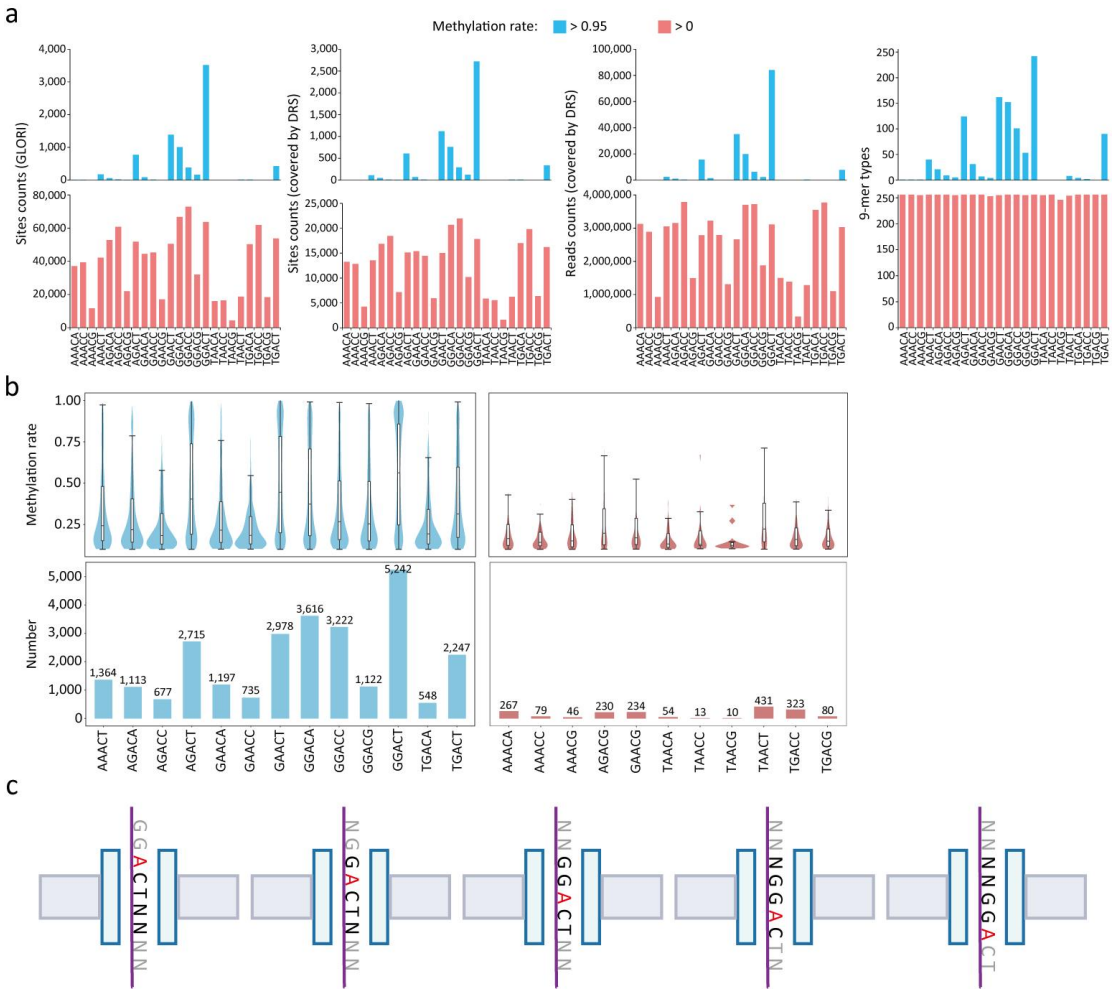
Supplementary information

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

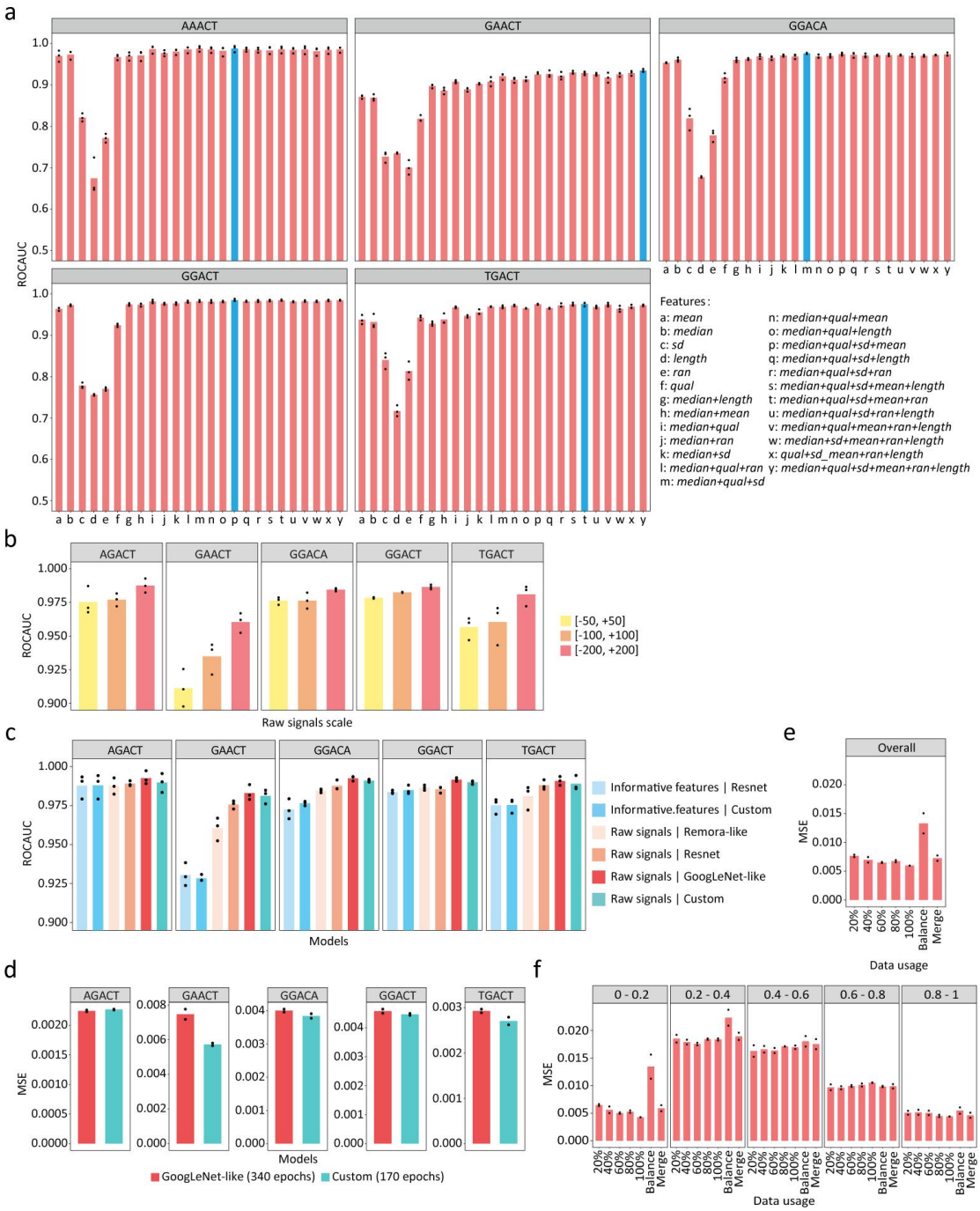
Ying-Yuan Xie[1,2,6], Zhen-Dong Zhong[1,3,6], Hong-Xuan Chen[1,2,6], Ze-Hui Ren[1,2], Yuan-Tao Qiu[1], Ye-Lin Lan[1], Fu Wu[1], Jin-Wen Kong[1], Ru-Jia Luo[1], Delong Zhang[4], Biao-Di Liu[1], Yang Shu[1], Feng Yin[5], Jian Wu[4], Zigang Li[5], Zhang Zhang[1,3,*], Guan-Zheng Luo[1,2,3,5*]

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 Sun Yat-sen University Institute of Advanced Studies Hong Kong, Science Park, Hong Kong SAR, 999077, China
3 Innovation Center for Evolutionary Synthetic Biology, Sun Yat-sen University, Guangzhou 510275, China
4 State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Agriculture, South China Agricultural University, Guangzhou 510642, China
5 Pingshan Translational Medicine Center, Shenzhen Bay Laboratory, Shenzhen 518118, China
6 These authors contributed equally
* Correspondence to: zhangzhang@mail.sysu.edu.cn, luogzh5@mail.sysu.edu.cn

