Transferable deep learning enables identification of multiple types of RNA modifications using nanopore direct RNA sequencing

You Wu[1], Wenna Shao[1], Mengxiao Yan[2], Yuqin Wang[2], Xiaofei Li[1], Brian D. Gregory[3], Jun Yang[2,4,*], Hongxia Wang[2,4,*], Xiang Yu[1,*]

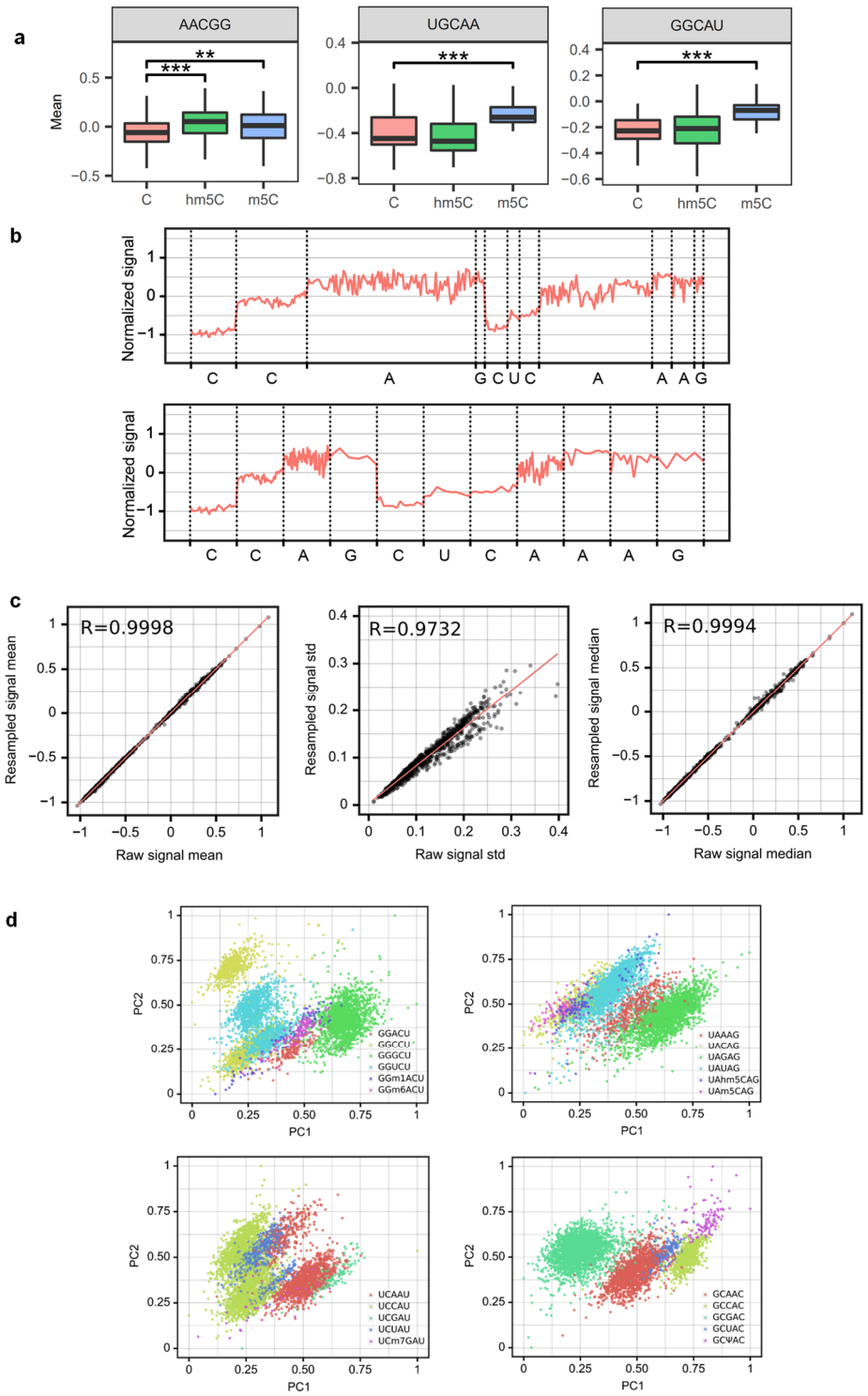
[1]Joint International Research Laboratory of Metabolic & Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

[2]Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Shanghai 201602, China

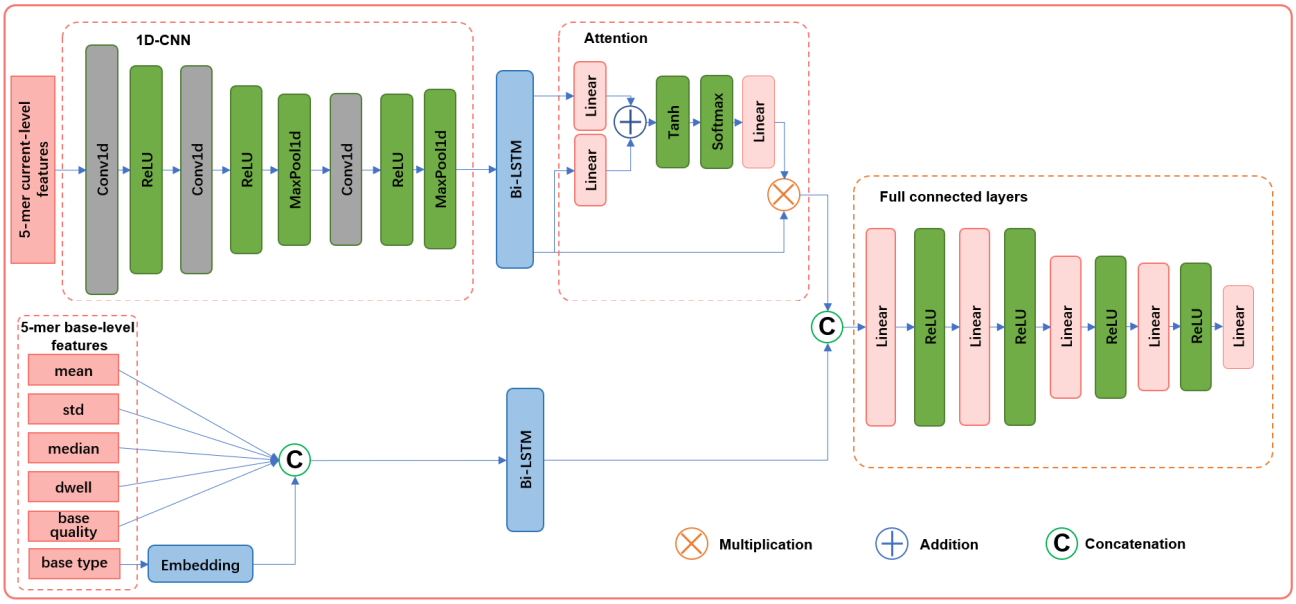
[3]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]Chenshan Scientific Research Center of CAS Center for Excellence in Molecular Plant Sciences, Shanghai 201602, China

