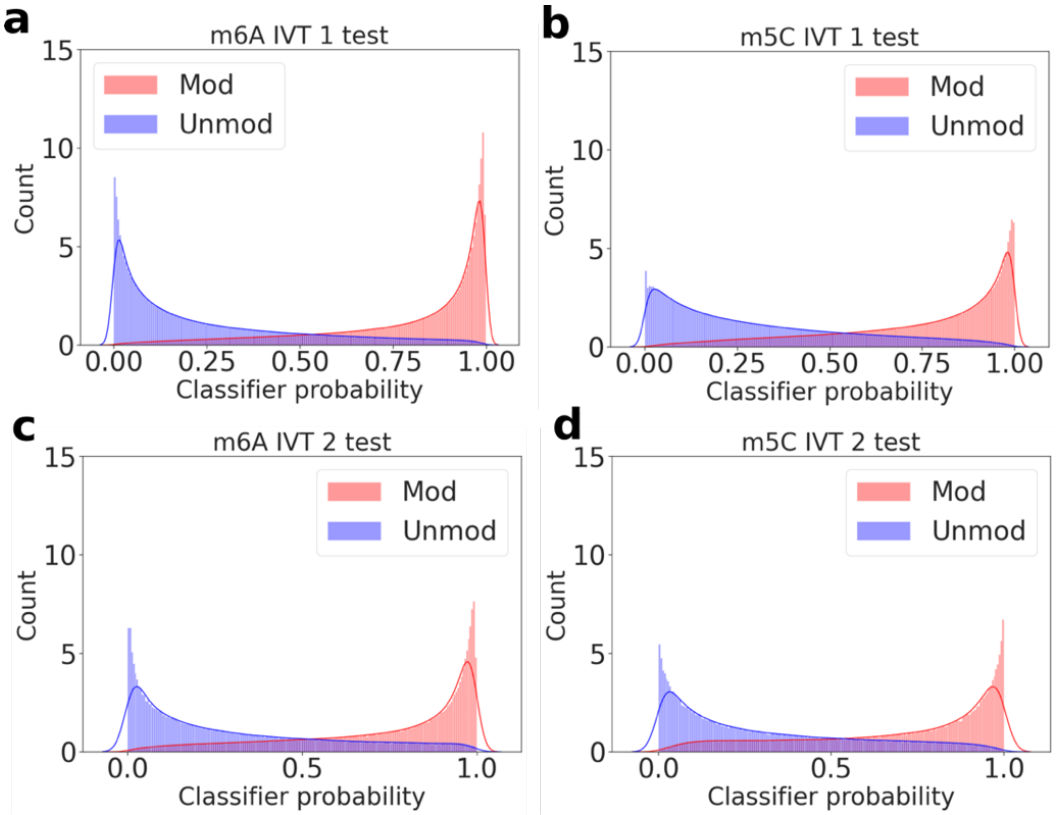
Supplementary Figure 1. **Nanopore DRS signals for sites harbouring modified and unmodified nucleotides. (a)** Characteristic profiles of multiple read signals from *in vitro* transcripts (IVTs) with N6-methyladenosine (m6A) (in red) or adenosine (in blue) (Unmod) located in the middle position. Shown are nanopore current signals from five consecutive nanopore translocation events, i.e., five overlapping 5-mers forming a fixed 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 its central nucleotide. Only the central nucleotide of each 5-mer is shown on the x axis. The line shows the mean values, and the shades indicate the standard deviation **(b)** Same as (a), but for read signals from IVTs containing 5-methylcytidine (m5C) (in red) and cytidine (in blue) (Unmod).



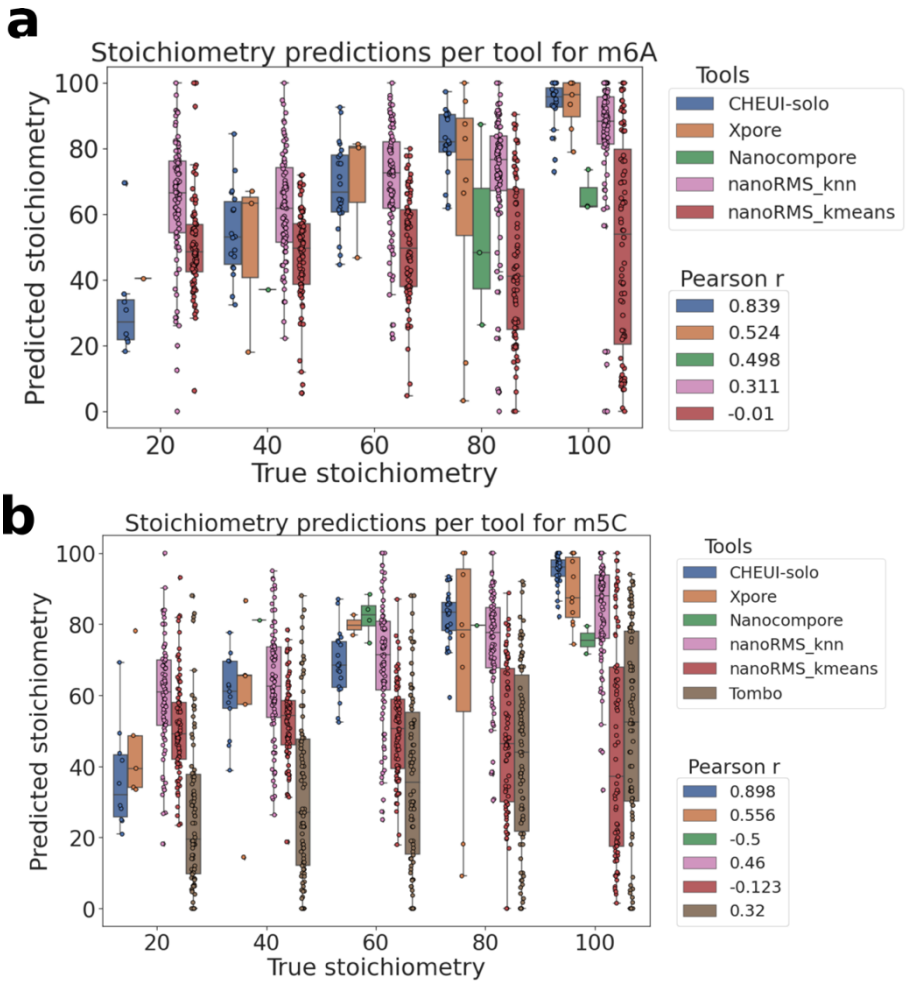
Supplementary Figure 2. Properties of the nanopore current signals and the effect of a distance vector in CHEUI-solo's metric performance. (a) Correlation of the values from the Nanopolish k-mer model (x-axis) (k=5) with the mean signal values calculated for the same k-mers (k=5) using sigsignalsncals from unmodified *in vitro* transcripts (IVTs) (y-axis). **(b)** Correlation of the values from the Nanopolish k-mer model (x-axis) (k=5) with the median signal values calculated for the same k-mers (k=5) using signals from unmodified IVTs (y-axis). **(c)** Distribution of the absolute differences (x-axis) between the values estimated from the unmodified IVT signals and those from the Nanopolish k-mer model. The line plot represents the fitting of the histogram to a density function. **(d)** Comparison of the performance of the CHEUI m6A model 1 with (blue) and without (red) using the distance vector as input. As described in the Methods section, the distance vector is calculated as the vector of absolute differences between the observed and Nanopolish k-mer model (k=5) values in each of the 100 components derived from the signals associated with each 9-mer in each individual read. CHEUI performance is assessed by the accuracy, precision, recall, and area under the receiver operating characteristic (ROC) curve (AUC) from the IVT test 1 set (x-axis).



