Supplementary Materials

MetaTX: deciphering the distribution of mRNA-related features in the presence of isoform ambiguity, with applications in epitranscriptome analysis

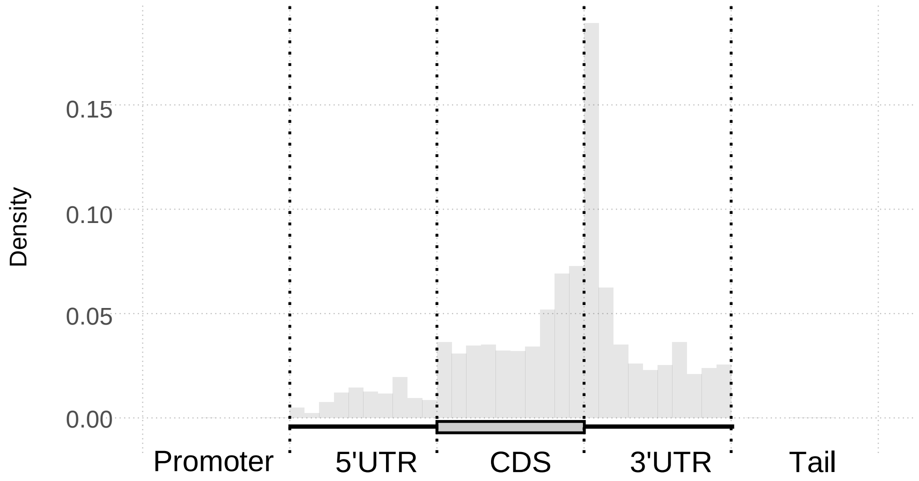

Figure S1. Bin-based distribution. Figure shows the distribution of m^{6}A sites reported by miCLIP technology [1] estimated by MetaTX method with smoothing.

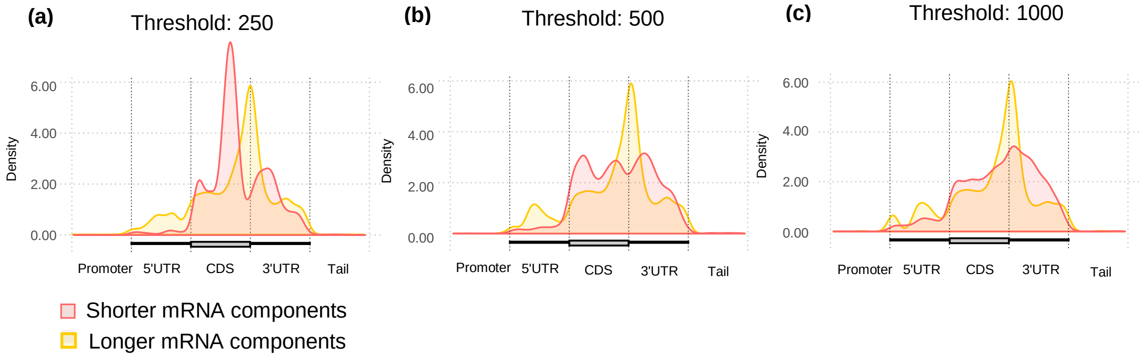

Figure S2. Comparison of m^{6}A pattern on transcripts of different lengths. Figure shows the distribution of m^{6}A sites on longer transcripts are likely to be more consistent our knowledge of m^{6}A site distribution. This is probably because longer transcripts can provide higher relative resolution with respect to the location on a standard gene model.

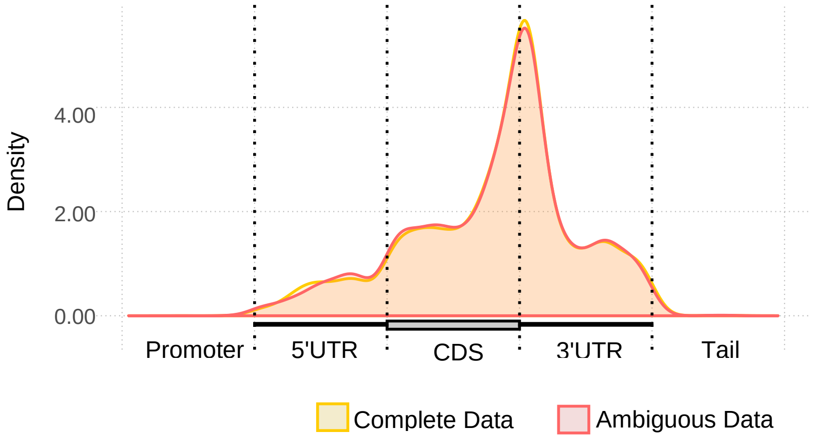

Figure S3. Distribution patterns inferred for only features with isoform ambiguity. In this test, we focused on only the m^{6}A sites overlapping with multiple isoform transcripts. There were a total of 9181 m^{6}A sites in the complete miCLIP dataset (Olarerin-George and Jaffrey, 2017), and 6712 of them are associated with multiple isoform transcripts of the same gene (or with isoform ambiguity). The yellow curve represents the distribution pattern estimated from complete dataset; while the red curves indicates the pattern inferred from only the m^{6}A sites with isoform ambiguity. The two distribution patterns are highly consistent with each other, suggesting that MetaTX method achieved very stable performance even when dealing with features with high degree of isoform ambiguity.