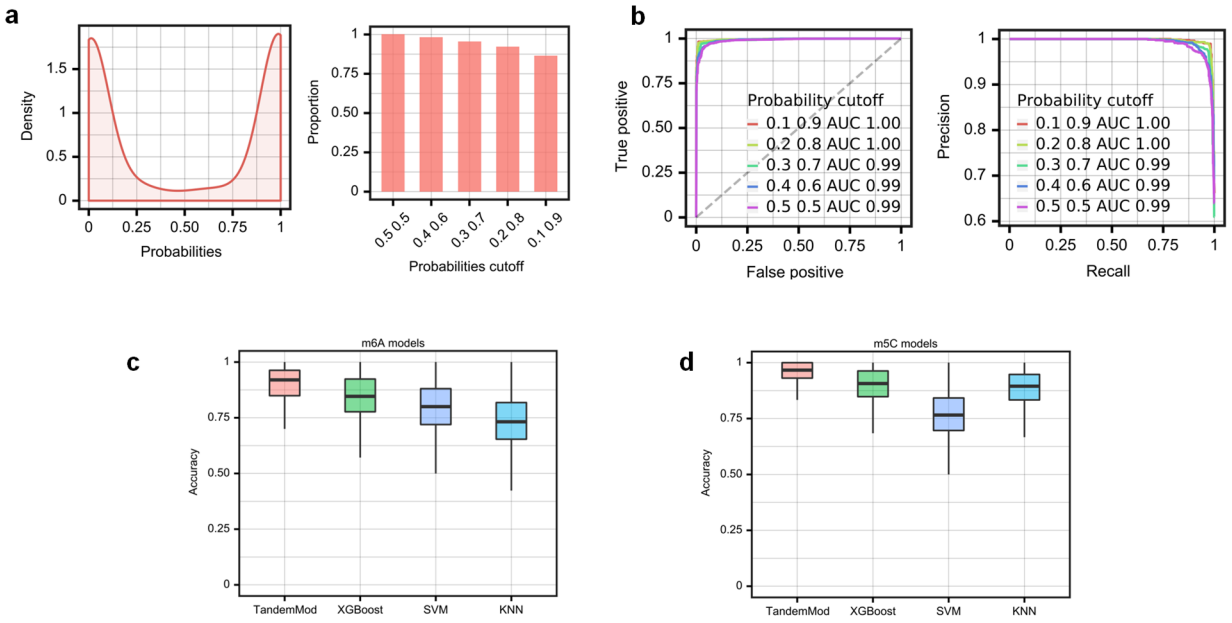
*Corresponding author: Xiang Yu, yuxiang2021@sjtu.edu.cn; Hongxia Wang, hxwang@cemps.ac.cn; Jun Yang, jyang03@cemps.ac.cn.



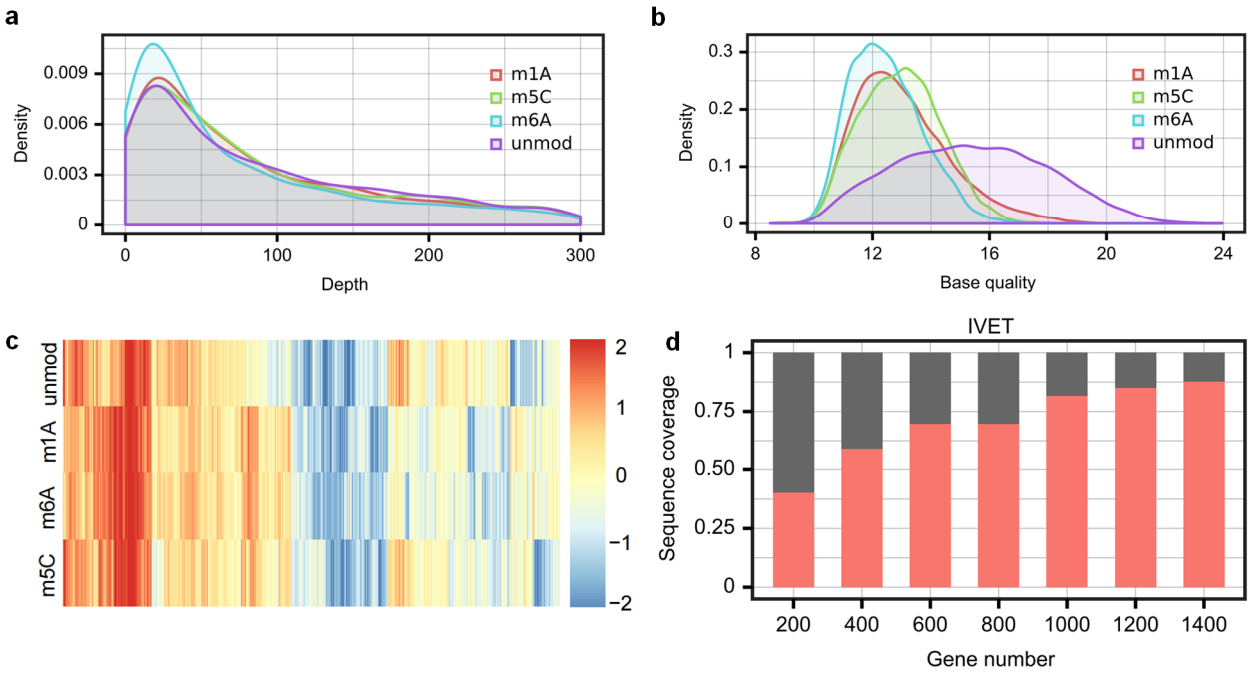
Supplementary Fig. 1: DRS features at base level and current level affected by 5-mer sequence contexts. a, Mean signatures of current signals at C, hm^{5}C and m^{5}C under different 5-mer sequence contexts. **b,** An example showing normalized current signals with raw signal width (top panel), and resampled current signals with equal width (bottom panel). **c,** Scatterplot of base-level features including mean, std and median between original signals and resampled signals. R value indicates the correlation