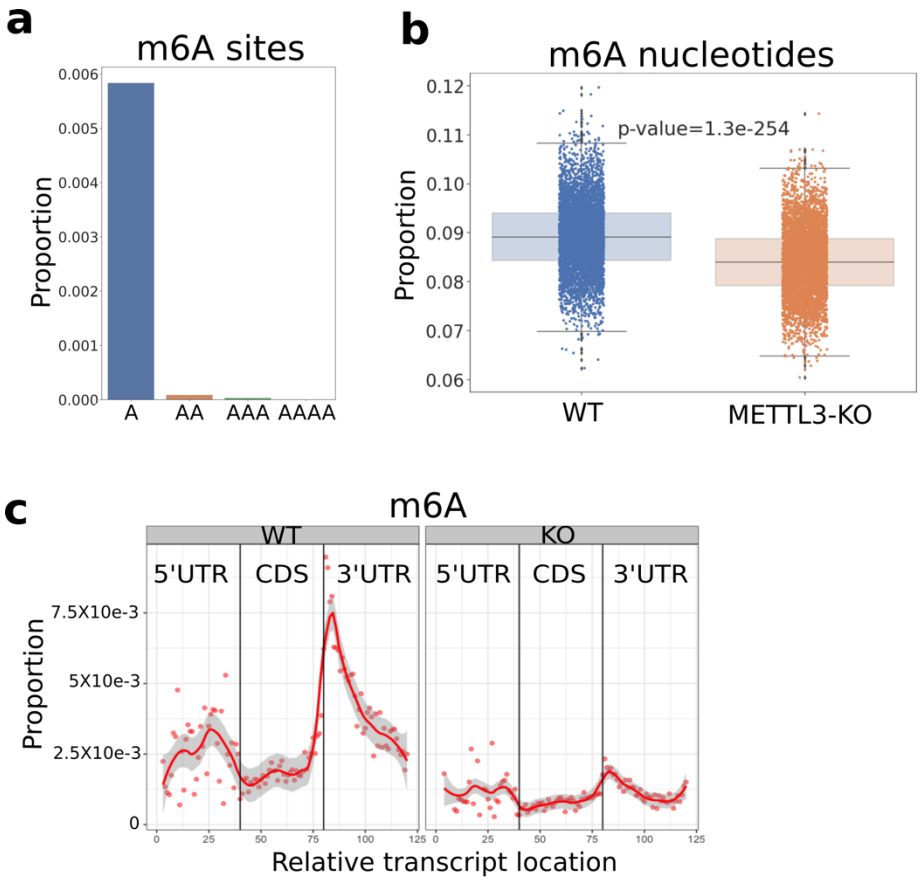
Supplementary Figure 3. Schematic view of CHEUI-solo neural network models. Both Model 1 and Model 2 in CHEUI-solo are deep convolutional neural networks (CNNs) defined as shown in the diagram and are based on the architecture employed in fast speech recognition algorithms [1]. The main body of the network is composed of 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.



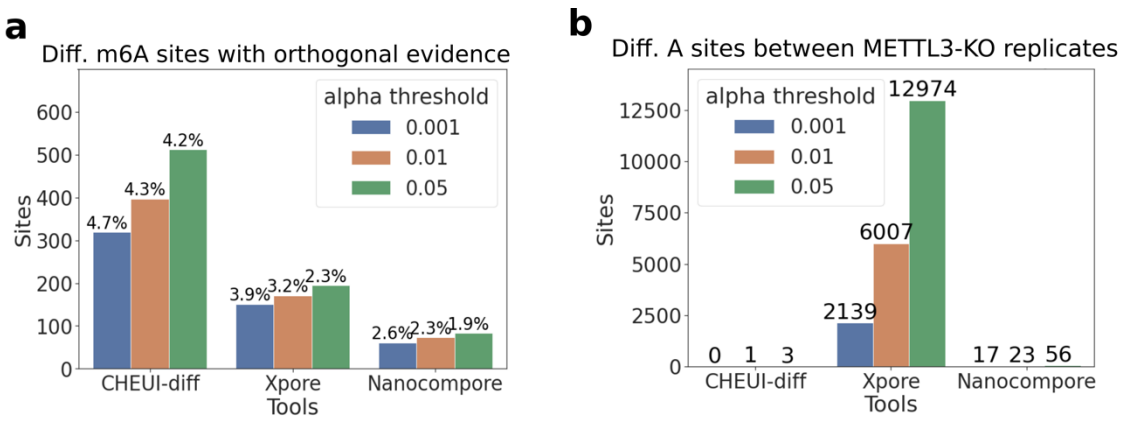
Supplementary Figure 4. Distribution of CHEUI's individual read probabilities. Shown are probability distributions returned by CHEUI-solo Model 1 for all the sites and all the reads in the IVT test 1 dataset, for the positive (Mod) and negative (Unmod) cases for m6A **(a)** and m5C **(b)**. **(c, d)**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



