

Supporting Information

Sorenson et al. 10.1073/pnas.1712312115

SI Materials and Methods

RNA Extraction, Library Prep, Data Processing, and Analysis. Total RNA was isolated from 50 to 60 seedlings (~50 mg fresh weight) using the Omega Bio-Tek E.Z.N.A. Plant RNA Mini Kit, including on-column DNase I digestion from four independent experiments per genotype, and subjected to ribosomal RNA depletion. RNA was submitted to the University of Utah Genomics core for further processing. Because we expect mRNA decapping mutants to accumulate capped deadenylated mRNAs, rRNA was removed using the Epicentre RiboZero Plant Leaf kit before library construction using the Illumina TruSeq Stranded Library Preparation kit. The samples were multiplexed so that one time course replicate of one genotype—a total of eight samples—were loaded onto a single lane. Raw and processed data for each gene are available at the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>), accession GSE86361.

The resulting sequences were aligned and processed using samtools version 0.1.19-44428cd, Bowtie2 version 2.2.3, TopHat2 v2.0.12, HTSeq. 0.6.1, and R version 3.1.0. FastQC was run on all 128 sequenced libraries to verify read quality. Sequence reads were aligned to the genome [The Arabidopsis Information Resource, release 10 (TAIR10)] using the following command in TopHat2: *tophat2-b2-very-sensitive-no-novel-juncs-max-multihits 1-library-type fr-firststrand*. Generated binary sequence alignment/map format (BAM) files were indexed and sorted using samtools before counting reads using HTSeq: *htseq-count -t gene -I ID-stranded = reverse*. Read counts were normalized to library size, and library size was scaled to 1 million [reads per million (RPM)]. For gene to gene comparisons, read counts were further normalized to transcript (representative splice variant for each gene) length and scaled to 1,000 nt [reads per thousand per million (RPKM)]. Further analyses were performed with in-house scripts using R and Bioconductor packages. Venn diagrams were generated using eulerAPE (1).

Decay Profile Normalization. To calculate transcript specific decay rates genome-wide, transcript read abundances (RPMs) were further normalized to (i) their respective mean T_0 values by genotype and (ii) a sample decay factor (genotype mean for each time point). Normalization to T_0 values converts values to fractional decreases and facilitates comparisons of RNA decay profiles. Decay factor normalization is necessary for the following reasons: As a majority of transcripts decay, the total pool of RNA decreases, and stable transcripts become a larger proportion of the library. This causes their RPM values to increase relative to T_0 even though their cellular levels may change very little or not at all. To adjust for this, the library values were scaled based on the assumption that the mean apparent fold increase of 30 stable reference genes reflects the total cellular decrease in the RNA pool. Reference genes were manually selected from the 500 most highly expressed genes based on their T_0 normalized decay profile, and include nuclear and organellar transcripts known to be stable: ATCG00490, ATCG00680, ATMG00280, ATCG00580, ATCG00140, AT4G38970, AT2G07671, ATCG00570, ATMG00730, AT2G07727, AT2G07687, ATMG00160, AT3G11630, ATMG00060, ATCG00600, ATMG00220, ATMG01170, ATMG00410, AT1G78900, AT3G55440, ATMG01320, AT2G21170, AT5G08670, AT5G53300, ATMG00070, AT1G26630, AT5G48300, AT2G33040, AT5G08690, and AT1G57720. T_0 and decay factor normalized decay profiles were used for modeling decay rates.

Modeling of mRNA Decay and Genotype Effects. We model the change in RNA concentration with the differential equation

$$\frac{dc}{dt} = -A(t)c, \quad [\text{S1}]$$

where $c(t)$ is the abundance of RNA at time t and $A(t)$ is the decay rate of the RNA. We assume that the RNA abundance is strictly decreasing because RNA synthesis is blocked by treatment with cordycepin, and set $c(0) = 1$ because the data were normalized to the initial time point for each transcript.

