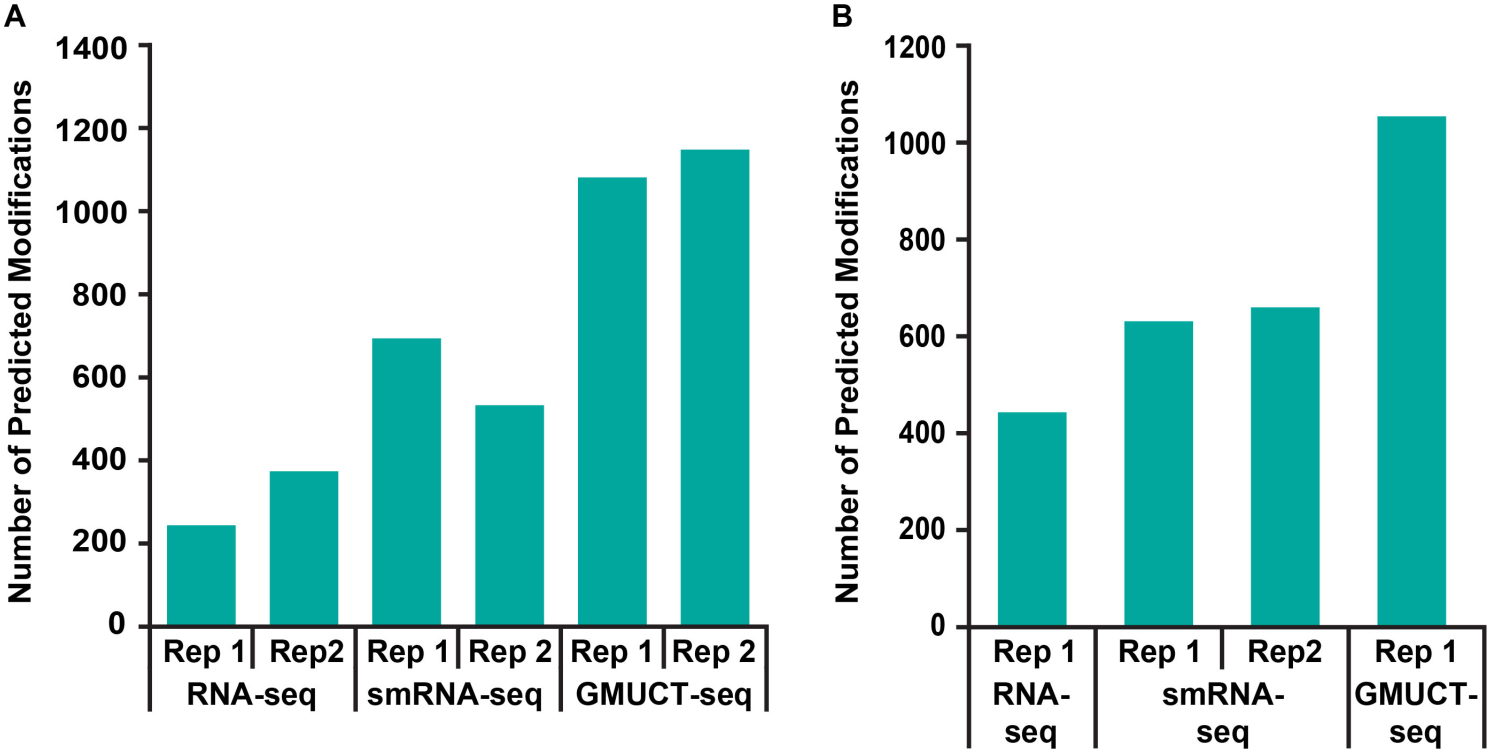
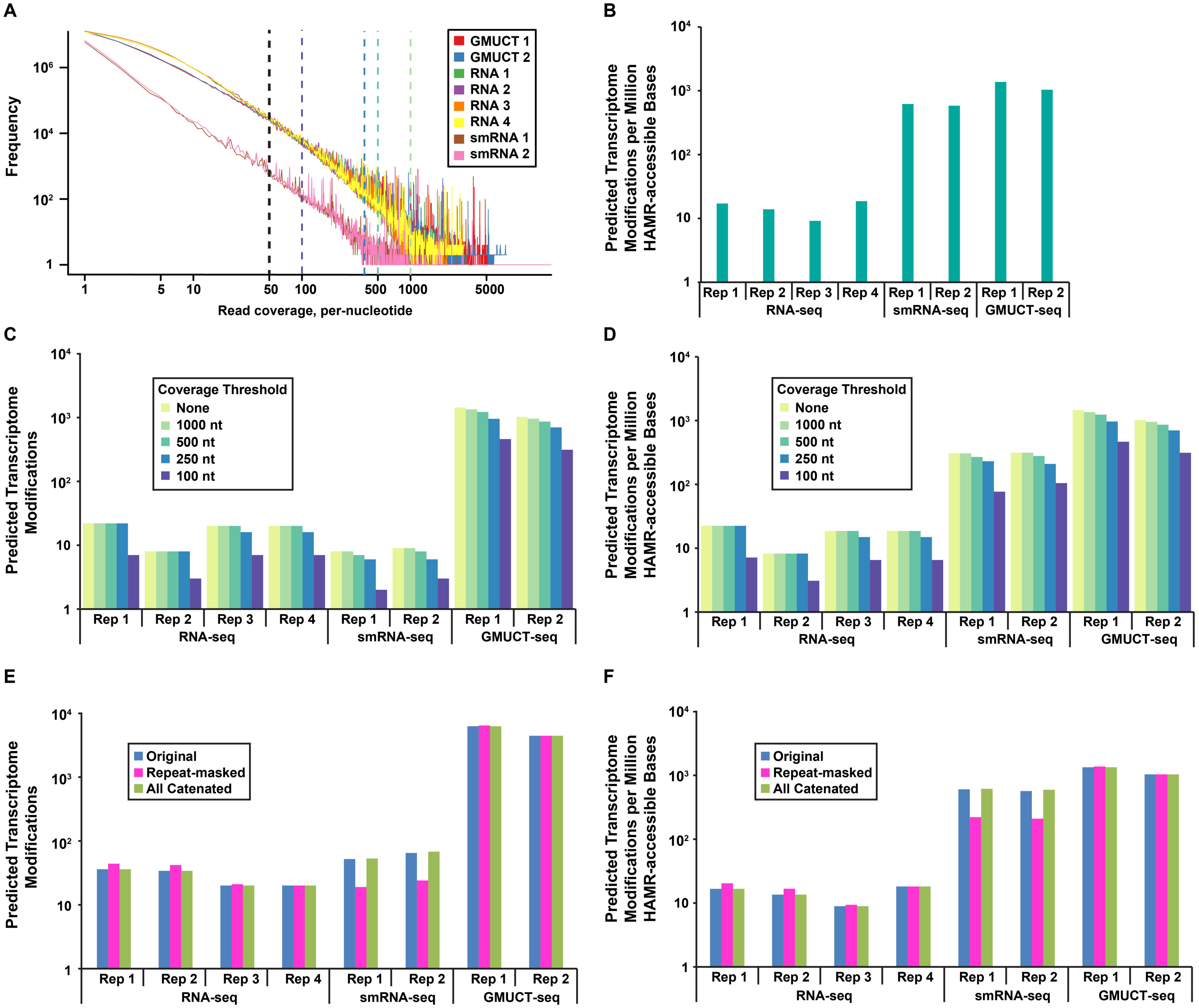
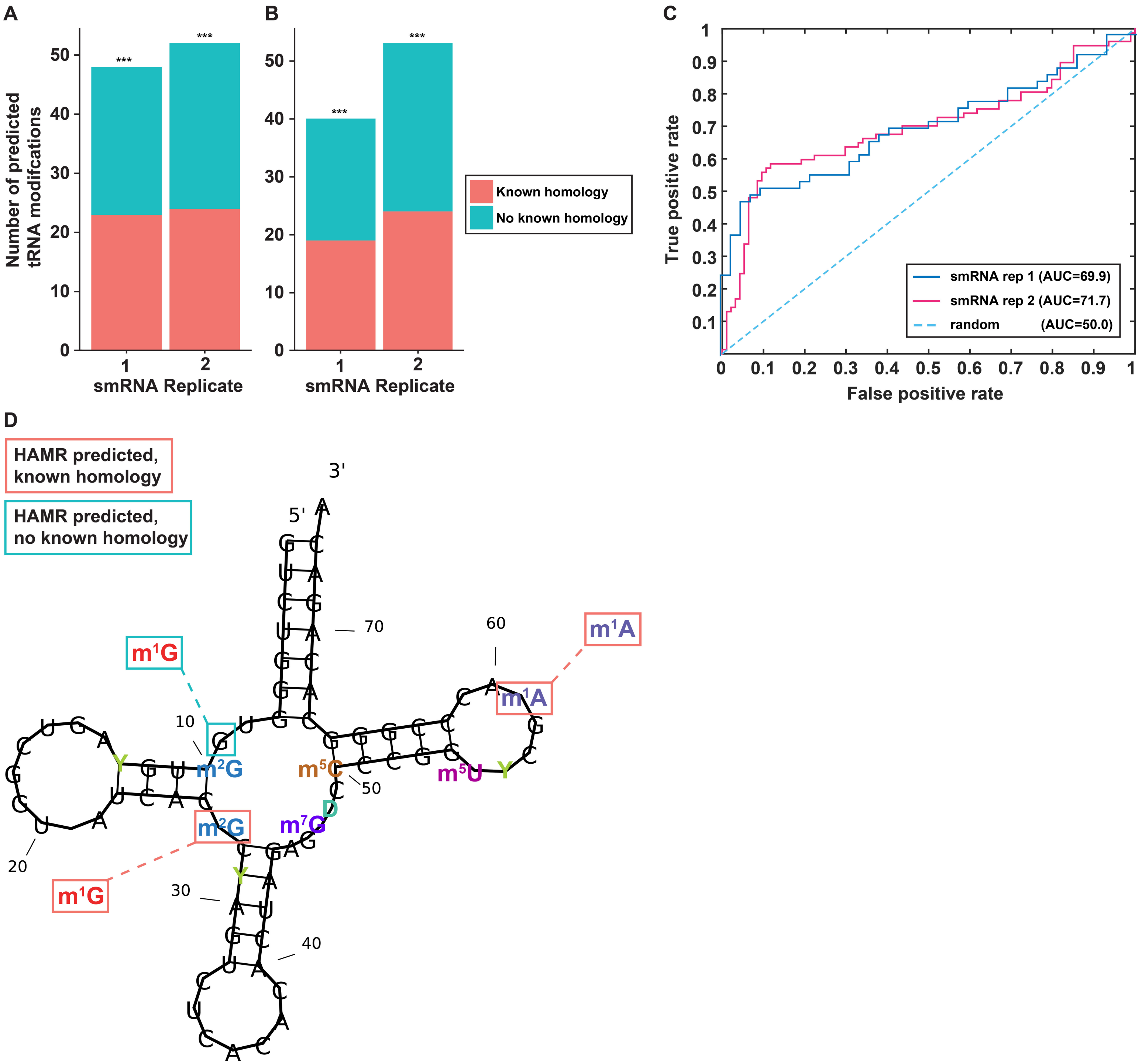
SUPPLEMENTAL FIGURES