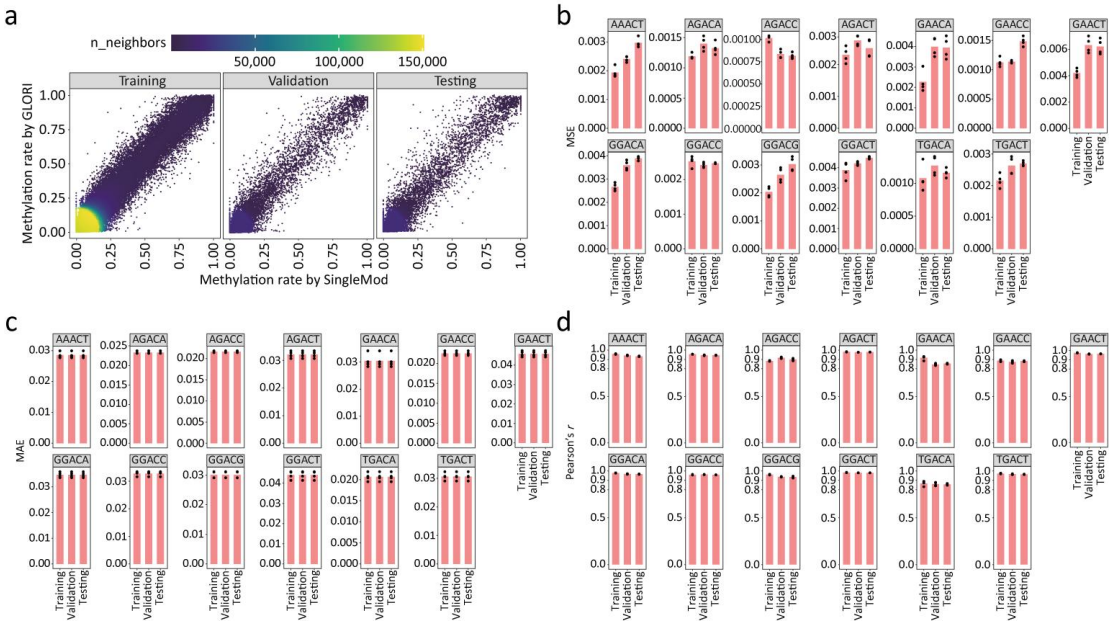
SUPPLEMENTARY FIGURES



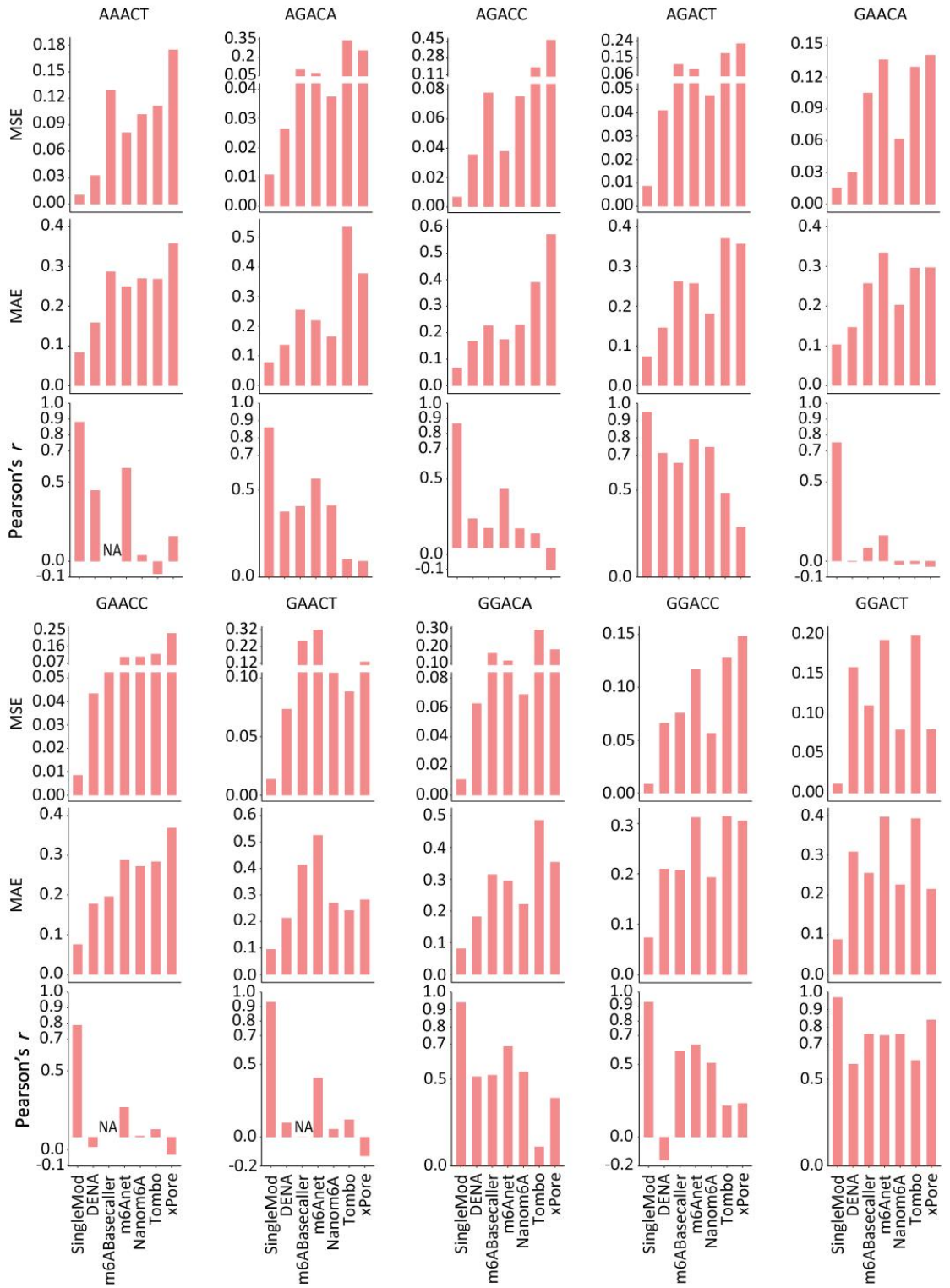
Supplementary Fig. 1|Data context faced during SingleMod development. **a** Statistics of GLORI and nanopore DRS data derived from HEK293T. From left to right: the number of m6A sites identified by GLORI; the number of these sites covered by DRS data; the number of DRS reads corresponding to these sites; the number of 9-mer types corresponding to these sites. **b** The distribution of methylation rates (top) and number (bottom) of m6A sites with methylation rate exceeding 0.1 in 24 DRACN motifs in HEK293T. Data was derived from GLORI results. Initially, SingleMod was separately trained using data from 13 predominant motifs (left). **c** Schematic illustration depicting the five locations of A base and corresponding 5mer in the nanopore during sequencing.



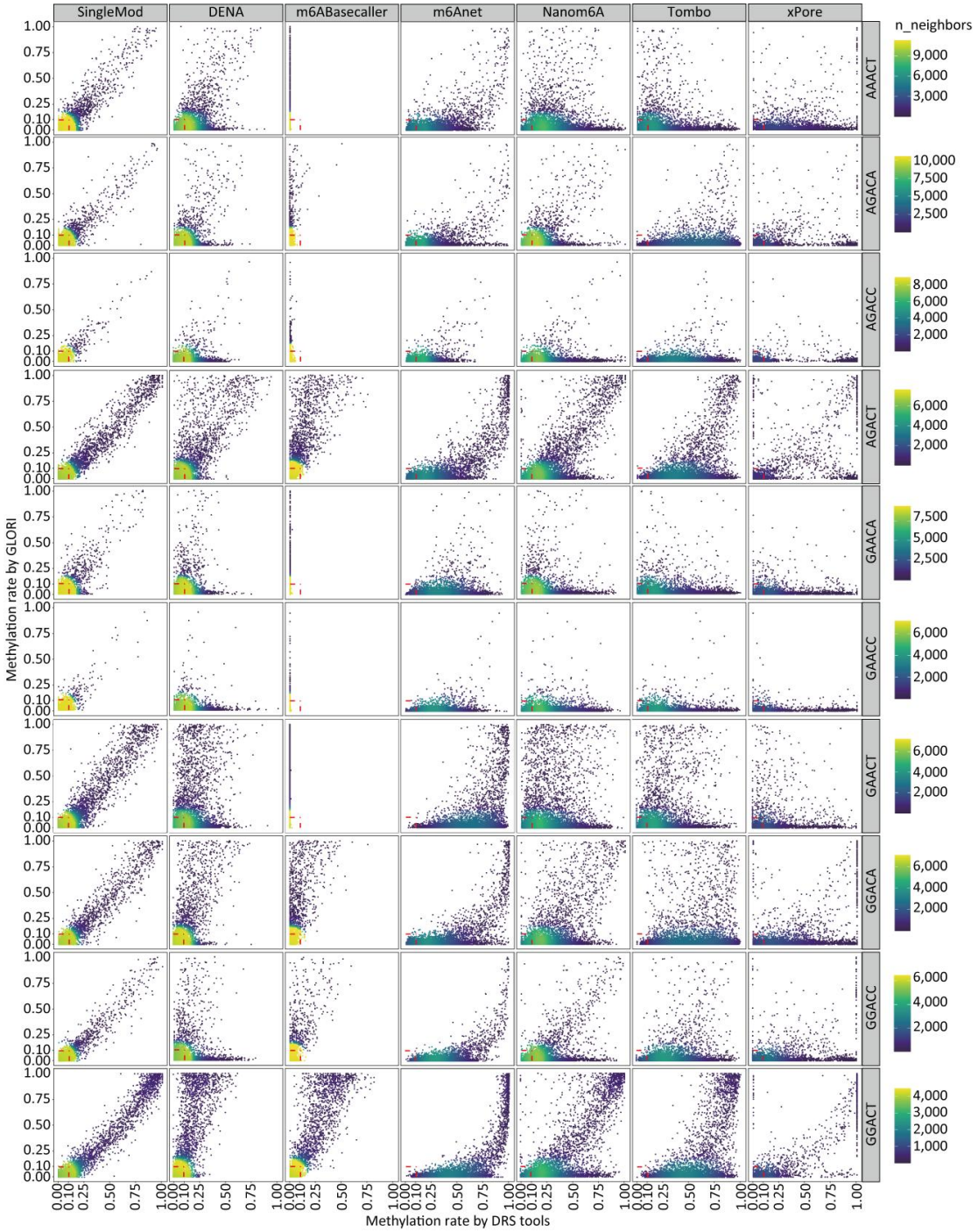
Supplementary Fig. 2|Optimization of input strategies and model architectures for SingleMod. a ROC AUC of classification models using different combinations of features across 5 motifs. **b** ROC AUC of classification models using raw signals within different length scales across 5 motifs. **c** ROC AUC of models using different network architectures and input strategies across five motifs. **d** MSE of SingleMod using two best network architectures in **b** across five motifs. **e-f** Examining SingleMod's overall performance (**e**) and its performance across various intervals (**f**), when incorporating varying proportion of low-methylation rate sites (<=0.05) for training.



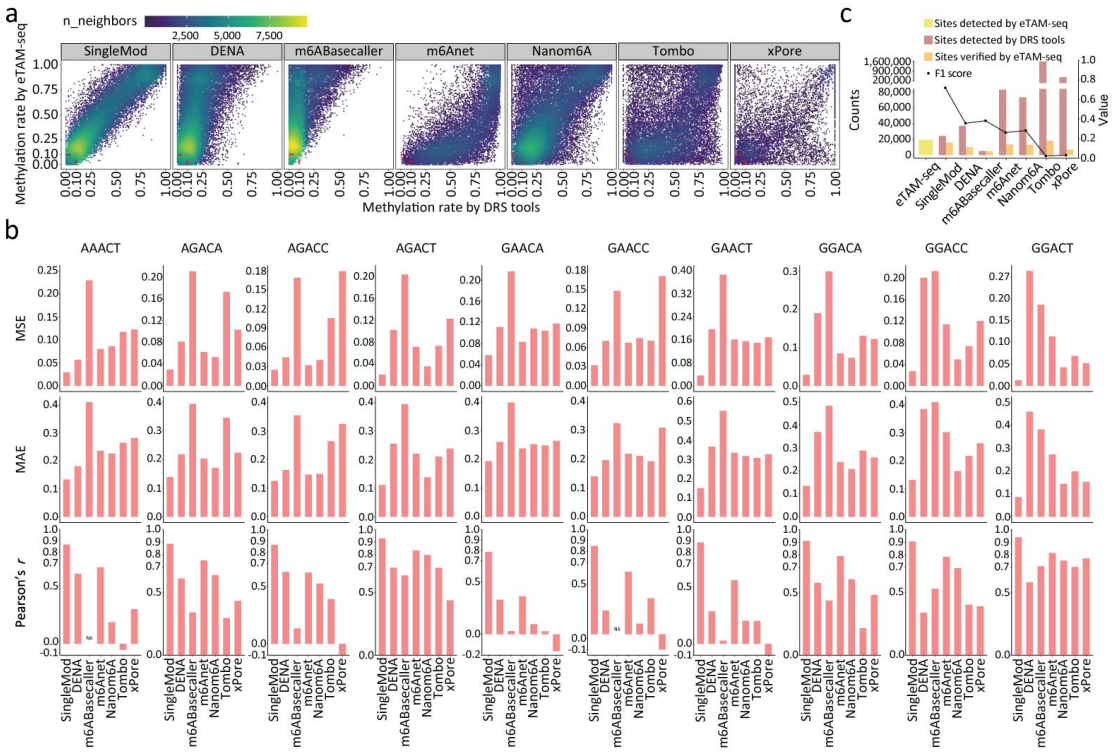
Supplementary Fig. 3|Performance of SingleMod on different datasets trained with combined HEK293T and HeLa data. a Scatter plots comparing the methylation rates predicted by SingleMod with GLORI benchmarks. **b-d** The performance of SingleMod in various motifs on different datasets.



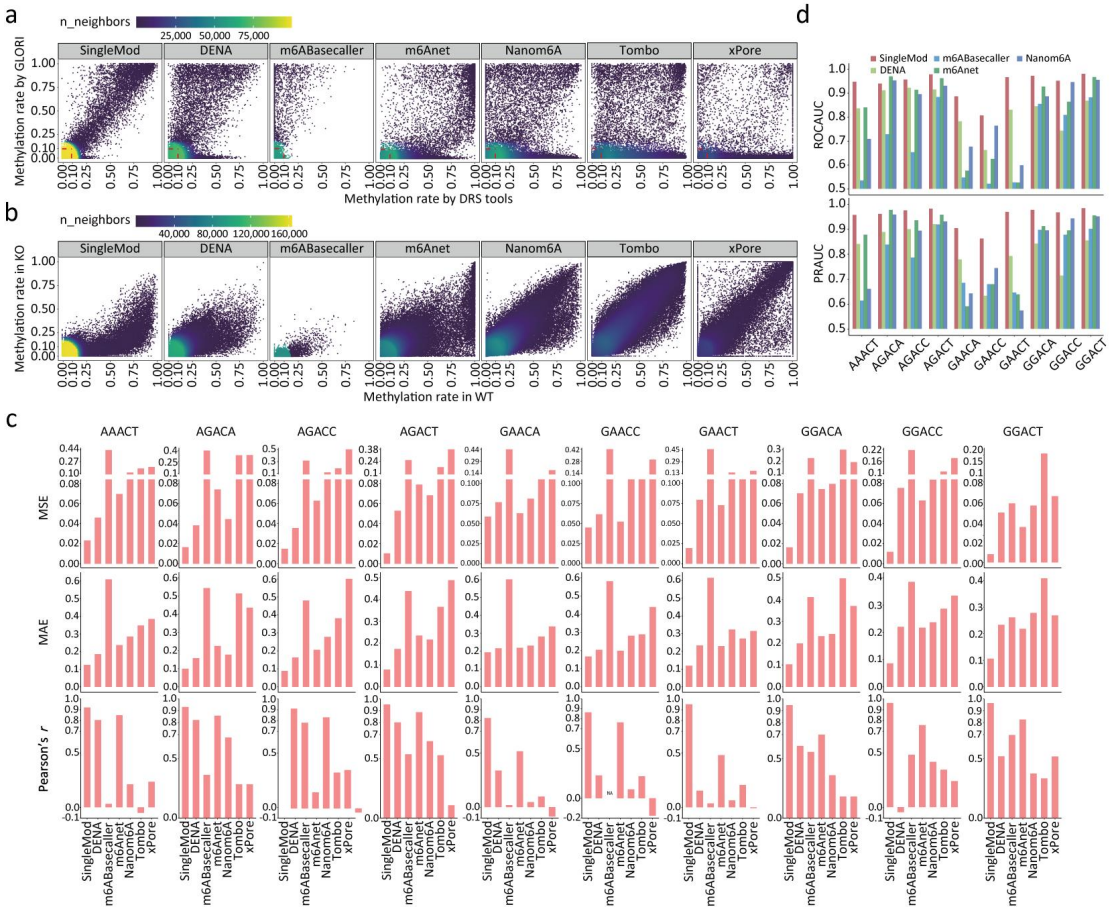
Supplementary Fig. 4|Comparison of the performance between SingleMod and other tools on *M. musculus* data as evaluated with GLORI benchmarks. The Pearson's r of m6ABasecaller in certain motifs was designated as "NA", because this tool predicted these motifs in all molecules to be unmethylated.



