Single-Molecule Direct RNA Sequencing Reveals the Shaping of Epitranscriptome Across Multiple Species

Ying-Yuan Xie[1,3], Zhen-Dong Zhong[1,3], Hong-Xuan Chen[1,3], Ye-Lin Lan[1], Fu Wu[1], Jin-Wen Kong[1], Ru-Jia Luo[1], Biao-Di Liu[1], Yang Shu[1], Feng Yin[2], Zigang Li[2], Zhang Zhang[1], Guan-Zheng Luo[1,2,*]

1 State Key Laboratory of Biocontrol, MOE Key Laboratory of Gene Function and Regulation, Guangdong Province Key Laboratory of Pharmaceutical Functional Genes, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

2 Pingshan Translational Medicine Center, Shenzhen Bay Laboratory, Shenzhen 518118, China

3 These authors contributed equally

* Correspondence to: luogzh5@mail.sysu.edu.cn

Abstract

The significance of RNA modification in gene regulation has been widely recognized. To understand the transcriptome-wide landscape and its underlying mechanisms, prevailing mapping strategies have been developed. However, these short-reads based techniques are primarily focused at the gene level, overlooking the nature of RNA as multiple copies within one cell. Third-generation sequencing (TGS) platforms provide direct RNA sequencing at the resolution of individual RNA molecules, offering the promise of detecting RNA modifications and RNA processing events simultaneously. In this study, we introduce SingleMod, a deep learning model tailored for the precise mapping of m6A modifications within single RNA molecules using nanopore direct RNA sequencing (DRS). We systematically dissect the transcriptome-wide m6A profiles in single-molecule and single-base resolution, refining our understanding of the genomics of m6A and revealing an additive mode through which m6A shapes the epitranscriptome. Through comparative analyses across diverse species, we identify three distinct m6A distribution patterns and unveil an exclusion-inclusion deposition mode that governs m6A biogenesis. Furthermore, we introduce a unified quantitative model that delineates this dual-mode in various species. This study pioneers single-molecule m6A landscape exploration across multiple species, deepening our understanding of m6A, including its genomics, biogenesis, mechanisms, and biological implications.

Introduction

In recent years, extensive research has highlighted m6A as the predominant RNA modification in eukaryotic mRNA[1, 2], exerting substantial influence over a broad spectrum of physiological and pathological processes due to its central role in regulating gene expression[3-5]. With the advent of Next-Generation Sequencing (NGS) platforms, advanced mapping techniques have been devised to reveal m6A landscapes across various tissues and cell types[6]. These studies have unveiled key

deposition patterns of m6A, notably a pronounced enrichment surrounding the stop codon and the presence of the consensus sequence DRACH[7, 8]. Nevertheless, the sequence specificity alone cannot explain the observed selectivity in m6A distribution, as m6A modifications are only found within a relatively small fraction of the available DRACH motifs, accounting for approximately 5%[8]. While the significance of m6A in eukaryotes has long been acknowledged, a systematic cross-species comparison of m6A landscapes across varying evolutionary distances is still lacking. Furthermore, the fundamental questions of whether a unified distribution pattern exists and if there are conserved mechanisms underpinning this pattern remain open in the field.

The distinctive distribution pattern of m6A has stimulated a widespread exploration of the underlying mechanisms and potential biological implications[9]. As a native decoration on RNA molecules, the m6A modifications are directly or indirectly involved in complex RNA processing events[10], which can influence the m6A biogenesis mechanism and shape its distribution pattern; conversely, these events can also be regulated by the deposition of m6A at specific positions. Recent studies have revealed a correlation between m6A biogenesis and the splicing machinery, proposing that the exon junction complex (EJC) may impede m6A deposition in nearby regions[11-13]. However, a comprehensive understanding of how this proposed mechanism contributes to observed distribution patterns, and whether it holds true across different species, remains largely unexplored. Given the potential variability in RNA processing and modifications at the individual RNA molecule level, the comprehensive characterization of m6A at the single-molecule resolution becomes indispensable. Such an approach is crucial for establishing precise patterns of m6A distribution and facilitating a deeper exploration of the intricate interplay between m6A modification and RNA processing, thereby providing insights into the underlying biogenesis mechanisms and shedding light on how RNA modifications regulate gene expression and influence the destiny of RNA molecules.

Presently, the predominant high-throughput sequencing platform in use is NGS[7, 8, 14-18], where the substrates to measure are short DNA fragments typically spanning a few hundred base pairs (bp). NGS-based RNA sequencing involves several essential steps, such as fragmentation, reverse transcription, and the subsequent alignment of sequencing reads to the genome. In this process, sequencing reads lose rich information about the original RNA molecule, including RNA modifications and the architecture of full-length spliced RNA. Given the profound heterogeneity observed among RNA molecules, conventional NGS-based techniques for mapping modifications have inherent limitations in fully unraveling the intricacies of the epitranscriptome at the molecular level. Therefore, there arises a compelling need for a more advanced approach, one that allows for the mapping of RNA modifications at the level of individual RNA molecules while concurrently preserving the associated RNA processing signatures. Such an approach would be essential for achieving a comprehensive and detailed understanding of the epitranscriptome.

The direct RNA sequencing (DRS) technique provide by Oxford Nanopore Technology (ONT) platform has revolutionized the field by enabling the sequencing of the native RNA molecule without the need for fragmentation and amplification[19, 20]. This technology holds immense potential for the detection of RNA modifications, particularly the most prevalent modification type m6A, within full-length RNA molecule[21]. However, most existing computational tools designed for extracting RNA modification information from DRS data determine the modification status (yes or no) for specific genomic sites, yet lack the ability to ascertain whether each RNA molecule corresponding to a site is indeed modified[21]. This limitation arises primarily due to the absence of standard training data, particularly information concerning the modification status of RNA molecules at specific m6A sites[22]. Existing tools like Epinano[23], nanom6A[24] and DENA[25] rely on artificially synthesized data or manually selected data for training, while MINES[26] and m6Anet[27] employ non-standard and non-quantitative site-level labels, resulting in reduced accuracy and limited generalization capability[21]. Consequently, there remains a significant gap in our ability to generate comprehensive single-molecule profiles of the transcriptome, which encompass precise m6A modifications and RNA processing signatures.

Here we introduce SingleMod, a deep learning model designed for the precise detection of m6A within single RNA molecules using nanopore DRS data. SingleMod is trained through a deep multiple instance regression framework, carefully tailored to harness the extensive methylation-rate labels provided by the golden standard GLORI method[14]. Our research encompasses the generation of m6A profile in human, providing a refined perspective on m6A genomics at the single-molecule and single-base level. In contrast to previous understanding, our findings unveil that m6A has a modest influence on alternative splicing and alternative polyadenylation, affecting only a small proportion of genes. However, these RNA processes, including alternative splicing and alternative polyadenylation, appear to reciprocally influence m6A deposition. Notably, we find that m6A exerts its function in promoting RNA degradation in a position-independent and additive manner. Furthermore, by conducting a comparative analysis of single-molecule m6A profiles from eight species, spanning mammals, fish, plants, protozoa, and algae, we've delineated three distribution patterns and established a dual deposition mode to elucidate these patterns and quantified the extent of the exclusion and inclusion effect present across each species. Our study establishes a unified deposition mode that spans large-scale multiple species, providing valuable insights into the biogenesis and regulatory mechanisms of m6A.

Result

Development of SingleMod for single-molecule m6A detection

To develop a method for detecting m6A within single RNA molecules, we initially planned to train a conventional classification model using reads derived from fully-methylated and -

unmethylated m6A sites identified by the GLORI method as positive samples and negative samples, respectively. However, we noticed that the majority of m6A sites exhibited low methylation rates, while the fully-methylated m6A sites were very limited (Extended Data Fig. 1a). Furthermore, the presence of an extra 2bp flanking the DRACH motif also influenced the current shift during nanopore sequencing (Supplementary Fig. 1a). Therefore, it became evident that the limited availability of fully-methylated m6A sites and the sequence diversity made training a conventional classification model unfeasible. To leverage the potential of the abundant m6A sites with low to intermediate methylation rates, we developed a deep multiple instance regression framework (Fig. 1a), which enables the inclusion of m6A sites with a broad range of methylation rates during training. Employing a deep neural network, this framework predicts the likelihood of each read being modified at a specific site and optimizes the network's parameters by minimizing the mean squared error (MSE) between the methylation rates calculated from the aggregated predicted statuses and GLORI benchmarks. By harnessing the power of regression strategy and deep learning, our novel framework, named SingleMod, is designed to achieve the detection of m6A at the single-molecule level using DRS data.

The model of SingleMod consists of two modules that are responsible for processing the features of each DRACN motif for every RNA molecule and providing the probability of m6A modification (Fig. 1b). In the first module, one-dimensional convolution is applied to six sets of informative features, with each set containing five features corresponding to the five positions of m6A within the nanopore (Extended Data Fig. 1b and Supplementary Fig. 1b). In recognition of the significant impact of variations in flanking sequences on current differences compared to m6A modification (Extended Data Fig. 1c), we annotated each feature with an embedding k-mer sequence (Fig. 1a-b). The second module directly employs one-dimensional convolution on the raw current signals, as the summarized current features may miss important representative information. Through iterative optimization of the convolution neural network structure, we arrived at the final model structure (Extended Data Fig. 1d-e). It is noteworthy that despite a significant proportion of sites exhibiting low methylation rates (Extended Data Fig. 1a and Supplementary Fig. 1c), including all sites for model training did not notably compromise the prediction performance for high methylation rate sites; instead, it substantially improved predictive accuracy for low methylation rate sites (Extended Data Fig. 1f and Supplementary Fig. 1d). As a result, we included all sites for model training.

Our model was separately trained using data from 13 predominant motifs, which collectively account for 78% of total m6A sites with methylation rate exceeding 0.1 in HEK293T cells (Supplementary Fig. 1c). During the training process, SingleMod achieved satisfactory evaluation metrics in both the training and validation sets (MSE: 0.0032/0.0044, MAE: 0.0356/0.0399, Pearson's r: 0.9637/0.9489) (Fig. 1c and Supplementary Fig. 2a). Importantly, SingleMod exhibited

consistent performance on testing set, with the methylation rates of sites calculated from the aggregation of single-molecule m6A predictions closely matching the benchmarks obtained from GLORI (MSE: 0.0043, MAE: 0.0398, Pearson's r: 0.9511) (Fig. 1c and Supplementary Fig. 2a). While the model achieved similar performance across most motifs, it excelled in its performance for AGACH, but displayed relatively weaker performance for GAACT and GAACA (Supplementary Fig. 2b-d). To assess the generality and robustness of SingleMod, we evaluated its performance on independent datasets from mouse embryonic stem cell line (mESC) and *Arabidopsis*, and compared it to other existing tools. SingleMod clearly outperformed other tools in both datasets, demonstrating its superiority in terms of overall performance and across individual motifs (Fig. 1d-f, Extended Data Fig. 2 and Supplementary Fig. 3-4). Furthermore, SingleMod accurately reflected changes in methylation rates following m6A writer knockout (KO) (Fig. 1g and Supplementary Fig. 4d), which is consistent with the nearly absent m6A signal in KO mESC and residual 10% in *Arabidopsis* mutant, as determined by mass spectrometry[21, 28]. In conclusion, SingleMod demonstrated high generality and robustness, and its performance surpassed that of any other existing tools.

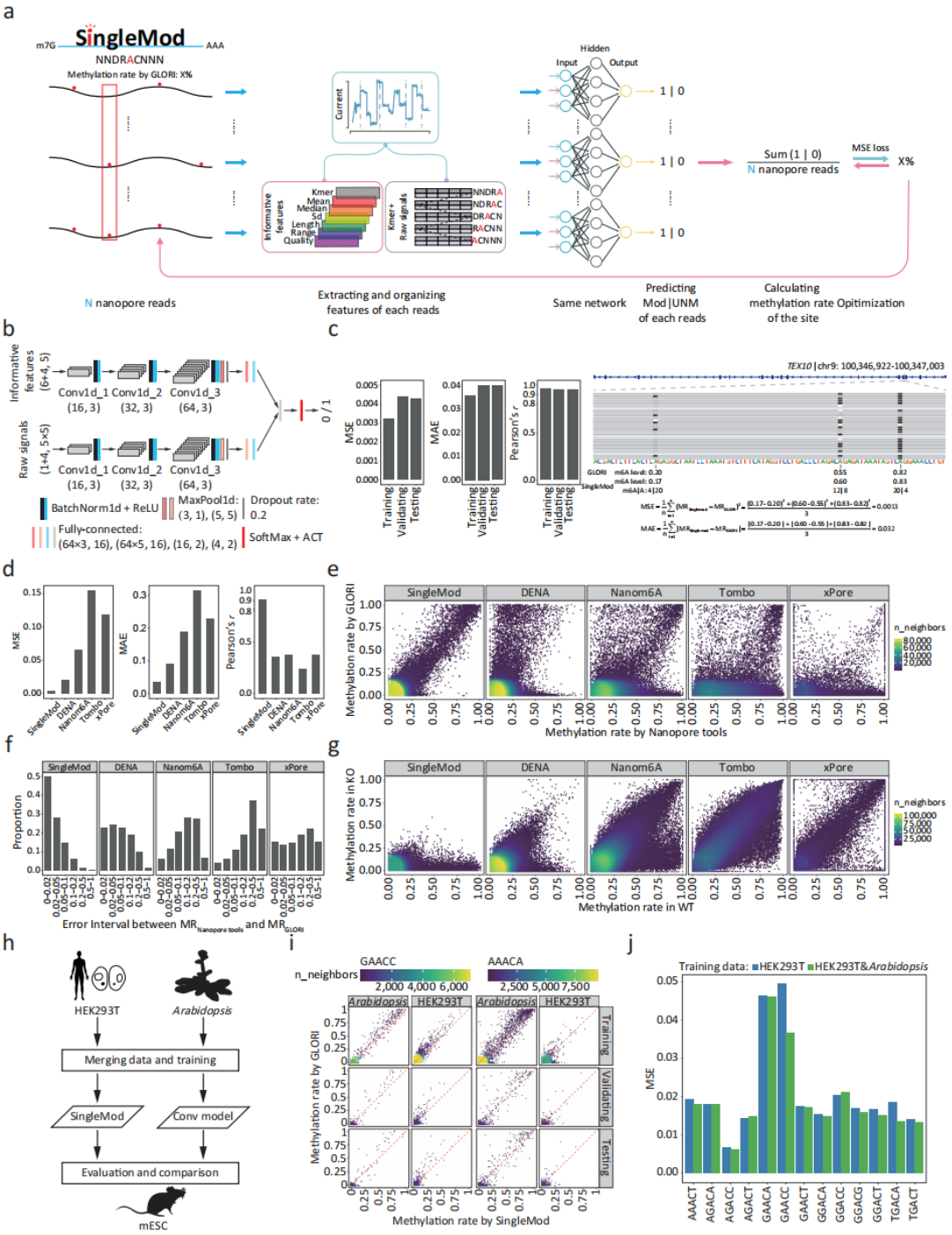


Fig. 1|Development and performance of SingleMod. a Methodological overview of SingleMod. **b** Architecture of the two modules within the SingleMod model. Input: (channels, length), Conv1d: (output_channels, kernel_size), MaxPool1d: (pool_size, stride), Fully-connected: (in_features, out_features). **c** Metrics used to evaluate SingleMod's performance on HEK293T datasets during training, validating, and testing, and an IGV snapshot Illustrates the calculation process of MSE and MAE. **d** Comparing the overall performance on mESC data between SingleMod and other tools. Data used in **e-g** were also from mESC. **e** Comparing the methylation rates predicted by nanopore tools to GLORI benchmarks. **f** Proportion of sites in different error intervals. **g** Comparing the

methylation rates between WT and KO samples predicted by nanopore tools. **h** Training and evaluation data of the final version of SingleMod and the conventional classification models (Conv model). **i** Methylation rate of sites within the GAACC or AAACA motif in different datasets from *Arabidopsis* or HEK293T, as detected by GLORI and SingleMod. **j** MSE on the middle methylation rate interval (0.4-0.6) for different motifs in mESC, comparing two versions of SingleMod. The versions trained on different data are labeled as "HEK293T" and "HEK293T&*Arabidopsis*".