coefficient. **d,** Scatter plots displaying the difference between the modified bases and the four canonical bases after PCA transform, with the two bases before and after the target base being the same.



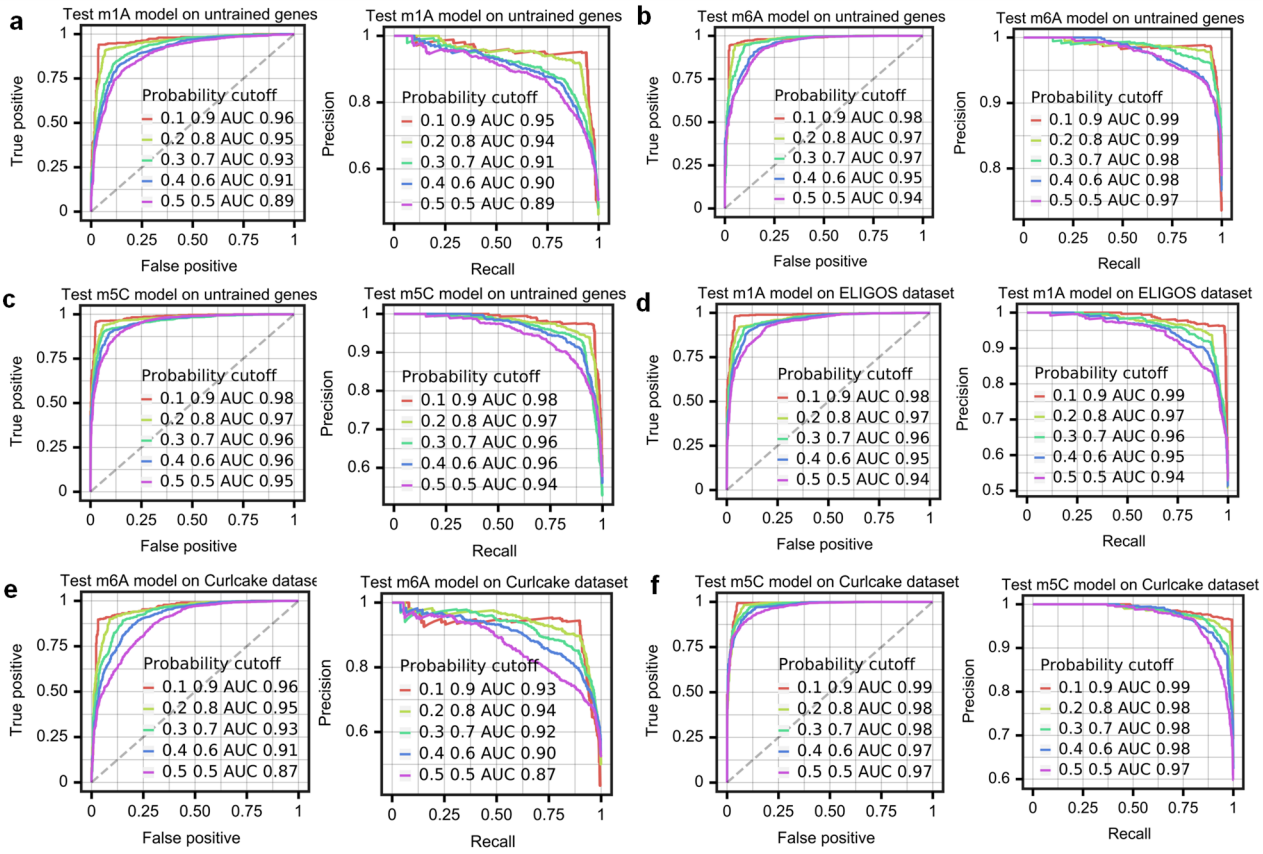
Supplementary Fig. 2: Detailed architecture of TandemMod model. TandemMod consists of 4 main components: a one-dimensional convolutional neural network(1D-CNN), a bi-directional long short-term memory (bi-LSTM) module, an attention mechanism and full-connected layers. The top layers of the model act as feature extractor that learns modification-related information and the bottom layers act as classifier that predict modification type.



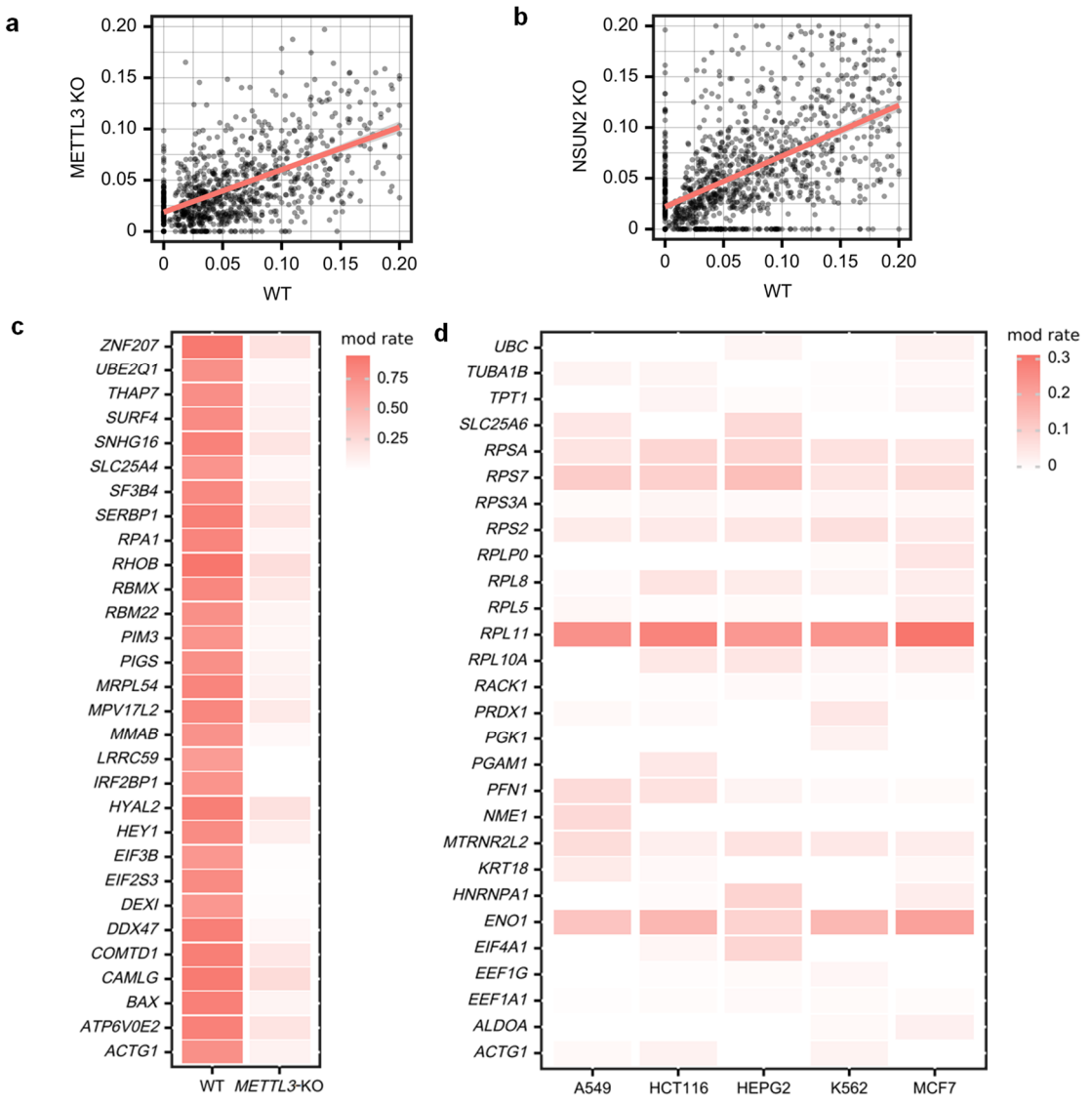