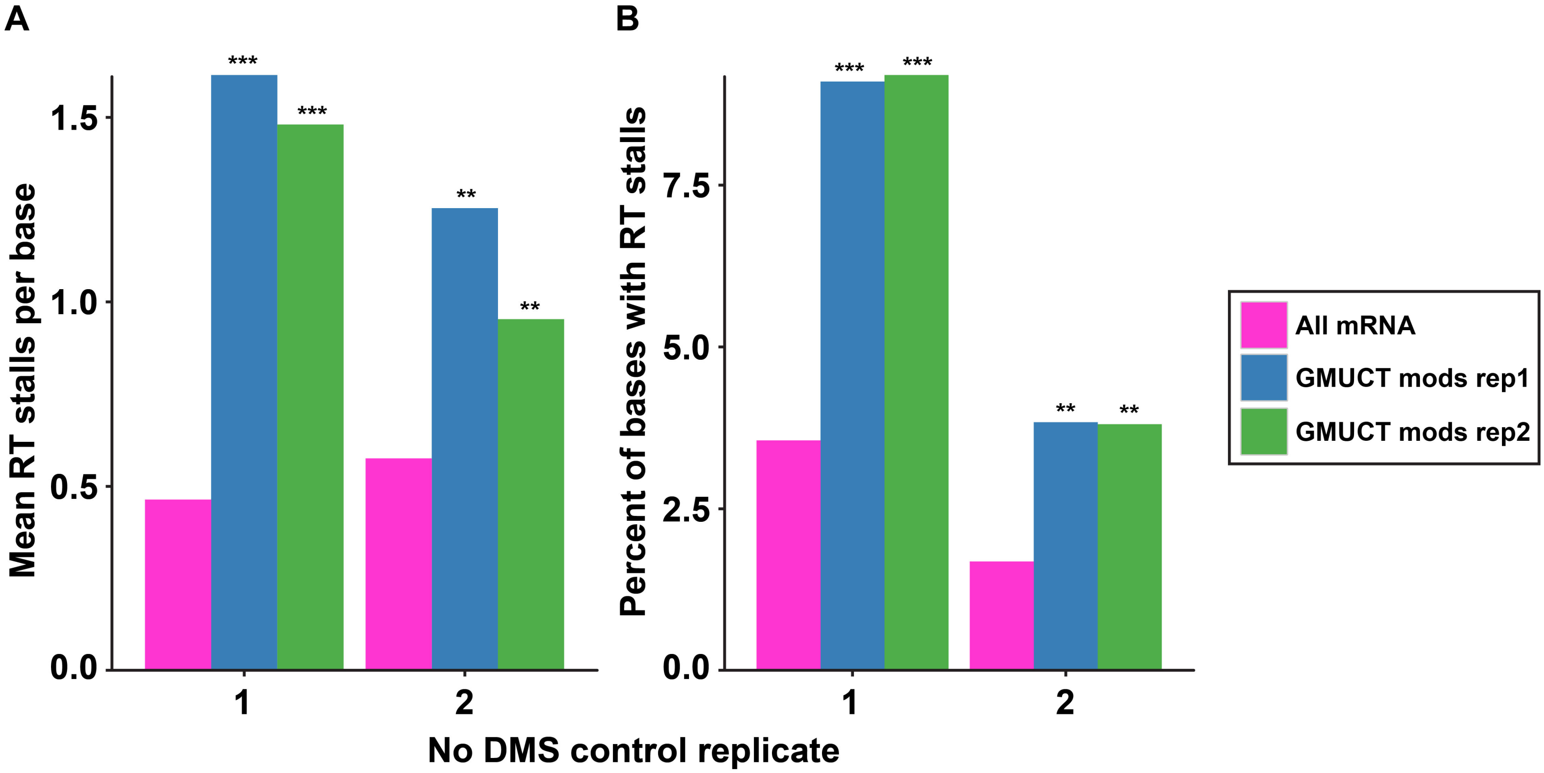
Supplemental Figure 1: HAMR-predicted modifications in two human cell lines. (A-B) Total number of HAMR-predicted modification sites from analyzing the three RNA-seq datasets (RNA-seq, smRNA-seq, and GMUCT) for HeLa (A) and HEK293T (B) cells.