The solution of Eq. S1 is

$$c(t) = e^{-\int A(t)dt}.$$

If the decay rate is constant, $A(t) = \alpha$, then $c(t)$ follows the exponential decay model

$$c(t) = e^{-\alpha t}. \quad [\text{S2}]$$

Because many transcripts asymptote to a nonzero value, we also considered a decreasing decay rate of the form $A(t) = \alpha e^{-\beta t}$. This defines the decaying decay model

$$c(t) = e^{-\frac{\alpha}{\beta}(1-e^{-\beta t})}. \quad [\text{S3}]$$

We assume that the parameters α and β for a given transcript could be different for each genotype but are the same for each replicate. The α and β parameters for each genotype are defined as follows: $\alpha_1 = \alpha_{\text{WT}}$ is the decay rate of WT genotype, $\alpha_2 = \alpha_{\text{sov}}$ is the decay rate of *sov* genotype, $\alpha_3 = \alpha_{\text{vcs}}$ is the decay rate of *vcs* genotype, and $\alpha_4 = \alpha_{\text{vcs.sov}}$ is the decay rate of *vcs sov* genotype. Similar labeling is applied to the β parameters. These values can vary independently, or different combinations can be constrained to be equal (Fig. S2A), creating a total of 240 models. For example, model 165 has α group 11 and β group 5, for which $\alpha_{\text{WT}} = \alpha_{\text{sov}} = a_1$, $\alpha_{\text{vcs}} = a_2$, $\alpha_{\text{vcs.sov}} = a_3$, $\beta_{\text{WT}} = \beta_{\text{sov}} = \beta_{\text{vcs}} = b_1$, and $\beta_{\text{vcs.sov}} = b_2$.

We index time with i , replicates with j , genotype with k , and models with l . The RNA abundance $m_{jk}(t_i)$ was measured at eight time points, $t_i \in (0, 7.5, 15, 30, 60, 120, 240, 480)$, where time is measured in minutes. With four genotypes, four replicates, and eight time points, there are $n = 128$ observations for each transcript.

We assume that errors around each model are normally distributed with mean 0 and variance σ_l^2 , $\varepsilon_l \approx N(0, \sigma_l^2)$, estimated separately for each model. The data were fit to the exponential decay model $[c(t_i; \alpha_k) = e^{-\alpha_k t_i} + \varepsilon_l]$ and the decaying decay model $[c(t_i; \alpha_k, \beta_k) = e^{-\frac{\alpha_k}{\beta_k}(1-e^{-\beta_k t_i})} + \varepsilon_l]$ by finding the values of the parameters that minimize the sum of the squared errors SSE_l . At these values, $\sigma_l^2 = \text{SSE}_l/n$,

$$\begin{aligned} \text{SSE}_l(\alpha; t, m) &= \sum_{k=1}^4 \sum_{j=1}^4 \sum_{i=1}^8 \varepsilon_l^2 \\ &= \sum_{k=1}^4 \sum_{j=1}^4 \sum_{i=1}^8 (m_{jk}(t_i) - c(t_i; \alpha_k))^2 \\ &= \sum_{k=1}^4 \sum_{j=1}^4 \sum_{i=1}^8 (m_{jk}(t_i) - e^{-\alpha_k t_i})^2 \end{aligned} \quad [\text{S4}]$$

or

$$\begin{aligned} \text{SSE}_l(\alpha, \beta; t, m) &= \sum_{k=1}^4 \sum_{j=1}^4 \sum_{i=1}^8 \varepsilon_i^2 \\ &= \sum_{k=1}^4 \sum_{j=1}^4 \sum_{i=1}^8 (m_{jk}(t_i) - c(t_i; \alpha_k \beta_k))^2 \\ &= \sum_{k=1}^4 \sum_{j=1}^4 \sum_{i=1}^8 \left(m_{jk}(t_i) - e^{-\frac{\alpha_k}{\beta_k} (1 - e^{-\beta_k t_i})} \right)^2. \end{aligned} \quad \text{[S5]}$$

Because the residuals, ε_i , are assumed to be normally distributed, this is equivalent to finding parameters such that the log likelihood (L) of the residuals is maximized with

$$L(\text{parameters; data}) = -\frac{n}{2} \ln(2\pi\sigma_l^2) - \frac{\text{SSE}_l}{2\sigma_l^2}.$$

For each transcript, we fit decay of the four genotypes to the 225 decaying decay models (Eq. S5) and 15 exponential decay models (Eq. S4). We found optimal parameter estimates by minimizing the SSE using the `slsqp()` function from the *nloptr* package in R (2–5), using linear equality constraints to distinguish the 240 models. The TMB package (6) was used to ensure accuracy and efficiency of the optimization calculations, given that some models have high dimensional parameter space.

We constrained the parameter estimates for the decaying decay model to the intervals $\alpha \in [0.0001, 0.75]$ and $\beta \in [0.001, 0.075]$ for the following reasons. When $\alpha = 0$ in Eq. S3, the parameter β is undetermined, and the exponential decay model captures this constant case. Thus, we constrain α away from zero when fitting this model. When β is large, $1 - e^{-\beta t} \approx 1$, meaning that β appears only in the combination α/β and cannot be identified separately from α . When β is small, $1 - e^{-\beta t} \approx -\beta t$ and $c(t) \approx e^{-\alpha t}$. In each of these cases, the behavior can be captured by the simpler exponential decay model.