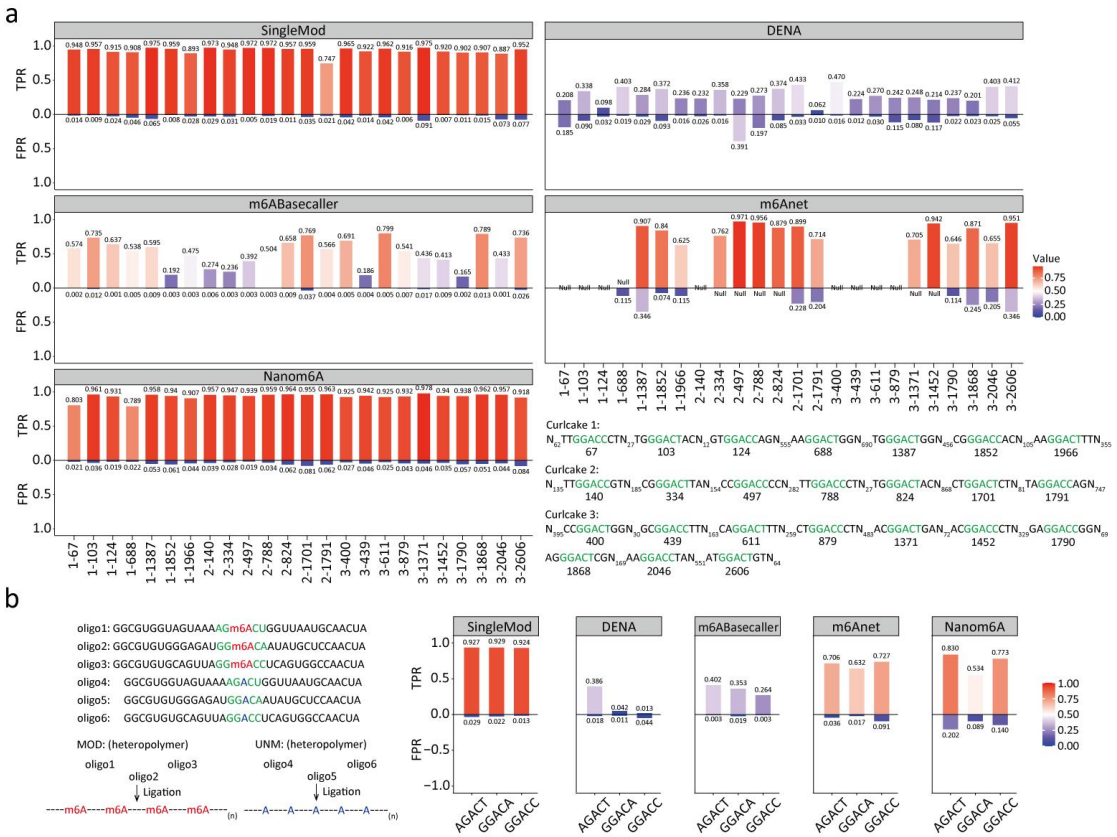
Supplementary Fig. 5|Scatter plots comparing the methylation rates in various motifs predicted by DRS tools with GLORI benchmarks on *M. musculus* data.



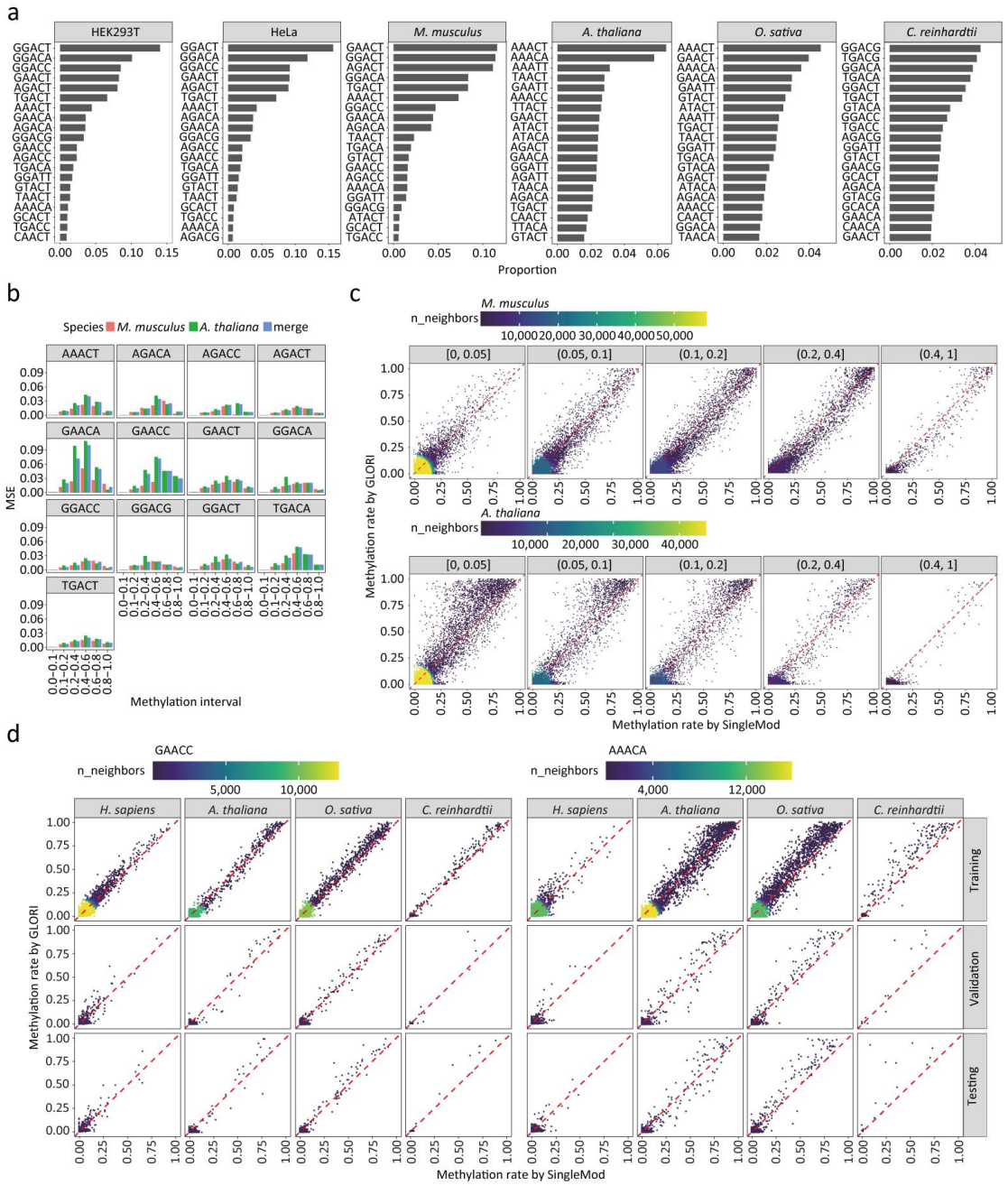
Supplementary Fig. 6|Comparison of the performance between SingleMod and other tools on *M. musculus* data as evaluated with eTAM-seq benchmarks. **a** Scatter plots comparing the methylation rates predicted by DRS tools with eTAM-seq benchmarks. **b** The performance of SingleMod and other tools on various motifs. The Pearson's r of m6ABasecaller in certain motifs was designated as "NA", because this tool predicted these motifs in all molecules to be unmethylated. **c** The number of identified m6A sites within the candidate list by DRS tools and eTAM-seq, and their overlap. F1 scores of site-level m6A predictions are also provided.



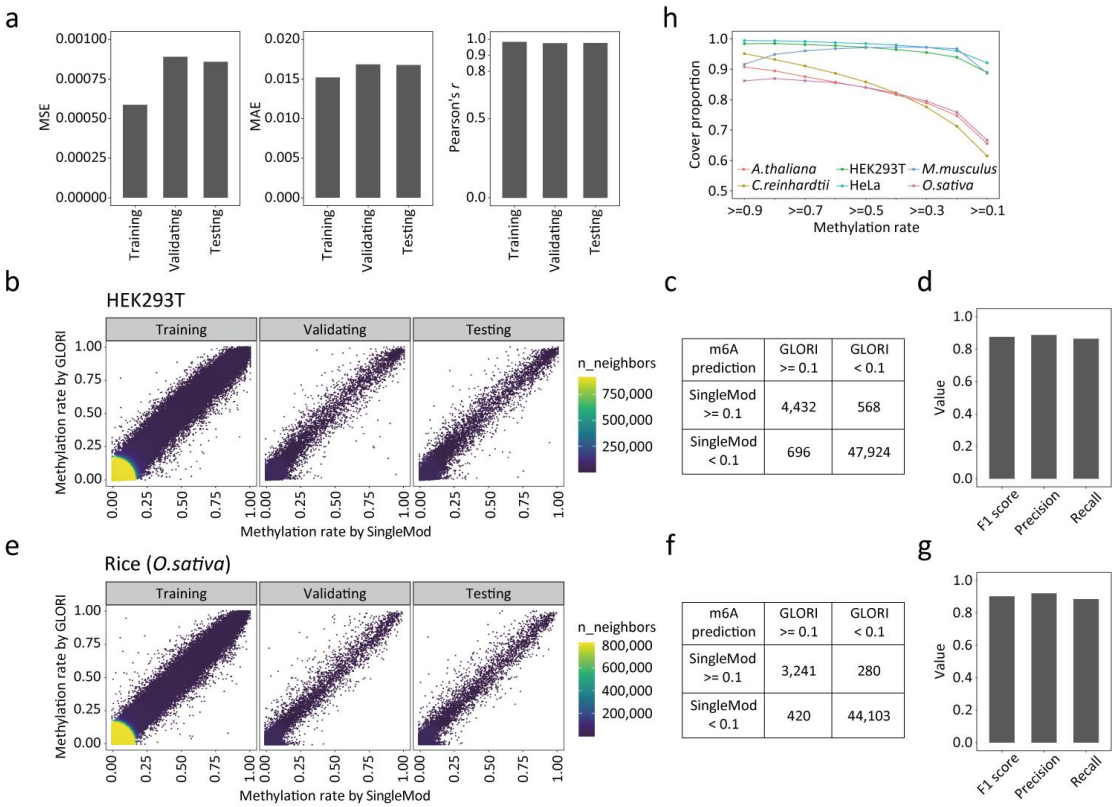
Supplementary Fig. 7|Comparison of the performance between SingleMod and other tools on *A. thaliana* data as evaluated with GLORI benchmarks. a Scatter plots comparing the methylation rates predicted by DRS tools with GLORI benchmarks. **b** Scatter plots comparing the methylation rates between WT and KO samples predicted by various DRS tools. **c** The performance of SingleMod and other tools on various motifs. **d** The AUC of ROC and PR curve for single-molecule m6A predictions in 10 common motifs. Fully-methylated and -unmethylated sites determined by GLORI in both *M. musculus* and *A. thaliana* were used.