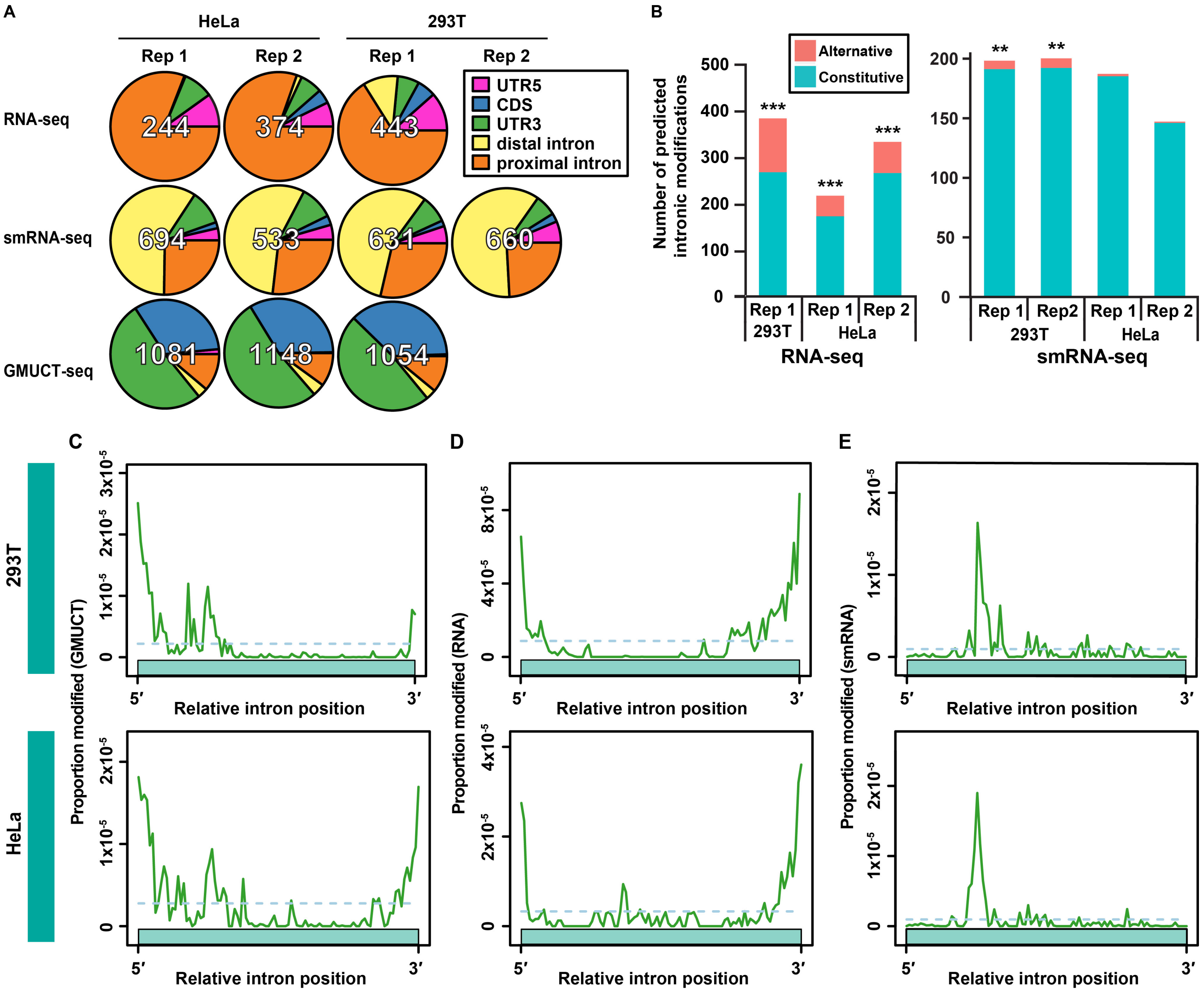
Supplemental Figure 2: Differences in number of HAMR-predicted modifications are not artifacts of differences in library preparation, overall size, or transcriptome coverage. (A) All *Arabidopsis* libraries were randomly down-sampled to the number of reads from the smallest library (~3 million), and a histogram of coverage at all TAIR10 mRNA transcriptome bases is plotted in log-log scale. The black dashed line indicates the 50x minimum coverage observed at a HAMR-predicted modification site (HAMR accessible bases), and colored dashed lines indicate various maximum coverage thresholds used in C and D. (B) Total number of HAMR modifications identified for each RNA-seq dataset were normalized to the number of HAMR accessible bases available from those experiments. (C) HAMR was rerun on down-sampled data, and modifications with greater than 100x, 250x, 500x, or 1000x coverage were excluded from the analysis. (D) Total number of HAMR modifications identified for each RNA-seq dataset after down-sampling were normalized to the number of HAMR accessible bases available from those experiments, and modifications with greater than 100x, 250x, 500x, or 1000x coverage were excluded from the analysis. (E) To exclude artifacts from mapping and read handling, HAMR was rerun on data from the three RNA-seq approaches that had been mapped to a repeat-masked (Smit, AFA, Hubley, R & Green, P. (2013) RepeatMasker Open-4.0. http://www.repeatmasker.org) TAIR10 transcriptome, and on RNA-seq and smRNA-seq data for which adapter-trimmed and untrimmed reads were concatenated in the same way that was done for GMUCT data (see methods). (F) The same analysis as in E in which the total number of HAMR modifications identified for each RNA-seq dataset were normalized to the number of HAMR accessible bases available from those experiments.



