

RESEARCH

Open Access



Comparative analyses suggest a link between mRNA splicing, stability, and RNA covalent modifications in flowering plants

Kyle Palos¹, Anna C. Nelson Dittrich¹, Eric H. Lyons², Brian D. Gregory³ and Andrew D. L. Nelson^{1*}

Abstract

Background In recent years, covalent modifications on RNA nucleotides have emerged as pivotal moieties influencing the structure, function, and regulatory processes of RNA Polymerase II transcripts such as mRNAs and lncRNAs. However, our understanding of their biological roles and whether these roles are conserved across eukaryotes remains limited.

Results In this study, we leveraged standard polyadenylation-enriched RNA-sequencing data to identify and characterize RNA modifications that introduce base-pairing errors into cDNA reads. Our investigation incorporated data from three Poaceae (*Zea mays*, *Sorghum bicolor*, and *Setaria italica*), as well as publicly available data from a range of stress and genetic contexts in *Sorghum* and *Arabidopsis thaliana*. We uncovered a strong enrichment of RNA covalent modifications (RCMs) deposited on a conserved core set of nuclear mRNAs involved in photosynthesis and translation across these species. However, the cohort of modified transcripts changed based on environmental context and developmental program, a pattern that was also conserved across flowering plants. We determined that RCMs can partly explain accession-level differences in drought tolerance in *Sorghum*, with stress-associated genes receiving a higher level of RCMs in a drought tolerant accession. To address function, we determined that RCMs are significantly enriched near exon junctions within coding regions, suggesting an association with splicing. Intriguingly, we found that these base-pair disrupting RCMs are associated with stable mRNAs, are highly correlated with protein abundance, and thus likely associated with facilitating translation.

Conclusions Our data point to a conserved role for RCMs in mRNA stability and translation across the flowering plant lineage.

Keywords RNA covalent modifications, RNA, Comparative transcriptomics, RNA decay, Angiosperms

*Correspondence:

Andrew D. L. Nelson
an425@cornell.edu

¹Boyce Thompson Institute, Cornell University, 533 Tower Road, Ithaca, NY 14853, USA

²School of Plant Sciences, University of Arizona, Tucson, AZ, USA

³Department of Biology, University of Pennsylvania, Philadelphia, PA, USA



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

There are over 100 RNA covalent modifications (RCMs) that are deposited on all classes of RNAs at various stages of their lifecycle [1]. RCMs, collectively referred to as the “epitranscriptome” [2, 3] are known to influence RNA stability, splicing, structure, intra- and intermolecular interactions, and translation [4–9]. While most RCMs have been observed on ribosomal and transfer RNAs, they have also been observed on messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs), albeit at lower levels [4, 8, 10]. Although pseudouridine is the most abundant modification among all classes of RNA [10], N^6 -methyladenosine (m^6A) is the most abundant internal mRNA RCM targeting approximately 30% of all mRNAs and is present at $\sim 0.4\%$ of all adenosine nucleotides in cellular RNAs [11, 12]. Additional internal RCMs, such as pseudouridine, N^1 -methyladenosine (m^1A), and 5-methylcytidine (m^5C), have been observed on mRNAs in a number of eukaryotic systems, although their functional significance on these RNA molecules is not always clear [13–15].

Much of what we do know about the epitranscriptome is drawn from studies on m^6A in mammalian and plant systems, but can likely be applied to many of the other modification classes. With the notable exception of pseudouridine, which results from the isomerization of uridine [16–19], RCMs are deposited through the enzymatic activity of highly conserved proteins called writers (e.g., methyltransferases), “interpreted”, or bound, by RNA binding proteins called readers, and removed by eraser enzymes (e.g., demethylases) [5, 20–22]. Arabidopsis and mammals have a similar m^6A frequency across the transcriptome ($\sim 1 m^6A$ per 2000 nucleotides), which is targeted to a similar consensus motif RRACH (R=G or A; H=U, A, or C), and has a similar 3' mRNA bias around the stop codon [11, 23]. In plants and mammals, pseudouridine has been found to localize primarily to the CDS and 5' UTR of mRNAs, is deposited on a wide array of mRNAs, and prefers the first U of a triplet codon (e.g., UUC/UCU/UUU) [17]. Motif preferences for m^5C can vary depending on the methyltransferase involved, but in humans are generally found on C/G rich regions of mRNAs [24]. While less is known about binding preferences of the other predominant RCMs, such as m^1A and m^3C , aberrant addition (or removal) has been observed as a hallmark for tumorigenesis [25, 26]. Importantly, loss of many of the writers, readers, and erasers for these RCMs is lethal or causes severe developmental defects in all tested systems [14, 27–30].

RCMs function through a variety of molecular mechanisms to regulate mRNA stability, structure, and splicing. RCMs induce changes in base-pairing properties and higher order structure, ostensibly allowing for increased flexibility in interactors, translational capacity, and

structure-associated stabilization [31]. There is a close association between splicing and m^6A deposition [23, 32], with recent studies revealing that m^6A is specifically excluded from splice junctions due to physical occlusion by the exon junction complex, a process that ultimately impacts cytoplasmic mRNA decay [33, 34]. Pseudouridine has also been implicated in mRNA splicing, stability, and translational efficiency [17, 35, 36], as well as the stability and maturation of lncRNAs such as the telomerase RNA [37, 38]. Other modifications, such as m^1A , m^3C , and m^5C are also believed to be important for mRNA stability, mostly due to phenotypes associated with loss of function mutants in their associated writer proteins [25, 35, 39, 40]. Thus, RCMs are both a ubiquitous and critical aspect of an RNA's lifecycle, but specific mechanistic details for many of them are still lacking.

Given the high frequency with which essential RNAs such as rRNAs, tRNAs and snoRNAs are modified, it has historically been difficult to use genetic approaches to monitor the RCM status of mRNAs and lncRNAs. As a result, most studies have relied either on transcriptome-wide antibody-based (m^6A), or modification-specific chemical-genomic approaches (e.g., bisulfite sequencing for m^5C ; [7, 10, 41, 42]). An alternative sequencing-based approach relies on the propensity for RCMs that occur at the interface between the canonical base-pairing edge (i.e., the Watson-Crick-Franklin, or WCF base-pairing edge) to be misidentified by reverse transcriptases during the cDNA synthesis step [43–47]. This misidentification results in non-random misincorporations or deletions at modified residues. Multiple algorithms, including HAMR and ModTect [15, 44], have been developed to infer modification status, and class, based on these “sequencing errors”. While this approach cannot detect RCMs outside of the WCF base-pairing edge, over 30 modification types can be identified, including m^1A , pseudouridine, m^3C , and m^5C [15, 44, 48, 49]. Importantly, where there is overlap (such as for m^5C , m^3C , and m^1A), the antibody, chemical-genomic, and sequencing error-based approaches are largely in agreement [14, 15]. Inferring modification status via HAMR or ModTect allows both for the repurposing of existing RNA-seq data, of which there are petabytes available in NCBI's SRA [50], and the side-by-side calculation of modification and transcript abundance.

Here we utilized both HAMR and ModTect to analyze new and publicly available RNA-seq in Arabidopsis and three agronomically important and closely related grasses: *Sorghum bicolor*, *Zea mays*, and *Setaria italica*. We used these diverse stress, tissue-atlas, and developmental datasets in an attempt to better understand what some of these less characterized RCMs might be doing in plants and how they might be conserved, both functionally and in terms of their target genes. We present

an in-depth comparative assessment of RCM dynamics across these plant species and uncover new insights into RNA splicing, RNA stability, and plant responses to stress. We demonstrate a level of conservation of target genes previously unseen for RCMs, and reveal that RCMs may participate in population-level variation in stress responses. Mechanistically, we identify a link between splicing and RCMs that is present in all examined systems. Finally, through the analysis of RNA decay pathways in *Arabidopsis*, we establish that RCMs are associated with unusually stable mRNAs, an aspect that may facilitate their continued translation, as we observe a link between modification state and protein abundance.

Results

Identification of post-transcriptional RNA modifications in three poaceae

To determine the degree to which the epitranscriptome is conserved at a gene level, we performed poly-A enriched RNA-sequencing on soil grown seedlings of *Zea mays*, *Sorghum bicolor*, and *Setaria italica* (Maize, Sorghum, and Setaria, respectively, going forward) two weeks after germination. Paired-end, 150 base-pair RNA-sequencing was performed on aboveground tissue for two biological replicates (~20 million reads per replicate; Fig. 1A). Following read mapping, modified sites were inferred using HAMR and ModTect (see **Methods**). Both of these algorithms take advantage of the propensity for certain reverse transcriptases to misinterpret ribonucleotides that are modified along the canonical base-pairing edge, and as a result, arrest, skip over the RCM, or incorporate non-reference nucleotides (i.e., SNPs) in the synthesized cDNA [51]. This misincorporation is non-random, and both HAMR and ModTect use the resulting population of SNPs to infer modification class. Importantly, both HAMR and ModTect disregard biallelic SNPs that could arise from natural variation, and instead focus on high coverage sites with all three non-reference nucleotides observed (HAMR) or incorporate information about both mismatches and deletions at the modified residue (ModTect). RNA-seq data were fed into both algorithms and sites that were predicted by both algorithms and both replicates were retained for subsequent analyses as high confidence RCMs. From this approach, 5,434–7,020 unique RCMs were identified in 1,944–2,542 transcripts in the three species (Fig. 1B, C, Additional File 1). These modifications represent the seven major classes detected by HAMR, with the m1A|m1I|ms2i6A class being the most common in all three species (approximately 33% of all identified RCMs; Fig. S1A). An examination of sequencing data across modified transcripts reveals a drop in coverage at the modification site, coinciding with a previously observed RT-arrest (Fig. S1B; [52]). Additionally, we find that mRNAs marked by RCMs are

present at a substantially higher abundance relative to non-modified mRNAs (modified median TPM=78.5, $n=2521$ mRNAs, not modified median TPM=10.2, $n=16,537$, $p<2.2e-16$, Fig. S1 E). Enriched gene ontology terms of modified RNAs in each Poaceae demonstrates significant over-representation in photosynthesis and cytoplasmic translation pathways (Fig. 1D). These data suggest that RCMs are targeted to conserved cohorts of photosynthesis and translation-associated RNAs in these grass seedlings.