We chose the precise intervals based on the first and last times of the measurements, 7.5 min and 480 min. To find the upper bound of $\alpha = 0.75$, we note that $c(7.5) = e^{-0.75(7.5)} \approx 0.004$, meaning that effectively all transcript has decayed before the first time point and that we cannot observe a higher decay rate from our data. For the exponential decay model, we constrain α to the interval $[0, 0.75]$. To find the lower bound of α in the decaying decay model, with $\alpha = 0.0001$, $c(480) = e^{-0.0001(480)} \approx 0.95$, a degree of decay sufficiently small to make it impossible to detect any slowing of the decay rate. In the decaying decay model, we thus constrain α to the interval $[0.0001, 0.75]$. To find the lower bound of the β , we observe that $1 - e^{-\beta t} \approx -\beta t$ if $\beta < 1/480 \approx 0.002$, making smaller values of β unobservable. To find the upper bound on β , we require the first four time points, up through $t = 60$, to give information on changes in the decay rate. To find the value of β where less than 99% of transcript has decayed by this time, we solve $e^{-60\beta} = 0.01$ to find that $\beta = -(1/60)\ln(0.01) \approx 0.077$. We thus constrain β to the interval $[0.001, 0.075]$.

We used 50 different starting conditions in the optimization of each gene–model pair ($18,674 \times 240 = 4,481,760$). These starting conditions were selected by explicitly finding the minimum of a discrete approximation of the sum of the squared errors function for the two simplest models: model 239, where α and β are the same for all genotypes, and model 240, where α is the same for all genotypes and $\beta = 0$. We selected the first 25 starting α values from a normal distribution with SD of 0.01 centered at the α corresponding to the minimum of model 239 for that gene. The additional 25 starting α values were selected from a normal distribution with SD of 0.01 centered at the α corresponding to the minimum of model 240 for that gene. The first 25 starting β values were selected from a normal

distribution with SD 0.01 centered at the β corresponding to the minimum of model 239 for that gene. To ensure complete search of the likelihood surface, the additional 25 starting β values were selected from a uniform distribution over all possible β values $[0.001, 0.075]$.

For each gene–model pair, we selected, from these 50 optimization results, those with a log likelihood within our tolerance, $\epsilon = 10^{-4}$, of the maximum, and defined J to be the number that lie within this set. We checked the consistency of these J parameter estimates by computing C_{tot} , the sum of a modified coefficient of variation over all eight parameters, where the modified coefficient of variation is defined as the ratio of the SD to the mean plus the tolerance ϵ . We add ϵ to the denominator to avoid large values created by low estimates of the parameter α . The parameters we report are the average of each parameter within this set, and we checked that these reported parameters gave an Akaike information criterion (AIC_c) within our tolerance (ϵ) of the AIC_c calculated with the maximum log likelihood for that gene–model pair. This difference was never larger than 2×10^{-5} for the model with lowest AIC_c .

After fitting each of the transcripts to all 240 models and finding the maximum log likelihood L_{max} , we compared models using the corrected AIC_c ,

$$\text{AIC}_C = -2L_{\text{max}} + 2p + \frac{2p(1+p)}{n-p-1},$$

where p is the number of parameters in the model and $n = 128$ is the number of observations. For example, with model 165, there are three α values, two β values, and σ_{165}^2 for a total of $P = 6$. We used the AIC_c because there are relatively few observations compared with the number of parameters estimated by each model (7).

For comparison of models for each gene, we consider models that lie within 2 AIC_c of the model with the minimum AIC_c . We are less confident in our choice of model if any model lying within 2 AIC_c of that model has a small value of J , such as $J \leq 5$, indicating that the algorithm might have missed the best value. This occurred for only 46 of the 18,674 genes, less than 0.3%. The estimated σ^2 provides another measure of goodness of fit for a model. We found that 1,381 genes (or 7.4%) had a $\sigma^2 > 0.065$ for the model with the minimum AIC_c . Because these have a SD of approximately $\sigma = 0.25$, a large proportion of the possible variation for measurements that range from 0 to 1, these genes were excluded from further analysis.

We selected the model with the minimum AIC_c for each transcript, and the averaged α and β parameter estimates were used in further analyses. We interpret α values as the baseline mRNA decay rates because they give the decay rate at $t = 0$ for each model.

R code for this analysis is available as an R package at <https://github.com/reedssorenson/RNAdecay>.