Enhancing generalization capability by integrating data from evolutionarily distant species

Because the training data contains a higher representation of low-methylated sites (Supplementary Fig. 5a), SingleMod showed better performance for sites with low methylation rates (Supplementary Fig. 5b-e). This is particularly advantageous for enhancing the overall performance of the model in practical applications, aligning with the more frequent occurrence of low-methylated sites in real samples. Furthermore, by categorizing 9-mers based on their overall methylation rates in training set, we found that our model showed reduced accuracy in predicting highly methylated sites using low-methylated 9-mers (Supplementary Fig. 5f). This challenge was more pronounced in *Arabidopsis*, mainly due to the prevalence of 9-mers exhibiting a species-specific modification spectrum. We speculate that although distinct features in the model ensured robust performance across various methylation rate intervals (Supplementary Fig. 5a-e, 6), the inherent sequence and methylation level preferences within a single species largely impair the generalization of our model. To address this, we combined datasets from HEK293T and *Arabidopsis* for training the final model, which was subsequently tested on mESC (Fig. 1h and Supplementary Fig. 7-8). By integrating *Arabidopsis* data, we successfully balanced the methylation rate distribution for specific motifs, such as GAACC. This in turn enabled us to obtain sufficient data to train additional models for two additional motifs, AAACA and TAACT (Fig. 1i). The final model outperformed previous version, which was solely trained on the HEK293T data (Fig. 1j). This strategy is expected to enhance the generalization of the model when applied to other species.

Integrating data from other species significantly increase the size of training data. We asked whether this strategy could also be effective when training conventional classification model (Fig. 1h). To assess this, we used the same neural network architecture as SingleMod but applied a classification framework to train conventional models, in which we progressively adjusted the ratio of negative to positive samples. In both balanced and imbalanced data scenarios, the SingleMod framework consistently outperformed conventional classification frameworks (Extended Data Fig. 3a). While progressively increasing the ratio of negative samples improved the overall performance of the conventional classification models, their performance towards high-methylation rate sites gradually declined (Extended Data Fig. 3b-c). In contrast, SingleMod exhibited satisfactory performance across different methylation rate sites. Furthermore, conventional classification models

performed poorly for motifs with very few fully-methylated sites available for training (Extended Data Fig. 3d-h). These results highlight the necessity and superiority of the SingleMod framework when dealing with methylation-rate labeled multi-instance data, particularly in real-world scenarios where training data is often limited.

Single-molecule m6A landscape in human cell line

Leveraging the high precision of SingleMod in detecting m6A at a single-molecule resolution, we are able to depict the m6A landscape in human and analyze its genomic features in a new perspective (Fig. 2a). In contrast to previous understanding generated from site-level m6A mapping methods[29], it is striking to note that approximately 60% of the mRNA molecules lack m6A modifications (Fig. 2b). However, a closer examination has unveiled that these unmodified molecules are predominantly derived from a small subset of genes, particularly those with the highest expression levels (Extended Data Fig. 4a-b). For the majority of genes or isoforms, a significant proportion of their corresponding RNA molecules have at least one m6A modification (Fig. 2c). Specifically, in more than 4,000 genes, which account for over 42% of the total detected genes, more than 90% of RNA molecules exhibit m6A modifications (Fig. 2c).

For a more in-depth quantitative exploration of single-molecule m6A mapping, we introduced two metrics to characterize m6A levels for each gene or isoform. These metrics calculate the average number of m6A sites per RNA molecule or per RNA molecule per 1,000 bp length (referred to as m6A number and m6A density, respectively). Remarkably, around 80% of genes or isoforms exhibit a m6A density value below 2 (Extended Data Fig. 4c). Nearly half of them, however, display an average of two or more m6A modifications per RNA molecule (Fig. 2d). It's intriguing to note that the heterogeneity in m6A numbers at the gene level is greater than that at the isoform level, as indicated by the larger standard deviation among distinct RNA molecules (Fig. 2d and Extended Data Fig. 4d). These findings underscore the significance of single-molecule m6A mapping, as gene-level quantification may not accurately represent the m6A levels within its isoforms (Extended Data Fig. 4e). In addition, we observed that higher m6A numbers coincide with an overall decrease in expression levels (Fig. 2e and Extended Data Fig. 4f). While the overall methylation rate of A bases within 15 DRACN motifs is 7%, this rate varies significantly among distinct motifs. Notably, 20% of GGACT motifs are modified, in contrast to only 3% of AGACC motifs (Fig. 2f). Following *METTL3* knockout, a general reduction in the methylation rate was observed for all motifs (Extended Data Fig. 4g).

Leveraging the advantages of long-read sequencing, we endeavored to redefine the distribution of m6A across full-length RNA molecules. Generally, m6A was found to be enriched toward the 3' end of mRNA, specifically within the region of relative positions 0.6-1 (Fig. 2g). To establish correlations between m6A sites and mRNA landmarks with biological significance, we annotated

each molecule and divided it into start codon, last junction, and stop codon regions. (Extended Data Fig. 4h). We observed that even though m6A exhibited an ascending and then descending trend after the last junction, the entire last exon remained the most susceptible region for m6A modification (Fig. 2g). Despite m6A displaying a general enrichment at the 3' end, it's noteworthy that there is a distinct, gene-body specific m6A distribution pattern. Notably, a substantial portion of genes displayed significant m6A levels at their 5' end and within the middle regions (Fig. 2h). In terms of absolute positioning, the methylation rate reached its peak and remained stable over an extended distance, either 200 bp from the last junction or 100 bp from the stop codon (Fig. 2i and Extended Data Fig. 4i). Consistently, although the modification potential at different positions was similarly round 0.07, the proportion of motifs modified in the last exon was elevated (Fig. 2j).

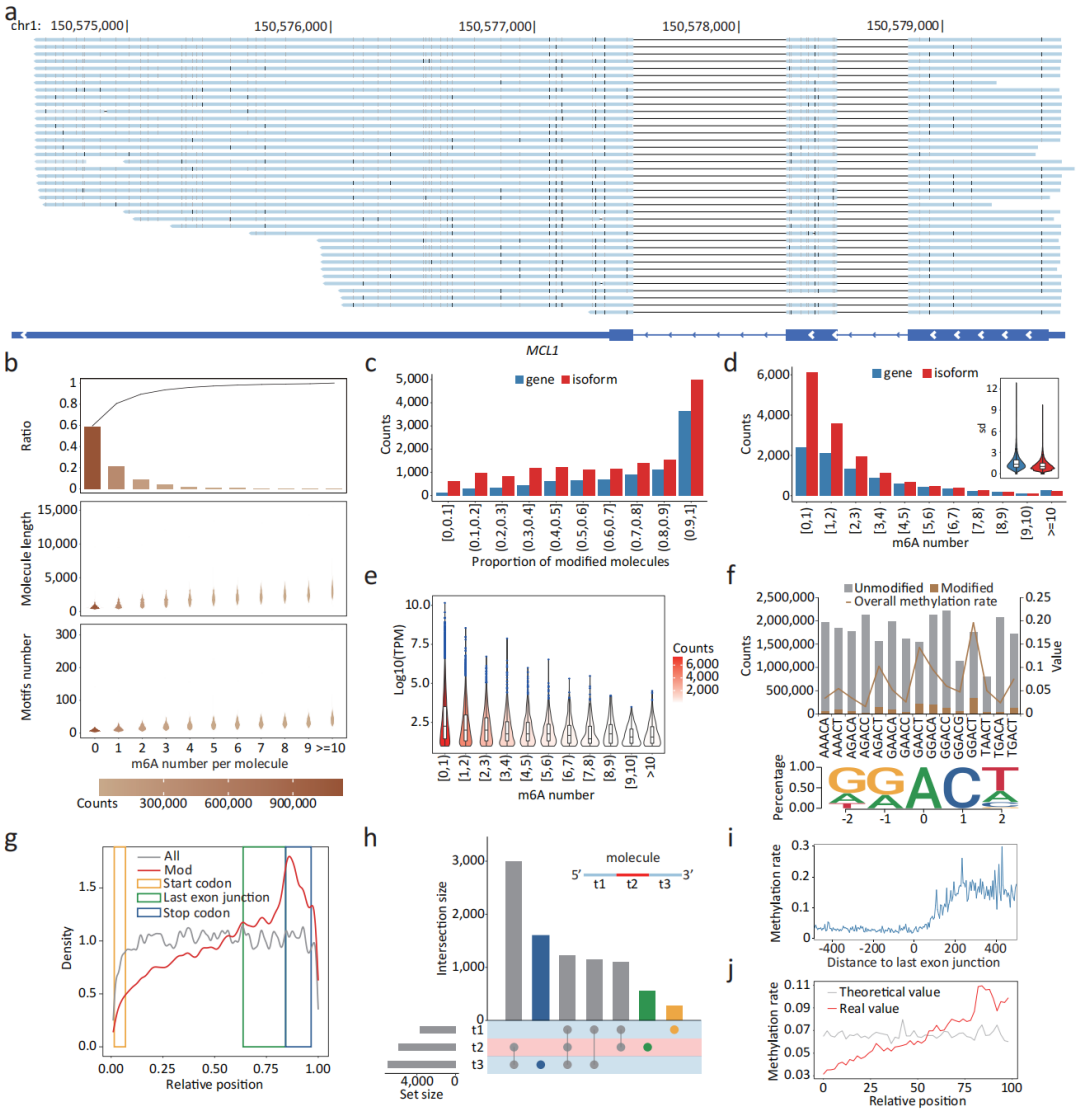


Fig. 2|Single molecule profiling in HEK293T by SingleMod. a A representative IGV snapshot illustrating the m6A modification status of each potential site on each individual molecule from gene *MCL1*. The gray and black vertical bars indicate A and m6A, respectively. **b** Ratio, length and DRACN motifs number of molecules with varying m6A number. **c** Counts of genes or isoforms

with varying proportion of modified molecules. **d** Counts of genes or isoforms with varying average m6A number per molecule, and the distribution of standard deviation in m6A number among molecules within each gene or isoform. **e** Distribution of expression levels of genes with varying m6A number. "TPM" indicates Transcripts Per Million. **f** Counts (top, bar) of unmodified and modified A bases within 15 DRACN motifs from all molecules and corresponding overall methylation rates (top, line). The composition of m6A flanking bases plotted by Seqlogo are shown below. **g** The distribution of relative position of background (All) and modified (Mod) A bases within their respective molecules. The relative regions of the start codon, last exon junction and stop codon are also marked. **h** Upset plot shows the counts of gene with different m6A distribution patterns. RNA molecules are divided into three equal intervals. If a gene's m6A modifications exceed 25% within any interval, it is deemed to have significant m6A modifications in that interval. **i** Overall methylation rate around the last exon junction. **j** Overall methylation rate at different relative positions.

Dissecting m6A-mediated RNA heterogeneity and regulation at the single-molecule level

Recent research demonstrated the tendency of m6A sites to cluster in "hot regions"[14, 17]. Supporting this notion, we observed a significantly shorter distance between adjacent m6A modifications on individual molecule (Extended Data Fig. 5a). Nevertheless, it remains unclear whether m6A enrichment occurs within certain RNA molecules or if it's randomly distributed among molecules within one cluster region. Our initial assessment of adjacent m6A sites demonstrated a probabilistic distribution rather than a concurrent pattern at the molecule level (Fig. 3a), implying that the presence of m6A at one motif is independent of the modification status of an adjacent motif. Consistently, RNA molecules with an expected random distribution of m6A counts were the most abundant, whereas those with no or complete modifications were rare (Fig. 3b). Consequently, very few RNA molecules had multiple m6A sites located in a cluster (Extended Data Fig. 5b). Furthermore, at the gene level, the m6A number across different molecules appeared fairly uniform, and m6A modifications were not concentrated at specific positions (Supplementary Fig. 9a-b). These findings indicate that within a gene, m6A is evenly distributed across multiple molecules and multiple sites, resulting in a diverse population of RNA molecules with varying m6A signatures. We further demonstrated that m6A modifications introduce substantial heterogeneity among molecules within a gene or isoform by measuring the probability of randomly selecting two identical molecules (Fig. 3c and Supplementary Fig. 9c).

The pronounced molecular heterogeneity introduced by m6A modifications has sparked our interest in understanding the mechanisms by which m6A exerts its functions. The most widely recognized function of m6A is promoting RNA degradation[30, 31]. Supporting for this, we observed a negative correlation between the change in m6A levels and gene expression following *METTL3*

knockout (Extended Data Fig. 5c). Furthermore, we directly observed a clear negative correlation between m6A levels and RNA stability, with the most notable correlation associated with m6A density (Fig. 3d-e and Supplementary Fig. 9d-e). Importantly, the strong correlation between m6A density and RNA stability persisted regardless of the specific positions of m6A modifications (Fig. 3f). These results imply that the function of m6A in promoting RNA degradation is position-independent and operates in an additive manner. Consistently, gene categories expected to have lower stability levels, such as lncRNA, exhibit higher m6A density (Fig. 3g).

As single-molecule profiling simultaneously captures m6A and other mRNA processing evens, we compared the splicing form and polyadenylation site (PAS) usage in WT and *METTL3* KO cells, aiming to further unravel the intricate role of m6A in mRNA biogenesis regulation. Contrary to previous understanding[32, 33], the vast majority of alternative splicing (AS) events remain unchanged in KO cells, with only a small fraction showing differences, primarily arising from a decrease in inclusion levels (Fig. 3h and Extended Data Fig. 5d). However, m6A appears to exert a more significant impact on alternative polyadenylation (APA) than AS, with 211 genes displaying some degree of differential APA, among which 161 exhibits a decrease in proximal PAS usage (Fig. 3h and Extended Data Fig. 5d). Although most of these differential AS and APA events show relatively modest changes, there are still some notable cases (Fig. 3i and Supplementary Fig. 9f). It's interesting to note that differential AS and APA events exhibit higher m6A levels in adjacent constant regions (Fig. 3j and Supplementary Fig. 9g), further supporting the role of m6A in these events. Moreover, m6A within alternative regions appears to facilitate the regulation of AS and APA, particularly in intron retention (Fig. 3j and Supplementary Fig. 9g).

To identify the specific m6A sites responsible for influencing mRNA processing, we conducted extensive multi-to-multi correlation analyses for each pair of m6A site and each AS or APA event (Fig. 3k). We successfully pinpointed m6A sites responsible for 15 differential AS and APA events, with the majority of these sites closely located in the regions where AS or APA takes place (Fig. 3l, Extended Data Fig. 5e-f and Supplementary Fig. 9h). Importantly, we found that certain events were regulated by multiple m6A sites, and these m6A sites appeared to work additively, independently of their specific locations (Fig. 3l-n and Extended Data Fig. 5e-f). These findings provide molecular-level evidence that the presence of flanking m6A sites inhibits the utilization of nearby splice sites, thereby promoting the retention of alternative regions (Fig. 3m). On the flip side, the flanking m6A sites facilitates the utilization of proximal polyadenylation sites in most cases (Fig. 3n). In many of these m6A-correlated events, RNA molecules with differing modification status at specific sites exhibit divergent trends in AS or APA (Fig. 3l, Extended Data Fig. 5g and Supplementary Fig. 9i). However, it is noteworthy that the majority of the m6A-correlated AS/APA events, even those linked to highly methylated sites, did not exhibit alterations in inclusion levels or PAS usage following *METTL3* KO (Fig. 3l, Extended Data Fig. 5g and Supplementary Fig. 9i). Conversely, this suggests

that the essence of these m6A-correlated events is that m6A deposition is influenced by the processes or outcomes of RNA processing, especially at the 3' end, where the selection of PAS in many genes can impact nearby m6A deposition (Extended Data Fig. 5h-i). Furthermore, in most cases, m6A modifications are more likely to be added when molecules choose distal polyadenylation sites (Extended Data Fig. 5h, 5j-k).