To determine if the conservation of RCM deposition goes beyond functional pathway and extends to the genes themselves, we next examined if RCMs are found on transcripts derived from orthologous genes using both sequence homology and synteny. Of the 1,912 possible modified orthologs (the smallest observed number of modified genes in *Setaria*), we observed 1,074 modified orthologs in one or both of the other species, and 474 orthologs that were modified in all three. This is significantly more orthologs than expected by chance, accounting for the number of annotated and expressed orthologs ($p<2.2e-16$ multi-set hypergeometric test, Fig. S1C). We then assessed whether orthologous mRNAs are marked by a similar number of RCMs across species, allowing us to understand whether the RCM density of the targeted RNA is a conserved feature. Indeed, for all three combinations of comparisons between two sampled species, there is a significant positive relationship between the number of RCMs deposited on orthologous mRNAs (Sorghum: Maize $r=.39$, Sorghum: *Setaria* $r=.43$, *Setaria*: Maize: $r=.32$, $p<2.2e-16$, Fig. S1D). These findings further demonstrate that the targeting and density of RCM deposition are conserved features in the sampled Grasses.

RCMs are dynamically deposited on mRNAs based on tissue and abiotic stress

Our previous analyses were limited to a single developmental time point across three species. Previous reports [14, 30, 53] suggest that RCMs may play important roles in post-transcriptional RNA regulation and thus would be dynamically deposited across development or environmental changes. Therefore, we chose to examine RCMs across diverse tissues and environmental contexts using RNA-seq from publicly available datasets [54–56]. We focused our efforts on Sorghum, first identifying RCMs in a large-scale tissue expression atlas by McCormick and co-authors [55] containing 137 sequencing samples across 48 tissues/stages/conditions (see **Methods**). From these data we identified 266,710 modifications on 6,805 unique transcripts, representing 19.3% of the expressed (TPM \geq 1) Sorghum transcriptome. To determine whether the same repertoire of transcripts are being targeted with RCMs as our seedling data from Fig. 1, or if our data were biased towards identifying a distinct subset

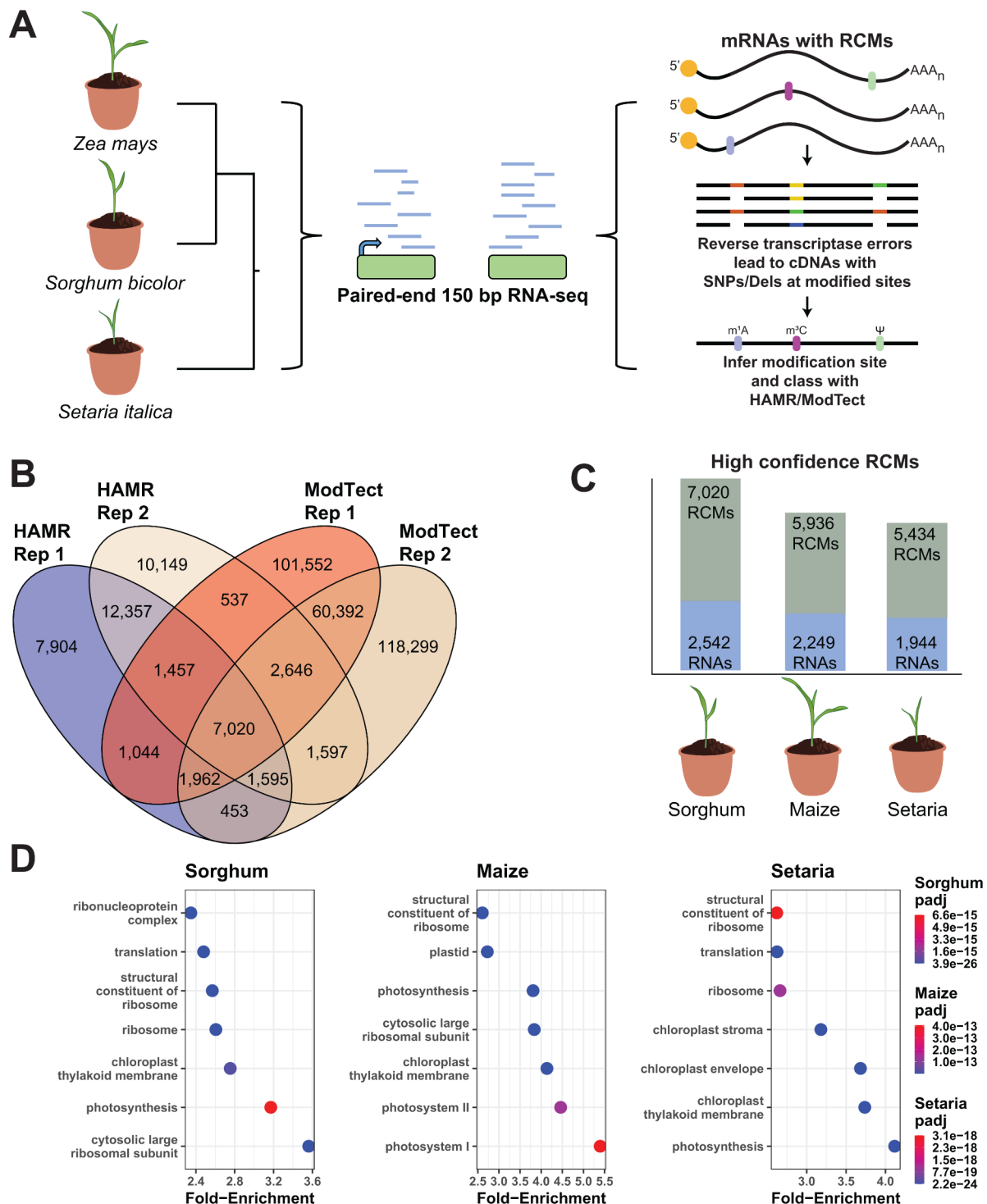


Fig. 1 RCM identification in three model Grasses. **(A)** Schematic of experimental design. Left and middle: Two biological replicates of grass seedlings from each species at a similar developmental timepoint were used to generate paired-end RNA-seq libraries. Right: Depiction of RCMs causing errors in reverse transcriptase nucleotide incorporation. Different modifications will result in a non-random pattern of SNPs or deletions at the modified site in the resulting cDNA. Machine Learning algorithms such as HAMR and ModTect can infer the modification class based on these errors. **(B)** Venn diagram showing the overlap of modified sites in the Sorghum transcriptome between biological replicates and two RCM detection algorithms. The union of all four categories ($n=7020$) was kept for downstream analyses. **(C)** Summarizing the number of modified sites and number of modified transcripts in all three sampled Grasses. **(D)** Enriched GO terms of the modified RNAs in **(C)**