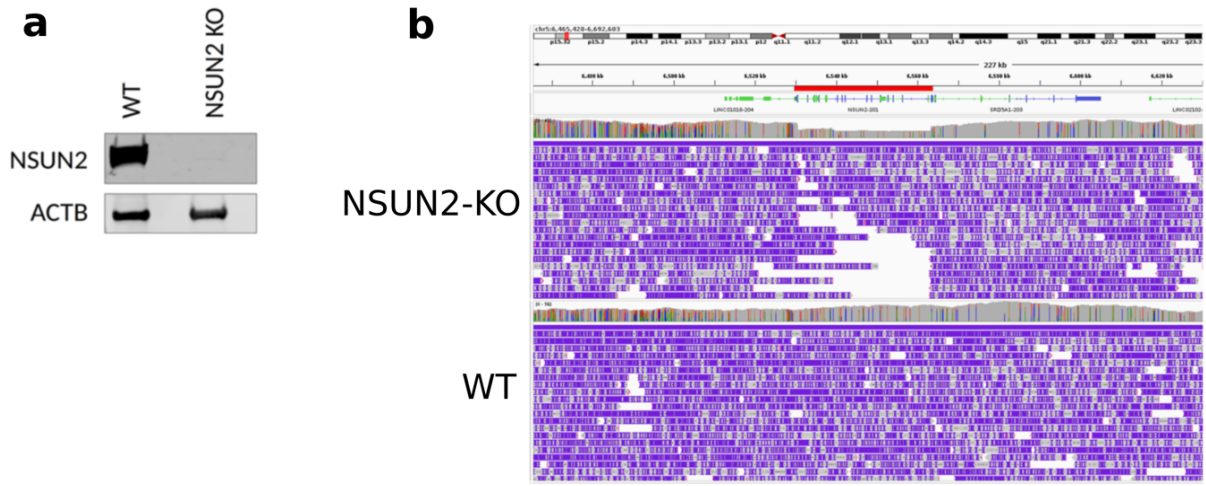
Supplementary Figure 5. Benchmarking the modification stoichiometry prediction by CHEUI in comparison to other available tools. We show the stoichiometry values predicted by each tool (y-axis) and the actual stoichiometry values of the controlled read mixtures (x-axis) for m6A **(a)** and for m5C **(b)**. Tool names and Person r values for the correlation with the true stoichiometry are given in the legends on the right.



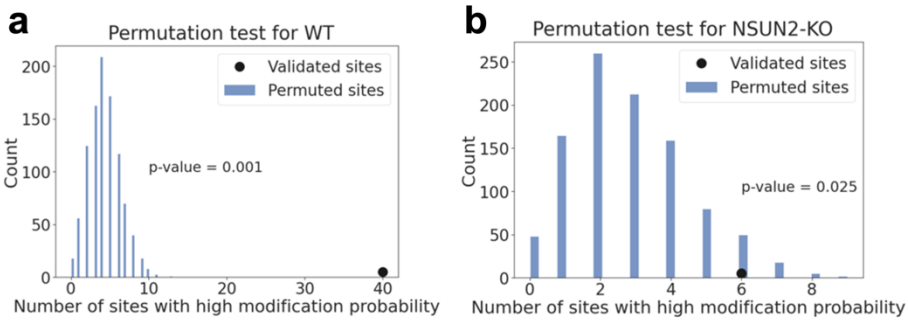
Supplementary Figure 6. CHEUI detects reduction of m6A in METTL3-KO cells. (a) We show the proportion of sites of the form A, AA, AAA, or AAAA, where all A positions were predicted as m6A. **(b)** Boxplots showing the proportions of modified A (m6A) over the total of A nucleotides (modified and unmodified) in the WT and METTL3-KO samples. Each dot indicates the proportion of individual nucleotides in individual reads predicted as modified by CHEUI-solo Model 1 (probability > 0.9), from a random pool of 5,000 individual nucleotides tested. The box plot represents the four quartiles of a distribution built from 5,000 random samples of those 5,000 nucleotides. The shaded boxes indicating the second and third quartiles, and the bottom and top bars representing the first and fourth quartiles. A non-parametric rank-sum test was used to test for statistical significance. **(c)** We show the proportion of significant transcriptomic sites over tested sites on mRNAs. Sites are mapped and rescaled into three equal-sized segments (each defined to be 40 positions) according to the region in the mRNA: 5' untranslated region (UTR), coding sequence (CDS), and 3' UTR. The proportions are shown for the WT sample (left panel) and the METTL3-KO sample (right panel).