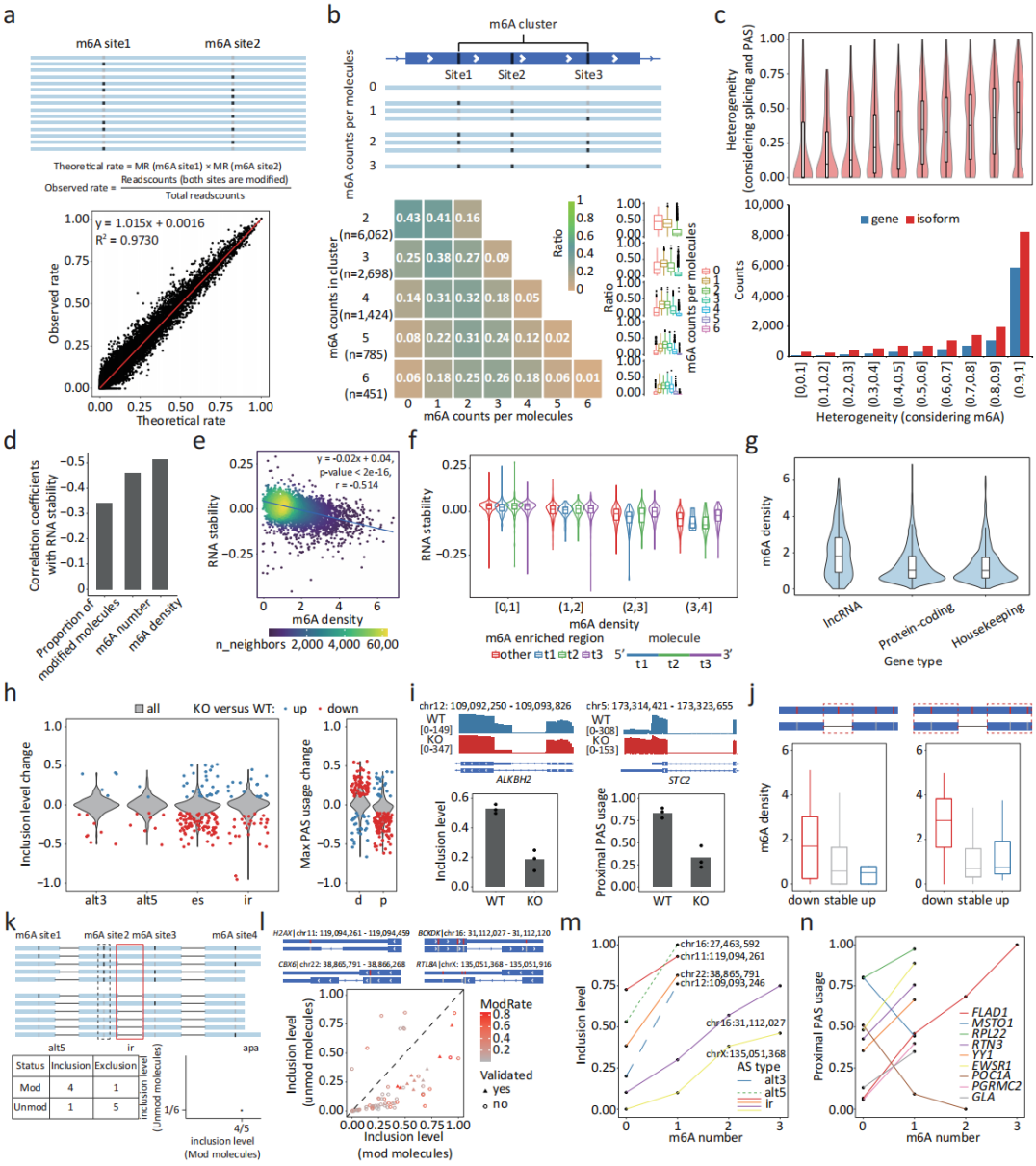


Fig. 3|m6A-mediated RNA heterogeneity and regulation profiled at single-molecule level in HEK293T. a Top: schematic illustration depicting the calculation of theoretical rate and observed rate of molecules where two adjacent A bases within 15 DRACN motifs are both modified. Bottom: comparison between these two values. **b** Schematic illustration (top) depicting reads displaying possible m6A patterns within an m6A cluster. Heatmap (bottom, left) and boxplot (bottom, right) show the ratio of molecules with varying m6A number within m6A cluster. **c** Heterogeneity among RNA molecules within a gene or isoform was represented by 1 - the probability of randomly

selecting two identical molecules. Bottom: gene or isoforms counts with varying molecules heterogeneity introduced by m6A. Top: distribution of molecular heterogeneity of corresponding genes only considering difference on splicing form and PAS selection. **d** Pearson's correlation coefficient between RNA stability and three m6A quantification metrics. **e** Correlation between RNA stability and m6A density. **f** Distribution of RNA stability of genes with varying m6A density and enriched region. **g** Distribution of m6A density from different gene types. **h** Distribution of change in inclusion level of AS events (left) and max PAS usage (right) following *METTL3* KO. "alt3", "alt5", "es", "ir", "d" and "p" indicate alternative 3' ss, alternative 5' ss, exon skipping, intron retention, distal PAS and proximal PAS, respectively. **i** Two examples of differential alt3 (left) and apa (right) events. **j** Comparing m6A density in alternative (left) and constant (right) regions among ir events with different status following *METTL3* KO. **k** Schematic illustration depicting the molecular correlation analysis between m6A sites and AS/APA events. **l** Bottom: dotplot comparing the inclusion levels between two classes of molecules based on modification statue of corresponding m6A site, as calculated in Fig.3k. Each dot represents significant m6A site-intron retention correlated event identified through correlation analysis. Methylation rates and whether the correlated events exhibited differential splicing following *METTL3* KO are also indicated (validated). Top: 4 differential ir events and responsible m6A sites. **m-n** Inclusion level (**m**) and proximal PAS usage (**n**) of molecule groups with varying numbers of responsible m6A.

Comparative analysis of single-molecule m6A profiles across multiple species reveals three distinct m6A patterns

The ability to map m6A with single-molecule precision has empowered us to conduct a comparative analysis of m6A distribution across various species. We employed SingleMod to analyze species with significant evolutionary distances, including mammals (HEK293, K562, mESC), fish (zebrafish), plants (*Arabidopsis*, Rice, *P. trichocarpa*), algae (*Chlamydomonas*), Protozoa (Toxoplasma), and Nematoda (*C. elegans*). Intriguingly, m6A displayed diverse motif preferences in different species, and this divergence could be utilized to distinguish their evolutionary relationships within distinct clusters (Fig. 4a). Additionally, the composition of flanking sequences surrounding m6A was highly consistent within each cluster (Extended Data Fig. 6a and Supplementary Fig. 10a). The *C. elegans* was excluded from subsequent analyses due to extremely low m6A levels in all motifs, aligning with previous findings that suggest the absence of m6A in *C. elegans* mRNA[34]. We also observed that different species displayed varying distributions of m6A numbers within single RNA molecules (Fig. 4b-c and Supplementary Fig. 10b). For example, zebrafish and Toxoplasma showed comparatively higher m6A number, which might be attributed to their longer RNA molecule lengths and unique motif compositions (Extended Data Fig. 6b and Supplementary Fig. 10c).

It's noteworthy that in all species studied, the relative location of m6A modifications on single RNA molecules displayed a distinct preference towards the 3' end (Fig. 4d). Consistently, the last exons contained 40% potential sites within 15 DRACN motifs, but accounted for significantly higher proportion of the total m6A modifications, ranging from 60% to 85% across different species (Fig. 4e and Extended Data Fig. 6c). This phenomenon was also reflected in the higher methylation rates and the increased proportion of modified last exons in these species (Fig.4f-g and Extended Data Fig. 6d-e). Nevertheless, the extent of this enrichment at the 3' end and the size of the enriched region varied among different species (Fig. 4d). For instance, in plants and Toxoplasma, m6A was sharply enriched in a small region near the RNA's terminal, while in *Chlamydomonas*, the enrichment region was larger and displayed a higher degree of enrichment compared to vertebrates. The enrichment observed in plants, Toxoplasma, and *Chlamydomonas* can likely be attributed to the confinement of m6A within a consistently delimited region within the last exon (Fig. 4h and Supplementary Fig. 10d). Despite these distinctions, vertebrates still exhibited a relatively higher proportion of m6A modifications in the first and middle exons compared to other species.

Recent studies have shown that m6A deposition is suppressed by Exon Junction Complex (EJC) within 200 bp from the junction in mammals[11-13]. This finding prompted us to conduct a comprehensive investigation into m6A modifications at different distances from the exon edge using full-length RNA sequencing. Remarkably, the m6A distribution patterns can be clearly classified into three categories, which align with the clustering results based on motif similarity (Fig. 4i and Extended Data Fig. 7a). In the first category, primarily observed in vertebrates, m6A is minimal near the edges of the first and middle exons, almost absent within 100 bp (Fig. 4i and Extended Data Fig. 7a). However, due to the short length of these exons, most of potential sites are situated near the edges (Extended Data Fig. 7b), resulting in low m6A levels in these exons. Within the last exon, the suppression effect primarily originates from its 5' edge, with the 3' terminal contributing to minor suppression (Fig. 4i and Extended Data Fig. 7a). Furthermore, the last exon is notably longer (Extended Data Fig. 7b), allowing it to possess a higher proportion of m6A modifications. In the second category, observed in plants and Toxoplasma, m6A is almost absent throughout the entire first and middle exons but is enriched in a region approximately 250 bp away from the 3' end (Fig. 4i and Extended Data Fig. 7a). The third category of distribution pattern is similar to the second category, with the difference that m6A on the last exon is only inhibited at the 5' end, while it maintains a stable and relatively high level in regions far from the 5' end (Fig. 4i and Extended Data Fig. 7a). Comparing the relative distribution of m6A modifications based on the varying lengths of the last exon further corroborated the observation of these three distinct patterns (Fig. 4j and Extended Data Fig. 6f, 8).

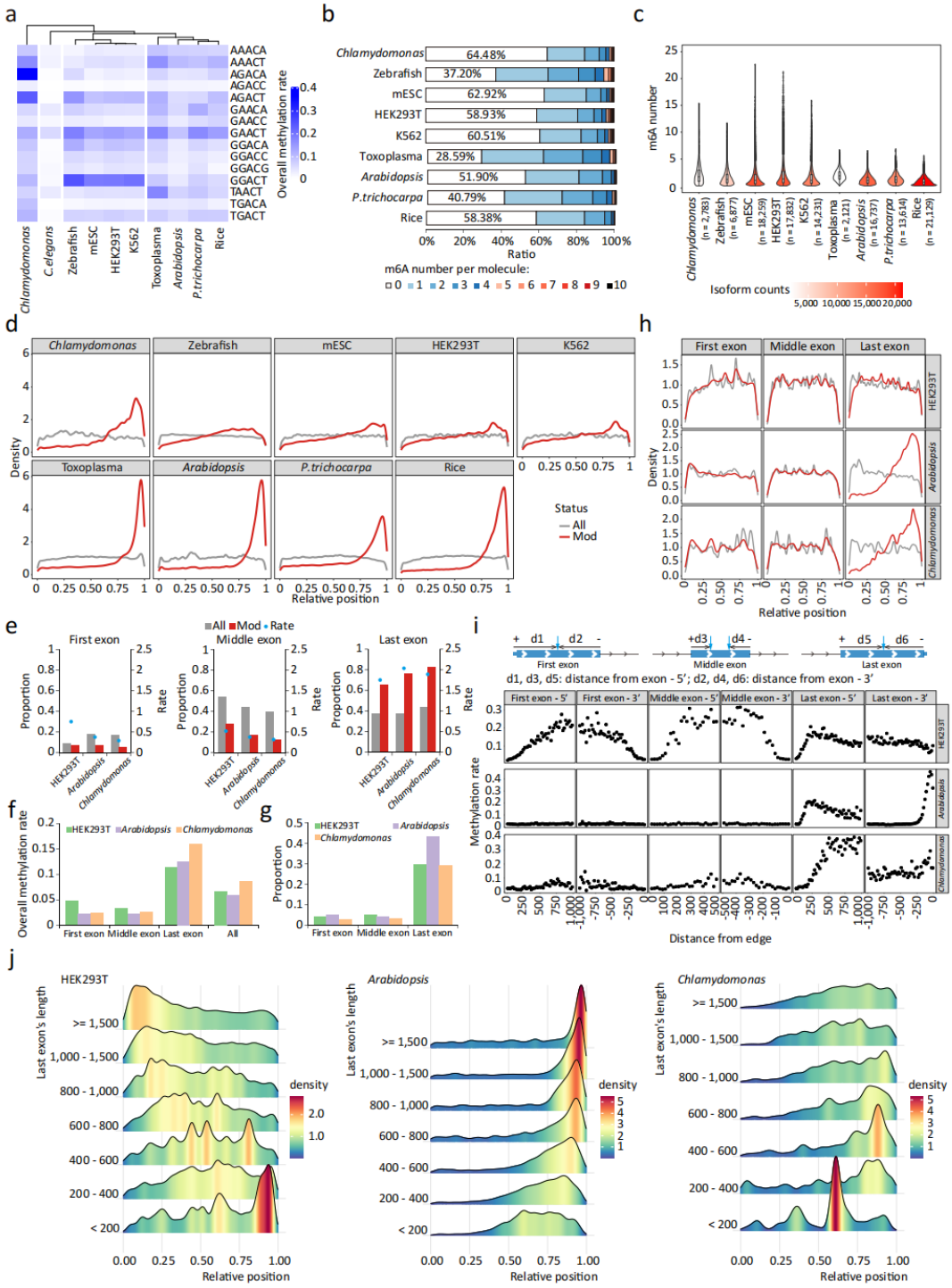


Fig. 4|Comparison of single-molecule m6A profiles across multi-species. a Overall methylation rate of distinct motifs across diverse species. **b** Ratio of molecules with varying m6A number. **c** Distribution of the average m6A number per molecule for different isoforms. **d** Distribution of relative position of background (All) and modified (Mod) A bases within their respective molecules. **e** Bar: proportions of background (All) and modified (Mod) A bases in distinct exons. Dot: the ratio between these two values. **f** Overall methylation rate within distinct exons. **g** Proportion of modified exons (containing at least one m6A). **h** The distribution of relative position of background (All) and

modified (Mod) A bases within their respective molecules in different exons. **i** Overall methylation rate at positions with varying distance from one exon edge. **j** The distribution of relative position of m6A within their respective molecules in last exons with varying lengths.

Exclusion-inclusion deposition mode shapes m6A distribution patterns

Next, we delved into the underlying mechanisms that shape the three categories of m6A distribution patterns across species. A well-established exclusion deposition model mediated by EJC has been recognized in mammals, wherein m6A undergoes suppression in specific regions and is added with intrinsic efficiency in regions unaffected by suppression[11-13]. To expand this model, we proposed that the intrinsic efficiency of the m6A deposition for each motif could be inferred from the overall methylation rate in middle exons regions that remain unaffected by EJC (Extended Data Fig. 9a). We then calculated this value as the baseline and compared it with methylation rates at other positions (Fig. 5a, Extended Data Fig. 9b, 10a and Supplementary Fig. 11). Our results have clearly recapitulated the EJC-mediated exclusion deposition pattern where m6A methylations are suppressed near exon edges. Moreover, m6A methylations at the 5' end of first exons were also suppressed over a broader span of 400 bp. Notably, the entire last exon was subjected to additional suppression, resulting in a methylation rate slightly below the baseline, which was particularly evident towards the 3' end (Extended Data Fig. 9k and Supplementary Fig. 12).

To quantify the exclusion effects, we compared the observed m6A counts to the hypothetical counts under default m6A writing efficiency. Given that the first and middle exons are relatively short, most potential m6A motifs are located within spans of cumulative exclusion effects from both edges, which resulted in much lower methylation rates (Extended Data Fig. 9c, 9g). These exclusions led to a marked reduction in m6A levels (Extended Data Fig. 9d, 9h), with an effect reaching 0.8 in both the first and middle exons (Extended Data Fig. 9e-f, 9i-j). Although exclusion in last exons also led to a notable reduction in m6A level with a 0.6 effect, the longer exon lengths and the moderate 3' end exclusion effect allowed for a higher retention of m6A level (Extended Data Fig. 9l-n). Remarkably, these effects were nearly identical between human and mouse, with a total 75% reduction in m6A levels following exclusion, while appearing somewhat lower in zebrafish (Fig. 5b, Extended Data Fig. 9o, 10b). These results supported the exclusion deposition model, which serves as the prevailing m6A biogenesis mechanism in vertebrates.

In the second pattern, m6A modifications occurred solely within specific regions in the last exon. We referred to this pattern as the inclusion deposition mode, where m6A is consistently added to certain defined regions. Supporting this, regardless of the exon length, m6A was observed within a fixed distance from the 3' end, spanning approximately 300 bp in *Arabidopsis* and a slightly broader range in other species (Fig. 5c, Extended Data Fig. 10g and Supplementary Fig. 13). In the third pattern, m6A were uniformly distributed within the last exon except the region near the 5' edge

(Fig. 5e and Supplementary Fig. 14). This pattern indicated that both the regions of m6A occurrence and suppression were clearly defined, suggesting a combination of inclusion and exclusion deposition mode. We conducted calculations for motif occurrences and the reference methylation rate within the defined inclusion span, as well as the background methylation rate outside this span (Extended Data Fig. 10c-f, 10h-i). The inclusion effect substantially enhanced the methylation rate compared to the background. Despite the minor proportion of motifs falling within the inclusion span, the inclusion effect contributed to nearly 60% of the total m6A level in the second pattern, while the exclusion effect was nearly negligible (Fig. 5d and Extended Data Fig. 10j). Remarkably, the establishment of *Chlamydomonas* pattern was orchestrated by an inclusion effect of 0.84, followed by an exclusion effect of 0.5 (Fig. 5f). Collectively, our results quantitatively depicted the shaping of m6A patterns by both exclusion and inclusion deposition effects and supported a unified dual-mode to explain the landscapes of m6A in multiple species (Fig. 5g).