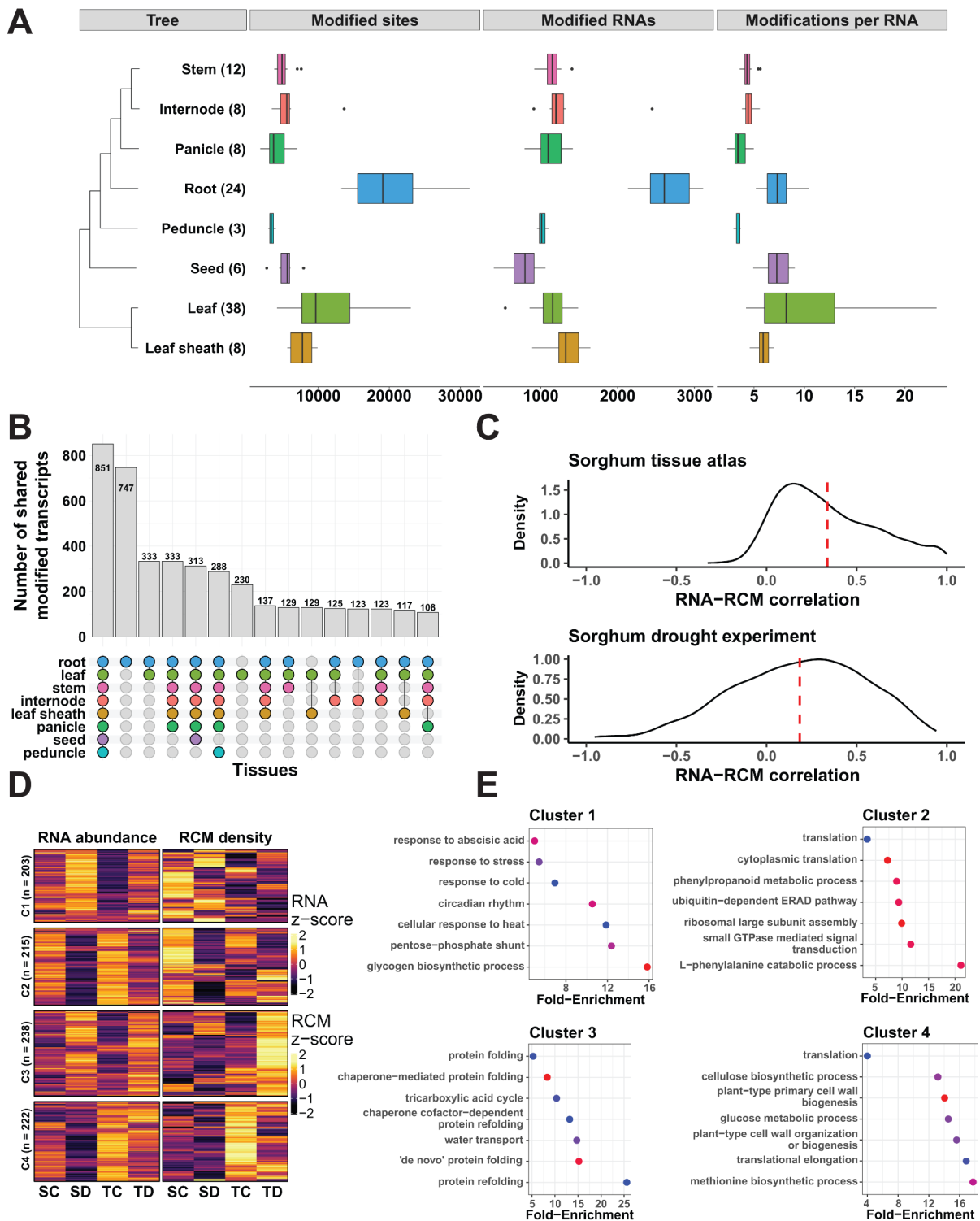


Fig. 2 (See legend on next page.)

of RNAs, we compared the modified RNAs in each dataset. We found 1,813 of the 2,542 modified Sorghum seedling mRNAs are also modified in the McCormick et al., tissue expression atlas. This overlap is substantially more

than expected by chance ($p < 2.2e-16$, Fig. S2 A) and likely reflects similarity in molecular processes within tissues examined.

(See figure on previous page.)

Fig. 2 RCM dynamics across development and during stress. **(A)** Summary of RCM findings from re-analysis of Sorghum tissue expression atlas [55]. Hierarchical clustering using the number of RCMs per transcripts as input. For ease of viewing, branches were collapsed based on tissue and modification similarity. See Fig. S2E for un-collapsed tree. Tree tip labels denote the broad tissue category with the number after the tissue representing the number of SRAs used to summarize that broad tissue. **Modified sites:** Boxplots showing the total number of RCMs per SRA in each broad tissue category. **Modified RNAs:** The total number of modified RNAs (genes) per SRA in each tissue category. **Modifications per RNA:** The RCM density per RNA per SRA in each broad tissue category. **(B)** Upset plot quantifying the shared modified RNAs across the broad tissue categories in **(A)**. **(C)** RCM-RNA abundance correlation distribution of all modified RNAs across the McCormick et al. tissue expression atlas [55] and the Varoquaux et al. [66] Sorghum drought experiment. The dashed red line shows the mean Pearson Correlation Coefficient of each experiment ($r = .34$ Tissue, $r = .18$ Drought). **(D)** Heatmap of RNA abundance and RCM changes in Varoquaux et al. re-analysis. Each row is a modified RNA that has been filtered for RNA abundance variability (see methods). Rows were clustered based on their abundance and RCM density values. The heatmap was grouped into four distinct clusters (C1–C4, n – values for each cluster shown) using k-means. SC = drought susceptible control, SD = drought susceptible treatment, TC = drought tolerant control, TD = drought tolerant treatment. At the time point analyzed, the drought susceptible genotype is RTx430 and the tolerant genotype is BTx642. **(E)** Enriched GO terms of the modified RNAs in each of the four clusters in **(D)**

We next aimed to characterize the tissue specificity of RCMs in the McCormick Sorghum tissue expression atlas. Using a modified calculation for Tau, a value typically used to calculate tissue specificity based on RNA abundance [57], we observed that most transcripts were modified in a very context-specific manner (Fig. S2B). The transcripts with the lowest Tau value, and therefore modified under the broadest context, were mostly associated with core cellular processes such as translation, whereas those transcripts modified in the most specific context were associated with mRNA maturation and abiotic stress responses (Fig. S2C, D). We also clustered transcripts based on the tissue type in which they were modified, the number of modifications identified, and the average number of modifications per transcript (Fig. 2A). In our analysis we observed a strong clustering based on tissue similarity, with seed and roots being notable exceptions. We find a strong bias for RCMs in this dataset towards root samples where both the number of RCMs and number of mRNA targets is substantially higher than other tissues (Fig. 2A and Fig. S2E). In contrast, seeds displayed the fewest number of modified transcripts, but the average number of modifications per transcript was very similar to that seen in leaf and root tissues where the number of modified sites and transcripts were much higher (Fig. 2A). We observed a large number of transcripts that are modified in all tissues ($n = 851$; Fig. 2B and Fig. S3 A), highlighting the existence of a core repertoire of RCM-targeted mRNAs. As expected, based on their presence across a broad tissue and developmental context, these mRNAs are enriched for terms associated with cytoplasmic mRNA translation (Fig. S3 B). However, the majority of transcripts were more restricted in terms of the tissue or developmental context under which they were modified and were enriched for more specialized GO terms. For instance, root-specific modified mRNAs were enriched for rhizosphere-associated terms [58] such as oxidation management, generic methylation, and aromatic compound biosynthesis (Fig. S3 C). Leaf-specific modified mRNAs were enriched for photosynthetic terms, whereas seed-specific mRNAs were enriched for lipid storage, ABA response, and cold acclimation terms

(Fig. S3 D, E). Importantly, the observed increase in context-specific modifications was not simply due to differences in the most abundant transcripts in each tissue. Indeed, we observed a low (although positive, $r = .34$) correlation between RNA abundance and modification levels (Fig. 2C; top) in the Sorghum tissue atlas. Thus, although there is a core set of modified Sorghum transcripts, most are targeted for RCMs in a context- or developmentally-specific manner.

The conservation of RCMs on orthologous genes in our grass seedling data suggest that the developmental and tissue contexts under which these marks are deposited might also be conserved. To address this contextual conservation, we examined RCMs in two publicly available *Arabidopsis thaliana* tissue atlases [54, 56]. Both atlases examined similar developmental stages, but did so under slightly different conditions (e.g., constant light and sterile MS media for Mergner et al. [54] vs. cycling light and soil for Klepikova et al. [56]). Importantly, while the Klepikova tissue atlas is primarily used by the community to examine transcript abundance in the Arabidopsis EFP browser [59], the work by Mergner et al. performed a paired assessment of transcript and protein abundance, which we used to examine the relationship between modifications and translation (see below). Due to differences in experimental design, we analyzed each of these atlases separately. A full breakdown of examined tissues, as well as total number of modifications identified, can be found in Additional File 2. Like Sorghum, clustering of Arabidopsis tissues by RCM density placed similar tissues together (Mergner et al.: Fig. S4A,B, Klepikova et al.: Fig. S5A, B). While Arabidopsis root tissues did not display a significantly elevated level of RCMs as in Sorghum, Arabidopsis seed transcripts had a reduced pool of very highly modified transcripts in both atlases. Similar to our observations in Sorghum, the seed-specific modified transcripts were enriched for lipid, nutrient, and ABA-response terms (Fig. S6A, B, C, D). Thus, as in Sorghum, the Arabidopsis epitranscriptome is diverse, highly context-specific, and appears to be associated with transcripts critical for cellular function.

Given the similar patterns of RCM abundance in Sorghum and Arabidopsis, we next examined whether orthologous mRNAs between Sorghum and Arabidopsis, which last shared a common ancestor ~150–250 million years ago [60], are targeted by RCMs. While we identified fewer mRNA targets of RCMs in both Arabidopsis expression atlases relative to the Sorghum expression atlas (Mergner: 1,324, Klepikova: 2,495 vs. 6,845 in McCormick Sorghum), the number of modified orthologous mRNAs (658 Mergner vs. McCormick; 1,180 Klepikova vs. McCormick) was significantly more than expected by chance for all possible mRNA combinations (Fig. S6E, F, $p = .001$ Hypergeometric test). We conclude that this class of RCMs target an ancient conserved repertoire of translation and photosynthetic related mRNAs.