Supplementary Fig. 3: Performance evaluation of TandemMod. a, The distribution of predicted modification probabilities by the TandemMod model on the Curlcake testing set (left panel) and the relationship between the remaining reads and the cutoff probabilities (right panel). **b,** Performance for m^5C identification model on Curlcake dataset using different cutoff values. **c,** Comparison of models trained on Curlcake m^5C dataset and tested on Curlcake testing set. **d,** Comparison of models trained on Curlcake m^6A dataset and tested on Curlcake testing set.



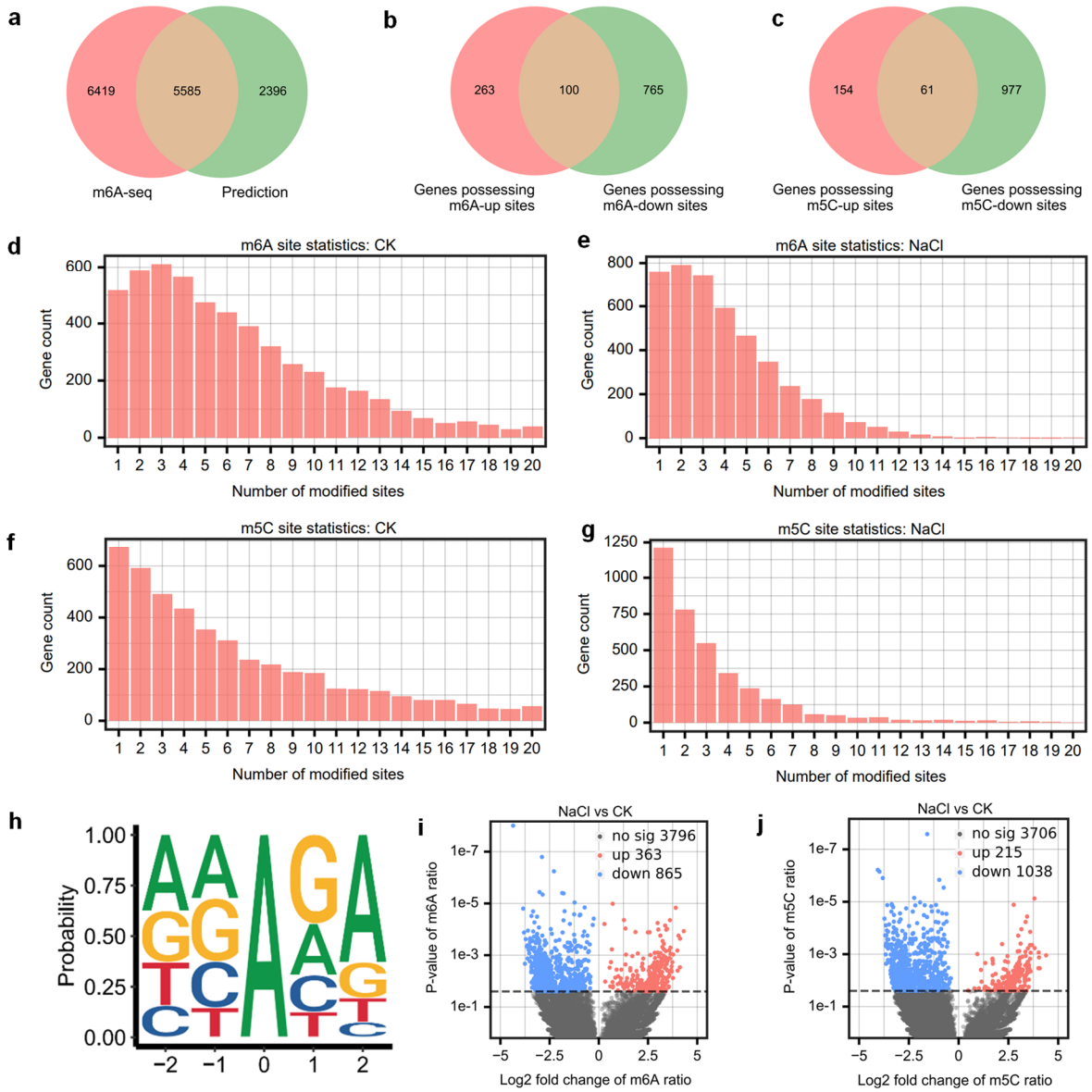
Supplementary Fig. 4: Statistics of direct RNA sequencing results from the IVET datasets. a, The sequencing depth of