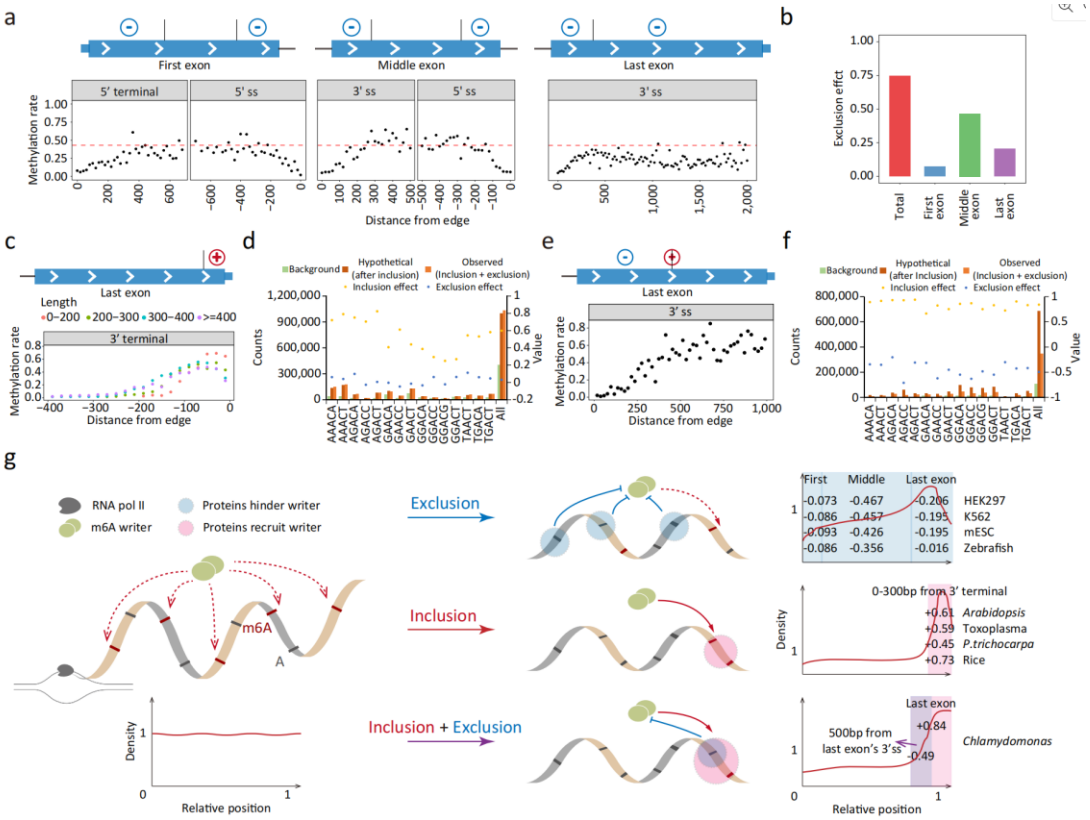


Fig. 5|m6A-pattern across multi-species shaped by exclusion and inclusion deposition mode. a Top: schematic representation of exclusion deposition mode in HEK293T. Bottom: m6A pattern shaped by exclusion deposition mode and comparing the overall methylation rate at positions with varying distances from one exon edge to hypothetical methylation rate (red dashed lines), using GGACA motif as a representative. For first and middle exons, only potential motifs with distance from opposite edge greater than setting were retained for calculations, eliminating the influence of opposite edge. **b** The total exclusion effect and the exclusion effect contributed by different exons. **c** Top: schematic representation of inclusion deposition mode in *Arabidopsis*. Bottom: m6A pattern

shaped by inclusion deposition mode and overall methylation rate at positions with varying distances from 3' terminal, using AAACT motif as a representative. **d** Counts of m6A modifications under diverse situations in *Arabidopsis*. Inclusion and exclusion effects are also calculated and marked. **e** Top: schematic representation of combination of inclusion deposition and exclusion mode in *Chlamydomonas.* Bottom: m6A pattern shaped by combination of inclusion and exclusion deposition mode and overall methylation rate at positions with varying distances from 3'ss of last exon, using AGACA motif as a representative. **f** Counts of m6A modifications under diverse situations in *Chlamydomonas*. Inclusion and exclusion effects are also calculated and marked. **g** Schematic representation illustrating the dual-mode (exclusion and inclusion) of m6A deposition and quantitatively depicting this mode in shaping m6A distribution pattern across different species.

Discussion

As a modification on RNA, m6A introduces a remarkable level of complexity to the cellular transcriptome. In the past decade, significant advancements in m6A research have unraveled insights into its genomics, functions, and regulatory mechanisms[5, 10]. However, as RNA exists in multiple copies within cells, the intrinsic heterogeneity among RNA molecules is challenging to completely capture using traditional NGS-based techniques. In this context, we introduce SingleMod, a cutting-edge approach that achieves the highest resolution and faithfully depicts m6A modifications on individual RNA molecules. Our method not only demonstrates the superior accuracy also showcases remarkable generalization capabilities. We have harnessed SingleMod to acquire single-molecule m6A profiles across multiple species with evolutionary distances. We anticipate that, with the growing prevalence of TGS platform[35], researchers will easily obtain single-base, single-molecule resolution m6A maps in various scenarios involving different species or diverse biological functions. Through DRS technology based on TGS platform, we are presented with a unique opportunity to obtain a holistic view of the transcriptome, including dimensions such as RNA splicing, polyadenylation, modifications, and more, that collectively contribute to the emerging field of epitranscriptomics.

In the context of multi-instance data labeled by methylation rates, the specific modification status of individual molecules remains uncertain. Given the scarcity of fully-methylated sites in actual samples[14], we have developed a multi-instance regression framework. This framework effectively leverages information from sites exhibiting intermediate modification levels, facilitating comprehensive model training. We believe that this innovative framework holds the potential for broader applications, extending to the detection of various RNA modifications and potentially even DNA modifications. This is especially pertinent, considering that most known RNA modifications exhibit low stoichiometry in native RNA. In practice, the specificity of m6A and the limited availability of data restricted our training of SingleMod to only 15 prominent motifs, although we

have introduced training data from mammal and plant. Consequently, this model can only make predictions regarding the modification status in certain motifs, rather than providing de novo detection for m6A modification in all A bases. Although m6A profiles from known species explicitly occur within DRACN motifs[6], employing our model in a species without prior knowledge may introduce bias, particularly in terms of quantitative assessment. We envision that this limitation can be surmounted by accumulating sufficient data across a wide range of species or by acquiring extensive data through synthetic methods in future research.

Leveraging precise and highest-resolution m6A mapping, we have conducted a comprehensive quantification of m6A levels. Comparing to previous studies, this reevaluation encompasses the proportion of modified molecules, the average number of m6A modifications per molecule, and the molecular m6A density, as well as the overall methylation rate of A bases in 15 DRACN motifs in poly(A) transcriptome. This endeavor has led to a revolutionary revision of previously reported quantitative findings[29]. In general, our findings suggest that m6A content on RNA is lower than previously assumed. Strikingly, more than 60% of RNA molecules in our study exhibited no m6A modifications. Upon closer examination, we discovered that these non-m6A RNA molecules primarily originate from a small subset of highly expressed genes, whereas half of genes had two or more m6A modifications in their corresponding RNA molecules. This investigation highlighted that within a given gene, m6A modifications tend to be distributed across multiple molecules and multiple sites, culminating in significant heterogeneity among these molecules. Intriguingly, our research indicates that m6A promotes RNA degradation, irrespective of its specific position, and this function operates in an additive manner. Furthermore, we found that adjacent m6A sites regulate AS and APA also in an additive manner. Confirming this hypothesis in future research holds significant implications for understanding the mechanisms by which m6A exerts its functions and determining target sites for potential functional disruption[36].

Our comparative analysis of single-molecule m6A profiles across multiple species has revealed three distinct m6A distribution patterns. In addition, we've established a dual-mode shaping of the m6A landscape, providing insights into the underlying mechanisms. In vertebrates, we identified a combination of the well-established EJC-mediated exclusion and exclusion effects at the 5' terminal and the last exon, contributing to a substantial 75% reduction in m6A levels. However, exact mechanisms driving these effects are yet to be fully elucidated. We speculate that cap-binding proteins and transcription termination complexes on RNA may play roles in this process. In the case of plants and Toxoplasma gondii, their distinct methylation patterns suggest interactions between their m6A writing machinery and unidentified factors, probably RNA binding proteins, pre-installed at the transcription termination and polyadenylation site. For *Chlamydomonas*, where inclusion spans the entire last exon, it is reasonable to infer that certain 3' UTR-preferred RBPs are involved. Nonetheless, *Chlamydomonas* also experiences exclusion at the 5' edge, probably through a

mechanism akin to the EJC-mediated exclusion observed in mammals.

Inspired by the well-established EJC-mediated exclusion, the generalized mechanisms of exclusion-inclusion deposition mode may involve intricate interactions between certain constitutive RNA-binding proteins or processing proteins and m6A writer. In the exclusion deposition mode, these proteins and the m6A writer are likely to mutually exclude each other at certain positions, resulting in specific sites refractory to m6A addition. Conversely, in the inclusion deposition mode, these proteins likely collaborate with the m6A writer, directing them to specific sites for m6A addition. It's important to note that a single species may possess multiple such proteins, contributing to the intricate native m6A patterns. While there could be several pathways governing RNA modifications[9], the presence of these proteins potentially establishes a predominant regulatory paradigm in the process of m6A biogenesis. We propose that the exclusion-inclusion deposition mode represents general principles underlying m6A biogenesis, as well as its regulatory mechanisms and interplay with other RNA processing evens, collectively shaping the intricate epitranscriptome across a diverse array of species.

Methods

Sample collection and RNA preparation

The mESC cells were provided by our collaborator (J. Chen from the Guangzhou Institutes of Biomedicine and Health). A detailed description can be found in a previous study[37]. The seeds of wild type *Arabidopsis* accession Col-0 were sown and the seedlings were harvested in accordance with the experimental conditions outlined in a previous publication[28]. Chamydomonas cells were collected by centrifugation and lysed in TRIzol reagent (Thermo Fisher Scientifific, 15596018). After extracting total RNA from the samples using TRIzol (Thermo Fisher Scientifific, 15596026), mRNA was isolated via a Dynabeads mRNA Purifification Kit (Thermo Fisher Scientifific, 61006) with genomic DNA removal using TuRBO DNase kit (Thermo Fisher, AM2238).

Data collection

The training and testing of SingleMod were conducted using nanopore direct RNA sequencing data and the corresponding GLORI data from three species (HEK293T, mESC and *Arabidopsis*). In addition, nanopore DRS data from seven other species (and K562) were included for the comparative analysis of multi-species single-molecule m6A profiles. Some of the data were generated by this study, while others were obtained from publicly available datasets:

1) HEK293T. Triplicate DRS data of WT and *METTL3* KO samples were collected from Ploy N. Pratanwanich et al., 2021[38] (ENA: PRJEB40872) and then merged respectively; GLORI data was from Cong Liu et al., 2022[14] (GEO: GSE210563).

2) mESC. DRS data of WT and *METLL3* KO samples were collected from Zhendong Zhong et al.,

2023[21] (GEO: GSE195618); additional DRS data of WT was collected from Piroon Jenjaroenpun et al., 2020[39] (GEO: SRP166020); two DRS data of WT were merged for SingleMod testing; GLORI data was generated by this study.

3) *Arabidopsis*. Duplicate DRS data of WT (Col-0) and KO (mutant defective in the function of a conserved m6A writer complex component *VIRILIZER*) samples were collected from Matthew T Parker et al., 2020[28] (ENA: PRJEB32782) and then merged respectively; GLORI data was generated by this study.

4) K562. DRS data were collected from Ploy N. Pratanwanich et al., 2021[38] (ENA: PRJEB40872).

5) Zebrafish. DRS data of zebrafish embryo sample were collected from Oguzhan Begik et al., 2022[40] (ENA: PRJEB53494).

6) Rice. DRS data of Rice root sample were collected from Feng Yu et al., 2023[41] (GSA: CRR521994).

7) *P.trichocarpa*. DRS data of *P.trichocarpa* stem sample were collected from Yubang Gao et al., 2021[24] (GEO: SRR12822922).

8) Toxoplasma. DRS data were collected from Dayana C Farhat et al., 2021[42]. Fast5 files were provided by Dr Christopher Swale.

9) *Chlamydomonas*. DRS data was generated by this study.

10) *C.elegans*. DRS data of adult C.elegans sample were collected from Nathan P.Roach et al., 2021[43] (ENA: PRJEB31791).

Nanopore direct RNA sequencing and data processing

Nanopore direct RNA sequencing was conducted following instructions provided by Oxford Nanopore Technologies (Oxford, UK) using DRS kits (SQK-RNA002) and MinION flowcells (FLO-MIN106D, R9.4.1 pore). For both the self-generated data and the public data, raw signals were basecalled using Guppy (v6.3.7) with default parameters and converted into fastq format. Reads that passed the quality threshold (7) and had a length greater than 25bp were then aligned to the reference genome (reference genome used in this study were listed in the Supplementary information) using Minimap2 (v2.22-r1101) with the parameter settings "-ax splice --secondary=no". Samtools (v1.10) was used to filter supplementary alignments and convert sam files to the bam format. Nanopolish[44] (v0.13.2) eventalign module was used to align raw signals to 5mer bases with the parameter settings "--scale-events --samples --signal-index". sam2tsv was used to extract bases quality and alignment results for each read from bam files. In-house python scripts were used to extract and organize current features and base quality features for each DRACN motif within each read. These features were then used as inputs for SingleMod for training or prediction.

GLORI experiments and data processing

GLORI experiments were conducted following instructions provided in Cong Liu et al., 2022[14], with an adjusted incubation temperature in deamination step of 25 °C. Subsequent to GLORI treatment, libraries were constructed using a one-pot RNA library preparation method described in Michael VanInsberghe et al., 2021. Briefly, 50 ng of GLORI-treated RNA sample was repaired using T4 Polynucleotide Kinase (NEB, #M0201S). Pre-adenylated 3'-adapter and 5'-adapter were sequentially ligated to the RNA using T4 RNA Ligase 2, truncated KQ (NEB, #M0373S) and T4 RNA Ligase 1 (NEB, #M0204S), respectively. cDNA synthesis was then performed using HiScript III 1st Strand cDNA Synthesis Kit (Vazyme, #R312-01). Finally, the library was amplified using Q5 High-Fidelity DNA Polymerases with GC enhancer (NEB, #M0491S) and subjected to size selection using a 6% polyacrylamide gel. The resulting library was sequenced on the Illumina NGS platform, with PhiX included to balance base bias. Two biological replicates were used for each species. Raw reads were preprocessed using cutadapt (v4.1), seqkit (v0.8.1) rmdup module and fastp (v0.21.0) for adaptor trimming, deduplication and quality control as well as UMI trimming, respectively. Then, the clean data were processed using the RNA-m5C pipeline developed by Huang T et al., 2022[45], with some modifications described below. The base conversion settings were A-to-G and T-to-C. We used Star (v020201) for alignment with key parameter settings "--outFilterMismatchNoverReadLmax 0.05 --outFilterMultimapNmax 20 --outFilterType BySJout --outFilterScoreMinOverLread 0.8". All A sites with signal greater than 0.9 (representing the ratio of reads with no more than 3 non-converted A to all reads) and coverage greater than 15 after discarding non-converted reads were retained for methylation rate calculation. Subsequently, sites with a difference in methylation rate between two replicates not exceeding 0.05 were kept, and their average methylation rate was used as the benchmark for SingleMod training and testing.