Bioinformatic Analysis. Gene annotation information and sequences were obtained from TAIR 10 (www.arabidopsis.org). Sequences were analyzed using R packages: *Biostrings*, *ggplot2*, *gplots*, *reshape2*, *GenomicAlignments*, *grid*, and *parallel*. GO classifications (“gene_association.tair” file accessed May 5, 2017) were obtained from the Gene Ontology Consortium (geneontology.org). Gene subsets modeled in distinct α subgroups were evaluated for GO enrichment using the *GOHyperGAll* R script (8).

Small RNA Experiment and Analysis. Total RNA was isolated using Invitrogen’s Plant RNA Reagent from all four genotypes of the genome-wide decay analysis at the same T_0 as the genome-wide mRNA decay experiment with three biological replicates. Isolated RNA was digested with RNase-free TURBO DNase followed by ammonium acetate and ethanol precipitation. Total RNA aliquots of 1 μg were submitted to the University of Utah Genomics Core

for small RNA library preparation (NEBNext Small RNA library prep for Illumina Sequencing) and sequencing on an Illumina HiSeq 2500. Flow cell reads were trimmed of adapter sequences, followed by filtering of reads shorter than 21 nt using cutadapt v. 1.8.3 with Python 2.6.6 (9). Bowtie v. 1.1.2 (10) was used to align reads to the TAIR10 genome scaffold with 0 mismatches and allowing only a single best alignment for any individual read (arguments: -k 1–best -n 0). Following alignment, sense and antisense reads of 21-, 22-, 23-, and 24-nt lengths were counted by gene and by exon with the Bioconductor package *GenomicAlignments*

(version 1.6.3). We used DESeq2 version 1.10.1 with a false discovery rate of 5% to quantify differential expression. We excluded mitochondrial- and chloroplast-annotated mRNAs, as well as annotated nuclear-encoded tRNAs (631), rRNAs (4), small nucleolar RNAs (71), snRNAs (13), miRNAs (178), and tasiRNAs (8), by searches for RNA types from TAIR10 via the “Bulk data retrieval” tool; this left a total of 27,591 transcripts that we classified as small RNAs and used for further analyses of differential expression. Raw and processed data for each gene are available at the Gene Expression Omnibus, accession GSE86361.

- Micallef L, Rodgers P (2014) eulerAPE: Drawing area-proportional 3-Venn diagrams using ellipses. *PLoS One* 9:e101717.
- Johnson SG (2011) The NLOpt Nonlinear-Optimization Package. Available at ab-initio.mit.edu/nlopt. Accessed September 15, 2017.
- Johnson SG (2014) The NLOpt Nonlinear-Optimization Package. Available at ab-initio.mit.edu/nlopt. Accessed September 15, 2017.
- Ypma J, Borchers HW, Eddelbuettel D, Ypma MJ (2014) Package “nloptr.” Available at <https://cran.r-project.org/web/packages/nloptr/index.html>. Accessed September 15, 2017.
- R Core Team (2015) *R: A Language and Environment for Statistical Computing* (R Found Stat Comput, Vienna).
- Kristensen K, Nielsen A, Berg CW, Skaug H, Bell BM (2016) TMB: Automatic differentiation and Laplace approximation. *J Stat Softw* 70:1–21.
- Burnham KP, Anderson DR (2003) *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach* (Springer, New York).
- Horan K, et al. (2008) Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol* 147:41–57.
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17:10–12.
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
- Narsai R, et al. (2007) Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. *Plant Cell* 19:3418–3436.