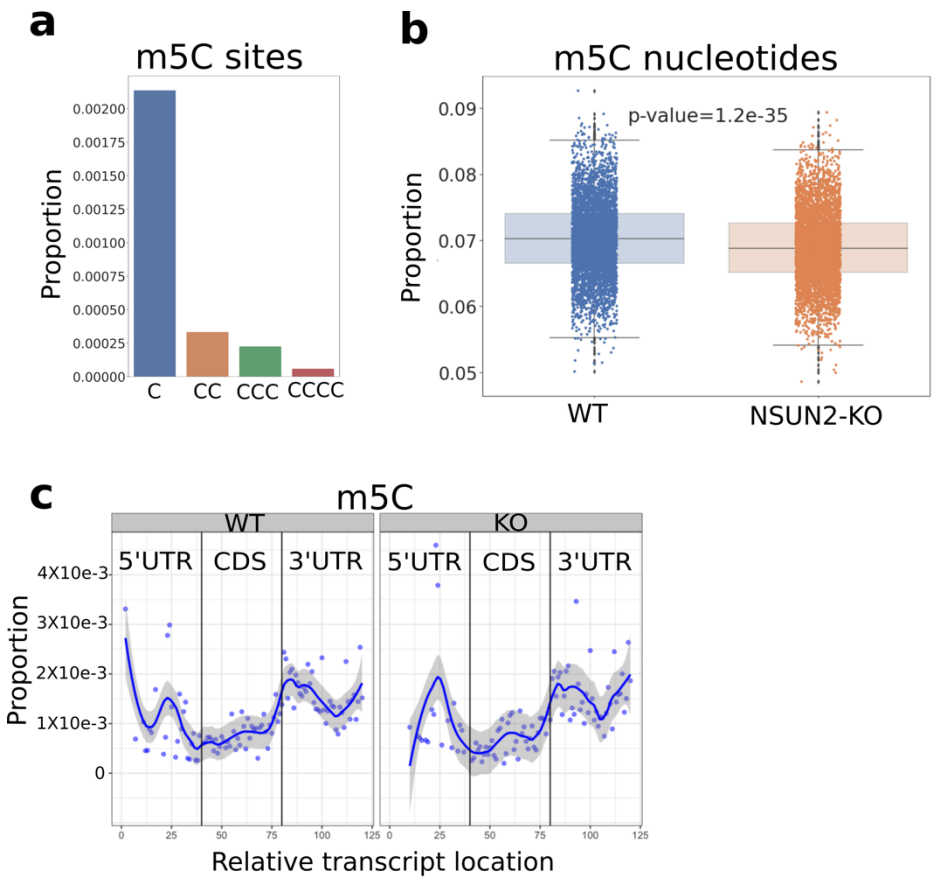
Supplementary Figure 7. Analysis of the differential modification site detection and quantification in METTL3-KO vs. wild-type cells by CHEUI in comparison with the other available tools. (a) The number of differentially modified m6A sites between HEK293 WT and METTL3-KO predicted by CHEUI-diff, Xpore 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.



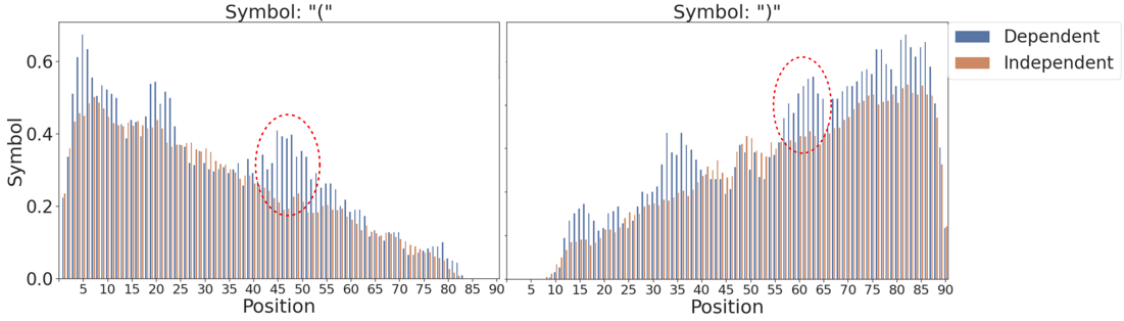
Supplementary Figure 8. Validation of the NSUN2 knock-out in HeLa cells. (a) Western blot with an NSUN2 antibody in WT and KO HeLa cells. Actin B (ACTB) signal is shown in each condition as a control. **(b)** IGV snapshot of the NSUN2 locus, indicating in red the region targeted for deletion. Shown are the Nanopore DNA sequencing reads for the NSUN2-KO sample (upper panel) and the WT sample (lower panel). The NSUN2-KO sample shows a decrease in coverage in the targeted region, with a possible incomplete KO in some of the clones sequenced.



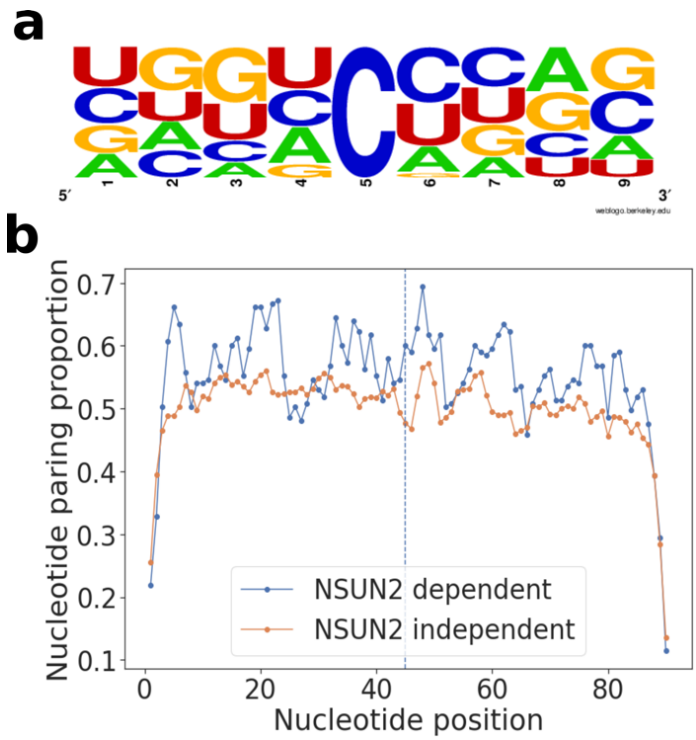
Supplementary Figure 9. Comparison of CHEUI m5C detections in wild-type and NSUN2-KO cells with calls derived from bisulfite RNA sequencing (bsRNA-seq). Bisulfite detection-based m5C sites are enriched in high CHEUI probabilities. The number of transcriptomic m5C sites predicted by bisulfite sequencing 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. Shown are the results using Nanopore signals from the WT cells **(a)** and from NSUN2-KO **(b)** HeLa cells.