Given the developmental differences in RCM deposition in both Sorghum and Arabidopsis, as well as reports associating RNA modifications with plant stress responses [9, 14, 61–65], we next determined if RCMs are associated with drought stress in Sorghum. We utilized a publicly available field-grown Sorghum transcriptome dataset that sampled well-watered and water-limited (i.e., drought treatment) Sorghum leaves and roots from drought tolerant and susceptible genotypes across weekly timepoints [66]. We focused on a post-flowering time point (week 10) where one genotype (BTx642) is considered drought tolerant and the other genotype (RTx430) is drought susceptible. Counterintuitively, we observed a shift towards a more negative correlation between RNA and RCM abundance during drought stress relative to the Sorghum tissue atlas (Fig. 2C, bottom). A heatmap comparing transcript and RCM abundance of the top 50% most variably expressed and modified root transcripts ($n=878$) between the two genotypes and treatments clustered into four groups (Fig. 2D). One of these clusters (Cluster 3) showed similar increases in transcript abundance levels between the two genotypes under water limiting conditions but showed an increase in RCMs specifically in the drought tolerant genotype (Fig. 2D). An examination of enriched GO terms revealed that Cluster 3 contained both heat shock proteins as well as water transport proteins (Fig. 2E), suggesting that RCMs may be associated with the drought response in the tolerant genotype.

RCMs accumulate near exon-exon junctions and are associated with splicing events

Thus far we have observed an association between RCMs and plant developmental and environmental transcriptional responses. To gain functional insight into RCMs, we analyzed their accumulation and distribution across mRNA topologies using our seedling RNA-seq data for all three species. RCMs were enriched across the CDS, with a bias towards the 3' CDS, and in the 3' UTR of

mRNAs (Fig. 3A). Like m⁶A [23], we also observed that RCMs are biased towards being deposited on abnormally long exons (median length of modified exons=361 nts, median unmodified=147 nts, median all exons=148 nts, Fig. 3B, $p < 2.2e-16$). Additionally, we observed a significantly higher proportion of expressed transcripts that are multi-exonic being targeted by RCMs, relative to mono-exonic transcripts (~11–14% vs ~5–8%, Fig. 3C, $p < 7.1e-15$). This was also the case for long non-coding RNAs (lncRNAs) where 4.8% of mono-exonic lncRNAs are marked by RCMs (17/355) and 8.4% of multi-exonic lncRNAs are targeted by RCMs (59/704, $p = .044$; Fig. S7A). A closer examination of distinguishing CDS features uncovered a dramatic buildup of RCMs on both 5' and 3' edges of exon-exon junctions (EJs) relative to start and stop codons (Fig. 3D). Thus, these data suggest that RCMs likely play a role in the regulation of RNA splicing on diverse transcript types.

This observation of RCMs preferentially occurring at EJs was initially made by Vandivier et al [14] on degrading mRNAs, whereas degrading mRNAs likely make up a small proportion of our dataset. Thus, these findings suggest a steady-state RCM enrichment at mRNA EJs. A transcriptome-wide 3' bias was observed for these EJ-enriched RCMs (Fig. 3E). This terminal EJ enrichment was not due to 3' sequencing bias that is often observed in poly-A enriched transcriptome datasets (Fig. S7B, C, D). Interestingly, genes that express multiple isoforms are more likely to be modified than single isoform transcripts with similar numbers of exons (Fig. 3F, $p < 2.2e-16$). However, we observe no significant difference in the buildup of modified sites at alternatively spliced junctions vs canonical splice sites (Fig. S8, see **Methods**). These data suggest that the increase in modification frequency at genes with more isoforms is likely due to the presence of more exon junctions. Thus, RCMs appear to be positively associated with splicing in plants.

RCMs are positively associated with stable and translating mRNAs

Because Mergner et al. [54] measured RNA and protein abundances from matched Arabidopsis samples they were able to correlate RNA: protein abundances across their samples. Therefore, we examined whether mRNAs marked with RCMs displayed a higher or lower RNA: protein correlation. A difference in RNA: protein correlation could suggest a RCM function in RNA stability and/or translation efficiency. mRNAs that are not marked with RCMs across the Mergner et al. atlas showed a lower median RNA: protein correlation ($n=3,361$, $r = .68$) compared to RCM marked mRNAs which showed a significantly higher RNA: protein correlation ($n=332$, $r > 0.758$, $p < .05$, Fig. S9A). Due to this positive correlation between transcripts harboring base-pair disrupting

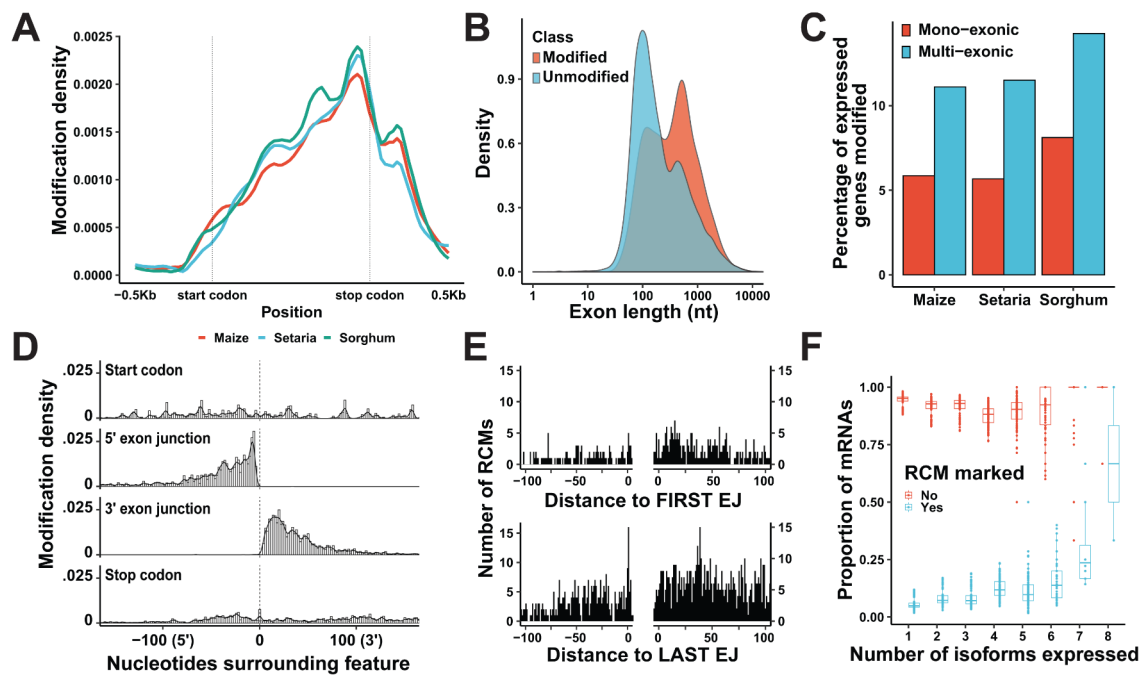


Fig. 3 RCMs accumulate at exon-exon junctions and are associated with splicing. **(A)** Metagen plot showing the genic distribution of grass RCMs identified in Fig. 1. Ten nucleotide windows are plotted where the signal is the sum of RCMs in that window and plotted as density. **(B)** Density distribution of exon lengths for RCM marked exons vs exons not marked by RCMs. (Student's t-test, $p < 2.2 \times 10^{-16}$). **(C)** Proportion of expressed transcripts that receive RCMs in grass seedlings between mono-exonic and multi-exonic transcripts. Multi-exonic transcripts are significantly more likely to be modified than mono-exonic (Chi-squared test, all p -values $< 7.1 \times 10^{-15}$). **(D)** Density of Sorghum seedling RCMs at start codons, stop codons, and 5'/3' exon-exon junctions. Density curves are plotted over a density histogram, both using a two-nucleotide window. The dotted line represents the first nucleotide of the start and stop codon, and the first nucleotide of the intron for 5'/3' exon junction panels. **(E)** Histogram quantifying the number of Sorghum seedling RCMs accumulating at terminal exon-exon junctions vs the first exon-exon junction. **(F)** Box Plot and dot plot overlaid quantifying the proportion of RNAs that are marked by RCMs or not in each SRAs in the McCormick et al Sorghum tissue expression atlas. The x-axis splits the data by the number of isoforms a gene is expressing. RCM proportion increases as the number of isoforms expressed increases ($p < 2.2 \times 10^{-16}$ one-way ANOVA)

modifications and their translation products, it is possible that these RCMs are positively influencing translation, either by reducing structural complexity or by stabilizing these transcripts.

To test for a relationship between RCMs and mRNA decay, we first examined publicly available transcriptomic data from *Arabidopsis* lines deficient in cytoplasmic mRNA decay pathways [67] by Sorenson and co-authors. Cytoplasmic mRNAs usually decay through three pathways: decapping (5' → 3'), the RNA exosome (3' → 5'), or an exosome independent 3' → 5' decay pathway. Decapping occurs through a multi-subunit complex that is scaffolded by VARICOSE in plants and metazoans (VCS, [68]). Meanwhile, the exosome independent 3' → 5' decay pathway occurs through SUPPRESSOR OF VARICOSE (SOV), which contributes to the decay of decapped RNAs and is also known as DIS3L2 in other Eukaryotes [69–71]. Sorenson et al took an RNA-seq approach to examine mRNA decay dynamics following transcriptional arrest (via cordycepin) in four *Arabidopsis* genotypes that vary in their cytoplasmic mRNA decay factors (wild-type, *sov* knockout, *vcs* knockout, and *sov/vcs* double knockout). If RCMs are primarily marking mRNAs for degradation (as

initially suggested by Vandivier et al [14]), then a buildup should be observed after transcriptional arrest in genetic backgrounds deficient for mRNA decay.