SingleMod

The framework of SingleMod. For a specific genomic position (A base), the m6A modification status for individual molecules is unknown. However, the methylation rate at that site, indicating the proportion of modified molecules, is available. For this type of data, we propose a multi-instance regression framework to achieve single-molecule m6A detection. For a specific DRACN motif, such as GGACT, the following conditions are known:

$$\mathrm{S} = \{\mathrm{S}_1, \mathrm{S}_2, \cdots, \mathrm{S}_\mathrm{n}\},\ S_i \text{ is one of the sites within GGACT motif in the genome}$$

$$\mathrm{Y} = \{\mathrm{Y}_1, \mathrm{Y}_2, \cdots, \mathrm{Y}_\mathrm{n}\},\ Y_i \text{ is the methylation rate of site } S_i$$

$$\mathrm{M} = \{\mathrm{M}_1, \mathrm{M}_2, \cdots, \mathrm{M}_\mathrm{n}\},\ M_i \text{ is the set of molecules at site } S_i$$

$$M_i = \{M_{i,1}, M_{i,2}, \cdots, M_{i,n}\},\ M_{i,j} \text{ is one molecule of set } M_i \text{ at site } S_i$$

$$F_i = \{F_{i,1}, F_{i,2}, \cdots, F_{i,n}\},\ F_{i,j} \text{ is the features of } M_{i,j}$$

The modification status $(R_{i,j})$ of molecule j at site i is unknown:

$$R_i = \{R_{i,1}, R_{i,2}, \cdots, R_{i,n}\},\ R_{i,j} \text{ is the modification status of } M_{i,j}$$

$$R_{i,j} = 1 \ if \ read \ j \ in \ site \ i \ is \ m6A \ modified, 0 \ otherwise$$

Given an initialized network N, the first step of our framework is to use network N to individually assess the modification status $(R_{i,j})$ of each molecule $M_{i,j}$ of M_i at site S_i:

$$p_{i,j} = N(F_{i,j}),\ p_{i,j} \text{is the probability of molecule j in site i being modified}$$

$$R_{i,j} = ATC\ (p_{i,j}), ATC(x) = \begin{cases} 1, & x < 0.5 \\ 0, & x \geq 0.5 \end{cases}$$

$$R_i = \left\{ATC\left(N(F_{i,1})\right), ATC\left(N(F_{i,2})\right), \cdots, ATC\left(N(F_{i,n})\right)\right\}$$

We obtain the prediction methylation rate of site S_i:

$$\hat{Y}_i = \frac{\sum_1^n R_{i,n}}{n}$$

To optimize the parameters of the network, we minimize the mean squared error (MSE) between Y_i and $\hat{Y}_i$ for all sites:

$$\text{minimize } MSE = \frac{\sum_1^n (Y_i - \hat{Y}_i)}{n}$$

The input features of SingleMod. For each molecule $M_{i,j}$ of M_i at site S_i, its features are composed of five parts corresponding to the five positions of A/m6A reside in the pore during sequencing:

$$Fs_{i,j} = \left\{Fs_{-2,i,j}, Fs_{-1,i,j}, Fs_{0,i,j}\ ,\ Fs_{1,i,j}, Fs_{2,i,j}\right\}$$

For example, assuming site S_i is within the GGACT motif, the five positions of A/m6A reside in the pore can be represented as NNGGA, NGGAC, GGACT, GACTN, and ACTNN:

$$Fs_{i,j} = \left\{Fs_{NNGGA,i,j}, Fs_{NGGAC,i,j}, Fs_{GGACT,i,j}\ , Fs_{GACTN,i,j}, Fs_{ACTNN,i,j}\right\}$$

We also annotate features with its sequence information for each position by adding a four-dimensional vector using a embedding layer in the network:

$$\mathrm{K}_{x,\mathrm{i,j}} = \mathrm{embedding}(\mathrm{sequence}_{x,\mathrm{i,j}}) = \left\{\mathrm{K}_{1,x,\mathrm{i,j}}, \mathrm{K}_{2,x,\mathrm{i,j}}, \mathrm{K}_{3,x,\mathrm{i,j}}, \mathrm{K}_{4,x,\mathrm{i,j}}\right\}, x \in \{-2, -1, 0, 1, 2\}$$

$$Fs_{x,i,j} = \left\{\mu_{x,i,j}, m_{x,i,j}, \quad \sigma_{x,i,j}, r_{x,i,j}, n_{x,i,j}, q_{x,i,j}, \mathrm{K}_{1,x,\mathrm{i,j}}, \mathrm{K}_{2,x,\mathrm{i,j}}, \mathrm{K}_{3,x,\mathrm{i,j}}, \mathrm{K}_{4,x,\mathrm{i,j}}\right\}, x \in \{-2, -1, 0, 1, 2\}$$

where $\mu_{x,i,j}, m_{x,i,j}, \sigma\ , R_{x,i,j}, n_{x,i,j}$ represent the mean, median, standard deviation, range and number of raw signals and $q_{x,i,j}$ represent the base quality.

We also directly utilized the raw signal, which may contain hidden information beyond statistical values. We divided the raw signals at each position into five parts and calculated their means, resulting in 25 new raw signals. Additionally, we annotated each raw signal with its corresponding sequence information:

$$Fr_{x,i,j} = \left\{r_{x,i,j}, \mathrm{K}_{1,x,\mathrm{i,j}}, \mathrm{K}_{2,x,\mathrm{i,j}}, \mathrm{K}_{3,x,\mathrm{i,j}}, \mathrm{K}_{4,x,\mathrm{i,j}}\right\}, x \in \{1, 2, \cdots, 25\}$$

$$Fr_{i,j} = \left\{Fr_{1,i,j}, Fr_{2,i,j}, \cdots, \quad Fr_{25,i,j}\right\}$$

Overall,the features $F_{i,j}$ of molecule $M_{i,j}$ of M_i at site S_i:

$$F_{i,j} = \left\{Fs_{i,j}, Fr_{i,j}\right\}$$

The network structure of SingleMod. We employed two independent modules to handle the statistical features $Fs_{i,j}$ and raw signals $Fr_{i,j}$, respectively. Both modules employed one-dimensional convolution as their core operation. Initially, we explored the model architecture using a traditional classification framework. Eventually, we finalized a three-layer one-dimensional convolution with a kernel size of 3. A dropout rate of 0.2 was applied after the final convolution layer, followed by two fully connected layers. Subsequently, the outputs of the two modules were merged using an additional fully connected layer. The probabilities $p_{i,j}$ of molecular modification were obtained through the Softmax function and further transformed into binary values (0 or 1) $R_{i,j}$ using a self-activation function.

Training and testing. We divided all the sites S into twenty intervals based on their methylation rate Y. Within each interval, the sites were randomly divided into training, validation, and testing datasets in an 8:1:1 ratio. The datasets from each interval were then merged to form the final training, validation, and testing datasets. During the training process, we set the batch size to 10 and employed the Adabound optimizer to minimize MSE between $\hat{Y}_i$ and $\mathrm{Y_i}$ for these 10 sites. We conducted training for 600 epochs and selected the model corresponding to the epoch with the minimum MSE on the validation set as the final model. We assessed the model's performance on the testing data using evaluation metrics including MSE, MAE, and Pearson's r. It is important to note that for both training and testing, only sites with a coverage of at least 20 reads were included. We trained models for different motifs separately, initially 13 motifs using data from HEK293T and finally 15 motifs using merged data from HEK293T and *Arabidopsis*. For each motif, we repeated the training process four times. The final models were selected based on their performance on independent testing data from mESC.

Compare SingleMod with existing tools

After obtain the models trained on HEK293T data, we evaluated their performance using data from mESC and *Arabidopsis* and compared them to existing tools. Nanom6A[24], Tombo[46], and Xpore[38] were used as described in Zhendong Zhong et al., 2023[21]. For DENA[25], the "LSTM_extract.py" script was used to extract feature with key parameter "--windows 2 2", and then "LSTM_predict.py" script was used to predict m6A methylation rates of all detected sites. Nanom6A, Tombo and DENA can provide site-specific methylation rates through single-molecule m6A detection, while Xpore estimates site-specific methylation rates through clustering. We compared SingleMod with these tools by assessing the accuracy of their predicted methylation rates against the GLORI benchmark. In addition, we also compared the methylation rates between WT and KO samples.

Compare SingleMod with conventional classification model

We adopted the same input features and network structure of SingleMod to develop conventional classification models, but optimized the parameters of network by minimizing cross entropy loss the using Adam optimizer. All molecule $M_{i,j}$ in site S_i with a methylation rate $Y_i > 0.95$ were treat as positive samples and labeled with 1, while all molecule $M_{i,j}$ in site S_i with a methylation rate < 0.05 were treat as negative samples and labeled with 0. We trained multiple conventional classification models using different ratios of negative samples to positive samples in the training data, including 1:1, 2:1, 4:1, 10:1, 50:1 and using all the negative samples. During training process, we set batch size to 500 and selected model corresponding minimum cross entropy loss on validation data. In addition, we trained balanced SingleMod models by sampling an equal number of sites from different methylation rate intervals as training data instead of using all available sites. We trained the above models on data from HEK293T and *Arabidopsis* and then compared them with SingleMod on data from mESC.

Single molecule m6A profile

m6A prediction at single molecule. The SingleMod predict module was used to predict m6A modifications in 15 motifs within each read, utilizing 15 individual SingleMod models corresponding to each motif. In-house Python scripts were employed to annotate m6A modifications onto each molecule in the BAM file, allowing for visual inspection of m6A modifications on each molecule in IGV.

Construct reference isoforms. bedtools (v2.26.0) was used to convert bam files to bed12 format. bedtools was used to merge 3' end of each reads with key parameter "-d 5" and identified transcription termination regions. flair[47] (v1.6.1) correct was used to corrects misaligned splice sites using genome annotations. Both corrected and inconsistent reads were retained and merged for subsequent use. flair collapse was used to generate intermediate files which were used to extract isoforms with key parameters settings "--gtf genome annotation --generate_map --keep_intermediate --isoform tss --trust_ends". We extracted the non-single-exon isoforms already existing in gtf and isoforms with novel transcription termination sites (TTS) from the flair.collapse.firstpass.bed with awk '$4 ~ /ENST/ && $10 > 1'. Novel isoforms were extracted from flair.collapse.isoforms.bed with awk '$4 !~ /ENST/ && $10 > 1'. Single-exon isoforms with novel TTS were identified by comparing termination regions identified from reads with the 3' end of single-exon isoforms in gtf. The above isoforms and the remaining isoforms in gtf were combined to construct the final reference annotations in bed12 format, with each isforms annotated with its host gene.

Read to isoforms assignment. trmap[48] (v0.12.6) was used to found overlapped reference isoforms for each reads, classifies their relationship based on exon match type and calculated the exon overlap ratio. In-house Python script was used to assign each reads to unique isoform and gene. Briefly, in

759 the read-isoform relationship, if "=" (exact match of intron chain) is present, the read is assigned to
760 the corresponding isoform. If "=" is not present but "c" (contained in reference) exists and the
761 overlap ratio is greater than 0.9, the read is assigned to the corresponding isoform. Lastly, if neither
762 "=" nor "c" is present but "k" (containment of reference) exists and the overlap ratio is greater than
763 0.9, the read is also assigned to the corresponding isoform. If multiple instances of "=", "c", or "k"
764 are present, the isoform with the highest overlap ratio is selected. For other cases, reads cannot be
765 assigned to an isoform. However, we assigned these reads to the host gene corresponding to the
766 isoform with the highest overlap ratio if overlap ratio was greater 0.9. Reads that could be
767 successfully assigned to isoform or gene were considered as full-length reads and used for following
768 analysis.
769 **m6A quantification.** We quantified m6A modifications at different levels. At the cellular level, we
770 analyzed the proportions of molecules with varying m6A numbers. At the gene or isoform level, we
771 calculated the percentage of modified molecules (those with at least one m6A modification), the
772 average number of m6A modifications per molecule (m6A number), and the average number of
773 m6A modifications per molecule per 1000 bp (m6A density). We calculated the counts per million
774 (CPM) for each gene or isoform to represent their expression levels. It is important to note that CPM
775 for direct RNA sequencing and transcripts per million (TPM) for short read sequencing serve as
776 comparable metrics.
777 **Sequence analysis of m6A.** 15 DRACN motifs from all molecules were counted and classified to
778 "Unmodified" and "Modified" based on their modification status. Overall methylation rate of each
779 motif was defined as the ratio of the number of "Modified" motifs to the total number of motifs.
780 Seqlogo was used to plotted the flanking sequence composition of m6A, which represented the
781 motif enrichment at single-molecule level.
782 **Relative position of m6A.** This value was obtained by calculating the ratio of the distance between
783 m6A position and the 5' end of the molecule (or exon) to the length of the molecule (or exon). To
784 define the region interval of start codon, last exon junction and stop codon, the relative position of
785 these landmarks were plotted firstly, and then we applied a sliding window manner to choose the
786 final region interval when the quantity occupied half and the window size was minimized. For gene-
787 specific m6A distribution, RNA molecules are divided into three equal intervals. If a gene's m6A
788 modifications exceed 25% within any interval, it is deemed to have significant m6A modifications
789 in that interval.
790 **Overall methylation rate at positions with distance from landmarks and in different exon**
791 **types.** 15 DRACN motifs from all molecules were counted and classified to "Unmodified" and
792 "Modified" based on their modification status. For overall methylation rate at positions with distance
793 to landmarks, these value were derived by calculating the proportion of modified motifs that share
794 the same distance (5 bp were grouped into a bin) from landmarks, relative to the total number of

motifs. The overall methylation rate for different exon types, is calculated as the proportion of modified motifs within responding exon type.

Relationship of adjacent m6A. To assess interaction between adjacent m6A, m6A sites pair (both coverage >= 50 and methylation rate >= 0.1) with distance between them less than 100 bp were used. We compared the theoretically probabilistic ratio to the observed ratio of molecules with both m6A modifications to the total number of molecules. If they are significantly unequal, it indicates that the modification status of one motif on a single molecule influences the status of adjacent motif.

Assess m6A cluster at single-molecule level. m6A clusters (with at least two m6a sites) were identified by merging m6A sites (coverage >= 10, methylation rate >= 0.1 and modified molecules >= 3) using bedtools merge with key parameter settings "-d 100". We then calculated the proportion of molecules with different numbers of m6A modifications within the cluster. We also counted the number of molecules containing m6A clusters with varying number of m6A modifications by merging adjacent m6A modifications with distance less than 100 bp within single RNA molecule.

RNA heterogeneity. To describe differences among molecules within a specific gene, we introduce the concept of heterogeneity. We quantify heterogeneity through the probabilistic sampling of two identical molecules within a gene, with equal likelihood. A higher probability signifies reduced heterogeneity. We explore two facets of molecular heterogeneity: one arising from splicing form variations or PAS selection variations (isoform-related), and the other arising from m6A modifications.

Overall methylation rate at positions with varying distance from one exon edge. 15 DRACN motifs from all molecules were counted and classified to "Unmodified" and "Modified" based on their modification status. These value were derived by calculating the proportion of modified motifs that share the same distance (20 bp were grouped into a bin) from exon edge. For the middle exons, when calculating the overall methylation rates toward one edge, only sites with a distance to the opposite edge greater than 200bp were retained for calculation.

m6A and RNA stability

RNA stability data were obtained from Qiushuang Wu et al., 2019[49]. To investigate the relationship between m6A levels and RNA stability, we simultaneously compared the RNA stability of genes with different proportion of modified molecules, average m6A modifications per molecule, and m6A density. To investigate whether the role of m6A in promoting RNA degradation is influenced by its precise location, we compared RNA stability among genes with different primary m6A contribution intervals, while keeping m6A density constant. Specifically, we divided the molecules into three intervals, and if a gene's m6A content within one of these intervals exceeded 60%, it was considered that the primary source of m6A for that gene was that particular interval. When

comparing m6A across different types of genes, we categorized protein-coding genes into regulatory genes and housekeeping genes, using the housekeeping gene set from the HRT Atlas v1.0 database[50].

m6A and RNA processing

Identify differential alternative splicing and alternative polyadenylation events upon METTL3 KO. Utilizing the aforementioned outcomes of the "Read to isoforms assignment" we assigned reads to genes if the read-isoform relationship was "=" or the overlap ratio was greater than 0.7 for each of the three replicates in both WT and KO samples. For each gene, we employed a modified version of call_diffsplice_events.py from Flair and in-house Python script es_detect.py to search for AS events and count reads with different splicing patterns for WT and KO samples. As for APA events, we counted reads with different PAS selections according to transcription termination regions constructed in "Construct reference isoforms". DRIMSeq[51] (3.17) was used to calculate inclusion level (PAS usage) and associated p-value. AS or APA events were considered differential between the WT and KO samples if the p-value < 0.01, and the change in inclusion level (max PAS usage change) >= 0.1.

Multi-to-multi correlation analyses between m6A sites and AS or APA events. Within WT samples, there are naturally contrasting molecules with different modification status at specific sites or different splicing patterns/PAS usage in specific AS or APA events, allowing us to study the relationship between m6A and AS or APA events. Specifically, concerning each m6A sites, we searched for AS (APA) events and counted reads with different splicing patterns (PAS selections) for modified and unmodified molecules. Consequently, for a specific correlation analysis between a m6A site and an AS or APA event, we constructed a 2×2 contingency table and conducted a Fisher's test to calculate significance, and if the p-value < 0.01, they were considered to be correlated.