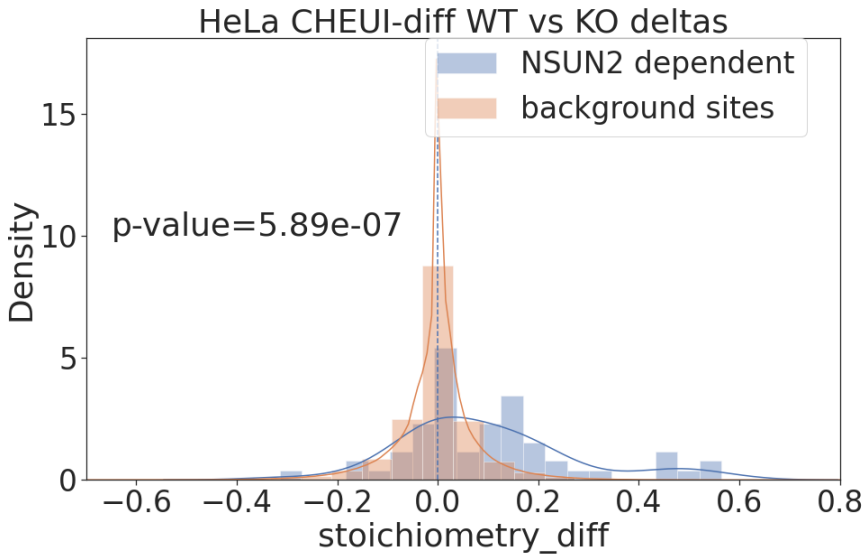
Supplementary Figure 10. CHEUI detects reduction of m5C in NSUN2-KO cells. (a) We show the proportion of sites of sites C, CC, CCC, or CCCC, where all C positions were predicted as m5C. **(b)** Boxplots showing the proportions of modified C nucleotides (m5C) over the total of C nucleotides (modified and unmodified) in the WT and NSUN2-KO samples. Each dot indicates the proportion of individual nucleotides in individual reads predicted as modified by CHEUI-solo Model 1 (probability > 0.9), from a random pool of 5,000 individual nucleotides tested. The box plot represents the four quartiles of a distribution built from 5,000 random samples of those 5,000 nucleotides. The shaded boxes indicating the second and third quartiles, and the bottom and top bars representing the first and fourth quartiles. Non-parametric rank sum test was used to test for the statistical significance. **(c)** We show the proportion of significant m5C transcriptomic sites over the total of tested sites on mRNAs. Sites are mapped and rescaled into three equal-sized segments (each defined to be 40 positions) according to the region in the mRNA: 5' untranslated region (UTR), coding sequence (CDS), and 3' UTR. The proportions are shown for the WT sample (left panel) and the NSUN2-KO sample (right panel).



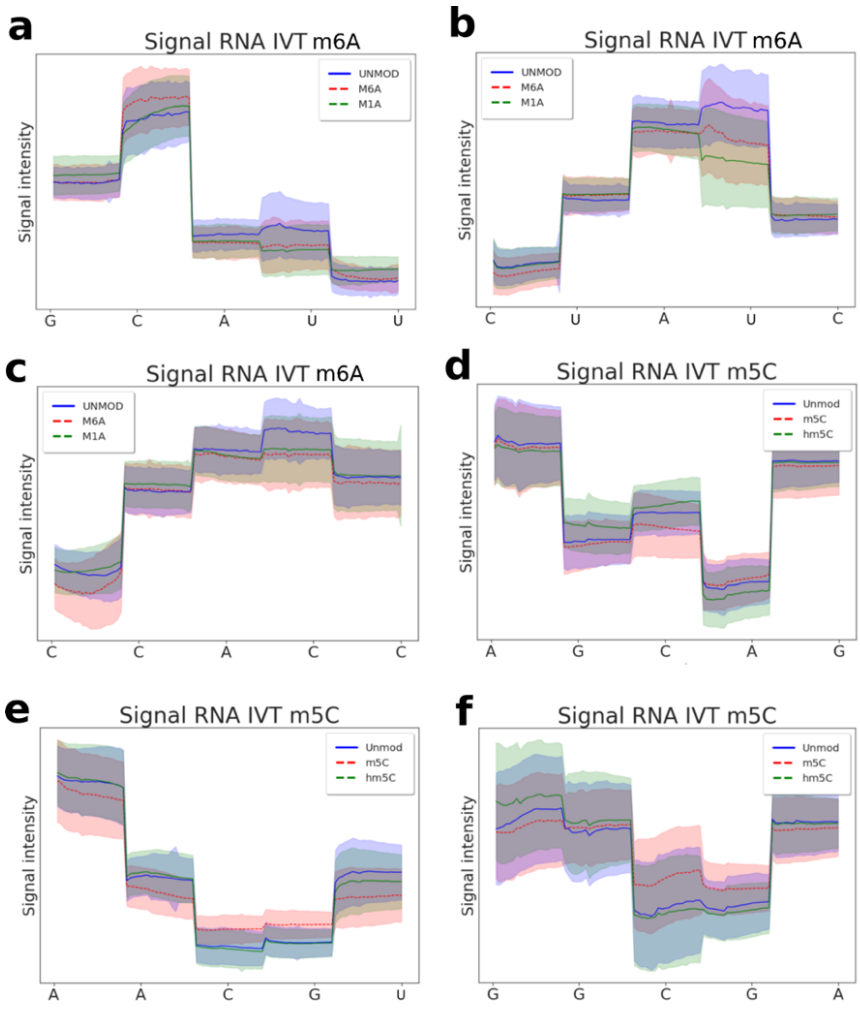
Supplementary Figure 11. Overview of predicted RNA secondary structure formation propensity associated with the m5C sites detected by CHEUI. Shown is the proportion of paired bases, indicated with the symbol '(' (left panel) or ')' (right panel) in the predicted secondary structures using 90 nt around (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.