The majority of modified transcripts at the beginning and end of the time series were mRNAs (Fig. 4A). As expected from arresting transcription, each genotype, with the exception of the *sov/vcs* double mutant, displayed a ~25–50% decrease in the total number of observable protein-coding transcripts eight hours post-treatment (see **Methods**; Fig. S9B). The number of observed RCMs increased in all genotypes after arresting transcription, as did the proportion of modified mRNAs (Fig. 4B and Fig. S9C), suggesting two possibilities: 1) that modification abundance increases with transcript age, or 2) that non-modified transcripts are degraded more quickly leading to a higher proportion of transcripts detectable over time containing RCMs.

To more directly test the association of RCMs with RNA degradation, we analyzed whether the pool of transcripts that are modified at time point 0 were still detectable and modified at subsequent time points. Surprisingly, mRNAs modified at time point 0 were nearly all detectable 8 h after transcription arrest while

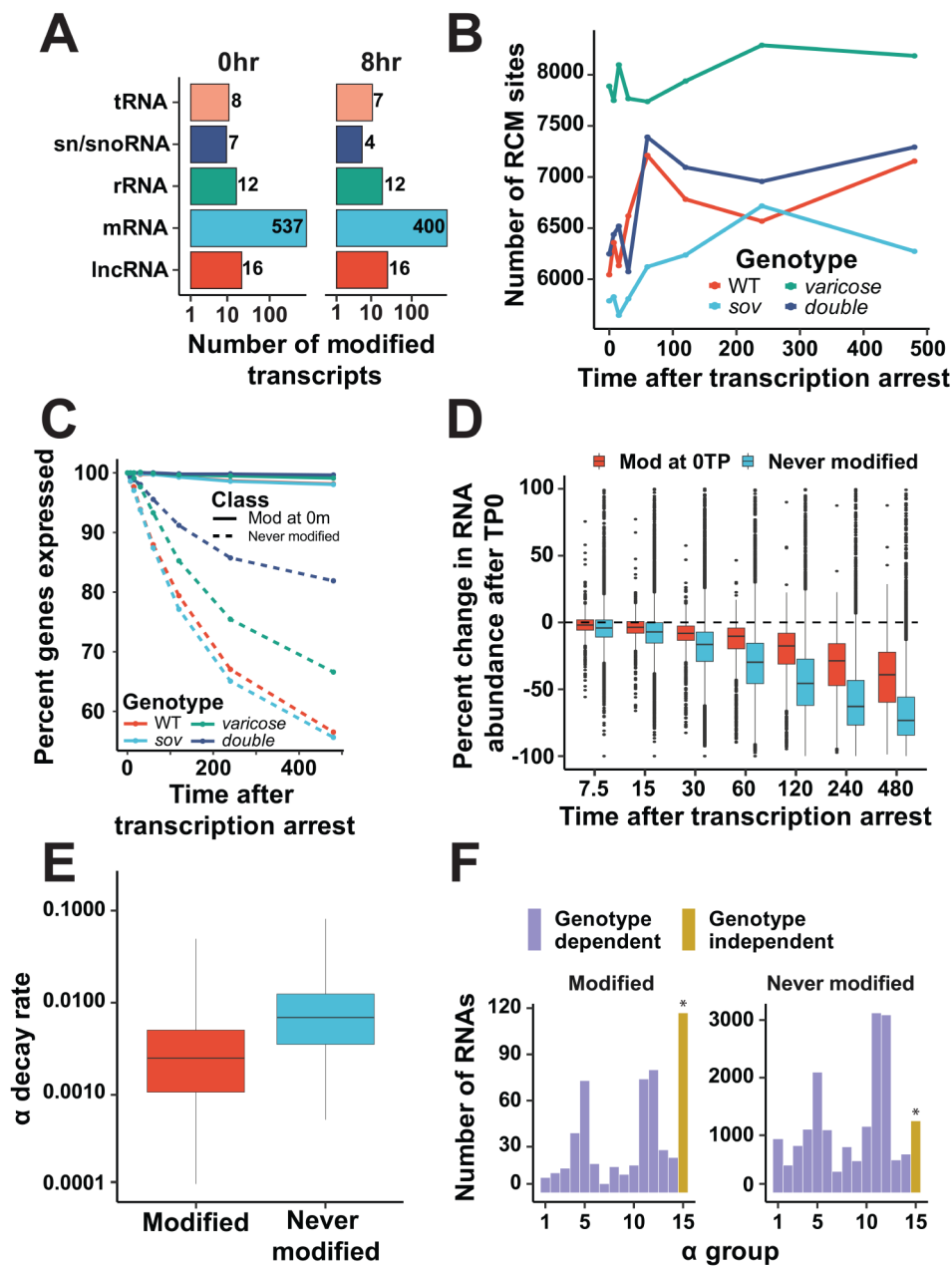


Fig. 4 RCMs mark mRNAs that degrade slowly. **(A)** Reanalysis of Sorenson et al. 2018 Arabidopsis dataset [67]. Distribution of Arabidopsis gene types that receive RCMs at time point 0 and 8 h after arresting transcription. tRNA = transfer RNA, sn/snoRNA = small nuclear/small nucleolar RNA, rRNA = ribosomal RNA, mRNA = messenger RNA, lncRNA = long noncoding RNA. **(B)** Number of Arabidopsis RCMs predicted at each time point after arresting transcription and in each RNA degradation genotype. **(C)** Change in number of detectable RNAs over time after arresting transcription. Solid lines represent mRNAs that have a RCM at time point 0 while dashed lines represent mRNAs that have no RCMs at any time point. Expressed genes are those that have a normalized expression value $>= 1$ (see methods). **(D)** Comparing RNA abundance change after transcription arrest at each time point. Red boxplots (left grouping) are mRNAs with an RCM at time point 0, blue boxplots (right group) are mRNAs that have no RCMs at any time point. Comparisons at time points 7, 15, and 30, $p > .05$, all time points starting at 60 min and afterwards, $p < .001$ by one-way ANOVA. **(E)** Boxplot showing the distribution of α -decay rates between RCM marked mRNAs at time point 0 and mRNAs that have no RCMs at any time point. **(F)** Modified figure from Sorenson et al. 2018 Fig. 1C showing the number of mRNAs that belong to different decay categories. Categories 1–14 are inconsequential for our conclusions. Group 15 represents genotype independent decay, (decay rates are not affected by the genotypes in the Sorenson et al. study). This category may represent targets of the exosome. * P-value between modified/never modified for Group 15 calculated by Chi-squared test

mRNAs that were not modified at time point 0 declined by more than 40% (Fig. 4C). These results indicate that RCMs mark mRNAs which degrade slower than the entire mRNA pool. Indeed, mRNAs that are not targeted by RCMs show a significantly larger magnitude of TPM decrease relative to mRNAs marked with RCMs at time point 0 (Fig. 4D). These results strongly suggest that RCM marked mRNAs degrade at a slower rate relative to mRNAs that are not marked by RCMs.

Sorenson et al. also modeled the initial decay rate of all mRNAs (alpha decay rate) based on read abundance after transcription arrest. Based on these values, mRNAs that are modified at time point 0 in wild-type show a significantly lower (slower) alpha decay rate relative to all other mRNAs (Fig. 4E). Sorenson et al. used the decay rates across genotypes to assign mRNAs to genotype-dependent RNA degradation pathways (Fig. 4F x-axis groups 1–14, see Fig. 1C from Sorenson et al.). For mRNAs that could not be assigned to the VARICOSE or SOV degradation pathways, Sorenson et al. hypothesized that they were likely targets of the RNA exosome (x-axis group 15; Fig. 4F). Interestingly, modified mRNAs are significantly biased towards being assigned to this group and thus are likely targets of the RNA exosome (Fig. 4F $p < 2.2e-16$, Pearson's Chi-squared test). While the importance of RCMs on rRNAs and tRNA stability and function has been known for decades, based on these data, we would argue that the HAMR/ModTect class of RCMs appear to be marks of Pol-II transcript stability, rather than marks of degradation.

RCMs are not associated with nonsense-mediated mRNA decay

To further investigate whether RCM marked mRNAs have an association with RNA degradation pathways, we next turned our focus to the Nonsense Mediated mRNA Decay (NMD) pathway. NMD is responsible for degrading aberrant mRNAs that have a premature termination codon. This is often recognized as a termination codon upstream of the exon junction complex (EJC, [72]). Given the accumulation of RCMs at exon-exon junctions and the recent report of RNA degradation intermediates accumulating near exon-exon junctions [73], we tested whether modified mRNAs were likely targets of NMD. SUPPRESSOR OF MORPHOLOGICAL DEFECTS ON GENITALIA7 (SMG7) is a critical component of early NMD signaling in most Eukaryotes [74]. Gloggnitzer and co-authors [75] performed RNA-seq in Arabidopsis with a loss-of-function *smg7* mutant in a genetic background avoiding the strong autoimmune response of knocking out NMD (*pad4*; [75]). We re-processed the RNA-seq data generated by Gloggnitzer et al. and identified RCMs from their data. The up-regulated mRNAs in the *smg7* genotype represent both direct and indirect targets

of NMD silencing. We observed strong statistical evidence that NMD targets are actually under-represented (that is, depleted) from mRNAs containing RCMs (5,275 mRNAs containing RCMs, 656 up-regulated mRNAs in *smg7*, overlap=103, $p < 2.2e-16$, Fig. S10). Furthermore, we identify no significant differences in RCM distribution at exon-exon junctions between *pad4* and *smg7-pad4* RCMs (Fig. S10B, $p = .986$), or between *smg7* RCMs in mRNAs up-regulated in *smg7* vs. those not differentially abundant in *smg7* (Fig. S10C, $p = .144$). In agreement with the *smg7* results which examined a single tissue, there is no significant overlap between predicted NMD targets and the RCMs identified in the Mergner et al. tissue atlas (Fig. S10D, $p = .304$). Collectively, these results suggest RCMs are not associated with the NMD pathway.