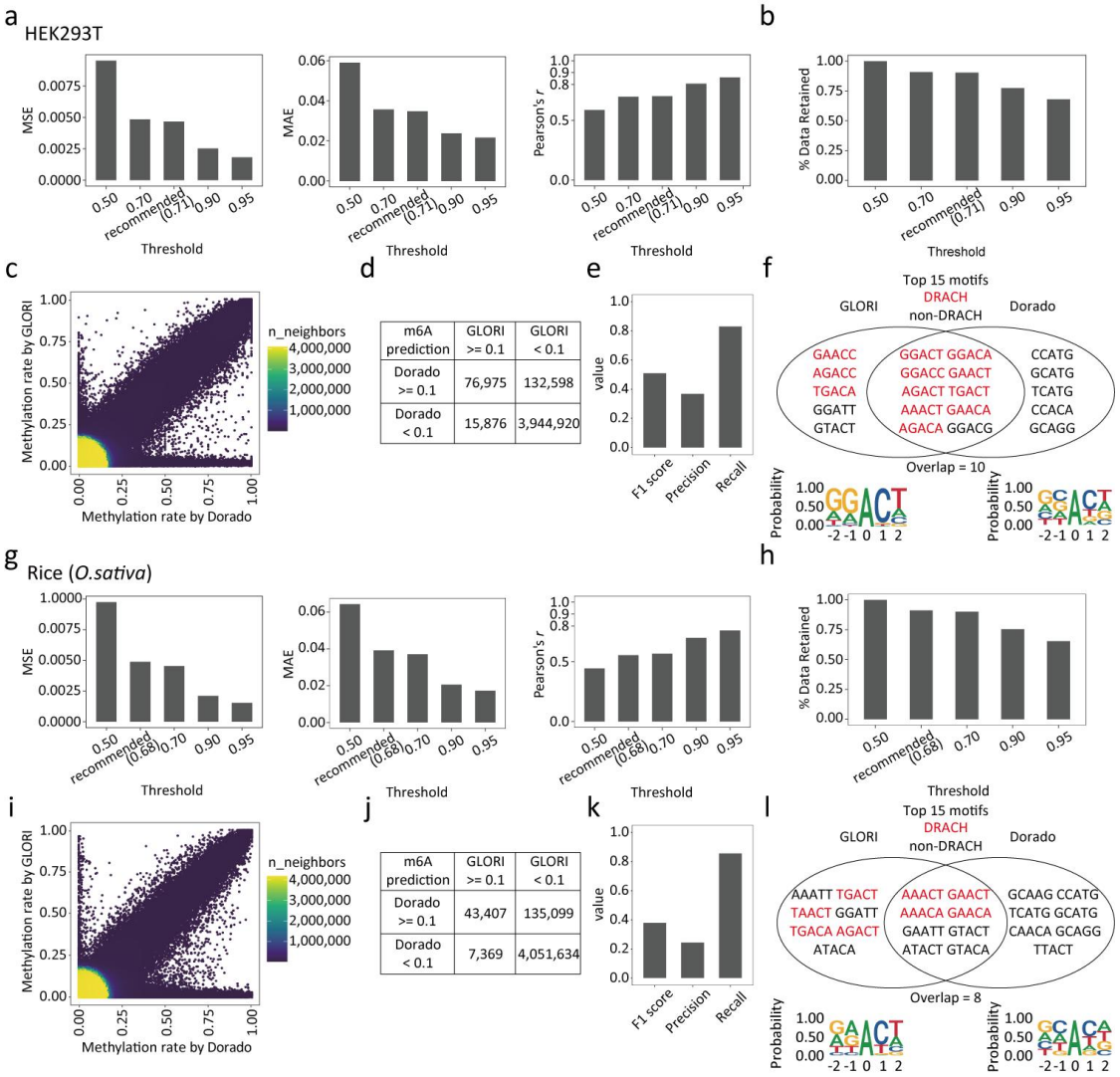
Supplementary Fig. 8|Comparison of SingleMod and other tools in single-molecule m6A prediction using synthetic RNA molecules. a-b Barplot comparing the TPR and FPR of single-molecule level m6A predictions between SingleMod and other tools, using DRS data from the IVT-derived (Curlcakes, **a**) or synthetic (SL-Oligomers, **b**) RNA molecules.



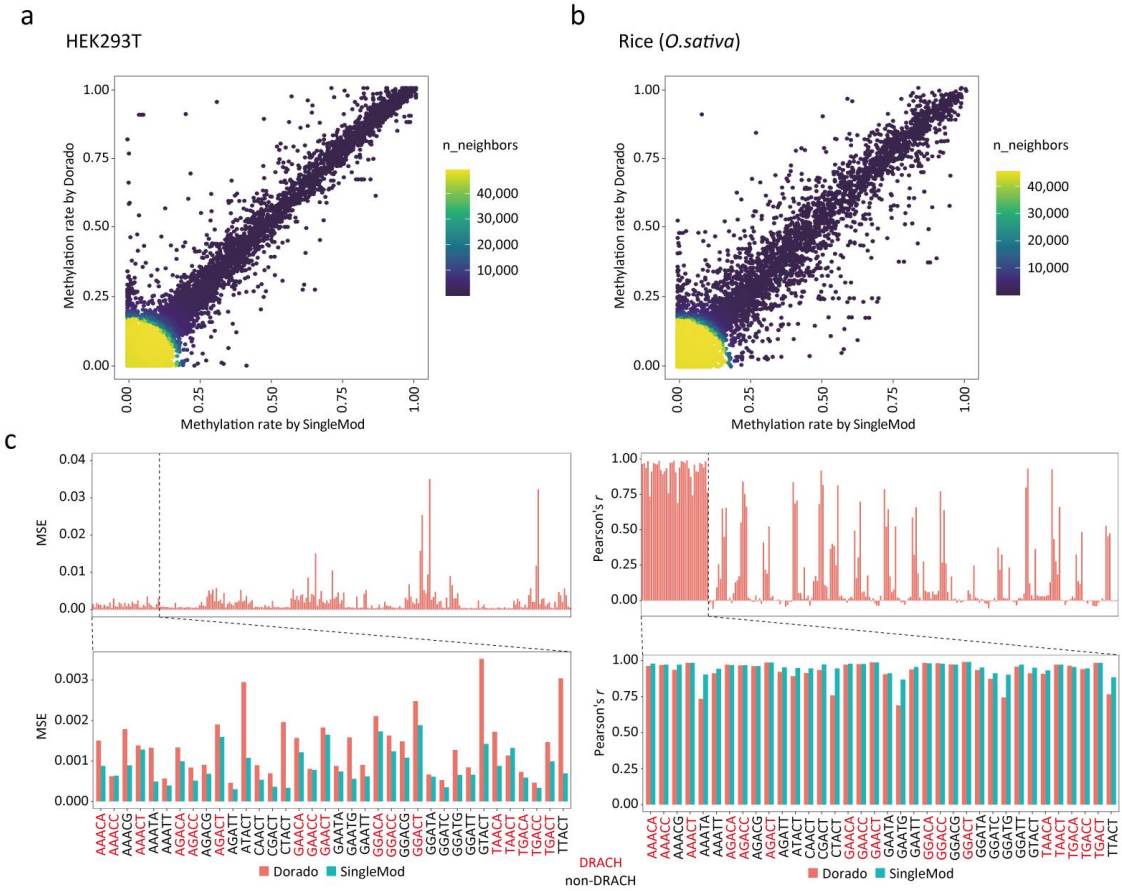
Supplementary Fig. 9|Further investigation on the performance of SingleMod and optimization of its generalization ability. a Top 20 motifs of m6A in various species. **b** MSE in varying methylation rate intervals when testing SingleMod in different species. "Merge" indicate merging data from two species. **c** Scatter plots comparing the methylation rates predicted by SingleMod with GLORI benchmarks for different 9-mers with varying weighted methylation rates in training set. **d** Methylation rate of sites within the GAACC (left) and AAACA (right) motif in different datasets, as detected by GLORI and SingleMod.



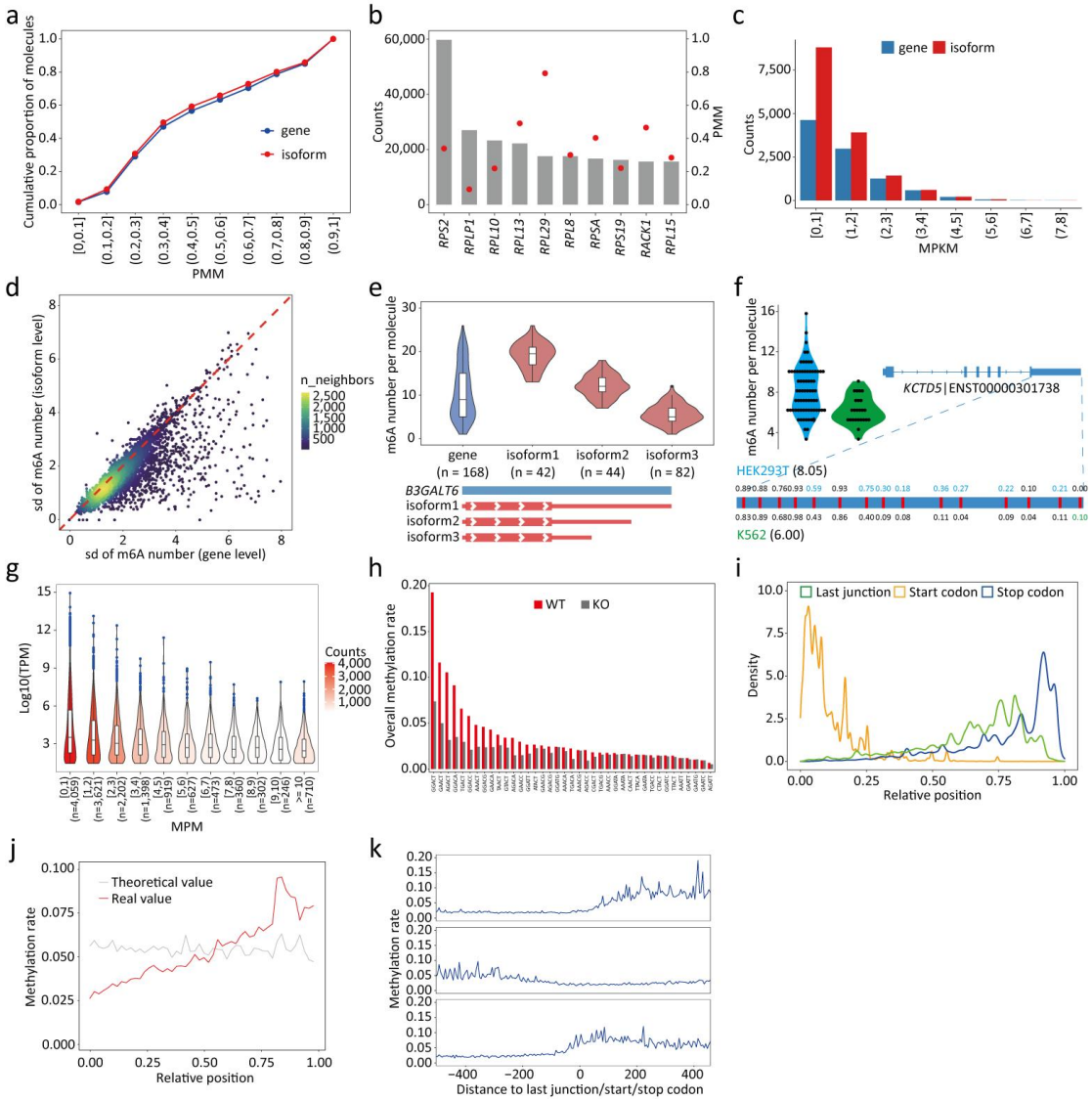
Supplementary Fig. 10|Performance of SingleMod trained with DRS data generated by the latest nanopore technology and sequencing kit (RNA004). a The performance of SingleMod in different datasets. **b** Scatter plots comparing the methylation rates predicted by SingleMod with GLORI benchmark in different datasets. Data used in **b-d** were from HEK293T. **c** The statistical results of the number of m6A sites with different methylation states predicted by SingleMod and GLORI. Data was from the testing set. **d** F1 scores, precision and recall of site-level m6A predictions by SingleMod. Data was from **c**. **e-g** Results identical to those of **b-d**, but the data came from *O.sativa*. **h** Proportion of m6A sites with varying methylation rates as determined by GLORI falling within RNA004-version SingleMod's motif scope.