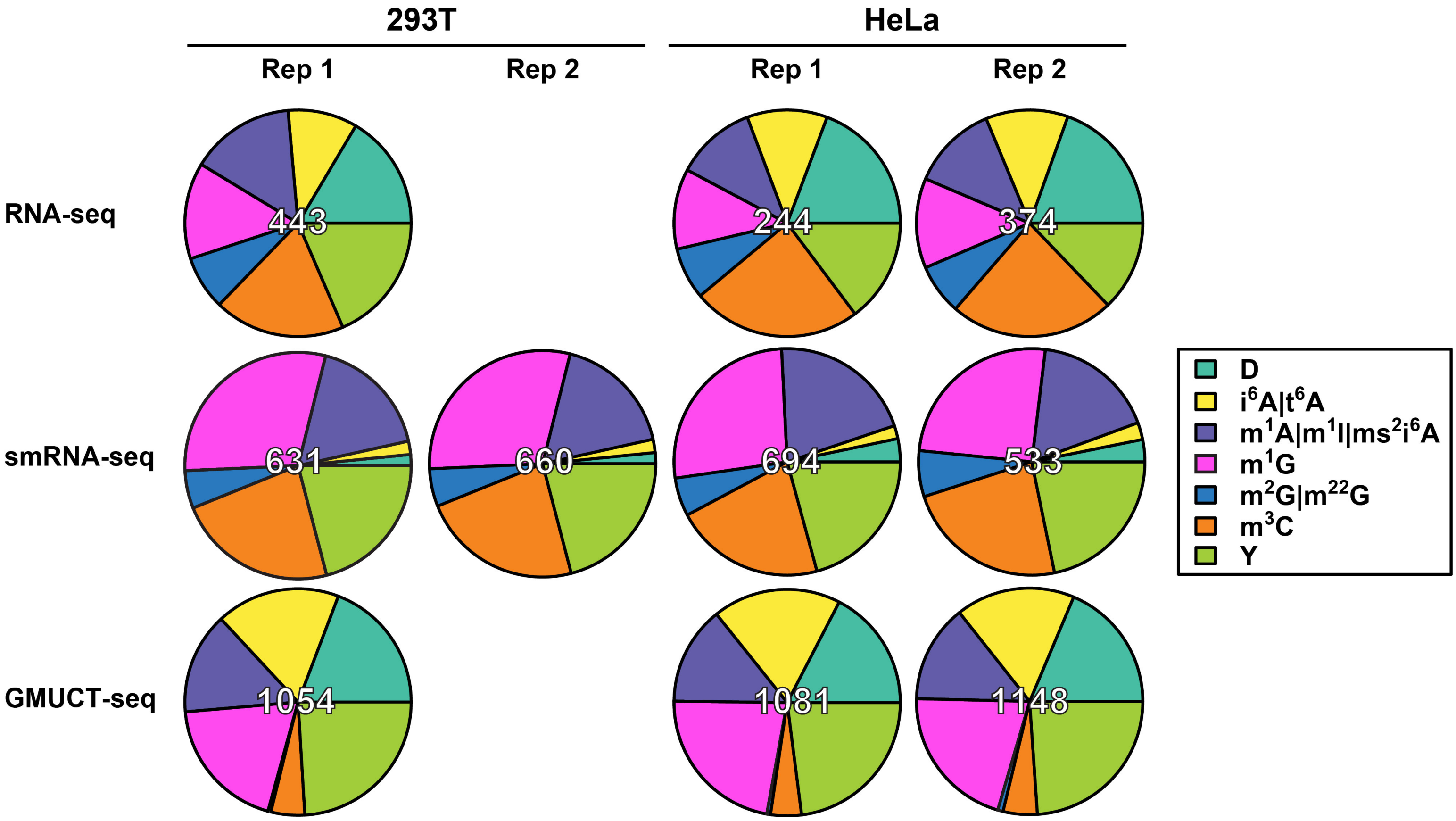
Supplemental Figure 3: HAMR captures a large proportion of known tRNA modification sites in the *Arabidopsis* transcriptome. HAMR modifications from (A) our smRNA sequencing data and (B) a previously published, tissue matched smRNA sequencing dataset (Li et al., 2014) are overlapped with known tRNA modifications, as determined by homology to yeast tRNAs. The total number of HAMR-predicted modifications are plotted on the y-axis. P-values were calculated by Fisher's exact test, over a background of all tRNA consensus bases (see methods). *** denotes p-value < $1x10^{-7}$. (C) Receiver operating characteristic curves for datasets from both replicates of our smRNA-seq experiments. AUC = area under curve. (D) An example tRNA, *tRNA-Val* (anticodon:CAC), with known modifications labeled as bold, colored letters across the structure backbone (black line). HAMR-predicted modification sites are labeled as known (red boxes) or novel (light blue boxes) with boxes across the structure backbone, while HAMR predicted modification types at those predicted nucleotide positions are shown as outlying boxes connected with dashed lines.