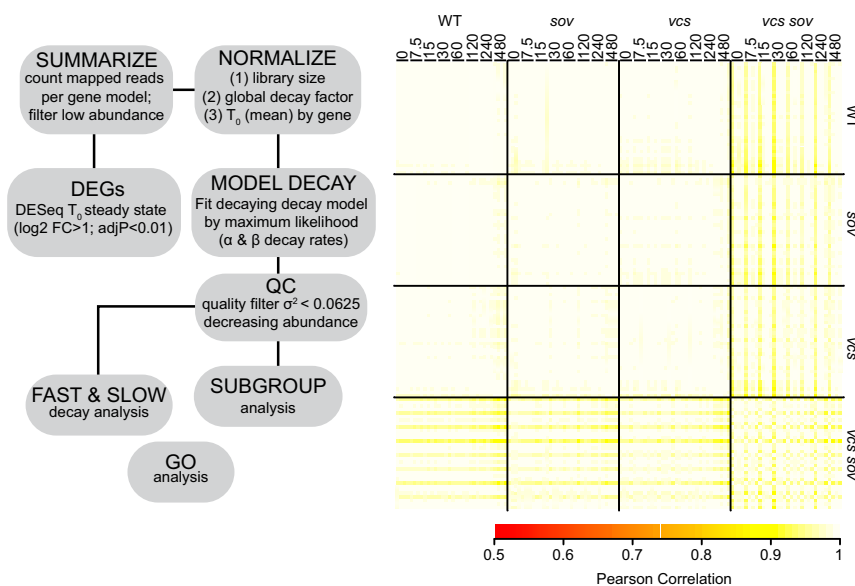


Fig. S1. Workflow and data quality. (A) Schematic of analysis workflow of RNA-Seq decay dataset. (B) Heatmap of pairwise sample Pearson correlation values of library-normalized read counts for 128 RNA sequencing libraries from four biological replicates of WT (VCS SOV), sov, vcs, and vcs sov genotypes from the time course.

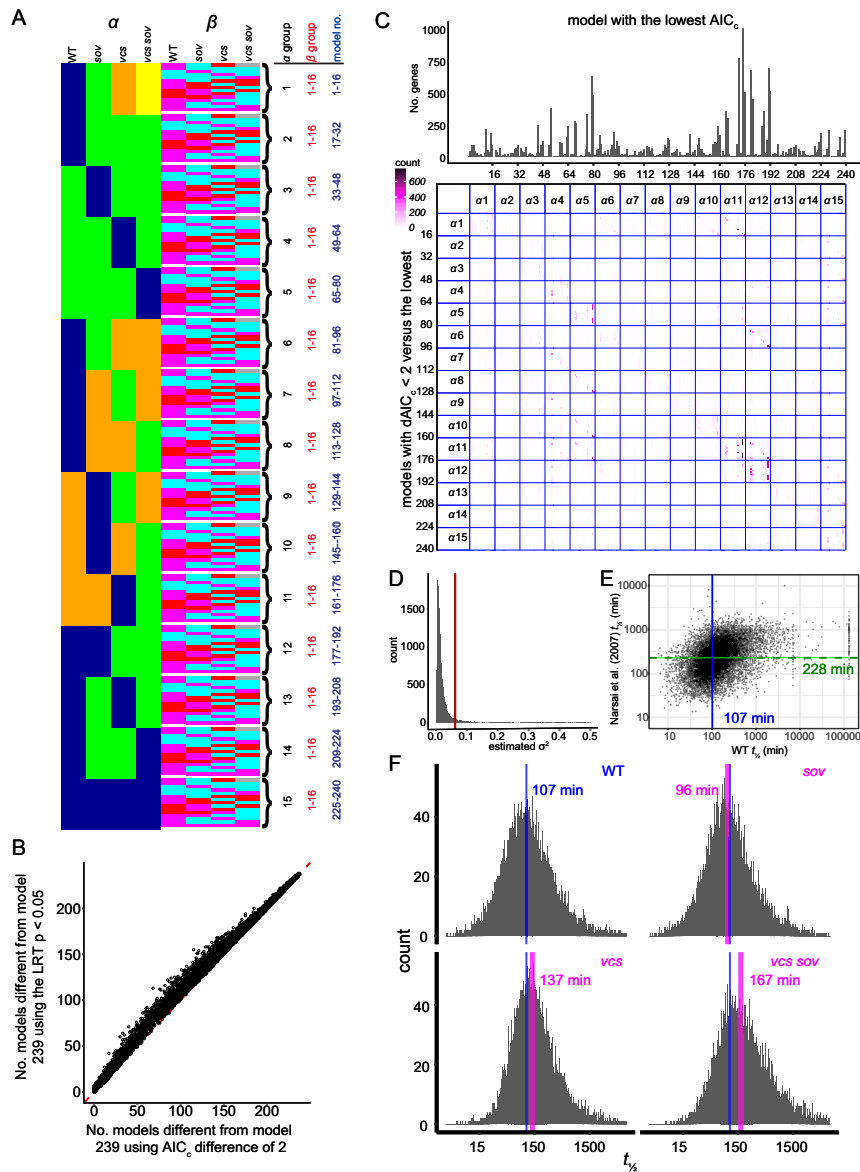
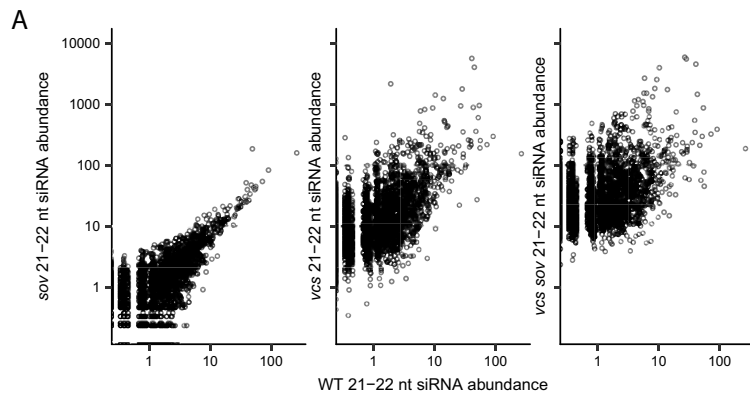