Quantification of exclusion and inclusion effect

Validation of exclusion deposition mode in vertebrates. We first calculated the overall methylation rate for each motif in regions that were not affected by EJC, specifically the middle-exon regions with distance from both edge greater than 200bp, to represent the intrinsic efficiency of m6A writer. Under the assumption that in exclusion deposition mode, the methylation rates at positions where exclusion is absent would stabilize at the efficiency of m6A writer, we then compared the overall methylation rate at positions with varying distance from one exon edge (20 bp were grouped into a bin) to the efficiency of m6A writer. When calculating the overall methylation rates toward the 5' edge of first exons and both edges of middle exons, to eliminate the influence of the opposite edge, only motifs with distance from the opposite edge greater than 200bp were retained. As for the 3' edge of first exons, only motifs with distance from the opposite edge greater than 400bp were retained.

Quantification of exclusion effect in vertebrates. We calculated overall methylation rate and counts of 15 DRACN motifs at positions with varying distance to both edges (20 bp were grouped into a bin). We then calculated the hypothetical m6A counts without exclusion by multiplying the number of motifs at each position by the efficiency of m6A writer for each motif and summing them all up. The exclusion effects were calculated according to the following formula:

$$exclusion\ effect = \frac{hypothetical\ m6A\ counts - observed\ m6A\ counts}{hypothetical\ m6A\ counts}$$

Determine the span of inclusion and exclusion. For plants-like pattern, we calculated the overall methylation rates for each motif at positions with varying distance from 3' terminal (20 bp were grouped into a bin) for last exon with different length. We determined the inclusion span to be within 300 bp from 3'edge for all plants and Toxoplasma, as outside this span the methylation rates are nearly 0. For *Chlamydomonas* pattern, we calculated the overall methylation rates for each motif at positions with varying distance from 5' edge of last exon (20 bp were grouped into a bin). We determined the whole last exon as the inclusion span and the first 500bp as the exclusion span, as when the distance from 5' edge reaches 500bp, the methylation rates would stabilize.

Determine the background methylation rates due to random deposition. Since the first and middle exons are not within the span of the inclusion effect seen in the plant-like pattern and *Chlamydomonas* pattern, we calculated the overall methylation rate in these regions as the background methylation rates, which represents the efficiency of random m6A deposition.

Determine the reference methylation rates with inclusion effect. For plants-like pattern, to eliminate potential exclusion effect from the 5' edge, we only used last exons with a length greater than 400bp to calculate the overall methylation rates for each motif at positions with varying distance from 3' edge (20 bp were grouped into a bin). These values were used as the reference methylation rates at different positions within inclusion span. For *Chlamydomonas* pattern, as the methylation rates stabilize outside the exclusion span, the overall methylation rates in regions with distance from 5' edge greater than 500bp were used as the reference methylation rates within inclusion span.

Quantification of inclusion and exclusion effects in plants, Toxoplasma and *Chlamydomonas*. The background counts of m6A modifications were calculated by multiplying the counts of motifs by background methylation rates for each motif and summing them all up. The theoretical counts of m6A modifications following inclusion were calculated by multiplying the counts of motifs by the reference methylation rates within inclusion span, and then adding the background m6A counts outside the inclusion span. The exclusion effect and inclusion effects were calculated using the following formula:

$$inclusion\ effect = \frac{hypothetical\ m6A\ counts - background\ m6A\ counts}{hypothetical\ m6A\ counts}$$

$$exclusion\ effect = \frac{hypothetical\ m6A\ counts - observed\ m6A\ counts}{hypothetical\ m6A\ counts}$$

References:

1. Roundtree, I.A., Evans, M.E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *CELL* **169**, 1187-1200 (2017).
2. Zaccara, S., Ries, R.J. & Jaffrey, S.R. Reading, writing and erasing mRNA methylation. *Nat Rev Mol Cell Biol* **20**, 608-624 (2019).
3. Yu, Q. et al. RNA demethylation increases the yield and biomass of rice and potato plants in field trials. *NAT BIOTECHNOL* **39**, 1581-1588 (2021).
4. Frye, M., Harada, B.T., Behm, M. & He, C. RNA modifications modulate gene expression during development. *SCIENCE* **361**, 1346-1349 (2018).
5. Jiang, X. et al. The role of m6A modification in the biological functions and diseases. *Signal Transduct Target Ther* **6**, 74 (2021).
6. Liang, Z. et al. m6A-Atlas v2.0: updated resources for unraveling the N6-methyladenosine (m6A) epitranscriptome among multiple species. *NUCLEIC ACIDS RES* (2023).
7. Meyer, K.D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *CELL* **149**, 1635-1646 (2012).
8. Dominissini, D. et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *NATURE* **485**, 201-206 (2012).
9. Huang, H., Weng, H. & Chen, J. The Biogenesis and Precise Control of RNA m(6)A Methylation. *TRENDS GENET* **36**, 44-52 (2020).
10. Yang, Y., Hsu, P.J., Chen, Y.S. & Yang, Y.G. Dynamic transcriptomic m(6)A decoration: writers, erasers, readers and functions in RNA metabolism. *CELL RES* **28**, 616-624 (2018).
11. Uzonyi, A. et al. Exclusion of m6A from splice-site proximal regions by the exon junction complex dictates m6A topologies and mRNA stability. *MOL CELL* **83**, 237-251 (2023).
12. He, P.C. et al. Exon architecture controls mRNA m(6)A suppression and gene expression. *SCIENCE* **379**, 677-682 (2023).
13. Yang, X., Triboulet, R., Liu, Q., Sendinc, E. & Gregory, R.I. Exon junction complex shapes the m(6)A epitranscriptome. *NAT COMMUN* **13**, 7904 (2022).
14. Liu, C. et al. Absolute quantification of single-base m(6)A methylation in the mammalian transcriptome using GLORI. *NAT BIOTECHNOL* **41**, 355-366 (2023).
15. Hu, L. et al. m(6)A RNA modifications are measured at single-base resolution across the mammalian transcriptome. *NAT BIOTECHNOL* **40**, 1210-1219 (2022).
16. Dierks, D. et al. Multiplexed profiling facilitates robust m6A quantification at site, gene and sample resolution. *NAT METHODS* **18**, 1060-1067 (2021).

17. Linder, B. et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *NAT METHODS* **12**, 767-772 (2015).

18. Xiao, Y.L. et al. Transcriptome-wide profiling and quantification of N(6)-methyladenosine by enzyme-assisted adenosine deamination. *NAT BIOTECHNOL* **41**, 993-1003 (2023).

19. Garalde, D.R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *NAT METHODS* **15**, 201-206 (2018).

20. Workman, R.E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *NAT METHODS* **16**, 1297-1305 (2019).

21. Zhong, Z.D. et al. Systematic comparison of tools used for m(6)A mapping from nanopore direct RNA sequencing. *NAT COMMUN* **14**, 1906 (2023).

22. Alfonzo, J.D. et al. A call for direct sequencing of full-length RNAs to identify all modifications. *NAT GENET* **53**, 1113-1116 (2021).

23. Liu, H. et al. Accurate detection of m(6)A RNA modifications in native RNA sequences. *NAT COMMUN* **10**, 4079 (2019).

24. Gao, Y. et al. Quantitative profiling of N(6)-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *GENOME BIOL* **22**, 22 (2021).

25. Qin, H. et al. DENA: training an authentic neural network model using Nanopore sequencing data of Arabidopsis transcripts for detection and quantification of N(6)-methyladenosine on RNA. *GENOME BIOL* **23**, 25 (2022).

26. Lorenz, D.A., Sathe, S., Einstein, J.M. & Yeo, G.W. Direct RNA sequencing enables m(6)A detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**, 19-28 (2020).

27. Hendra, C. et al. Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *NAT METHODS* **19**, 1590-1598 (2022).

28. Parker, M.T. et al. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A modification. *ELIFE* **9** (2020).

29. He, P.C. & He, C. m(6) A RNA methylation: from mechanisms to therapeutic potential. *EMBO J* **40**, e105977 (2021).

30. Murakami, S. & Jaffrey, S.R. Hidden codes in mRNA: Control of gene expression by m(6)A. *MOL CELL* **82**, 2236-2251 (2022).

31. Wang, X. et al. N6-methyladenosine-dependent regulation of messenger RNA stability. *NATURE* **505**, 117-120 (2014).

32. Wei, G. et al. Acute depletion of METTL3 implicates N (6)-methyladenosine in alternative intron/exon inclusion in the nascent transcriptome. *GENOME RES* **31**, 1395-1408 (2021).

33. Xiao, W. et al. Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. *MOL CELL* **61**, 507-519 (2016).

34. Sendinc, E., Valle-Garcia, D., Jiao, A. & Shi, Y. Analysis of m6A RNA methylation in Caenorhabditis elegans. *CELL DISCOV* **6**, 47 (2020).

35. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K.F. Nanopore sequencing technology, bioinformatics and applications. *NAT BIOTECHNOL* **39**, 1348-1365 (2021).

36. Chen, X. et al. Targeted RNA N(6) -Methyladenosine Demethylation Controls Cell Fate Transition in Human Pluripotent Stem Cells. *Adv Sci (Weinh)* **8**, e2003902 (2021).

37. Liu, J. et al. The RNA m(6)A reader YTHDC1 silences retrotransposons and guards ES cell identity. *NATURE* **591**, 322-326 (2021).

38. Pratanwanich, P.N. et al. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *NAT BIOTECHNOL* **39**, 1394-1402 (2021).

39. Jenjaroenpun, P. et al. Decoding the epitranscriptional landscape from native RNA sequences. *NUCLEIC ACIDS RES* **49**, e7 (2021).

40. Begik, O. et al. Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cDNA sequencing. *NAT METHODS* **20**, 75-85 (2023).

41. Yu, F. et al. Identifying RNA Modifications by Direct RNA Sequencing Reveals Complexity of Epitranscriptomic Dynamics in Rice. *Genomics Proteomics Bioinformatics* (2023).

42. Farhat, D.C. et al. A plant-like mechanism coupling m6A reading to polyadenylation safeguards transcriptome integrity and developmental gene partitioning in Toxoplasma. *ELIFE* **10** (2021).

43. Roach, N.P. et al. The full-length transcriptome of C. elegans using direct RNA sequencing. *GENOME RES* **30**, 299-312 (2020).

44. Simpson, J.T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *NAT METHODS* **14**, 407-410 (2017).

45. Huang, T., Chen, W., Liu, J., Gu, N. & Zhang, R. Genome-wide identification of mRNA 5-methylcytosine in mammals. *NAT STRUCT MOL BIOL* **26**, 380-388 (2019).

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

47. Tang, A.D. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *NAT COMMUN* **11**, 1438 (2020).

48. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9** (2020).

49. Wu, Q. et al. Translation affects mRNA stability in a codon-dependent manner in human cells. *ELIFE* **8** (2019).

50. Hounkpe, B.W., Chenou, F., de Lima, F. & De Paula, E.V. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *NUCLEIC ACIDS RES* **49**, D947-D955 (2021).

51. Nowicka, M. & Robinson, M.D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res* **5**, 1356 (2016).

Data availability

Sequencing data generated in this work can be found in GEO with the accession number GSE246632. Token for reviewer evaluation is onepemaqdtavlez

Acknowledgements

This work was supported by the Ministry of Science and Technology of China (2022YFC3400400, 2022YFA0912900 and 2019YFA0802203), National Natural Science Foundation of China (92253202, 32271499, 32270644, and 32100461), and Shenzhen Bay Scholars Program.

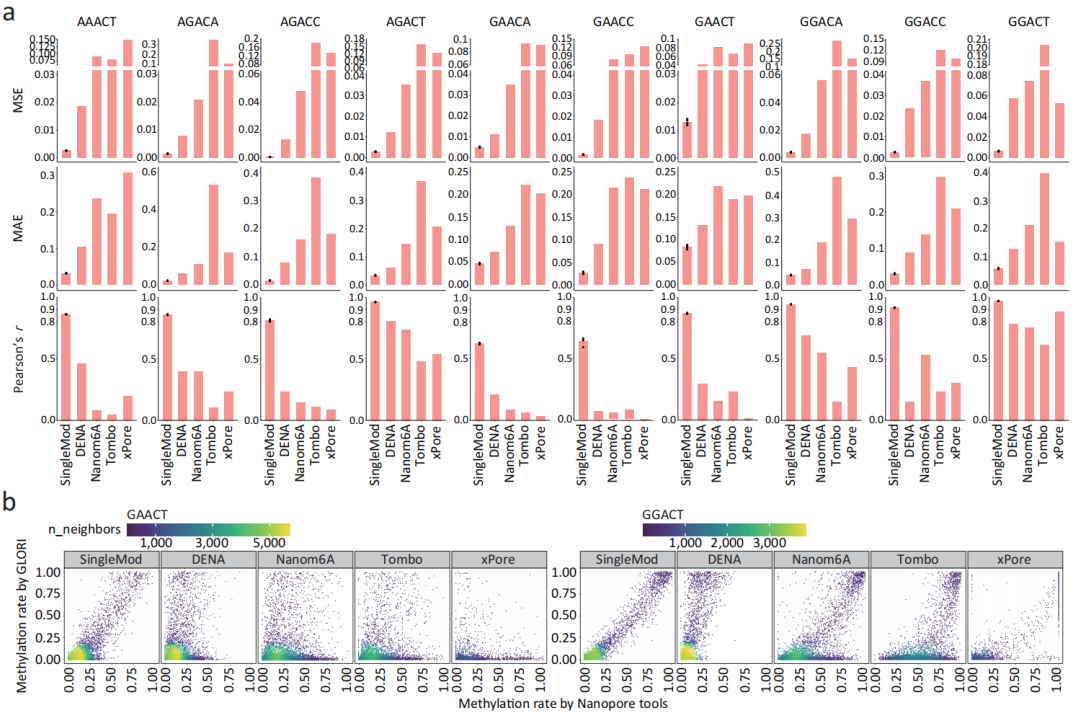
Author Contributions

G.Z.L. conceived the project; Y.Y.X. and Z.D.Z. analyzed the data with the assistance from Y.L.L., B.D.L., and J.W.K.; Y.Y.X., Z.D.Z., Z.Z., and G.Z.L. wrote the manuscript; H.X.C. conducted the experiments with the assistance from F.W., R.J.L., and Y.S.; Y.Y.X., Z.D.Z., F.Y., Z.L., Z.Z., and G.Z.L. revised the manuscript. All authors reviewed the results and approved the final version of the manuscript.

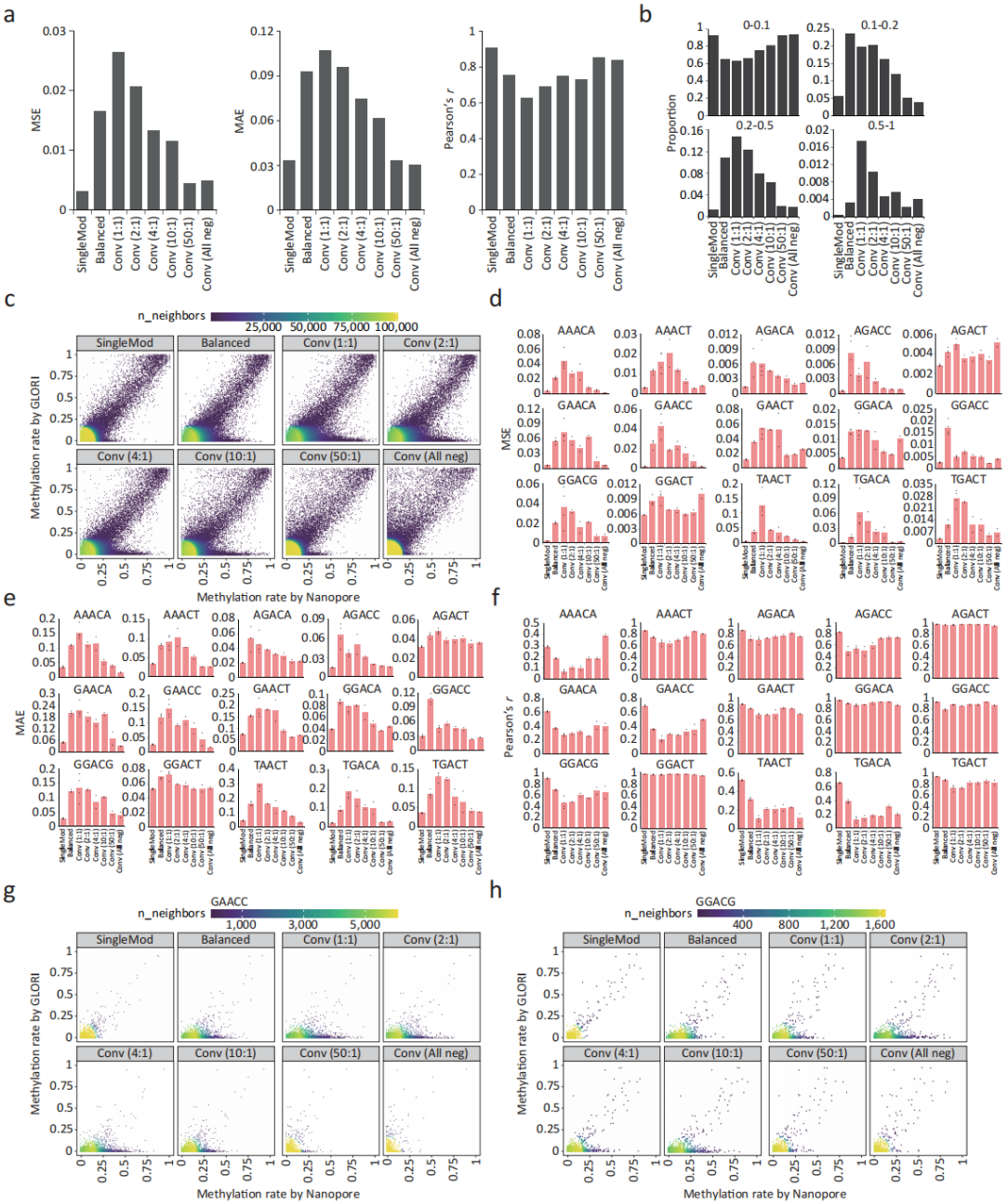
Competing Interests

The authors declare no competing financial interests.