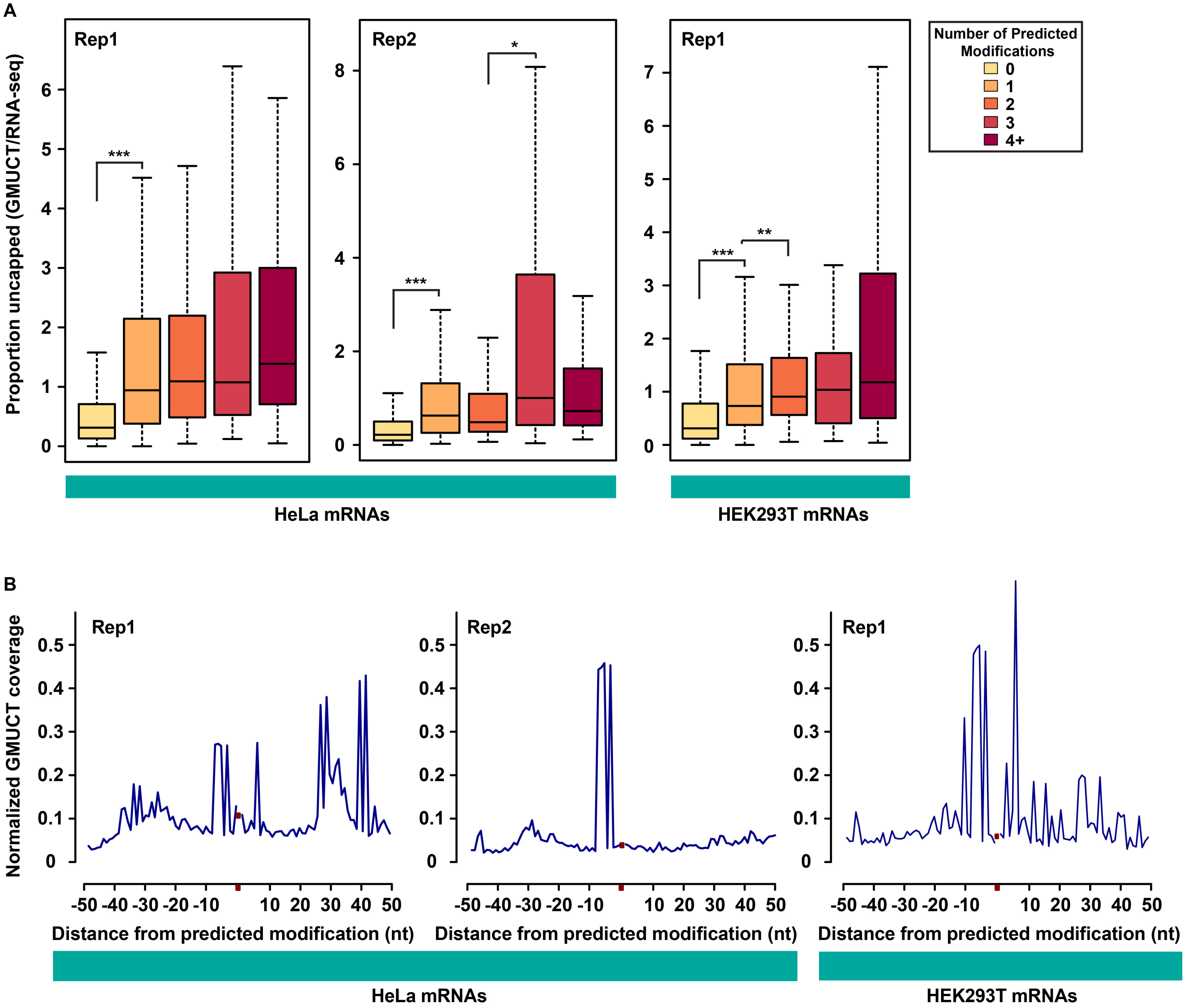
Supplemental Figure 4: Sites of HAMR-predicted modifications are enriched in reverse transcriptase (RT) stalls. RT stalls from no DMS control experiment datasets for Structure-seq (Ding et al., 2014) are tabulated across all mRNA bases (magenta bars), and across mRNAs predicted to contain modifications based upon GMUCT sequencing (blue and green bars). (A) The mean RT stalls per base and (B) the percent of bases with any number of RT stalls are plotted. Significance was determined for A with a Wilcoxon Rank Sum test (mean RT stalls per base) and for B with a Fisher's exact test (percent of bases with RT stalls) over a background of all mRNA bases. ** denotes p-value < 1x10^{-20} and *** denotes p-value < 1x10^{-50}.