Discussion

In this study we used a bioinformatic workflow consisting of HAMR and ModTect [14, 15, 44] to predict RCMs in diverse species across the flowering plant lineage in order to clarify RCM distribution and putative function on mRNAs. Of note, because these algorithms rely on a certain minimum depth of expression in order to statistically call a transcript as being modified, and we tended towards a conservative definition of when to call an RCM (present at the same site in bio-replicates based on both algorithms), making conclusions about low abundance transcripts can be problematic. Despite these caveats, we observed a substantial number of transcripts whose modification status changes along developmental or environmental gradients. In perhaps the most extensive comparative analysis to date, we demonstrated that the RCMs detected by HAMR and ModTect (e.g., those that fall along the WCF base-pairing edge) are found on a large, yet contextually discrete set of transcripts. From these diverse analyses we believe that a number of conclusions can be made that provide lessons as to the function of RCMs in eukaryotes.

RCMs are not found on all expressed transcripts, nor are they always on the most abundant transcripts, suggesting a contextual specificity. Indeed, there appear to be two classes of transcripts that receive RCMs in plants. There are a core set of transcripts associated with photosynthesis and cellular metabolism that are modified in each tested tissue or developmental context. These transcripts tend to be abundantly modified and are the targets of all of the major base-pair disrupting modifications. Due to their critical functions, these genes are conserved, as is their modification status, across the plant lineage. The other class of modified transcripts receive RCMs in a more context-specific manner. These transcripts may be expressed in multiple tissues but are generally targeted by modifications in a more limited subset of conditions. Functional enrichment suggests that these

transcripts are modified in response to environmental or developmental cues.

We observed similar patterns of RCM deposition on all transcripts, with a bias towards longer exons and transcripts containing multiple exons. In contrast to m⁶A, these RCMs are enriched at exon splice junctions. Interestingly, while we see an increase in RCM abundance on transcripts with multiple isoforms, we see no association between RCM status and alternative splicing or intron retention, suggesting that modifications are more of a general marker of splicing rather than a driver of isoform selection. We would propose that these modifications are deposited on the base-pairing edge of the newly transcribed mRNA to reduce structure at exon junctions in order to facilitate splicing, whereas m⁶A is occluded from these sites due to its interference with binding of the splicing machinery [33, 34]. This model would further suggest that, like m⁶A, these other RCMs are deposited co-transcriptionally and thus their writers should appear in complexes with transcriptional and epigenetic machinery. In support of this hypothesis, the m⁵C methyltransferases NSUN3 and DNMT2 are known to interact with hnRNPk for Pol-II recruitment and modulating chromatin state in response to stress [76].

Previous work in *Arabidopsis* comparing RCM abundance in degrading and total mRNA populations found evidence for RCMs as a marker for mRNA degradation [14] whereas others have suggested that loss of RCMs leads to unstable mRNAs [10]. In an attempt to resolve these conflicting models, we turned to an elegant time-series RNA-seq experiment in *Arabidopsis* examining the impacts of mutants in the mRNA decay pathway on RNA turnover [67]. We used these data to assess the relationship between degradation and RCM status. To our surprise, we found that RCMs are more likely to mark extremely stable mRNAs, rather than those that are rapidly degrading. Our finding that most of the RCMs fall on transcripts associated with core cellular processes fits with the notion that these transcripts would also be more stable, a finding reported by Sorenson et al. and others [77]. Interestingly, we also observe a positive relationship between RCM abundance and protein abundance in the *Arabidopsis* tissue atlas, further suggesting that RCMs, as a whole, have a positive influence on mRNA stability. Importantly, we cannot definitively say that degrading RNAs are not marked, just that the RCMs we can monitor are enriched on mRNAs with long half-lives. A key difference between our work and that of Vandivier et al. [14] is in the depth of sequencing, both for the grass seedling experiments that we performed and for the additional datasets we reprocessed, and so further examination of the connection between these base-pair disrupting RCMs and RNA degradation is warranted.

Conclusions

In this work, we used a comparative transcriptomic approach to better understand the conservation and function of a poorly studied class of RNA modifications. Grass seedlings contained similar cohorts of modified transcripts, a finding that spurred us to look in other datasets that incorporated additional stress or developmental timepoints in both *Sorghum* and *Arabidopsis*. Again, we observed a conserved, core set of modified transcripts, but these expanded datapoints allowed us to uncover context specificity to the epitranscriptome, suggesting that some, but not all, transcripts are modified at key steps in their functional lifecycle. Attempts to better understand function of these base pair disrupting modifications revealed that they are likely important for splicing, translation, and intriguingly, for RNA stability. Similarities across the Angiosperm lineage, along with supporting literature in metazoans, suggests that these modifications, like m⁶A, are a deeply conserved aspect of RNA biology for which we still know very little.

Methods

Plant growth

Seed from *Zea mays* (acc. B73), *Sorghum bicolor* (acc. BTx642), and *Setaria italica* (acc. Yugu1) were sown onto damp soil approximately 2 cm below the surface, and stratified for one week in the dark at 4 °C. Seeds were then transferred to a growth cabinet with lights configured to a long-day cycle (16 h light, 8 h dark). All species' seeds germinated within a day of each other and were allowed to grow for two weeks once germinated. Seedling shoot tissue was collected and instantly frozen with liquid nitrogen.

RNA-sequencing library preparation

RNA from two biological replicates in each species was isolated using TriZol per the manufacturer's instructions (ThermoFisher #15596026). Paired-end and strand specific sequencing libraries were generated for each species using the YourSeq FT Strand-Specific mRNA library prep kit (Active Motif #23001). Libraries were sequenced on a Novogene Novaseq platform.

Read processing and modification identification

Raw sequencing reads were trimmed for adapters and low-quality bases using fastp (v 0.23.4, [78]) with default settings for paired end reads. Trimmed reads were then used to call HAMR modifications using the PAMLINC workflow (<https://github.com/chosenobih/pamlinc>). Briefly, this workflow maps reads to each species' reference genome (obtained from Ensembl Plants in January 2022) using BowTie2 (v 2.2.5, [79]) and retains unique alignments that don't overlap exon-junctions or deletions.

The HAMR statistical model is then run on these filtered alignment files (BAM files) to predict RCMs.

In addition, a similar algorithm, ModTect (v 1.7.5, [15]), was used to call RCMs. For ModTect prediction, trimmed reads were mapped to the same Ensembl Plants genomes and associated annotation files as above using STAR (v 2.7.10b, [80]). The following non-default STAR arguments were used for mapping: `--outFilterIntronMotifs RemoveNoncanonical`, `--alignIntronMax 10000`, `--outSAMstrandField intronMotif`. Raw BAM outputs were then used as input for ModTect with the following arguments: `--minBaseQual 30`, `--readlength 150`, and `--regionFile genes.bed` which is a four-column file of all gene regions in the annotation file.

For sequencing data generated in this study, RCM sites were retained that overlapped both ModTect and HAMR predictions among both sequencing replicates (Fig. 1B). For public sequencing data, RCMs were identified using only ModTect and RCMs were retained that were common among three or more experiments (i.e., three or more distinct SRAs).

RNA abundance measurement

Trimmed fastq files were used to measure RNA abundance using Salmon (v 1.10.0, [81]) in “quant” mode against a decoy aware transcriptome index. The tximport (v 1.28, [82]) R package was then used to import Salmon quant files and generate gene or transcript level transcript per million values, specifying the following tximport option: `countsFromAbundance = “lengthScaledTPM”`. RNAs used for downstream RCM and expression analyses were those with at least one transcript per million in one experiment. For the reanalysis of the Varoquaux et al., dataset in Fig. 2D, root tissue was chosen for comparison. Transcripts whose RNA abundance did not vary in the root control vs. drought treatments were removed by filtering out the bottom 50% of transcripts based on their abundance variability measured by median absolute deviation. The frequency of RCM density for every transcript (number of RCMs in a condition / total number of RCMs across all root samples) was computed and then transcripts without RCMs in any dataset was removed. For the reanalysis of Sorenson et al. 2018, we used RPKM (reads per kilobase per million mapped) values that the authors provide in their supplemental data which accounts for stable transcripts containing an increasing proportion of sequencing reads.

Alternative splicing

Two approaches were used to infer alternative splicing. First, genes that express multiple isoforms can be inferred from Salmon [81] using the tximport [82] R package as described above. Specifying the tximport “txOut” argument gives isoform level quantifications. Isoforms expressed above one TPM were retained. Genes

expressing more than one isoform were inferred to be undergoing alternative splicing. Second, we examined statistically significant changes in splicing patterns between experimental conditions using EventPointer [83].

GO term enrichment

GO terms were downloaded for each species from Ensembl Plants. GO term enrichment was performed using the clusterProfiler (v 4.8.2, [84]) R package, specifically the “enricher” function was used, specifying: (1) the input gene set, (2) a p-value cutoff of 0.05, (3) Benjamini and Hochberg multiple testing correction, (4) a background gene set (usually, all expressed genes (TPM >= 1)), (5) a minimum and maximum gene-set size of 10 and 500, respectively, and (6) a q-value cutoff of 0.05. Fold enrichment was calculated using the following functions: `parse_ratio(GeneRatio) / parse_ratio(BgRatio)`.