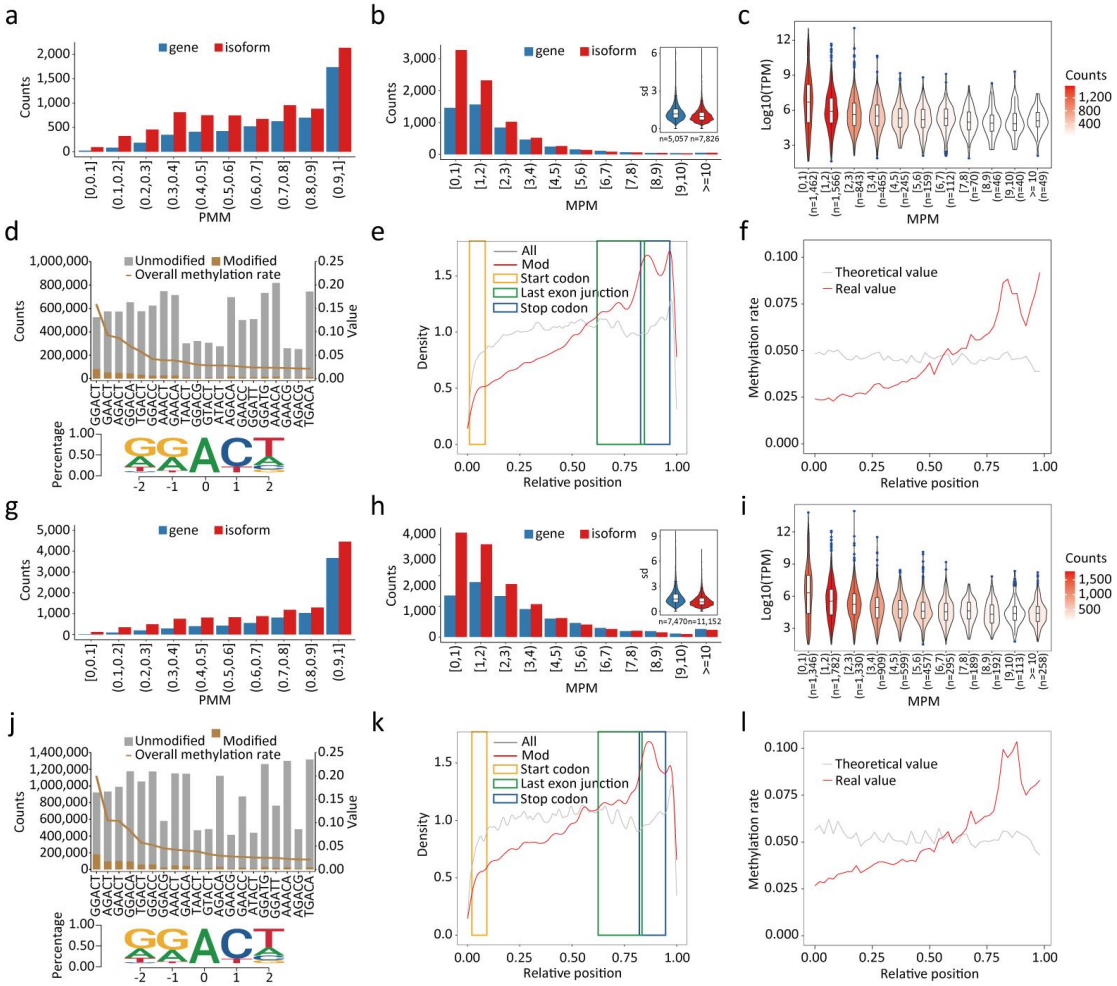
Supplementary Fig. 11|Performance of Dorado on RNA004 data. a-b The performance of Dorado (**a**) as evaluated with GLORI benchmarks and the ratio of data retained (**b**) when using different modkit filtering thresholds. When using Modkit to extract Dorado prediction results, if both the probability of the canonical base and the modification are below the specified filtering threshold, the prediction is classified as a failure and filtered out. Data used in **a-f** were from HEK293T. **c** Scatter plots comparing the methylation rates predicted by Dorado with GLORI benchmark. Results in **c-f** were generated using a Modkit filtering threshold of 0.9. **d** The statistical results of the number of m6A sites with different methylation states predicted by Dorado and GLORI. **e** F1 scores, precision and recall of site-level m6A predictions by Dorado. **f** Top 15 motifs harboring the most m6A sites identified by GLORI and Dorado. The composition of m6A flanking bases plotted by Seqlogo was shown below. **g-l** Results identical to those of **a-f**, but the data came from *O.sativa*.



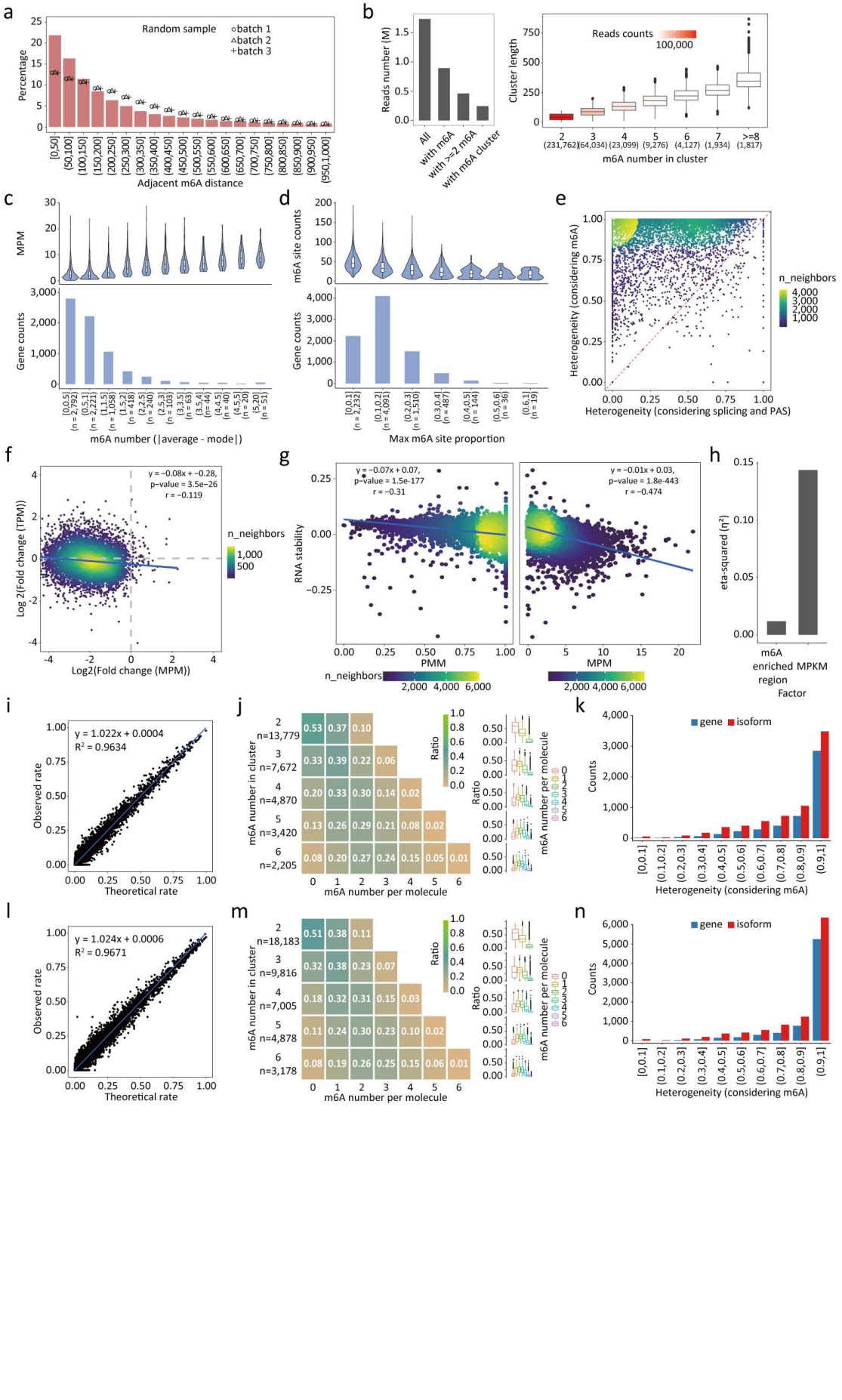
Supplementary Fig. 12|Comparison of the performance between Dorado and SingleMod (RNA004). a-b Scatter plots comparing the methylation rates predicted by Dorado versus SingleMod on HEK293T (**a**) and *O.sativa* (**b**) data. **c** The performance of Dorado on various motifs, with a focus on the comparison with SingleMod across common motifs.



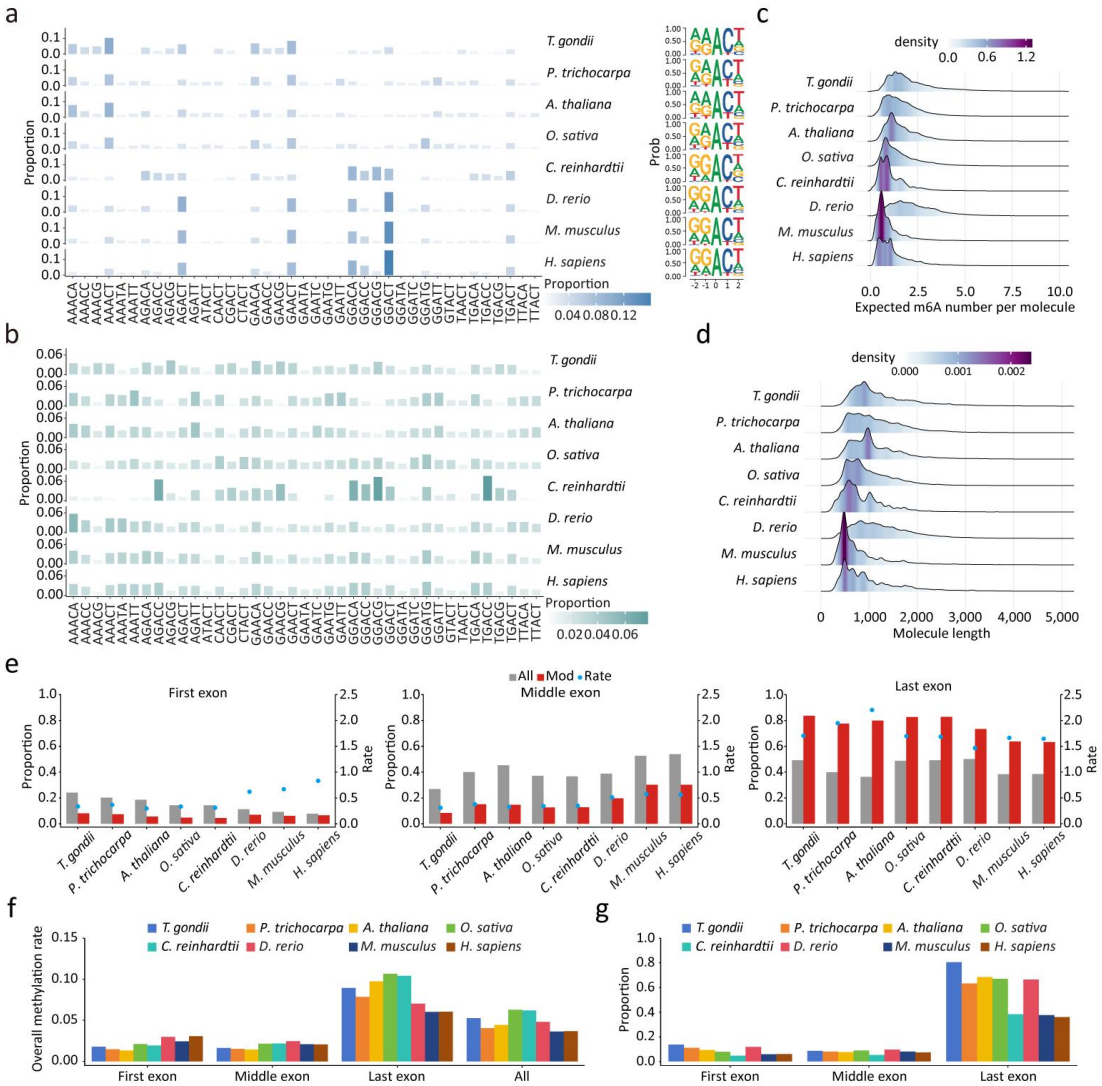
Supplementary Fig. 13|Additional results for the analysis of HEK293T single-molecule m6A profile. a Cumulative proportion of molecules from gene groups with varying PMM. **b** Molecules count (bar) and PMM (dot) for the top 10 highest expressed genes. **c** Counts of genes or isoforms with varying MPKM. **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 isoforms. **f** Comparing the distribution of m6A number in all molecules of a representative isoform (top), the MPM of this isoform, and the methylation rates of each site (bottom) between HEK293T and K562. **g** Distribution of expression levels of isoforms with varying MPM. "TPM" indicates Transcripts Per Million. **h** Overall methylation rate of various motifs in WT and *METTL3* KO cells. **i** The distribution of relative position of start codon, last junction and stop codon within their molecules. **j-k** Overall methylation rate for DRACH motifs at different relative positions (**j**) and around last exon junction, start codon, and stop codon (**k**).