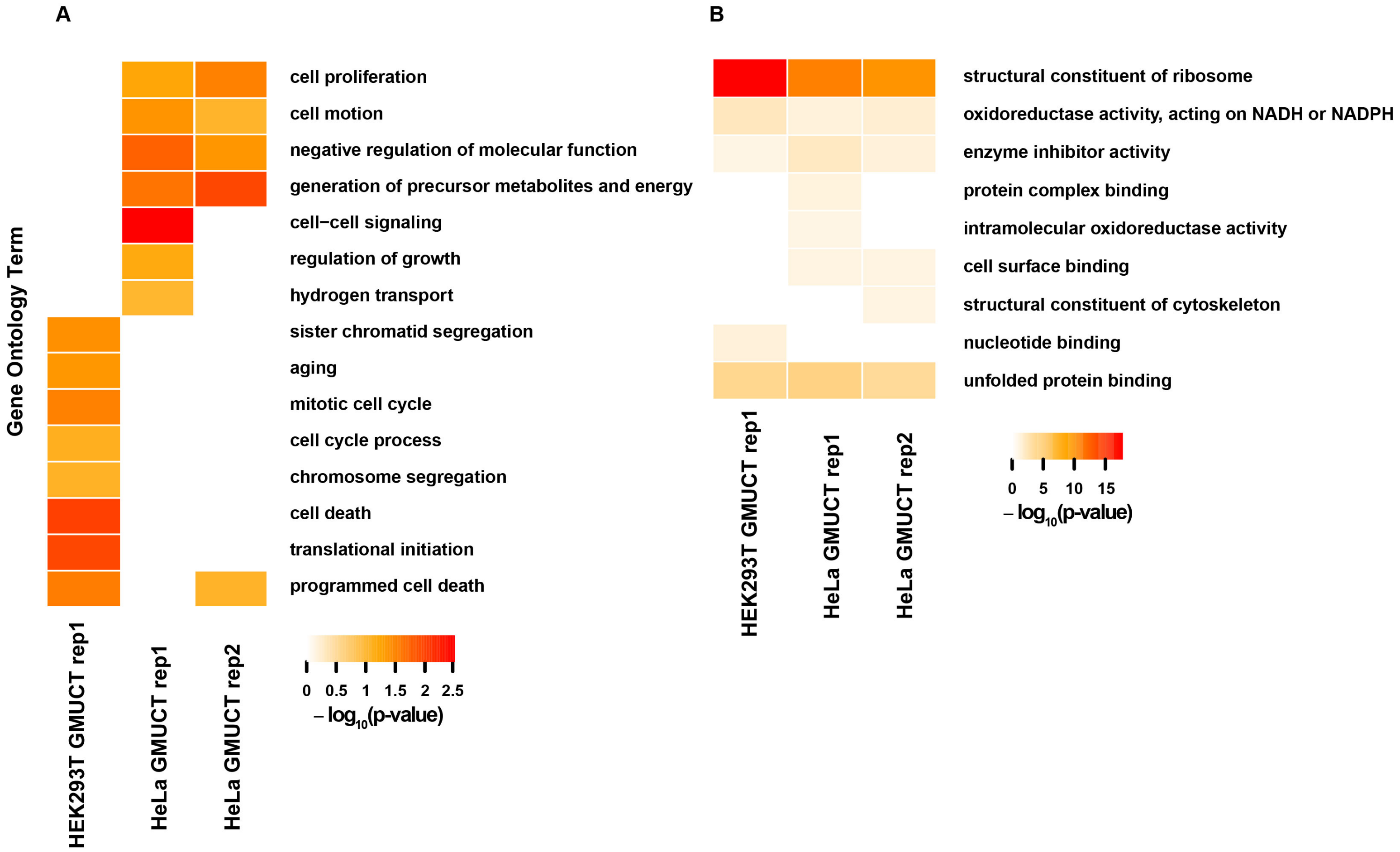
Supplemental Figure 5: HAMR-predicted modifications in two human cell lines mark uncapped and alternatively spliced transcripts. (A) The relative transcript location of predicted modification sites in mRNAs. Modifications that lie outside of mRNAs were excluded from this analysis. Intronic modification sites are proximal if within 500 nucleotides (nt) of a known constitutive or alternative splice donor/acceptor site, and distal if further than 500 nt from these sites. (B) Localization of HAMR-predicted modification sites identified using RNA-seq (left) and smRNA-seq (right) datasets within alternative compared to constitutive introns as annotated in hg19. Enrichment was calculated with a Fisher's exact test. ** denotes p-value < $1x10^{-10}$ and *** denotes p-value < $1x10^{-50}$. (C-E) Relative position of intron-localized HAMR-predicted modification sites using the data from (C) GMUCT, (D) RNA-seq, and (E) smRNA-seq plotted across the length-normalized average of all hg19 introns.