Fig. S2. Modeling parameter color map and output analysis. (A) Color map indicating α and β groups for maximum likelihood models. Identical box colors within a model represent equal values. White boxes indicate the absence of β in a model. (B) Model comparison statistics compared. Model 239 was compared against all other models for each gene using the Likelihood Ratio Test (LRT; model 239 is nested in all other models) and the AIC_c . The number of significantly better models using an AIC_c difference of >2 and LRT $P < 0.05$ are compared. Fewer different models using the AIC_c supports its use as a rigorous statistical approach. (C) Selected model distribution is displayed as a histogram (Upper). Selected models were the models with the lowest AIC_c . Similarly scoring models (i.e., the AIC_c difference with the lowest model was <2) were tallied for all genes and presented as a heatmap matrix (Lower) and depicts the alternative model frequency of association for each selected model. This reveals that competing models tended to be in the same α group, or, if the α -group deviated, in the same β group. (D) Histogram of modeled σ^2 estimates from the selected model for each gene. Vertical red line demarcates $\sigma^2 = 0.0625$ used as a quality cutoff. (E) Comparison of mRNA half-life found in this study to that of Narsai et al. (11). (F) Log-normal RNA half-life distributions in decay mutants as histograms. Vertical blue lines represent the median half-life (107 min) of WT (VCS SOV), and magenta lines and numbers represent the median half-life of each respective decay mutant.



B

tasiRNA-targeting locus	tasiRNA-targeted mRNA	mRNA Half-life	
		WT	sov
<i>TAS1a</i>	AT1G12775	99.2	99.2
<i>TAS2</i>	AT1G62930	102.7	102.7
	AT1G63130	83.9	83.9
<i>TAS1b</i>	AT1G50055	56.5	56.5
<i>TAS1c</i>	AT2G27400	41.0	41.0
	AT2G39675	56.6	56.6
	AT2G39681	41.5	41.5
<i>TAS3a</i>	AT3G17185	41.6	41.6
<i>TAS3a</i> , <i>TAS3b</i> , <i>TAS3c</i>	AT2G33860	69.3	53.1
	AT5G60450	131.3	102.9
	AT5G62000	120.9	101.4

C

tasiRNA locus	DESeq2 average normalized expression and log2 fold change					
	21 nt			22 nt		
	WT	sov	Log2 fold	WT	sov	Log2 fold
<i>TAS1a</i>	10921	7046	-0.63	2034	1589	-0.35
<i>TAS2</i>	29007	21096	-0.45	23436	20946	-0.16
<i>TAS1b</i>	5310	4854	-0.13	1077	968	-0.15
<i>TAS1c</i>	36661	29719	-0.30	6770	5080	-0.41
<i>TAS3a</i>	4331	2492	-0.79	516	380	-0.43
<i>TAS3b</i>	4	2	-1.06	1	1	-0.83
<i>TAS3c</i>	30	24	-0.34	12	5	-1.20

Fig. S3. Small RNA abundances and tasiRNA sensitivity analysis. (A) Distribution of read counts for all identified 21- to 22-nt siRNAs for *sov*, *vcs*, and *vcs sov*, relative to the WT. (B) Decay rate comparison of tasiRNA targets in WT and *sov* mutants; similar decay rates suggest that SOV mutants do not show greater sensitivity to siRNAs. (C) Comparison of tasiRNA abundance in WT and *sov* mutants.