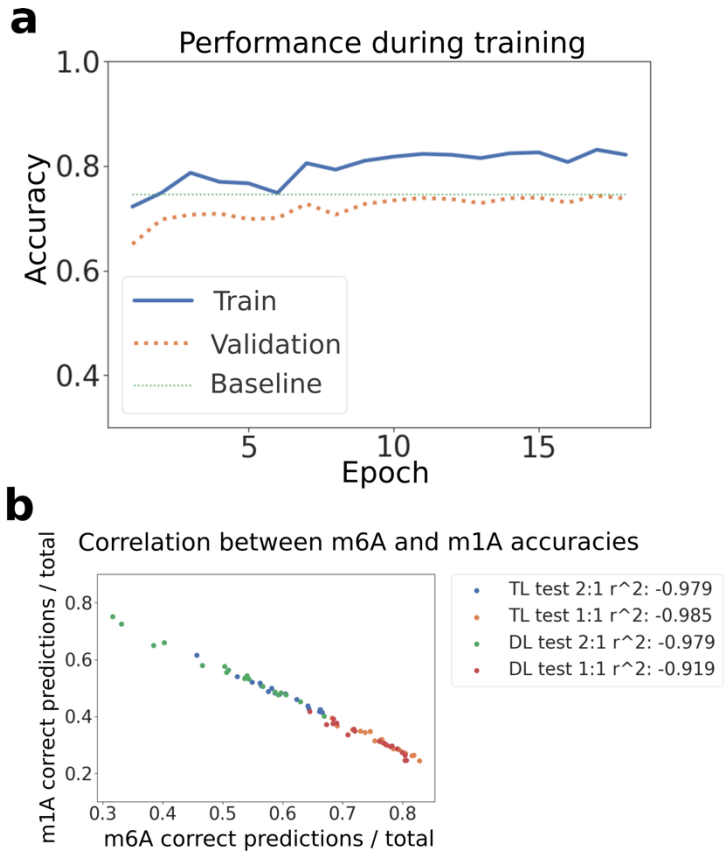
Supplementary Figure 12. Sequence motifs and RNA secondary structure predictions in m5C sites. (a) Sequence motifs for NSUN2-independent sites defined as significant sites by CHEUI-solo in the NSUN2-KO sample (1,841 transcriptomic sites). **(b)** Proportion of base-paired positions along 90 nt centered 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.



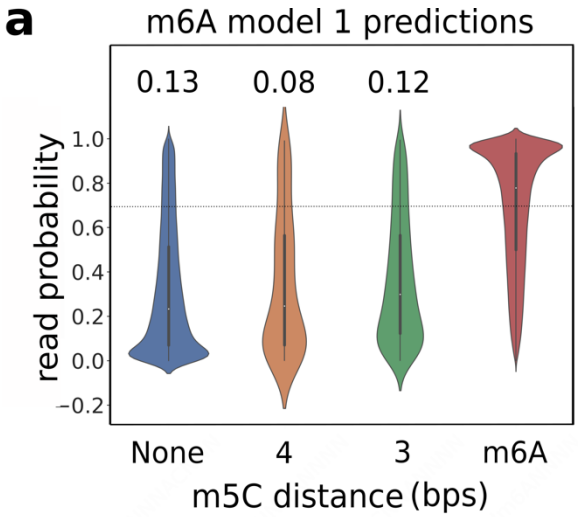
Supplementary Figure 13. Distribution of the stoichiometry difference reported by CHEUI-diff in NSUN2-dependent and background sites. Blue bars indicate the stoichiometry difference between WT and NSUN2-KO samples in sites that were reported previously to be NSUN2-dependent by Huang et al. 2019. Orange bars indicate the stoichiometry difference between WT and NSUN2-KO sites that were not in the NSUN2-dependent list. A non-parametric test using rank-sum was performed to compare the two distributions.



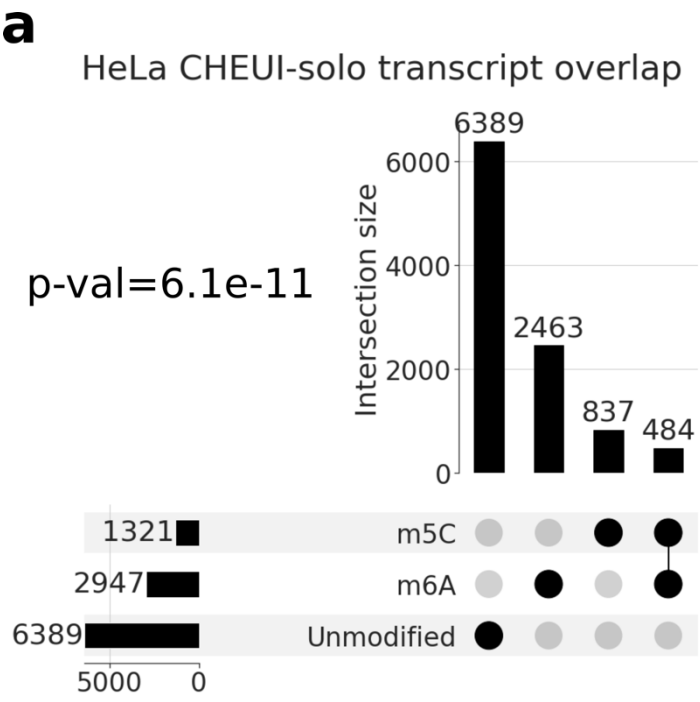
Supplementary Figure 14. Visualization of nanopore current signals for sites harbouring different A and C modifications against non-modified sites. Nanopore current signals from five consecutive nanopore translocation events, i.e., five overlapping 5-mers forming a fixed 9-mer. The signal from each overlapping 5-mer is scaled over the central nucleotide of that 5-mer to twenty values and plotted atop the position of its central nucleotide. Only the central nucleotide of each 5-mer is shown on the x axis. The line shows the mean values, and the shades indicate the standard deviation. Three different sequence contexts are shown exemplifying typical signal profiles for A (Unmod), m6A, and m1A (**a**, **b**, **c);** and for C (Unmod), m5C, and hm5C **(d, e, f)**.