Data analyses

All analyses and statistical testing was performed using the R programming language (v 4.2.0, [85]). Tests for over or underrepresentation were performed using the phyper function (i.e., a hypergeometric test) in R. Correlations between RNA abundance and RCMs were performed by collecting the RNA abundance values and RCM density values (number of RCMs per transcript) from each sequencing experiment of interest. Pearson Correlation Coefficients were then calculated from the RNA abundance and RCM density using the “cor” function in R. To measure statistical significance of mean-separation between groups, the Student’s t-test was used for two group comparisons, while one-way Anova was used for more-than-two group comparisons.

All visualizations were generated in R from various packages including: ggplot2 (v 3.4.4, [86]), ggvenn (v 0.1.10, [87]), enrichPlot (v 1.20.0, [88]), complexUpset (v 1.3.3, [89]), complexHeatmap (v 2.16., [90]), and the ggtree package (v 3.8.2, [91]).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-024-05486-7>.

Supplementary Material 1: Additional File 1: Overview of RCMs identified in all three grass seedlings.

Supplementary Material 2

Supplementary Material 3: Additional File 2: Summary of RNA-seq datasets used and the number of RCMs identified across the tissue expression atlases.

Supplementary Material 4

Supplementary Material 5

Acknowledgements

We would like to acknowledge support from NSF IOS PGRP 2023310 (to ADLN, EHL, and BDG), NSF MCB 2120131 (to ADLN), and NSF-DBI-1743442 to EHL. We

would like to thank Drs Duke Pauli (University of Arizona) and Susan Schroeder (Oklahoma University) for insightful comments during the development of the project. We would like to thank Drs. Rebecca Murphy (Centenary College of Louisiana) and Daryl Morishige for providing useful insight into the Sorghum tissue atlas. Finally, we would like to thank members of the Gregory and Nelson labs for constructive feedback.

Author contributions

KP, ACND, EHL, BDG, and ADLN developed the project. ACND grew plants, isolated RNA, and performed RNA-seq. KP analyzed data, EHL, BDG, and ADLN provided feedback on data analysis. KP and ADLN wrote the manuscript. All authors edited and approved the manuscript.

Funding

We would like to acknowledge support from NSF IOS PGRP 2023310 (to ADLN, EHL, and BDG), NSF MCB 2120131 (to ADLN), and NSF-DBI-1743442 to EHL.

Data availability

The accession numbers of RNA-sequencing data generated in this study can be found at <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1116564>. The scripts and intermediate files for analyses are available at: <https://github.com/kylepalos/comparative-PRM-paper>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 23 May 2024 / Accepted: 5 August 2024

Published online: 12 August 2024

References

- Cappannini A, Ray A, Purta E, Mukherjee S, Boccaletto P, Moafinejad SN, et al. MODOMICS: a database of RNA modifications and related information. 2023 update. *Nucleic Acids Res.* 2024;52:D239–44.
- Baumer ZT, Erber L, Jolley E, Lawrence S, Lin C, Murakami S, et al. Defining the commonalities between post-transcriptional and post-translational modification communities. *Trends Biochem Sci.* 2024;49:185–8.
- Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE. The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* 2012;13:175.
- Gilbert WV, Nachtergaele S. mRNA regulation by RNA modifications. *Annu Rev Biochem.* 2023;92:175–98.
- Lewis CJT, Pan T, Kalsotra A. RNA modifications and structures cooperate to guide RNA-protein interactions. *Nat Rev Mol Cell Biol.* 2017;18:202–10.
- Roy B. Effects of mRNA modifications on translation: an overview. In: McMahon M, editor. *RNA modifications: methods and protocols*. New York, NY: Springer US; 2021. pp. 327–56.
- Prall W, Ganguly DR, Gregory BD. The covalent nucleotide modifications within plant mRNAs: what we know, how we find them, and what should be done in the future. *Plant Cell.* 2023;35:1801–16.
- Sharma B, Prall W, Bhatia G, Gregory BD. The diversity and functions of plant RNA modifications: what we know and where we go from Here. *Annu Rev Plant Biol.* 2023;74:53–85.
- Kramer MC, Janssen KA, Palos K, Nelson ADL, Vandivier LE, Garcia BA, et al. N6-methyladenosine and RNA secondary structure affect transcript stability and protein abundance during systemic salt stress in *Arabidopsis*. *Plant Direct.* 2020;4:e00239.
- Sun H, Li K, Liu C, Yi C. Regulation and functions of non-m6A mRNA modifications. *Nat Rev Mol Cell Biol.* 2023;24:714–31.
- Luo G-Z, MacQueen A, Zheng G, Duan H, Dore LC, Lu Z, et al. Unique features of the m6A methylome in *Arabidopsis thaliana*. *Nat Commun.* 2014;5:5630.
- Perry RP, Kelley DE, Friderici K, Rottman F. The methylated constituents of L cell messenger RNA: evidence for an unusual cluster at the 5' terminus. *Cell.* 1975;4:387–94.
- Wiener D, Schwartz S. The epitranscriptome beyond m6A. *Nat Rev Genet.* 2021;22:119–31.
- Vandivier LE, Campos R, Kuksa PP, Silverman IM, Wang L-S, Gregory BD. Chemical modifications mark alternatively spliced and uncapped Messenger RNAs in *Arabidopsis*. *Plant Cell.* 2015;27:3024–37.
- Tan K-T, Ding L-W, Wu C-S, Tenen DG, Yang H. Repurposing RNA sequencing for discovery of RNA modifications in clinical cohorts. *Sci Adv.* 2021;7.
- Kannan K, Nelson ADL, Shippen DE. Dyskerin is a component of the *Arabidopsis* telomerase RNP required for telomere maintenance. *Mol Cell Biol.* 2008;28:2332–41.
- Sun L, Xu Y, Bai X, Bai X, Zhu H, Dong H, et al. Transcriptome-wide analysis of pseudouridylation of mRNA and non-coding RNAs in *Arabidopsis*. *J Exp Bot.* 2019;70:5089–600.
- Rintala-Dempsey AC, Kothe U. Eukaryotic stand-alone pseudouridine synthases - RNA modifying enzymes and emerging regulators of gene expression? *RNA Biol.* 2017;14:1185–96.
- Zhang M, Zhang X, Ma Y, Yi C. New directions for Ψ and m1A decoding in mRNA: deciphering the stoichiometry and function. *RNA.* 2024;30:537–47.
- Yang Y, Hsu PJ, Chen Y-S, Yang Y-G. Dynamic transcriptomic m6A decoration: writers, erasers, readers and functions in RNA metabolism. *Cell Res.* 2018;28:616–24.
- Flamand MN, Tegowski M, Meyer KD. The proteins of mRNA modification: writers, readers, and Erasers. *Annu Rev Biochem.* 2023;92:145–73.
- Wilkinson E, Cui Y-H, He Y-Y. Roles of RNA modifications in Diverse Cellular functions. *Front Cell Dev Biol.* 2022;10:828683.
- Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature.* 2012;485:201–6.
- Li M, Tao Z, Zhao Y, Li L, Zheng J, Li Z, et al. 5-methylcytosine RNA methyltransferases and their potential roles in cancer. *J Transl Med.* 2022;20:214.
- Chen X, Li A, Sun B-F, Yang Y, Han Y-N, Yuan X, et al. 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. *Nat Cell Biol.* 2019;21:978–90.
- Barbieri I, Kouzarides T. Role of RNA modifications in cancer. *Nat Rev Cancer.* 2020;20:303–22.
- Zhong S, Li H, Bodi Z, Button J, Vespa L, Herzog M, et al. MTA is an *Arabidopsis* messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell.* 2008;20:1278–88.
- Yang W, Meng J, Liu J, Ding B, Tan T, Wei Q, et al. The N1-Methyladenosine methylome of *Petunia* mRNA. *Plant Physiol.* 2020;183:1710–24.
- Cui X, Liang Z, Shen L, Zhang Q, Bao S, Geng Y, et al. 5-Methylcytosine RNA methylation in *Arabidopsis thaliana*. *Mol Plant.* 2017;10:1387–99.
- David R, Burgess A, Parker B, Li J, Pulsford K, Sibbritt T, et al. Transcriptome-wide mapping of RNA 5-Methylcytosine in *Arabidopsis* mRNAs and noncoding RNAs. *Plant Cell.* 2017;29:445–60.
- Decatur WA, Fournier MJ. rRNA modifications and ribosome function. *Trends Biochem Sci.* 2002;27:344–51.
- Zhou KI, Shi H, Lyu R, Wyllder AC, Matuszek Z, Pan JN, et al. Regulation of co-transcriptional Pre-mRNA splicing by m6A through the low-complexity protein hnRNPG. *Mol Cell.* 2019;76:70–e819.
- Uzonyi A, Dierks D, Nir R, Kwon OS, Toth U, Barbosa I, et al. Exclusion of m6A from splice-site proximal regions by the exon junction complex dictates m6A topologies and mRNA stability. *Mol Cell.* 2023;83:237–e517.
- Yang X, Triboulet R, Liu Q, Sendinc E, Gregory RI. Exon junction complex shapes the m6A epitranscriptome. *Nat Commun.* 2022;13:7904.
- Karikó K, Muramatsu H, Welsh FA, Ludwig J, Kato H, Akira S, et al. Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol Ther.* 2008;16:1833–40.
- Eyler DE, Franco MK, Batool Z, Wu MZ, Dubuke ML, Dobosz-Bartoszek M, et al. Pseudouridylation of mRNA coding sequences alters translation. *Proc Natl Acad Sci U S A.* 2019;116:23068–74.
- Zhang Q, Kim N-K, Feigon J. Architecture of human telomerase RNA. *Proc Natl Acad Sci U S A.* 2011;108:20325–32.
- Kim N-K, Theimer CA, Mitchell JR, Collins K, Feigon J. Effect of pseudouridylation on the structure and activity of the catalytically essential P6.1 hairpin in human telomerase RNA. *Nucleic Acids Res.* 2010;38:6746–56.
- Anderson BR, Muramatsu H, Jha BK, Silverman RH, Weissman D, Karikó K. Nucleoside modifications in RNA limit activation of 2'-5'-oligoadenylate