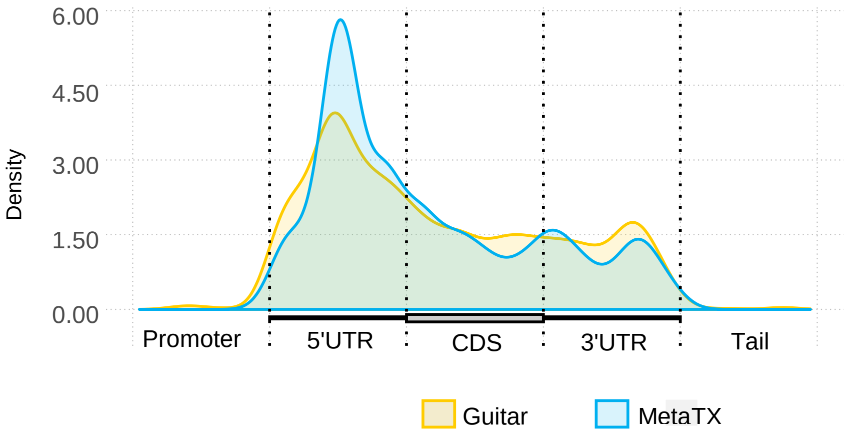

Figure S4. Distribution of human m^{5}C RNA methylation sites. Stronger enrichment pattern of m^{5}C is reported at 5’UTRs by MetaTX compared to that reported by Guitar [2] method, which did not correct isoform ambiguity. The data was generated from BS-seq with an improved protocol [3].