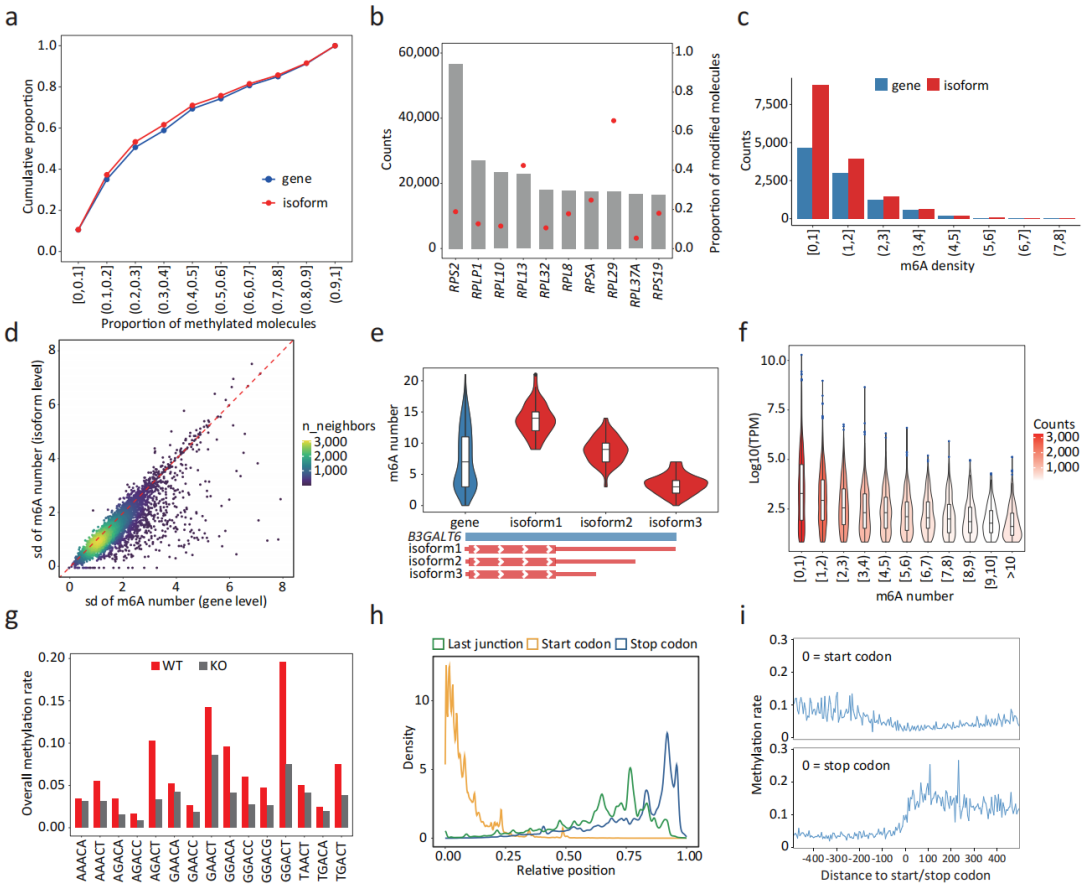
Extended Data Fig. 1|The development of SingleMod. a Statistics of GLORI and nanopore DRS data derived from HEK293T. From left to right: the number of m6A sites identified by GLORI; the number of these sites covered by DRS data; the number of DRS reads corresponding to these sites; the number of 9-mer types corresponding to these sites. **b** Comparison of features employed for training between m6A (M) and A (U) at different positions within 9-mer CGGGACTGA. **c** Comparing the influence of flanking bases and m6A modification on the current shift. **d-e** Using conventional classification framework to explore the optimal network architecture which is then used by SingleMod. **f** Examining the performance of SingleMod across diverse intervals when incorporating different proportions of low-methylation rate sites (<=0.05).



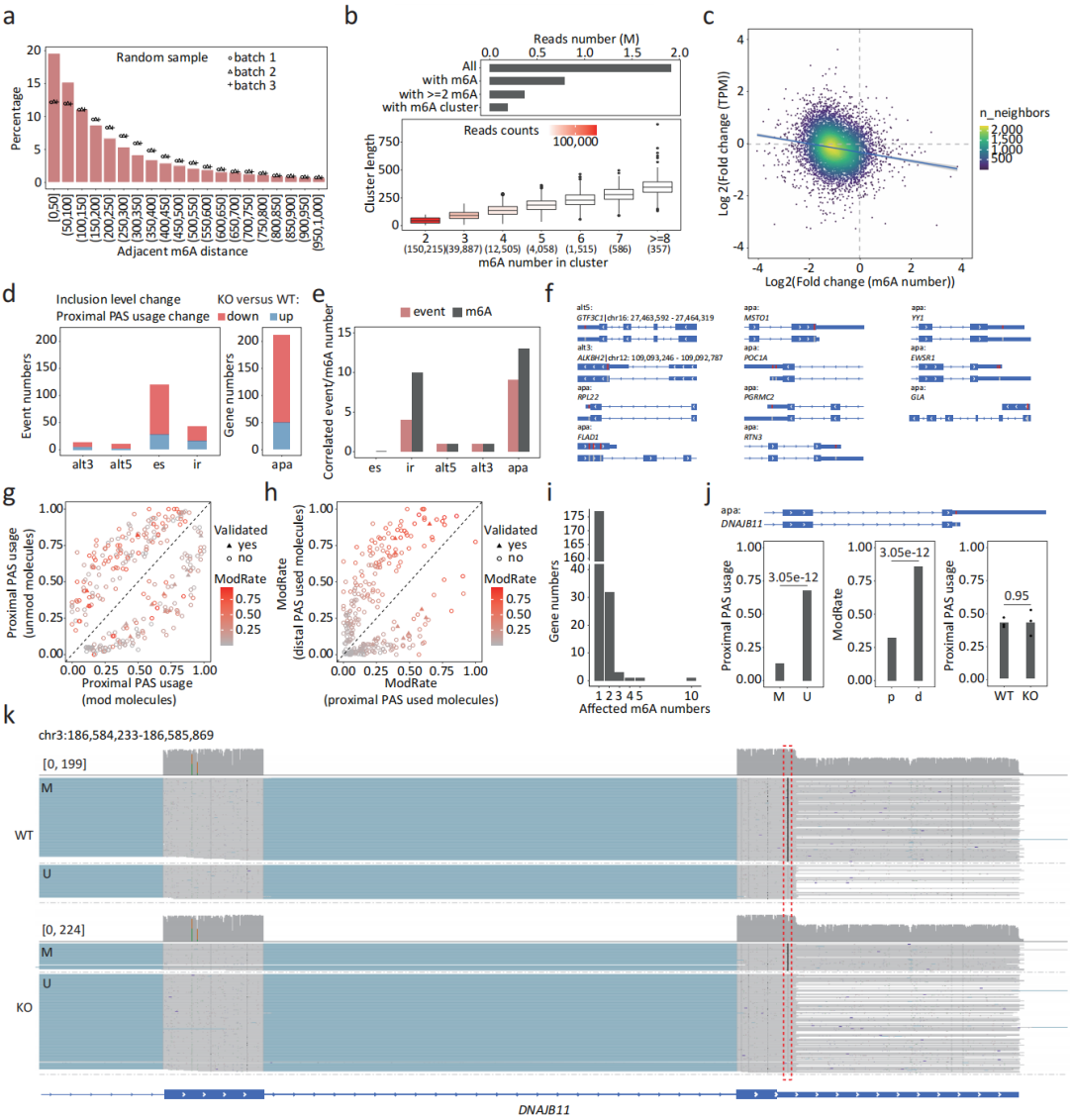
Extended Data Fig. 2|Comparison of the performance between SingleMod and other tools on mESC data. a The performance of SingleMod and other tools on various motifs. **b** Comparing the methylation rates predicted by nanopore tools to GLORI benchmarks. Two motifs GAACT and GGACT were used as representative.



Extended Data Fig.3|Comparison of the performance between SingleMod and conventional classification models on mESC data. a The overall performance of the final version of SingleMod (combining data from HEK29T and *Arabidopsis* for training), a balanced-form SingleMod (Balanced) and multiple conventional classification models (Conv) with varying ratio of negative to positive samples during training. **b** Proportion of sites in different error intervals. **c** Comparing the methylation rates predicted by different models to GLORI benchmarks. **d-f** The performance of SingleMod and conventional classification models in various motifs. **g-h** Comparing the methylation rates predicted by different models to GLORI benchmarks. Two motifs GAACC and GGACG were used as representative.

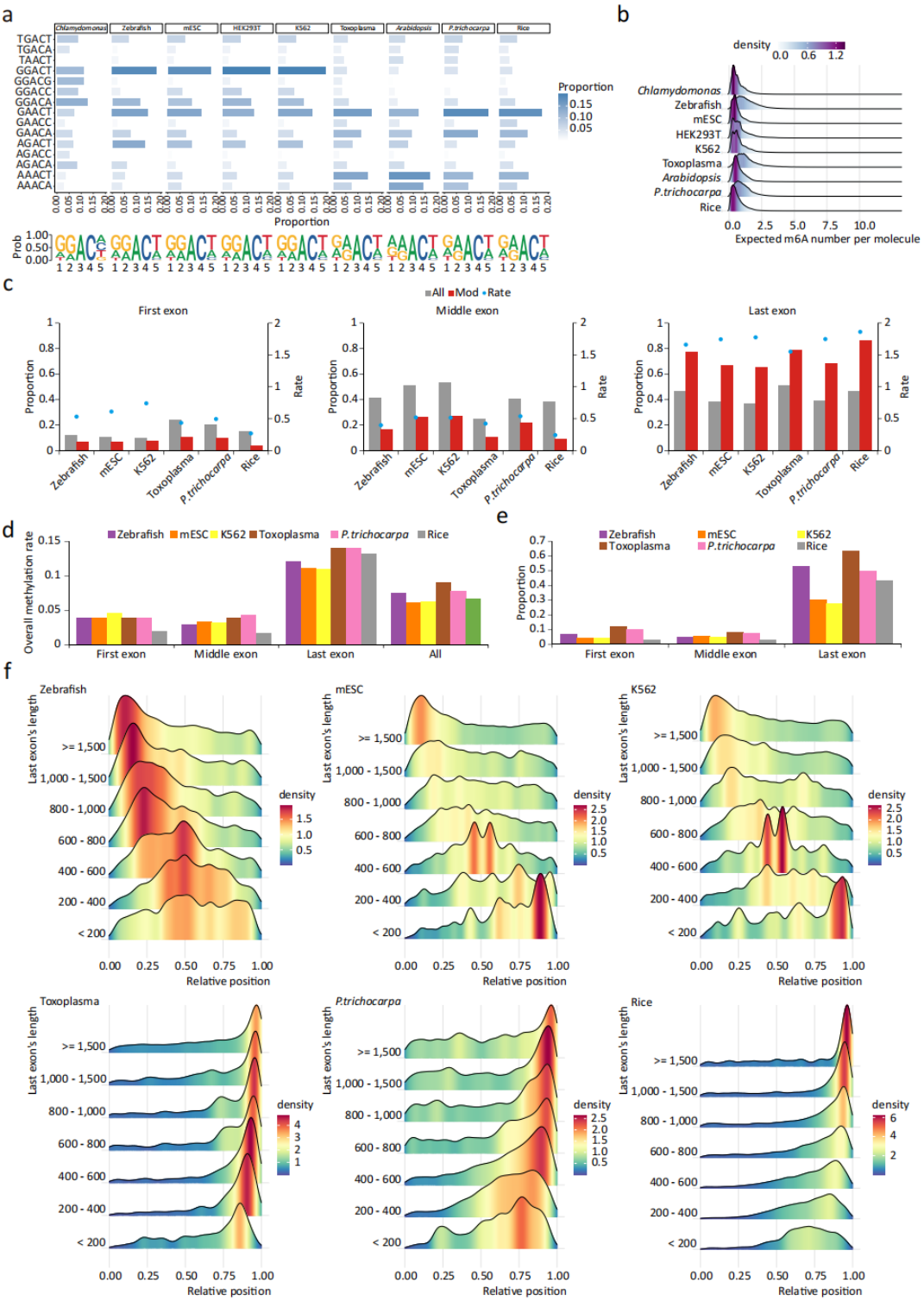


Extended Data Fig. 4|Single-molecule m6A profiles in HEK293T. a Cumulative proportion of molecules of gene groups with varying proportion of modified molecules. **b** Total molecules (bar) and proportions of modified molecules (dot) for the top 10 highest expressed genes. **c** Counts of genes or isoforms with varying m6A density. **d** Comparing the standard deviation in m6A number among molecules between gene and isoform level (the mean standard deviation of multiple isoforms). **e** Distribution of m6A number in each single molecule from gene *B3GALT6* and its three types of isoforms. **f** Distribution of expression levels of isoforms with varying m6A number. "TPM" indicates Transcripts Per Million. **g** Overall methylation rate of different motifs in WT and KO samples. **h** The distribution of relative position of start codon, last junction and stop codon within their molecules. **i** Overall methylation rate around start codon (top) and stop codon (bottom).



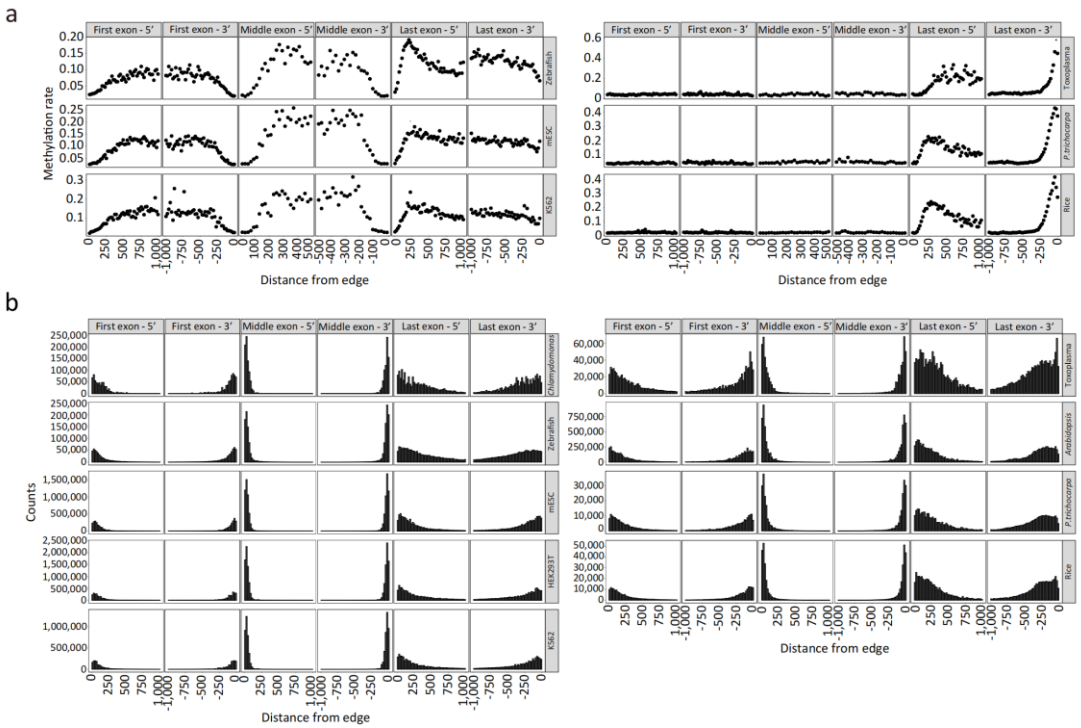
Extended Data Fig. 5|Interplay between m6A and RNA processing in HEK293T. a Percentage of different distances between two adjacent m6A in individual molecule. **b** Top: number of different types of reads. Bottom: length of clusters with varying m6A numbers and their corresponding reads numbers. **c** Correlation between changes in gene expression and m6A level following *METTL3* KO. **d** Numbers of differential AS/APA events following *METTL3* KO. **e-f** The overlap between differential AS/APA events and corresponding responsible m6A sites identified through correlation analysis. **g-h** Dotplot comparing the proximal PAS usage (**g**) or methylation rate (**h**) between two classes of molecules, as calculated in Fig.3k. Each dot represents significant m6A site-APA correlated event identified through correlation analysis. Methylation rates of corresponding m6A and whether the correlated events exhibited differential APA following *METTL3* KO are also indicated (validated). **i** Summarizing significant m6A site-APA correlated events assuming that m6A deposition is influenced by the processes or outcomes of PAS selection. **j-k** A representative gene *DNAJB11*. *A* significant m6A site-APA correlated events was detected (**j**, left and middle), however, proximal PAS usage remained unaltered following *METTL3* KO (**j**, right). In the IGV snapshot

which provided a visual representation of this circumstance (**k**), reads were grouped based on the modification status of the correlated m6A site (red box).