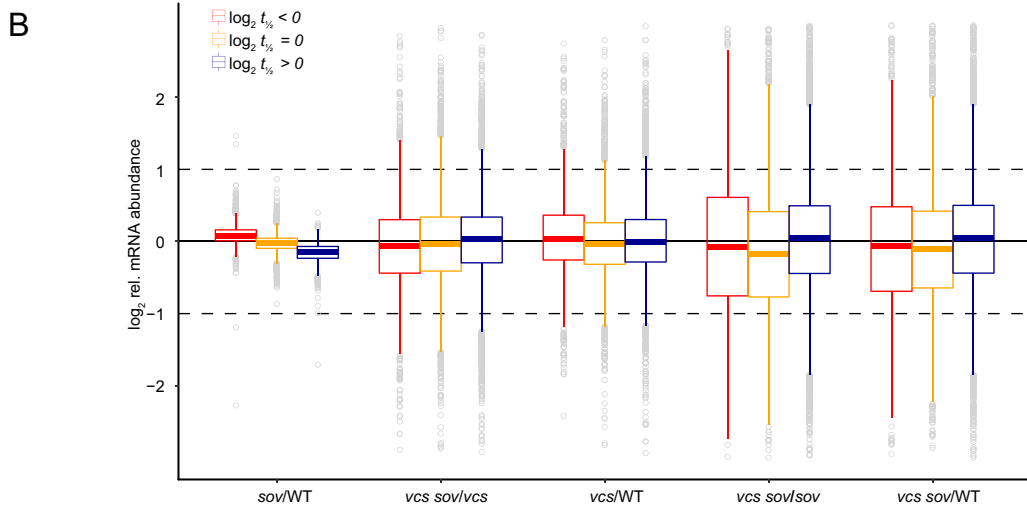
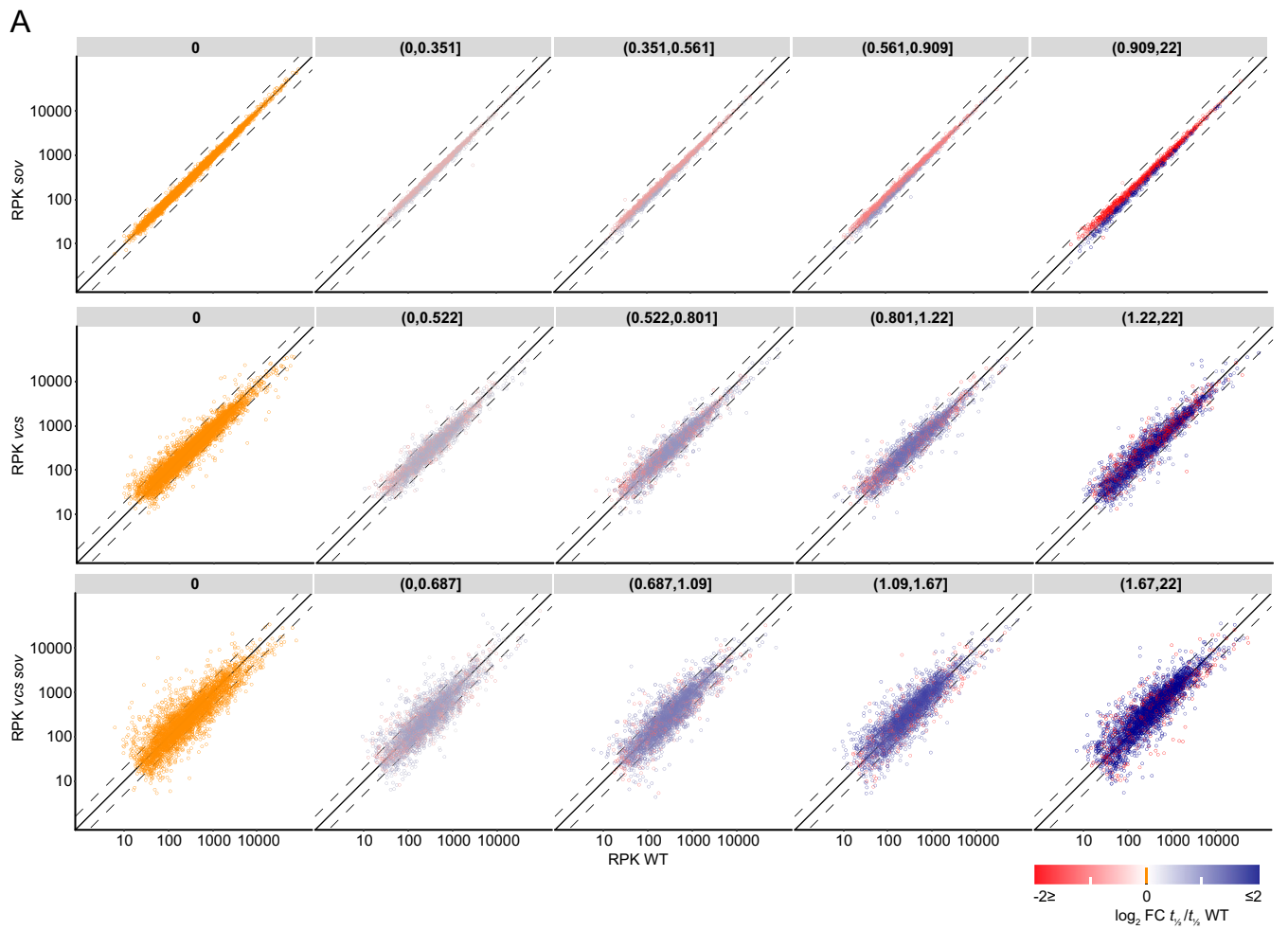