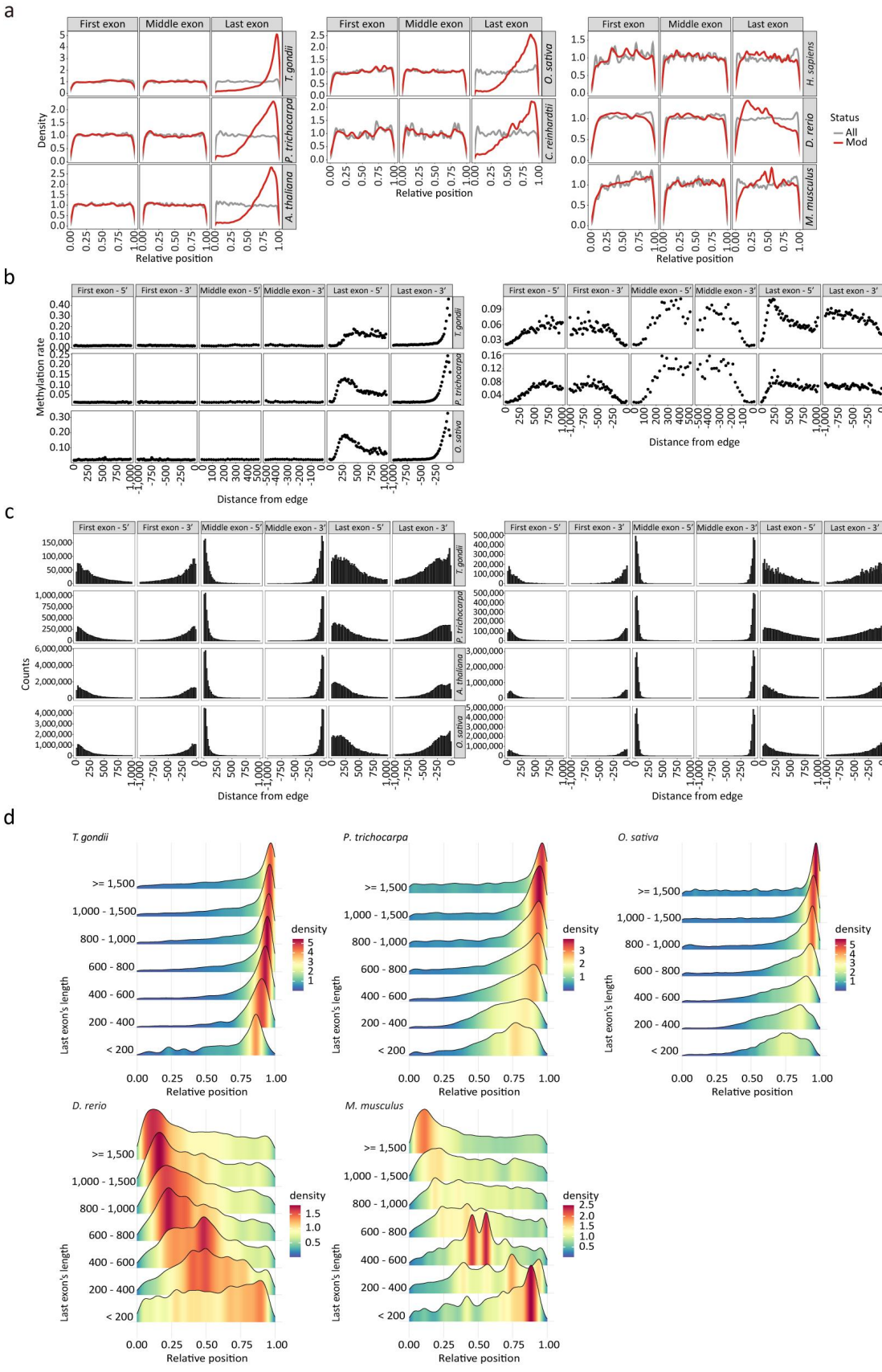
Supplementary Fig. 14|Additional results for the analysis of HeLa and K562 single-molecule m6A profile. **a** Counts of genes or isoforms with varying PMM. Data used in **a-f** were from HeLa. **b** Counts of genes or isoforms with varying MPM, and the distribution of standard deviation in m6A number among molecules within each gene or isoform. **c** Distribution of expression levels of genes with varying MPM. **d** Counts (top, bar) of unmodified and modified A bases within top 20 motifs for methylation rate from all molecules and corresponding overall methylation rates (top, line). The composition of m6A flanking bases plotted by Seqlogo are shown below. **e** 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. **f** Overall methylation rate at different relative positions. **g-l** Results identical to those of **a-f**, but the data came from K562.


a
Random sample
batch 1
batch 2
batch 3
Percentage
Adjacent m6A distance
b
Reads number (M)
All
with m6A
with >=2 m6A
with m6A cluster
Cluster length
Reads counts
100,000
m6A number in cluster
c
MPM
Gene counts
m6A number (|average - mode|)
d
m6A site counts
Gene counts
Max m6A site proportion
e
Heterogeneity (considering m6A)
Heterogeneity (considering splicing and PAS)
n_neighbors
f
y = −0.08x + −0.28,
p−value = 3.5e−26
r = −0.119
Log 2(Fold change (TPM))
Log2(Fold change (MPM))
n_neighbors
g
y = −0.07x + 0.07,
p−value = 1.5e-177
r = −0.31
y = −0.01x + 0.03,
p−value = 1.8e-443
r = −0.474
RNA stability
PMM
MPM
h
eta-squared (η²)
m6A
enriched
region
MPKM
Factor
i
y = 1.022x + 0.0004
R² = 0.9634
Observed rate
Theoretical rate
j
m6A number in cluster
m6A number per molecule
Ratio
k
gene
isoform
Counts
Heterogeneity (considering m6A)
l
y = 1.024x + 0.0006
R² = 0.9671
Observed rate
Theoretical rate
m
m6A number in cluster
m6A number per molecule
Ratio
n
gene
isoform
Counts
Heterogeneity (considering m6A)

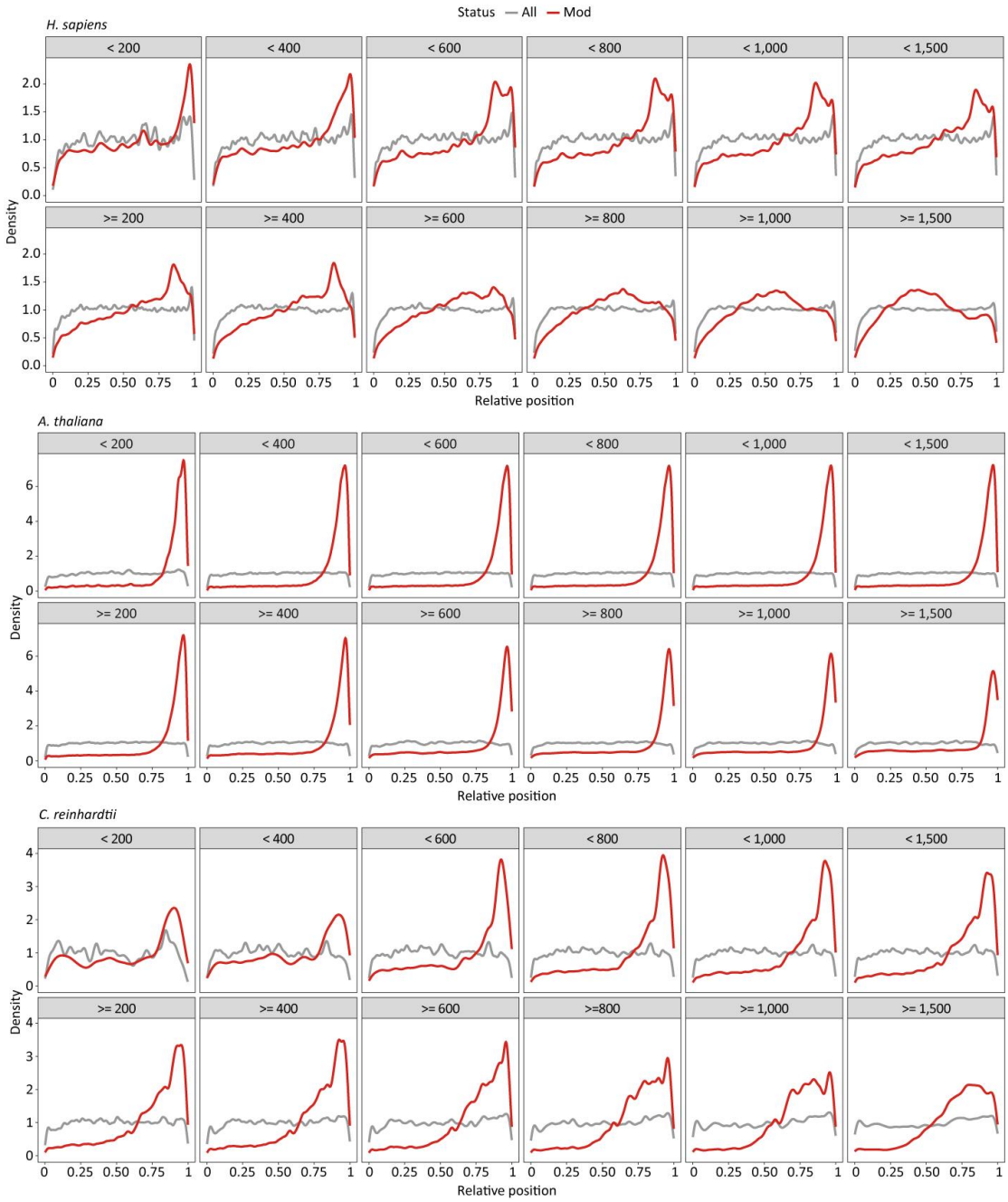
Supplementary Fig. 15|Additional results for m6A-mediated RNA heterogeneity and regulation profiled at single-molecule level in human cell lines. **a** Percentage of different distance (nt) between two adjacent m6A on individual RNA molecule. **b** Left: number of different types of molecules. Right: length of m6A clusters with varying m6A numbers and their corresponding molecules numbers. **c** Counts of genes with varying difference between the mode and the mean of m6A number among molecules in a gene. Data used **a-h** were from HEK293T. **d** Counts of genes with varying max m6A site proportion (bottom) and distribution of m6A sites counts in corresponding genes (top). Max m6A site proportion indicates the proportion of m6A modifications in the site of the highest methylation rate, relative to the total number of m6A modifications in a gene. **e** Scatter plots comparing the molecular heterogeneity introduced by m6A and by difference on splicing form and PAS selection for all genes. **f** Correlation between changes in gene expression and m6A level following *METTL3* KO. **g** Correlation between RNA stability and PMM (left) or MPM (right). **h** The importance of MPKM and m6A enriched region in explaining the variance of RNA stability were quantified using eta-squared (η^2), which measure the proportion of total variance explained. **i** Comparison of theoretical rate and observed rate of molecules where two adjacent A bases within all detected motifs are both modified. Data used **i-k** were from HeLa. **j** Heatmap (bottom, left) and boxplot (bottom, right) show the ratio of molecules with varying m6A number within m6A cluster. **k** Gene or isoforms counts with varying molecules heterogeneity introduced by m6A. **l-n** Results identical to those of **i-k**, but the data came from K562.