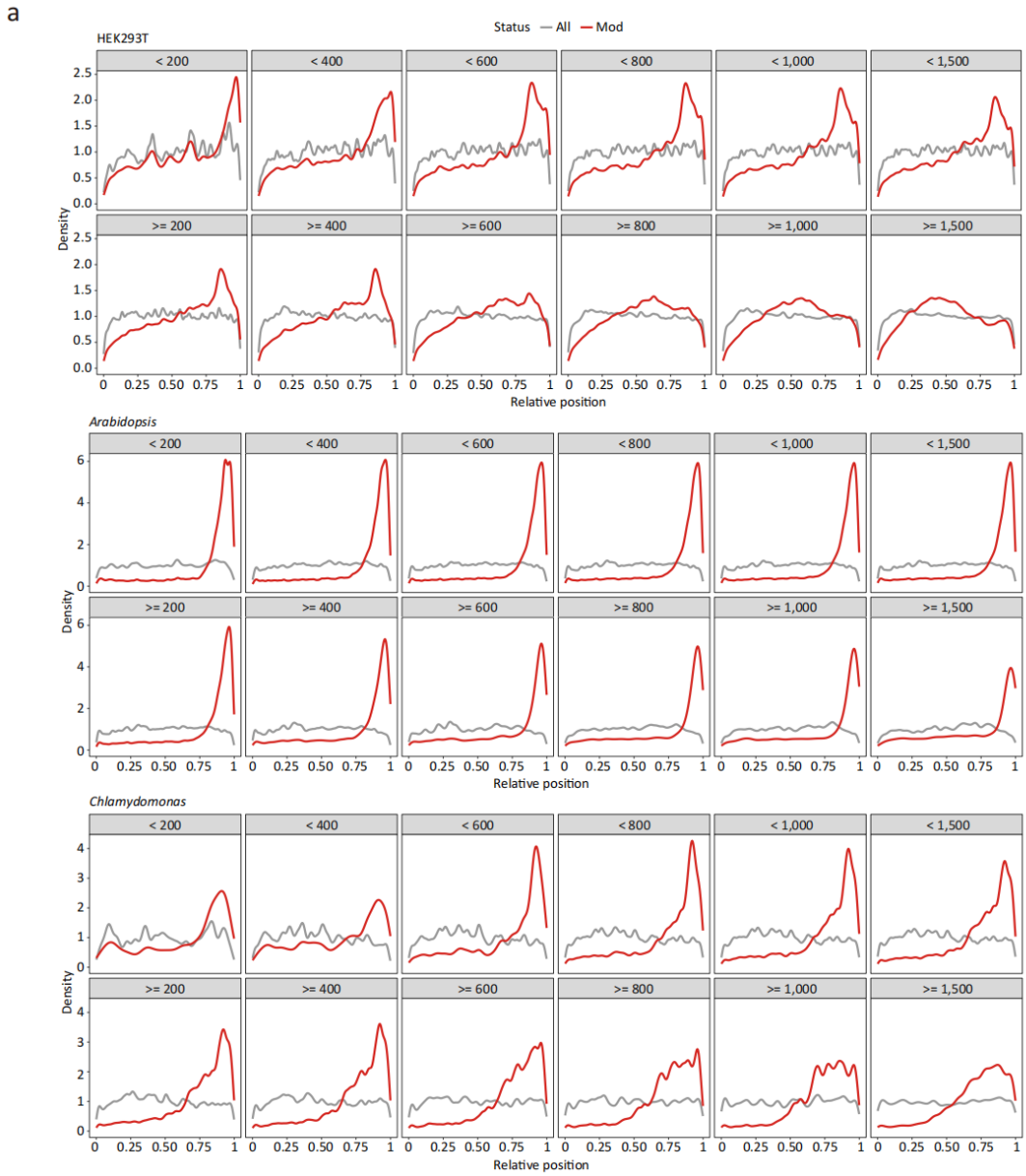


Extended Data Fig. 6|Single-molecule m6A profiles across multi-species. a Proportion of modified A bases within various motifs. The composition of m6A flanking bases plotted by Seqlogo are shown below. **b** Distribution of the expected number of m6A modifications per molecule. The expected number was calculated by summing the overall methylation rate of corresponding on a molecule. **c** Bar: proportions of background (All) and modified (Mod) A bases in distinct exons. Dot: the ratio between these two values. **d** Overall methylation rate within distinct exons. **e**

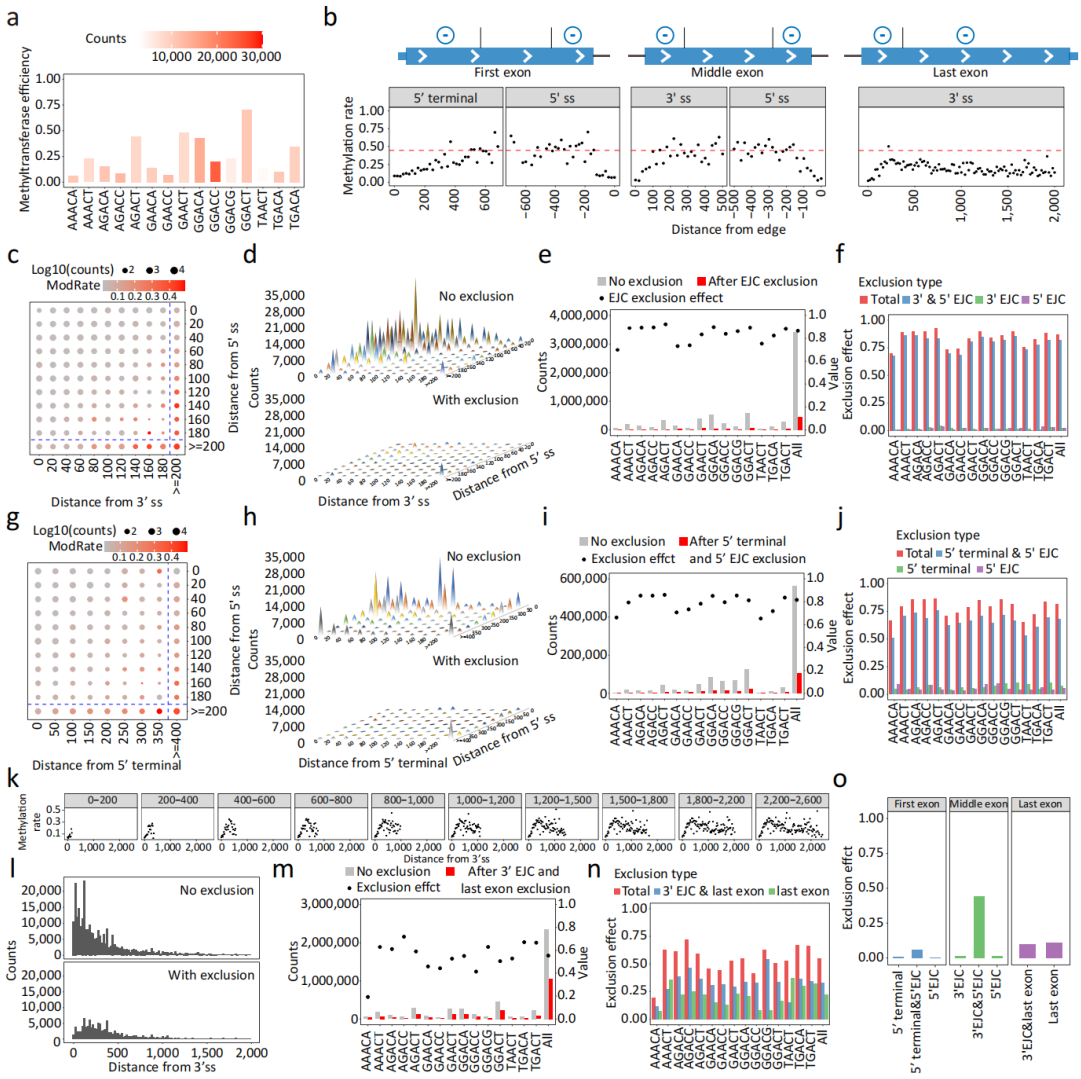
Proportion of modified exons (containing at least one m6A). **f** Relative distribution of m6A sites within last exons with varying exon lengths.



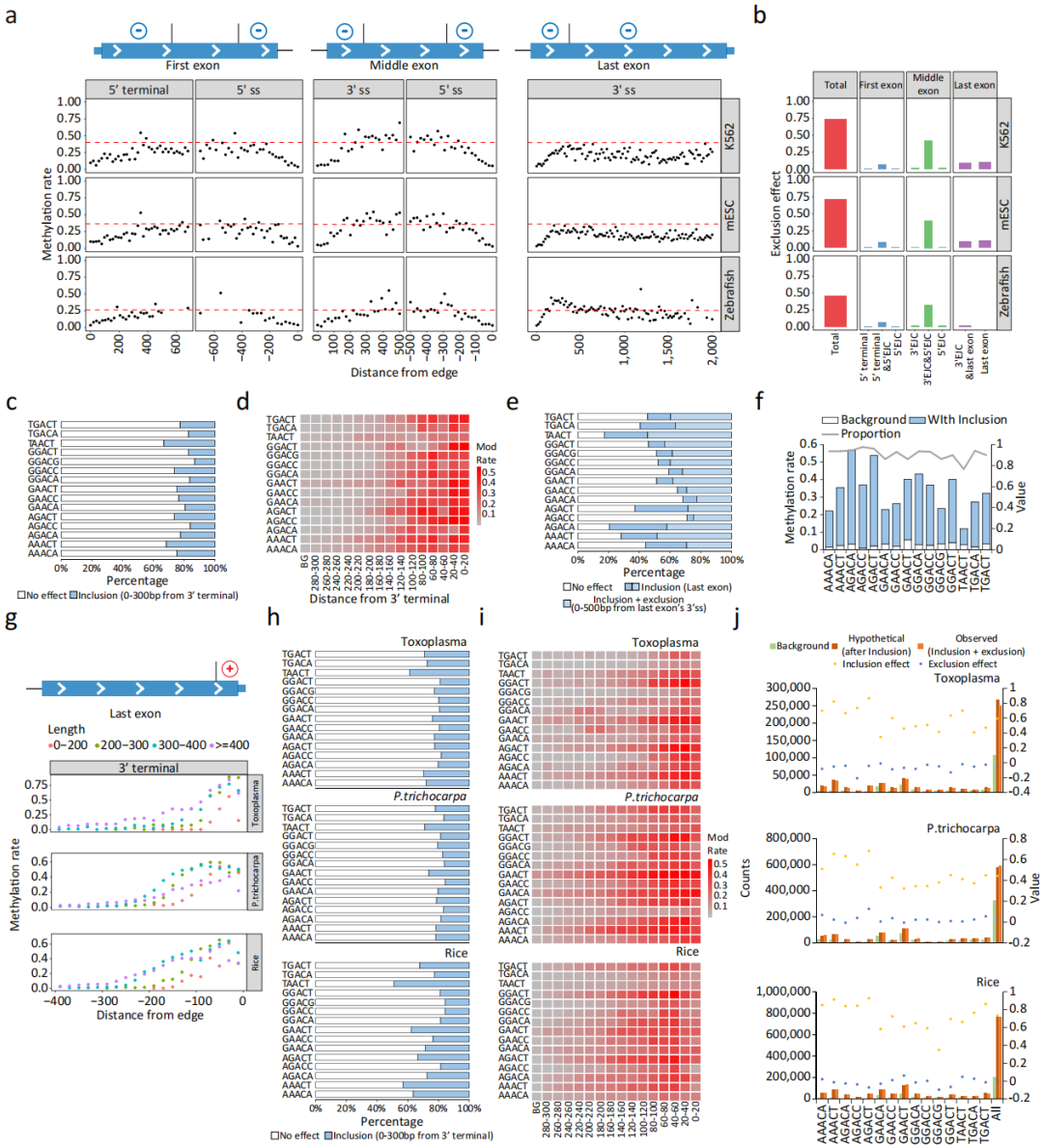
Extended Data Fig. 7|Overall methylation rate (a) and counts of 15 DRACN motifs (b) with varying distance from one exon edge.



Extended Data Fig. 8|Distribution of relative position of background (All) and modified (Mod) A bases within their respective molecules with different last exon length.



Extended Data Fig. 9|Quantitative analysis of exclusion deposition mode in HEK293T. a Intrinsic efficiency of m6A deposition for different motifs. **b** The same as Fig. 5a, but using AGACT motif as a representative. **c, d, g, h** Overall methylation rate and counts of motifs at positions with varying distance from both edges, and corresponding hypothetical counts (**d** and **h**, top) and observed counts (**d** and **h**, bottom) of m6A modifications, in the middle (**c, d**) and first (**g, h**) exons, using GGACA motif as a representative. **e, i** Bar: hypothetical counts and observed counts of m6A modifications in the middle (**e**) and first (**i**) exons for all motifs. Dot: exclusion effect. **f, j** Exclusion effects from different edges or their combination in the middle (**f**) and first (**j**) exons. **k-l** Overall methylation rate at positions with varying distances from 3'ss in last exon with different lengths (**k**), and corresponding hypothetical counts (**l**, top) and observed counts (**l**, bottom) of m6A modifications, using GGACA motif as a representative. **m** Bar: hypothetical counts and observed counts of m6A modifications in the last exon for all motifs. Dot: exclusion effect. **n** Exclusion effects from different edges or their combination in the last exons. **o** Overall exclusion effects contributed by different exons and different edges.



Extended Data Fig. 10|Quantitative analysis of exclusion and inclusion deposition mode across multi-species. a Top: schematic representation of exclusion deposition mode in vertebrates. Bottom: m6A pattern shaped by exclusion deposition mode and comparing the overall methylation rate at positions with varying distances from one exon edge to hypothetical methylation rate (red dashed lines), using GGACA motif as a representative. **b** The total exclusion effect and the exclusion effect contributed by different exons and edges in different vertebrates. **c** Percentage of motifs within and outside selection span. Data used in **c-d** are derived from *Arabidopsis*. **d** Overall methylation rate within inclusion span with varying distances from 3' terminal for various motifs. "BG" indicates background methylation rate outside inclusion span. **e** Percentage of motifs located in inclusion span, combination of inclusion and exclusion span, and background span (no effect). Data used in **e-f** are derived from *Chlamydomonas*. **f** Bar: overall methylation rate with and without (background) inclusion effect. Line: contribution of inclusion effect to overall methylation rate. **g** Top: schematic

representation of inclusion deposition mode in plants and *Toxoplasma*. Bottom: m6A pattern shaped by inclusion deposition mode and overall methylation rate at positions with varying distances from 3' terminal, using AAACT motif as a representative. **h** The same as (**c**) but in other plants and *Toxoplasma.* **i** The same as (**d**) but in other plants and *Toxoplasma.* **j** Counts of m6A modifications under diverse situations in plants and *Toxoplasma*. Inclusion and exclusion effects are also calculated and marked.