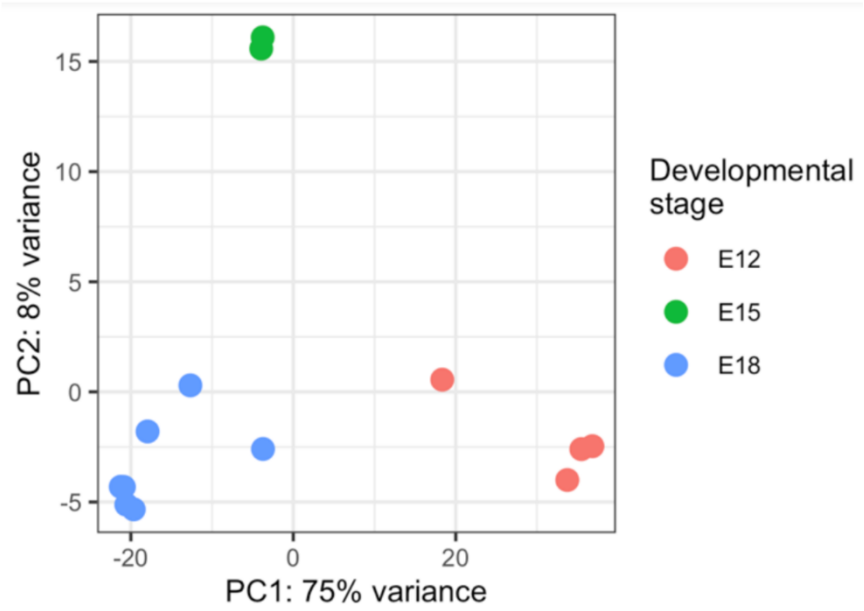
Supplementary Figure 15. Accuracy metrics returned by CHEUI-solo Model 1 upon retraining with other modifications as negatives. We retrained CHEUI-solo using the signals for m1A and all other modifications different from m6A as part of the negative set for the training. **(a)** Accuracy (y axis) of CHEUI 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).



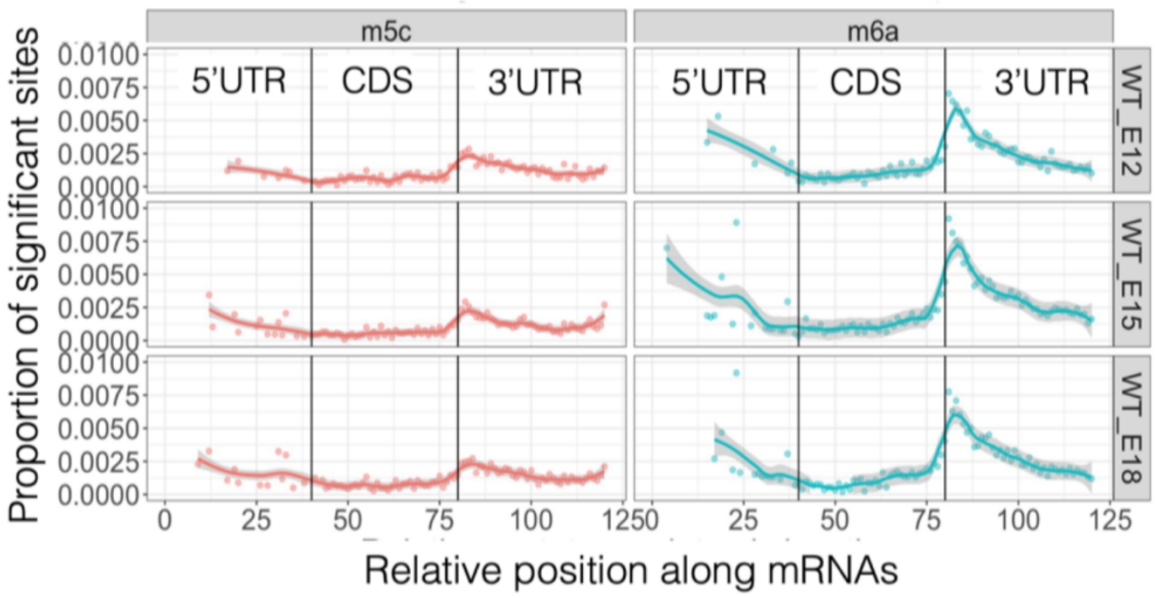
Supplementary Figure 16. CHEUI's prediction probability of m6A at individual read level. IVT2 read signals at 9-mers were centered at A (NNNNANNNN) with various configurations: 9-mers with no m6A (None), 9-mers with m5C present 2 or 4 nucleotides from the central A, and 9-mers with a modified middle A (m6A). We could not find any 9-mers NNNNANNNN in the IVT test 2 dataset with only one m5C at distances 1 or 2 from the middle A.