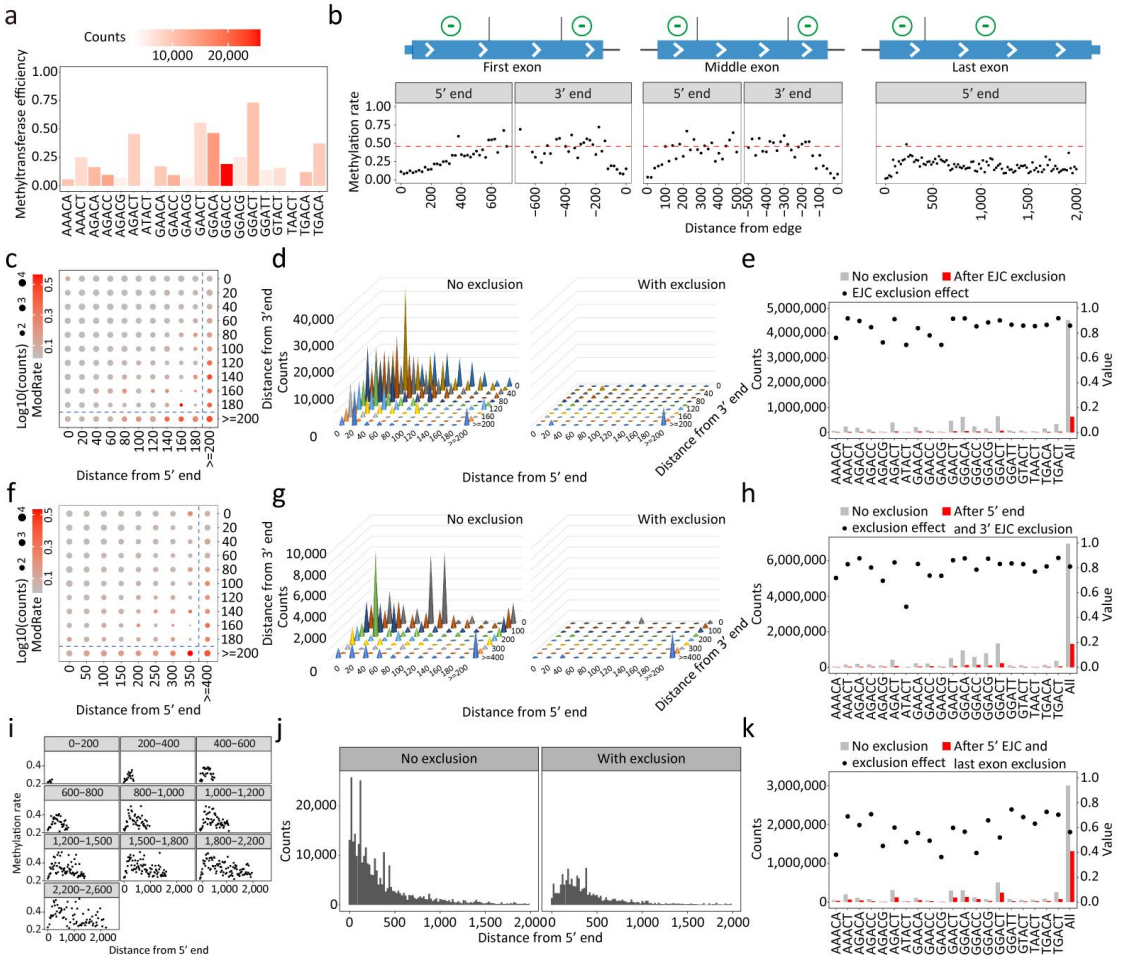
Supplementary Fig. 16|Comparative analysis of m6A sequence and exon type preferences across various species. a Proportion of modified A bases within various motifs to all modified A bases. The composition of m6A flanking bases plotted by Seqlogo was shown on the right. **b** Proportion of candidate A bases within various motifs. **c** Distribution of the expected number of m6A modifications per molecule. The expected m6A number was calculated by summing up the overall methylation rates of each motif occurrence on a molecule. **d** Distribution of molecule lengths. **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 in distinct exon types. **g** Proportion of modified exons (containing at least one m6A).


a
First exon
Middle exon
Last exon
T. gondii
P. trichocarpa
A. thaliana
O. sativa
C. reinhardtii
H. sapiens
D. rerio
M. musculus
Density
Relative position
Status
All
Mod
b
First exon - 5′
First exon - 3′
Middle exon - 5′
Middle exon - 3′
Last exon - 5′
Last exon - 3′
Methylation rate
Distance from edge
c
Counts
Distance from edge
d
T. gondii
P. trichocarpa
O. sativa
D. rerio
M. musculus
Last exon's length
>= 1,500
1,000 - 1,500
800 - 1,000
600 - 800
400 - 600
200 - 400
< 200
density
Relative position

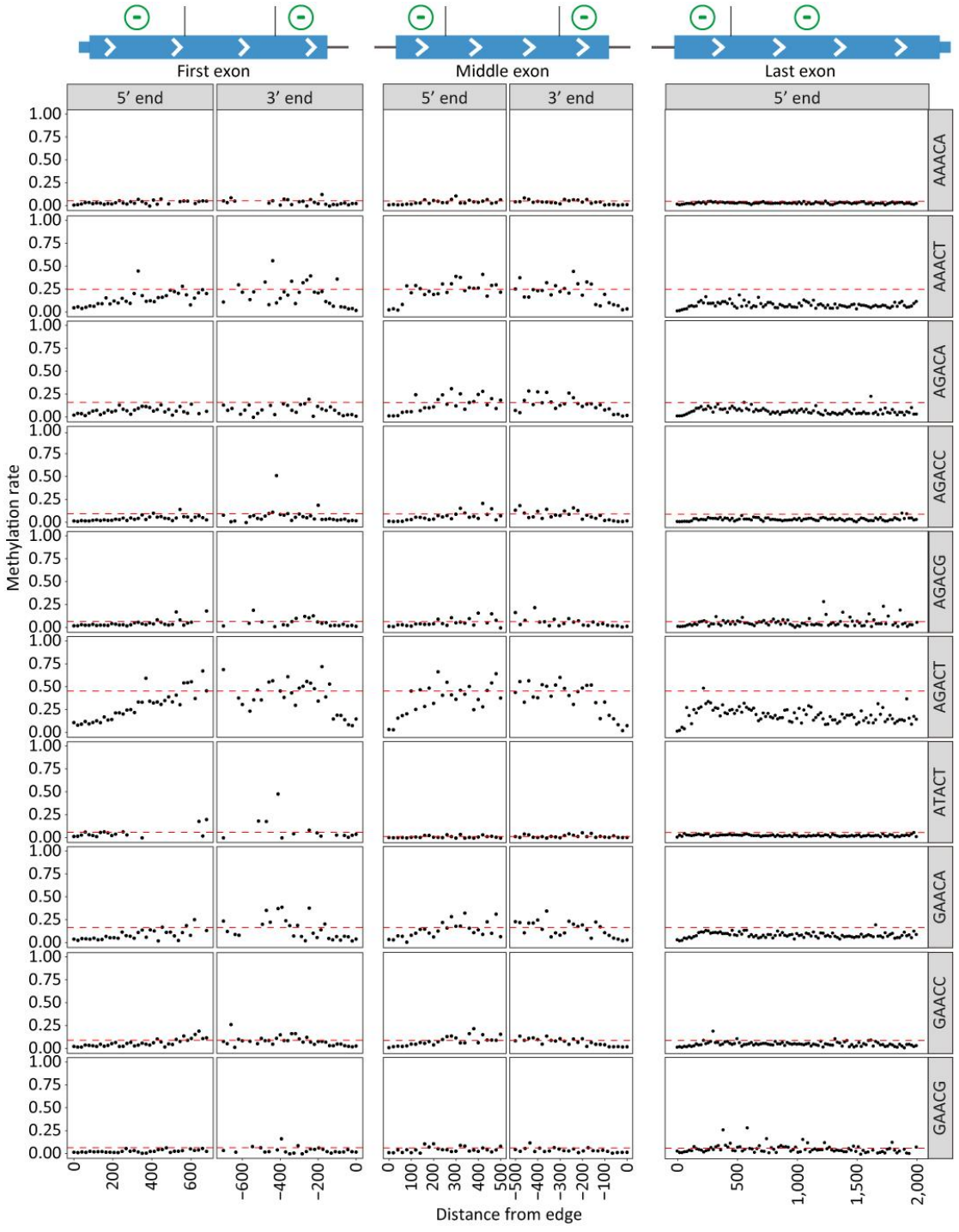
Supplementary Fig. 17|Comparative analysis of m6A distribution across various species. a The distribution of relative position of background (All) and modified (Mod) A bases within their respective molecules in different exon types. **b-c** Overall methylation rate (**b**) and counts of motifs (**c**) with varying distance from one exon edge. **d** Distribution of relative position of m6A in the last exons with varying lengths.

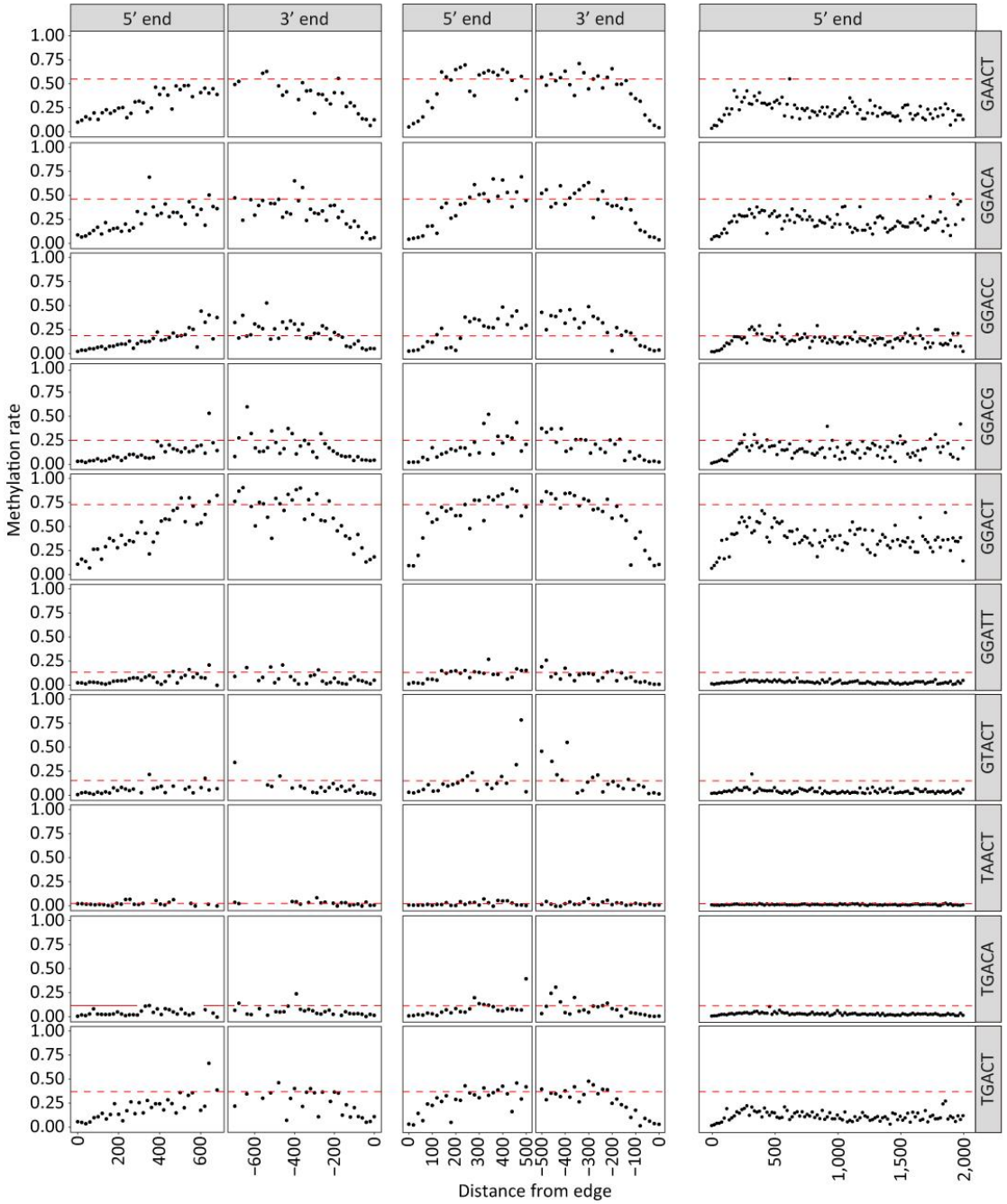


Supplementary Fig. 18|Distribution of relative position of background (All) and modified (Mod) A bases within their respective molecules with different last exon length.