Supplemental Figure 6: HAMR predicts a variety of known and novel modification types in the human transcriptome. Distribution of the specific identities of HAMR-predicted modification sites, as determined by a nearest-neighbor classification approach trained on known tRNA modifications from *Saccharomyces cerevisiae*.



Supplemental Figure 7: Human RNAs with HAMR-predicted modifications have higher levels of uncapped transcripts. (A) Distribution of the proportion of uncapped transcripts (total GMUCT reads per transcript normalized to total RNA-seq reads) for protein-coding mRNAs. Modifications in noncoding RNAs were too sparse to test. P-values were calculated with a Wilcoxon Rank Sum test; * denotes p-value < 0.05, ** denotes p-value < 0.001, *** denotes p-value < $1x10^{-5}$. (B) Averaged GMUCT coverage profiles 50 bp up- and downstream of all predicted mRNA modification sites, normalized to RNA-seq read abundance. Red dots indicate the position of the predicted modification, and are plotted within 50 bp up- and downstream flanking regions. Modifications within 50 bp of the mRNA 5' or 3' ends were given correspondingly shorter flanking regions.



Supplemental Figure 8: Human transcripts with HAMR-predicted modifications encode proteins with coherent functions. (A) Biological process and (B) molecular function Gene Ontology (GO) terms are reported if they are significantly enriched (FDR < 0.05), over a background of all “HAMR accessible transcripts” with at least 10 uniquely mapping reads. Analyses were performed using the DAVID package (Huang, Sherman, and Lempicki, 2009). Furthermore, terms are only reported if they are separated from their ancestor term by no more than two parents, as determined by a depth first search as previously described (Vandivier et al., 2013). Lack of color denotes lack of significance.

Supplemental Table 1: HAMR correctly classifies a portion of homology-based predicted tRNA locus modification sites. Family-based predicted tRNA loci in Arabidopsis were intersected with HAMR machine learning-based predictions.

A. Arabidopsis smRNA Replicate 1.

tRNA family	Relative Start	Relative Stop	HAMR-predicted Modification	Actual Modification	Correct?
AT_Ala_TGC_consensus_0	25	26	m1G	m2,2G	N
AT_Arg_ACG_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y
AT_Arg_TCT_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y
AT_Asn_GTT_consensus_0	26	27	m1G	m2,2G	N
AT_Gly_GCC_consensus_0	8	9	m1G	m1G	Y
AT_Leu_CAA_consensus_0	64	65	Y	m5U	N
AT_Leu_TAA_consensus_0	25	26	m1G	m2,2G	N
AT_Leu_TAG_consensus_0	26	27	m1G	m2,2G	N
AT_Leu_TAG_consensus_0	64	65	m1A|m1I|ms2i6A	m1A	Y
AT_Lys_CTT_consensus_0	25	26	m1G	m2,2G	N
AT_Lys_CTT_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y
AT_Lys_TTT_consensus_0	56	57	m1A|m1I|ms2i6A	m1A	Y
AT_Met_CAT_consensus_0	8	9	m1G	m1G	Y
AT_Phe_GAA_consensus_0	25	26	m1G	m2,2G	N
AT_Pro_TGG_consensus_0	32	33	D	xU	N
AT_Pro_TGG_consensus_0	56	57	m1A|m1I|ms2i6A	m1A	Y
AT_Ser_AGA_consensus_0	25	26	m1G	m2,2G	N
AT_Ser_GCT_consensus_0	66	67	m1A|m1I|ms2i6A	m1A	Y
AT_Trp_CCA_consensus_0	56	57	m1A|m1I|ms2i6A	m1A	Y
AT_Val_AAC_consensus_0	58	59	m1A|m1I|ms2i6A	m1A	Y
AT_Val_CAC_consensus_0	26	27	m1G	m2G	N
AT_Val_CAC_consensus_0	58	59	m1A|m1I|ms2i6A	m1A	Y
AT_Val_TAC_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y