- synthetase and increase resistance to cleavage by RNase L. *Nucleic Acids Res.* 2011;39:9329–38.
40. Yang Y, Wang L, Han X, Yang W-L, Zhang M, Ma H-L, et al. RNA 5-Methylcytosine facilitates the maternal-to-zygotic transition by preventing maternal mRNA decay. *Mol Cell.* 2019;75:1188–e20211.
41. Zhang Y, Lu L, Li X. Detection technologies for RNA modifications. *Exp Mol Med.* 2022;54:1601–16.
42. Schaefer M, Pollex T, Hanna K, Lyko F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.* 2009;37:e12.
43. Woodson SA, Muller JG, Burrows CJ, Rokita SE. A primer extension assay for modification of guanine by Ni(II) complexes. *Nucleic Acids Res.* 1993;21:5524–5.
44. Ryvkin P, Leung YY, Silverman IM, Childress M, Valladares O, Dragomir I, et al. HAMR: high-throughput annotation of modified ribonucleotides. *RNA.* 2013;19:1684–92.
45. Ebhardt HA, Tsang HH, Dai DC, Liu Y, Bostan B, Fahlman RP. Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.* 2009;37:2461–70.
46. Findeiss S, Langenberger D, Stadler PF, Hoffmann S. Traces of post-transcriptional RNA modifications in deep sequencing data. *Biol Chem.* 2011;392:305–13.
47. Motorin Y, Muller S, Behm-Ansmant I, Branlant C. Identification of modified residues in RNAs by reverse transcription-based methods. *Methods Enzymol.* 2007;425:21–53.
48. Todkari IA, Chandrasekaran AR, Punnoose JA, Mao S, Haruehanroengra P, Beckles C, et al. Resolving altered base-pairing of RNA modifications with DNA nanoswitches. *Nucleic Acids Res.* 2023;51:11291–7.
49. Chiu NHL, Simpson JH, Wang H, Tannous BA. A theoretical perspective of the physical properties of different RNA modifications with respect to RNA duplexes. *BBA Adv.* 2021;1:100025.
50. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The sequence read Archive: a decade more of explosive growth. *Nucleic Acids Res.* 2022;50:D387–90.
51. Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat Rev Genet.* 2017;18:275–91.
52. Hauenschild R, Tserovski L, Schmid K, Thüring K, Winz M-L, Sharma S, et al. The reverse transcription signature of N1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res.* 2015;43:9950–64.
53. Wang Y, Li S, Zhao Y, You C, Le B, Gong Z, et al. NAD⁺-capped RNAs are widespread in the *Arabidopsis* transcriptome and can probably be translated. *Proc Natl Acad Sci U S A.* 2019;116:12094–102.
54. Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A, et al. Mass-spectrometry-based draft of the *Arabidopsis* proteome. *Nature.* 2020;579:409–14.
55. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 2018;93:338–54.
56. Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J.* 2016;88:1058–70.
57. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform.* 2017;18:205–14.
58. Matilla MA, Espinosa-Urgel M, Rodríguez-Herva JJ, Ramos JL, Ramos-González MI. Genomic analysis reveals the major driving forces of bacterial life in the rhizosphere. *Genome Biol.* 2007;8:R179.
59. Sullivan A, Purohit PK, Freese NH, Pasha A, Esteban E, Waese J, et al. An eFP-Seq browser for visualizing and exploring RNA sequencing data. *Plant J.* 2019;100:641–54.
60. Murat F, Armero A, Pont C, Klopp C, Salse J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet.* 2017;49:490–6.
61. Sharma B, Govindan G, Li Y, Sunkar R, Gregory BD. RNA N6-Methyladenosine affects Copper-Induced oxidative stress response in *Arabidopsis thaliana*. *Noncoding RNA.* 2024;10.
62. Prall W, Sheikh AH, Bazin J, Almeida-Trapp M, Crespi M, et al. Pathogen-induced m6A dynamics affect plant immunity. *Plant Cell.* 2023;35:4155–72.
63. Govindan G, Sharma B, Li Y-F, Armstrong CD, Merum P, Rohila JS, et al. mRNA N6-methyladenosine is critical for cold tolerance in *Arabidopsis*. *Plant J.* 2022;111:1052–68.
64. Cheng Q, Wang P, Wu G, Wang Y, Tan J, Li C, et al. Coordination of m6A mRNA methylation and gene transcriptome in rice response to cadmium stress. *Rice.* 2021;14:62.
65. Hu J, Cai J, Park SJ, Lee K, Li Y, Chen Y, et al. N6-Methyladenosine mRNA methylation is important for salt stress tolerance in *Arabidopsis*. *Plant J.* 2021;106:1759–75.
66. Varoquaux N, Cole B, Gao C, Pierroz G, Baker CR, Patel D, et al. Transcriptomic analysis of field-droughted sorghum from seedling to maturity reveals biotic and metabolic responses. *Proc Natl Acad Sci U S A.* 2019;116:27124–32.
67. Sorenson RS, Deshotel MJ, Johnson K, Adler FR, Sieburth LE. *Arabidopsis* mRNA decay landscape arises from specialized RNA decay substrates, decapping-mediated feedback, and redundancy. *Proc Natl Acad Sci U S A.* 2018;115:E1485–94.
68. Deyholos MK, Cavaness GF, Hall B, King E, Punwani J, Van Norman J, et al. VARICOSE, a WD-domain protein, is required for leaf blade development. *Development.* 2003;130:6577–88.
69. Zhang W, Murphy C, Sieburth LE. Conserved RNaseII domain protein functions in cytoplasmic mRNA decay and suppresses *Arabidopsis* decapping mutant phenotypes. *Proc Natl Acad Sci U S A.* 2010;107:15981–5.
70. Lubas M, Damgaard CK, Tomecki R, Cysewski D, Jensen TH, Dziembowski A. Exonuclease hDIS3L2 specifies an exosome-independent 3'-5' degradation pathway of human cytoplasmic mRNA. *EMBO J.* 2013;32:1855–68.
71. Malecki M, Viegas SC, Carneiro T, Golik P, Dresseira C, Ferreira MG, et al. The exonuclease Dis3L2 defines a novel eukaryotic RNA degradation pathway. *EMBO J.* 2013;32:1842–54.
72. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.* 2016;44:1483–95.
73. Lee W-C, Hou B-H, Hou C-Y, Tsao S-M, Kao P, Chen H-M. Widespread exon Junction Complex footprints in the RNA degradome mark mRNA degradation before steady state translation. *Plant Cell.* 2020;32:904–22.
74. Unterholzner L, Izaurralde E. SMG7 acts as a molecular link between mRNA surveillance and mRNA decay. *Mol Cell.* 2004;16:587–96.
75. Gloggnitzer J, Akimcheva S, Srinivasan A, Kusenda B, Riehs N, Stampfl H, et al. Nonsense-mediated mRNA decay modulates immune receptor levels to regulate plant antibacterial defense. *Cell Host Microbe.* 2014;16:376–90.
76. Cheng JX, Chen L, Li Y, Cloe A, Yue M, Wei J, et al. RNA cytosine methylation and methyltransferases mediate chromatin organization and 5-azacytidine response and resistance in leukaemia. *Nat Commun.* 2018;9:1163.
77. Li Y, Yi Y, Lv J, Gao X, Yu Y, Babu SS, et al. Low RNA stability signifies increased post-transcriptional regulation of cell identity genes. *Nucleic Acids Res.* 2023;51:6020–38.
78. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018;34:i884–90.
79. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9:357–9.
80. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21.
81. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017;14:417–9.
82. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* 2015;4:1521.
83. Ferrer-Bonsoms JA, Gimeno M, Olaverri D, Sacristan P, Lobato C, Castilla C, et al. EventPointer 3.0: flexible and accurate splicing analysis that includes studying the differential usage of protein-domains. *NAR Genom Bioinform.* 2022;4:lqac067.
84. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16:284–7.
85. Team RC. Others. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>. 2016.
86. Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2016.
87. Yan L. ggvenn. Draw Venn Diagram by ggplot2. R Package Version. 2021;19.
88. Yu G. Enrichplot: visualization of functional enrichment result. R package version. 2021;1.
89. Krassowski M. ComplexUpset. ComplexUpset. 2020.
90. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.* 2016;32:2847–9.
91. Yu G. Using ggtree to visualize data on Tree-Like structures. *Curr Protoc Bioinf.* 2020;69:e96.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.