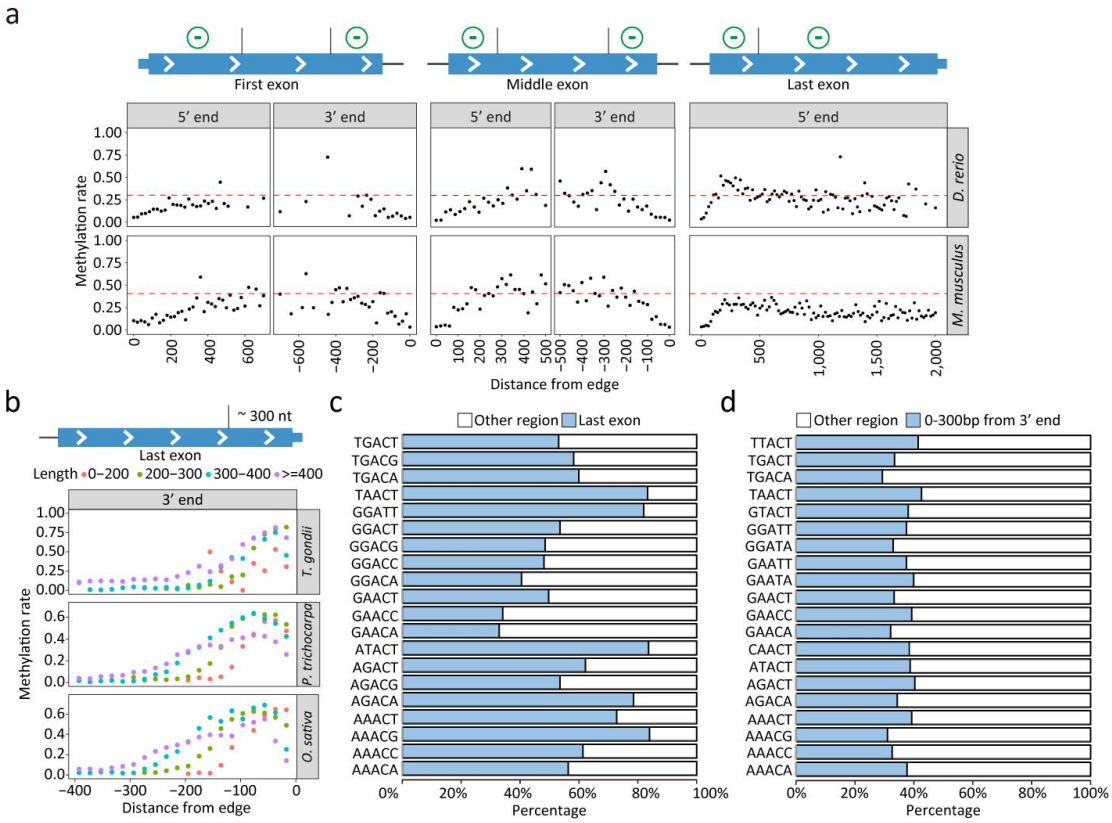
Supplementary Fig. 19|Quantitative analysis of exclusion deposition model in HEK293T. a Intrinsic efficiency of m6A deposition for top 20 motifs. **b** The same as **Fig. 6a**, but using AGACT motif as a representative. **c, f** Overall methylation rate and counts of motifs at positions with varying distance from both edges. **d, g** Hypothetical m6A content if without exclusion effect (left) and observed m6A content (right). **e, h** Bar: hypothetical m6A content and observed m6A content for various motifs. Dot: exclusion effect. **c-e** and **f-h** are results for middle and first exons, respectively. **c, d, f, g** use GGACA motif as a representative. **i-j** Overall methylation rate at positions with varying distances from 5' end of last exon with different lengths (**i**), and corresponding hypothetical m6A content (**j**, left) and observed m6A content (**j**, right), using GGACA motif as a representative. **k** Bar: hypothetical m6A content and observed m6A content in the last exon for all motifs. Dot: exclusion effect.


First exon
Middle exon
Last exon
5′ end
3′ end
Methylation rate
1.00
0.75
0.50
0.25
0.00
AAACA
AAACT
AGACA
AGACC
AGACG
AGACT
ATACT
GAACA
GAACC
GAACG
0
200
400
600
−600
−400
−200
0
100
200
300
400
500
−500
−400
−300
−200
−100
0
500
1,000
1,500
2,000
Distance from edge

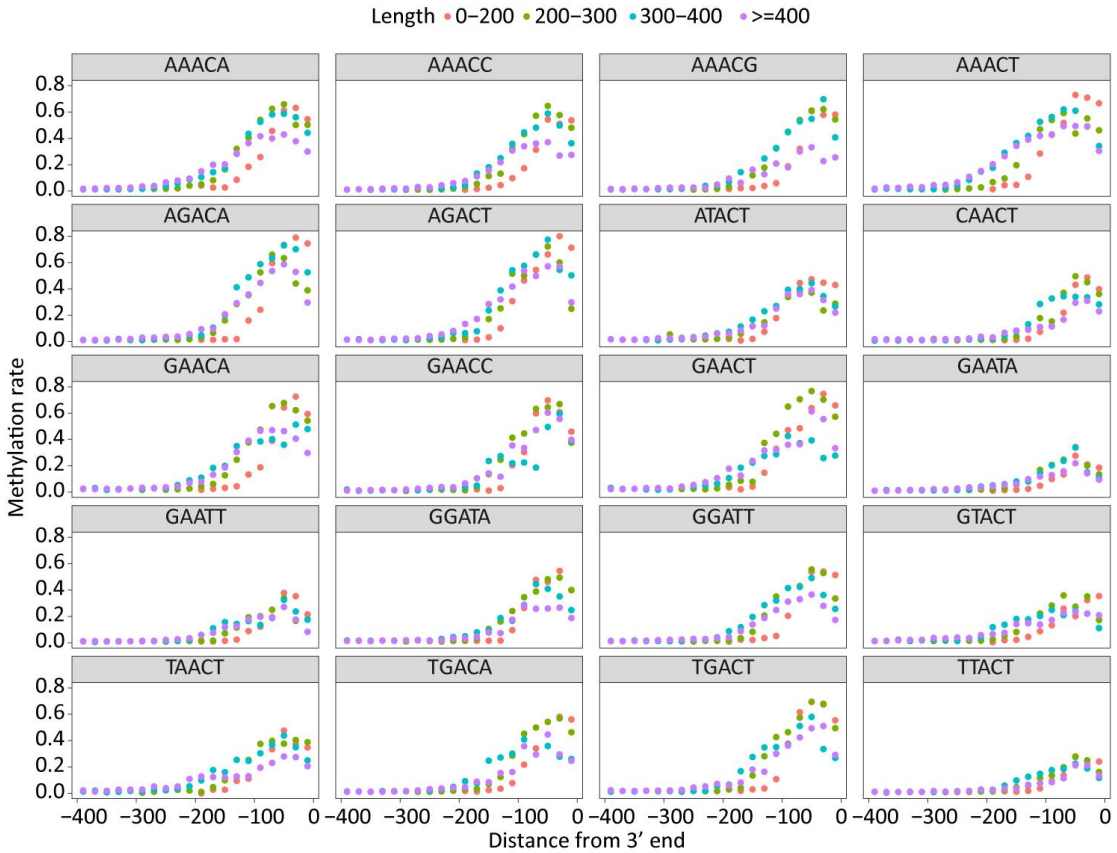


Supplementary Fig. 20|m6A-pattern in HEK293T. Top: schematic representation of exclusion deposition model in HEK293T. Bottom: m6A pattern shaped by exclusion deposition model and comparing the overall methylation rate at positions with varying distances from one exon edge to hypothetical default efficiency (red dashed lines).

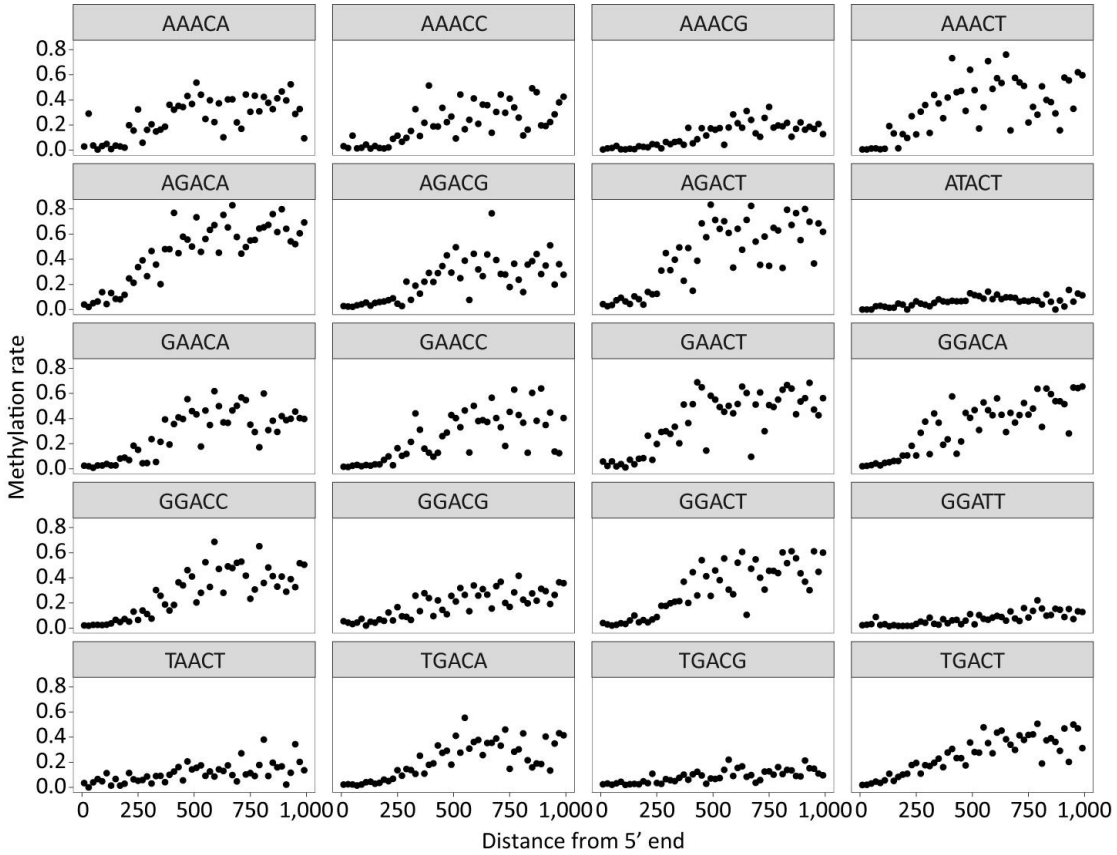
Supplementary Fig. 21|Overall methylation rate at positions with varying distances from 5' end of last exon with different lengths. Data are derived from HEK293T.



Supplementary Fig. 22|Quantitative elucidation of m6A distribution patterns across multi-species. a Top: schematic representation of exclusion deposition mode in vertebrates. Bottom: m6A pattern shaped by exclusion deposition model 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** Top: schematic representation of the defined m6A deposition region in plants and protozoa. Bottom: their m6A patterns and overall methylation rate at positions with varying distances from 3' terminal, using AAACT motif as a representative. **c** Percentage of motifs within and outside the defined m6A deposition region in *A. thaliana*. **d** Percentage of motifs within and outside the last exon in *C. reinhardtii*.



Supplementary Fig. 23|m6A pattern and overall methylation rate at positions with varying distances from 3' end of last exon. Data are derived from *A. thaliana*.



Supplementary Fig. 24|m6A pattern and overall methylation rate at positions with varying distances from 5' end of last exon. Data are derived from *C. reinhardtii*.