the four IVET datasets. **b,** The mean base quality of the four IVET datasets. **c**, Heatmap displaying the transcribed mRNA abundance across the four IVET samples. **d**, The barplot showing the 9-mer sequence coverage across different numbers of genes in the IVET dataset.



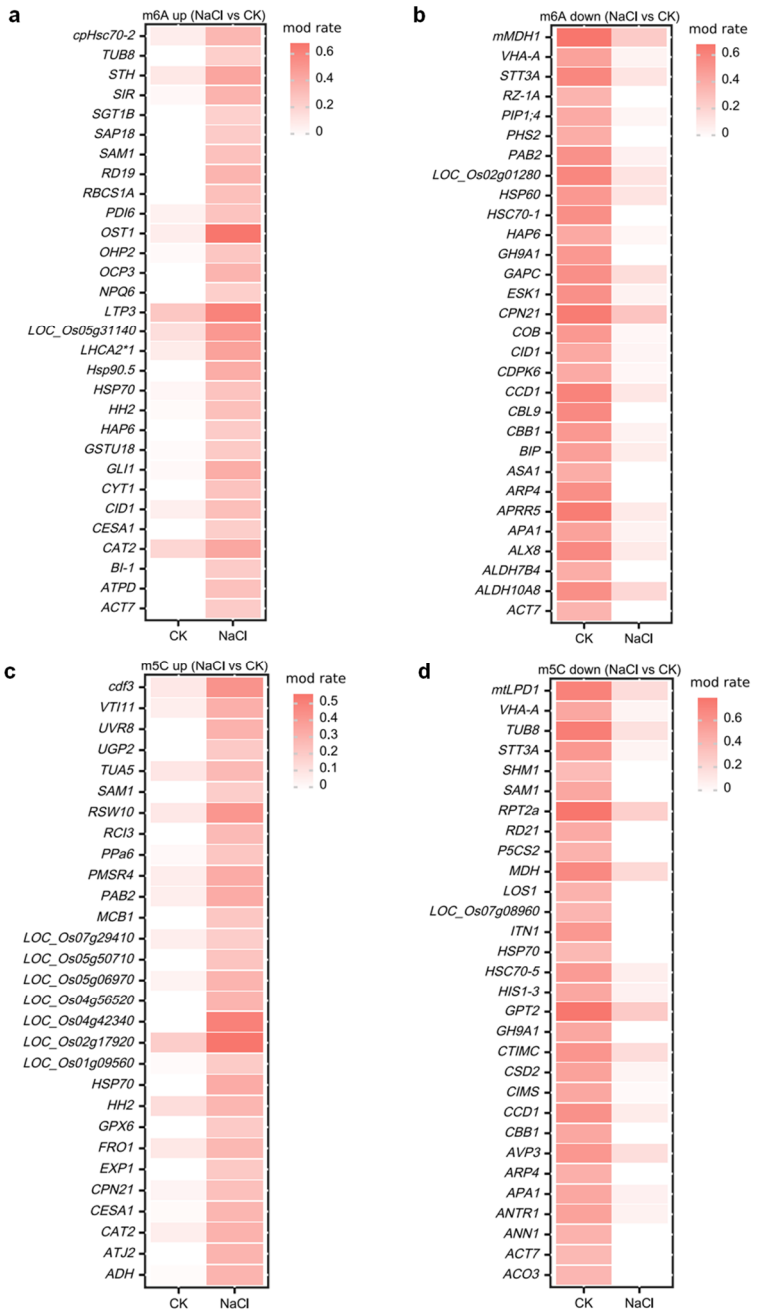
Supplementary Fig. 5: Performance evaluation of TandemMod models trained on IVET datasets. a-c, Performance evaluation of the m^1A model (**a**), m^6A model (**b**) and m^5C model (**c**) trained on the IVET training sets and tested on the IVET independent genes. **d,** Performance evaluation of the m^1A model trained on the IVET training set and tested on ELIGOS dataset. **e,** Performance evaluation