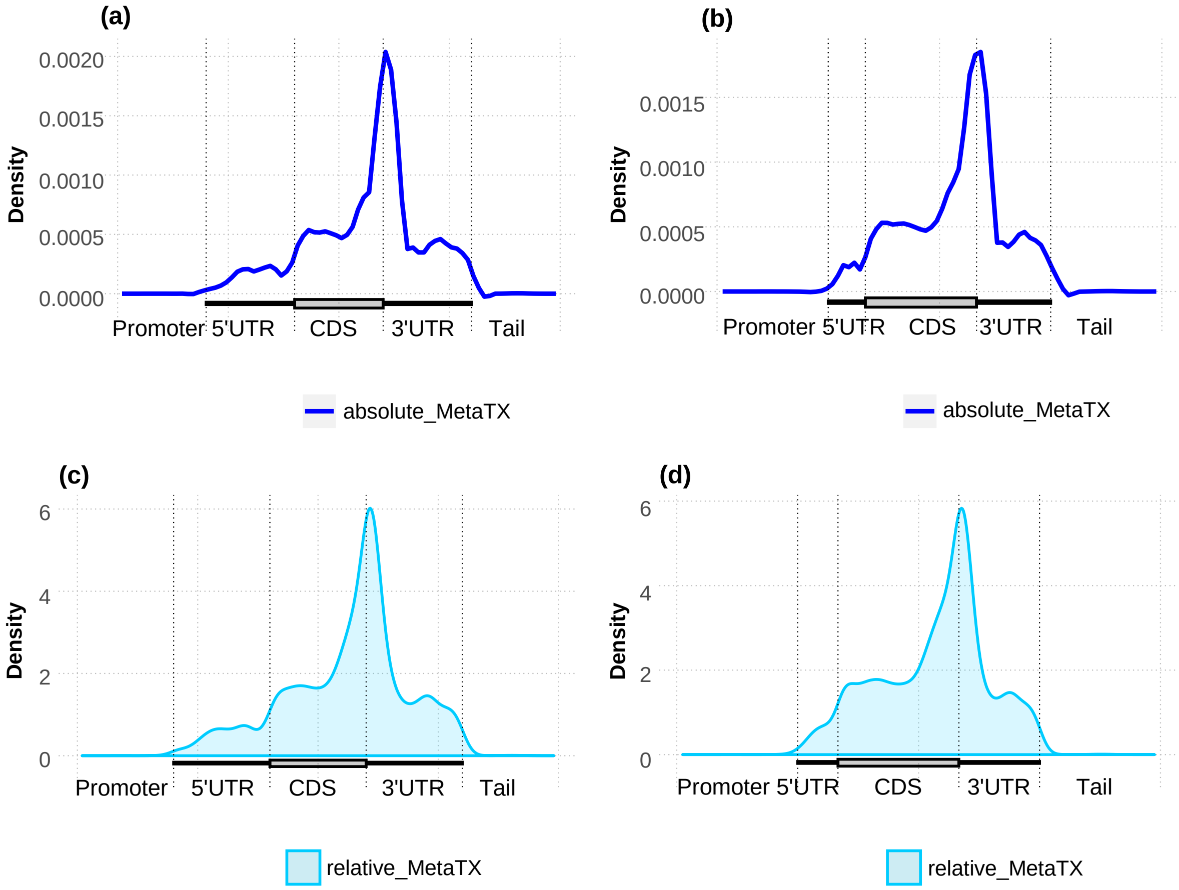

Figure S5. Visualization of m^{6}A pattern with customized relative length of RNA components. m^{6}A pattern is visualized with the promoter/5'UTR/CDS/3'UTR/tail ratio of 1:1:1:1:1 (a, c) and 3:1:3:2:3 (b, d). MetaTX R package supports two different ways of visualization. The 'absolute' method (a, b) provides absolute density (with the unit: number of features per bp exon transcript), which will not be affected by the relative length of different RNA components defined by the user. The 'relative' method (c, d) provides probability density function (with the area under the curve equals to 1), which can be affected by the relative length of different RNA components specified by user. Please note that the feature density reported by MetaTX is always length-normalized (see Equation 13 of the main manuscript) as suggested by its unit: number of features per nucleotide of mature transcript. We do not consider other possible unit such as number of features per component (5'UTR, CDS and 3'UTR). This is because 5'UTR, CDS and 3'UTR are substantially different in length, especially among different genes, and thus should not be considered equivalent. The option for customizing relative lengths (ratio) of different components supported by MetaTX package is for visualization purpose only. It helps to generate metagene plot where the lengths of different gene components complying with the convention or people's common understanding. It does not affect the inference procedure or its results.

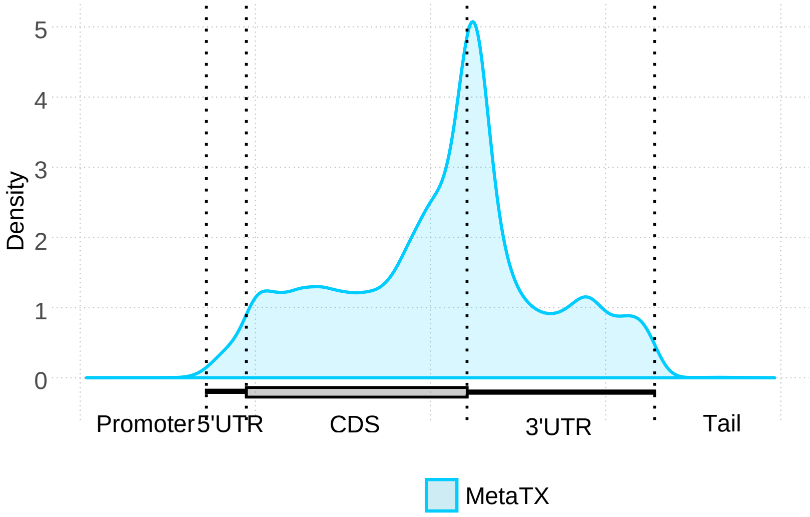

Figure S6. Visualization using true relative length of RNA components. The pattern is visualized with the real length of different RNA components estimated from the gene annotation used in the analysis. Users can choose this mode with the Boolean argument 'trueRelativeLength' of the MetaTXPlot function. People often assume that CDS is much longer than 3'UTR, while in reality the average length of the two are quite similar according to the transcriptome databases such as RefSeq and UCSC. The 5'UTR is indeed shorter.

Reference

1. Olarerin-George AO, Jaffrey SR: **MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites**. *Bioinformatics* 2017, **33**(10):1563-1564.
2. Cui X, Wei Z, Zhang L, Liu H, Sun L, Zhang S-W, Huang Y, Meng J: **Guitar: An R/Bioconductor Package for Gene Annotation Guided Transcriptomic Analysis of RNA-Related Genomic Features**. *BioMed research international* 2016, **2016**:8367534-8367534.
3. Huang T, Chen W, Liu J, Gu N, Zhang R: **Genome-wide identification of mRNA 5-methylcytosine in mammals**. *Nature structural & molecular biology* 2019, **26**(5):380-388.