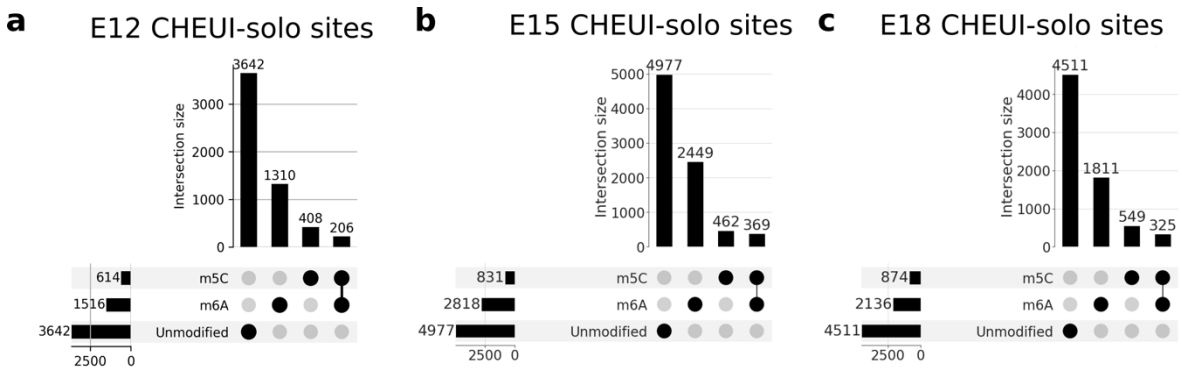
Supplementary Figure 17. The coincidence of m6A and m5C transcript sites in HeLa RNAs detected by CHEUI-solo. UpsSet plot for the number of transcripts containing m6A and m5C sites, transcripts containing only one of the modifications, and transcripts containing none. The Fisher's test p-value shows the association of m6A and m5C sites relative to all the tested sites. Only cases where m6A and m5C are at distance 5 or larger were considered.



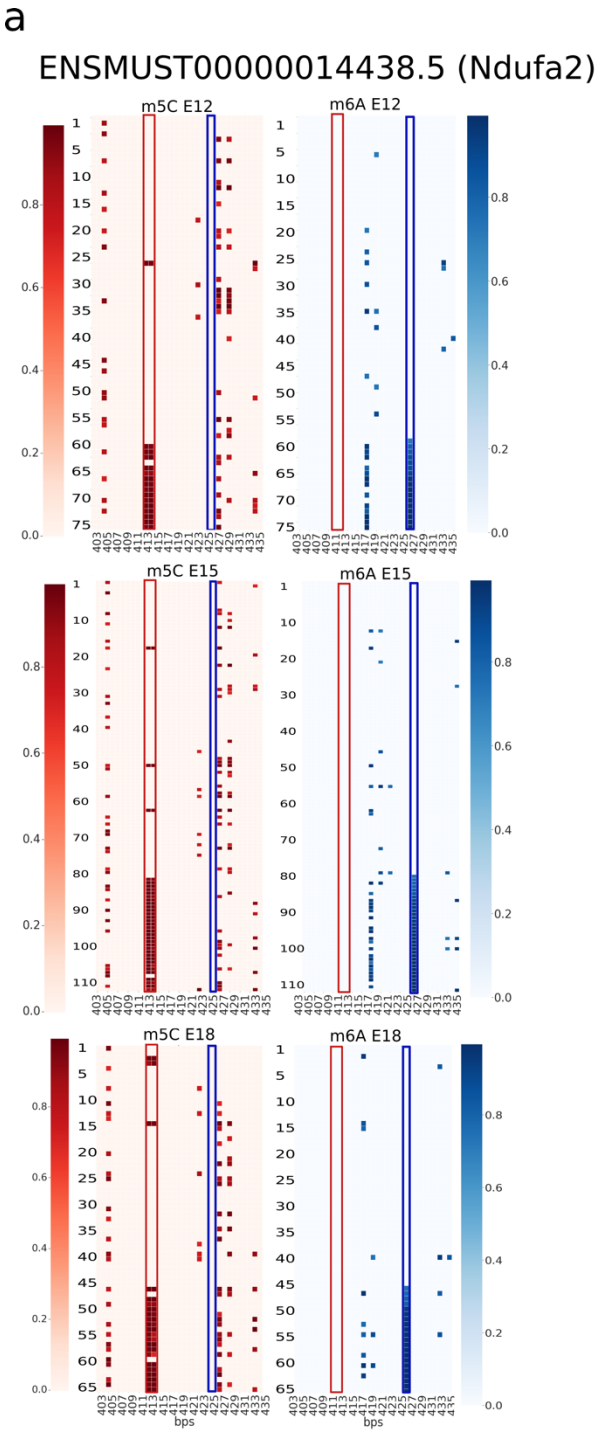
Supplementary Figure 18. Assessment of nanopore DRS replicates of developing mouse pre-frontal cortex. Principal Component Analysis (PCA) plot of the samples sequenced at stages E12, E15 and E18. The PCA plot was generated using log-normalised gene counts from DESeq2.



Supplementary Figure 19. Methylation profiles of mRNAs across the three stages of mouse embryonic pre-frontal cortex development reported by CHEUI-solo. Shown are the proportion of significant m5C (left panels) and m6A (right panels) transcriptomic sites over the total of tested sites on mRNAs. Sites are separated into three equal sized segments (of length 40nt) according to the region in the mRNA: 5' untranslated region (UTR), coding sequence (CDS), and 3' UTR. The proportions are shown for the E12, E15, and E18 pre-frontal embryonic mouse samples.



Supplementary Figure 20. CHEUI-solo identifies co-occurrence of m6A and m5C sites on mRNAs in mouse embryonic pre-frontal cortex. UpSet plot for the number of protein-coding transcripts from the mouse Ensembl v104 (GRCm39) annotation containing m6A and m5C sites, only one of the modifications, or none, in the three studied embryonic stages, E12, E15, and E18. Only cases with m6A and m5C at a distance of 5 or larger were considered. In all three cases, the observed co-occurrence of m6A and m5C was significantly higher than expected (Fisher's exact 2-sided test p-value < 0.000150).



Supplementary Figure 21. Conservation of m6A and m5C co-occurrence in mouse embryonic pre-frontal cortex across developmental time-points. We show m6A and m5C modifications co-occurring in the same RNA reads at 13nt of relative distance. Read IDs are represented with numbers (y-axis) and the position along the transcript is represented on the x-axis. The color scales represent CHEUI's detection probabilities at the read level for m6A (in blue) and m5C (in red). Only predictions with probability > 0.7 are shown in each read. The upper panel shows the m6A and m5C modifications in E12, the middle panel in E15, and the lower in E18.

References

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