B. Arabidopsis smRNA Replicate 2.

tRNA family	Relative Start	Relative Stop	HAMR-predicted Modification	Actual Modification	Correct?
AT_Ala_AGC_consensus_0	33	34	m1A|m1I|ms2i6A	I	N
AT_Ala_TGC_consensus_0	25	26	m1G	m2,2G	N
AT_Arg_ACG_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y
AT_Arg_CCT_consensus_0	25	26	m1G	m2,2G	N
AT_Arg_TCT_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y
AT_Asn_GTT_consensus_0	26	27	m2G|m22G	m2,2G	Y
AT_Gly_GCC_consensus_0	8	9	m1G	m1G	Y
AT_Leu_TAA_consensus_0	25	26	m1G	m2,2G	N
AT_Leu_TAG_consensus_0	26	27	m1G	m2,2G	N
AT_Leu_TAG_consensus_0	64	65	m1A|m1I|ms2i6A	m1A	Y
AT_Lys_CTT_consensus_0	25	26	m1G	m2,2G	N
AT_Lys_CTT_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y
AT_Lys_TTT_consensus_0	56	57	m1A|m1I|ms2i6A	m1A	Y
AT_Met_CAT_consensus_0	8	9	m1G	m1G	Y
AT_Phe_GAA_consensus_0	25	26	m1G	m2,2G	N
AT_Pro_TGG_consensus_0	32	33	D	xU	N
AT_Pro_TGG_consensus_0	56	57	m1A|m1I|ms2i6A	m1A	Y
AT_Ser_AGA_consensus_0	25	26	m1G	m2,2G	N
AT_Ser_GCT_consensus_0	66	67	m1A|m1I|ms2i6A	m1A	Y
AT_Trp_CCA_consensus_0	56	57	m1A|m1I|ms2i6A	m1A	Y
AT_Val_AAC_consensus_0	58	59	m1A|m1I|ms2i6A	m1A	Y
AT_Val_CAC_consensus_0	26	27	m1G	m2G	N
AT_Val_CAC_consensus_0	58	59	m1A|m1I|ms2i6A	m1A	Y
AT_Val_TAC_consensus_0	57	58	m1A|m1I|ms2i6A	m1A	Y

Supplemental Table 2: Primer sequences used for RT-qPCR.

Target	Primer
AT1G43170 forward	TGGGCACAGCATTTGAGTGA
AT1G43170 reverse	ACTGCTTAGCGTACCCAGTG
AT4G25080 forward	CCCAGGGCCATCAAAAGCTA
AT4G25080 reverse	TCCAGCCGACTTTACCCAAC
AT4G25080 forward (additional primer set)	TCGTGGAAGACATGCAGATTC
AT4G25080 reverse (additional primer set)	GTTTGTACAGACCGTCCTCCT
AT1G04410 forward	GCTGCAATCATCAAGGCGAG
AT1G04410 reverse	TGGAAACGAACGTACCCCTC
AT1G04410 forward (additional primer set)	ACAACAGGGCTTTGGGACAG
AT1G04410 reverse (additional primer set)	GACAGGCTTCTCTCCAGACG
AT1G15220 forward	CAACACGAGCCCGAAGAGT
AT1G15220 reverse	AGAAAGTGAACGACTGAGGCT
AT1G28330 forward	GCGGAAGATCAGGTCACCAT
AT1G28330 reverse	TGGGGTGTTTGCAGGTTGTA
AT1G28330 forward (additional primer set)	TAAAGACGCTCCTCCACACG
AT1G28330 reverse (additional primer set)	GAGCAGCAGTAAGGTGGTGA
AT2G15580 forward	GAGAAACTTGACGGAGCAGC
AT2G15580 reverse	TGTACGTGGTGGGATTCTCAG
AT3G15353 forward	CTGTGCTGACAAGACCCAGT
AT3G15353 reverse	CTCCTGAGTCTCGACGATGT
AT4G08620 forward	CCCGGAATCTTGATCATCC
AT4G08620 reverse	CGGCATGCCATATTCCTTAG
AT3G21170 forward	TGAGGCAGGGTCGTCTTATC
AT3G21170 reverse	CACGCCACTGGTGATATTTG
AT1G66850 forward	GCCATCAAAGCCGAAGACAC
AT1G66850 reverse	ACGCAGGGTTCTTAGCGAAA
AT3G20865 forward	GGAGTCTCCAGCACCATCAC
AT3G20865 reverse	GAAGAGCCAAGAAGGCGGAG
AT5G39420 forward	CAAGGAGATTGGGCGGTTCT
AT5G39420 reverse	CCAACTTCTGGAACGCCTCT
AT4G31070 forward	CTGAAGGGTTTGGTGTCGGA
AT4G31070 reverse	CTGTGAAGCCATTGGTCCCT
tRNA-Arg (anticodon: AGT) forward	CCGCGTGGCCTAATGGATAA
tRNA-Arg (anticodon: AGT) reverse	GATCACGGTGGGACTCGAAC
tRNA-Trp (anticodon: CCA) forward	GATCCGTGGCGCAATGGTAG
tRNA-Trp (anticodon: CCA) reverse	TGAACCCGACGTGAATCGAA
tRNA-ala (anticodon:AGC) forward	GGGGATGTAGCTCAGATGGT
tRNA-ala (anticodon:AGC) reverse	TGGAGATGCGGGGTATCG

ADDITIONAL SUPPLEMENTAL FILES

Supplemental Files 1 and 2 must be downloaded separately.

Supplemental File 1: Homology-based prediction of *Arabidopsis* tRNA family modification sites. Families of tRNA loci in *Arabidopsis* were collapsed to consensus sequences, and yeast modifications were lifted over based upon sequence homology. Table is in BED format, with the following columns from left to right: tRNA family consensus sequence, 0-based start, 1-based stop, modification type (MODOMICS short name), supporting yeast sequence, strand (not applicable). Note that modifications are duplicated when supported by multiple yeast sequences, and should be collapsed with a tool such as Bedtools merge (http://bedtools.readthedocs.org/en/latest/content/tools/merge.html) before use in analysis.

Supplemental File 2: Homology-based prediction of *Arabidopsis* tRNA locus modification sites. Family-based predicted tRNA loci in *Arabidopsis* were assigned to all loci of each corresponding family. Table is in BED format, with the following columns from left to right: *Arabidopsis* chromsome, 0-based start, 1-based stop, modification type (MODOMICS short name), *Arabidopsis* transcript name, strand.