of the m^6A model trained on the IVET training set and tested on the Curlcake dataset. **f,** Performance evaluation of the m^5C model trained on the IVET training set and tested on the Curlcake dataset.



Supplementary Fig. 6: Identification of m^{6}A sites across different tissues in human. a, Scatter plot showing decreased m^{6}A level in *METTL3*-KO HEK293T sample compared to WT sample. **b**, Scatter plot showing decreased m^{5}C level in *NSUN2*-KO HeLa sample compared to WT sample. **c**, Heatmap of top 30 genes with differentially m^{6}A-modified rates between *METTL3*-KO and WT samples. **d**, Heatmap showing m^{6}A rates of randomly selected genes across 5 human cell lines.



Supplementary Fig. 7: Identification of m^{6}A and m^{5}C sites across different conditions in rice. a, Venn diagram showing gene with TandemMod-predicted m^{6}A sites and genes detected by m^{6}A-seq. **b**, Venn diagram of genes containing m^{6}A sites with increased and decreased modification rate. **c**, Venn diagram showing of genes containing m^{5}C sites with increased and decreased modification rate. **d-e,** Statistics of genes possessing different number of m^{6}A-modified sites in CK (**d**) and NaCl-treated sample (**e**). **f-g,** Statistics of genes possessing different number of m^{5}C-modified sites in CK (**f**) and NaCl-treated sample (**g**). **h**, Sequence motif of m^{6}A-modified sites enriched in rice control sample in addition to DRACH. **i**, Volcano plot showing differentially m^{6}A-modified genes. **j**, Volcano plot showing differentially m^{5}C-modified genes. Chi-square test was performed and sites with Chi-square p-value less than 0.05 were considered as differentially modified.



Supplementary Fig. 8: Differentially m^{6}A-modified and m^{5}C-modified genes identified by TandemMod in NaCl-treated rice as compared to the control. a, Genes with increased m^{6}A modification rate in NaCl-treated sample. **b**, Genes with decreased m^{6}A modification rate in NaCl-treated sample. **c**, Genes with increased m^{5}C modification rate in NaCl-treated sample. **d**, Genes with decreased m^{5}C modification rate in NaCl-treated sample.