Fig. 54. RNA decay feedback. (A) Comparison of transcript abundance (RPK) from T_0 samples of *sov* (Upper), *vcs* (Middle), and *vcs sov* (Lower) relative to WT, and separated by the magnitude of the decay rate difference. The decay rate differences are labeled in the gray headers, with orange (at left) showing RNAs with identical decay rates, and increasing decay rate magnitudes toward the right. Magnitude range is given as $|\log_2(t_{1/2}/t_{1/2} \text{ WT})|$; square brackets, inclusive; parentheses, exclusive. RNA decay rates (relative to WT) shown by color, with faster depicted in red, and slower depicted in blue. (B) Relative RNA abundances for RNAs with faster, equal, and slower decay rates, relative to the WT. Near-WT RNA abundance in *sov* requires VCS. Dashed lines, $|\log_2 \text{ rel. abundance}| > 1$; solid line, equal expression.

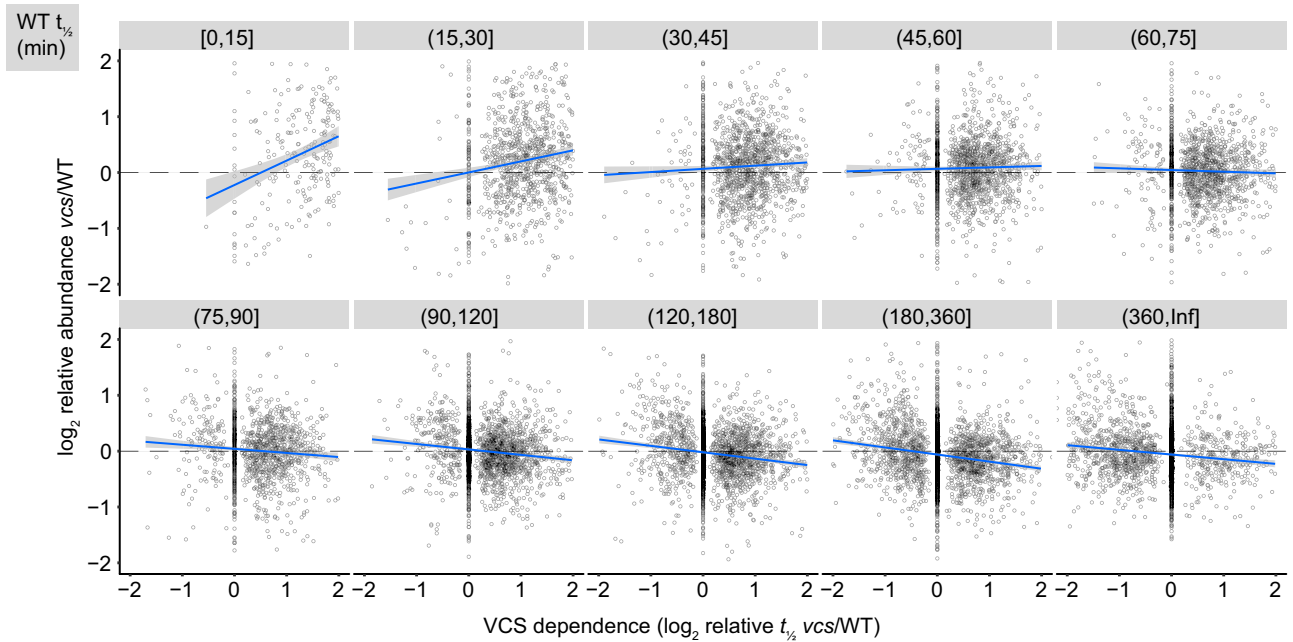


Fig. 55. Fast-decaying VCS-dependent RNAs show abundance to increase in decay mutants. Comparisons of VCS dependence and *vcs* log₂ relative transcript abundance from T_0 samples. RNAs were binned by their WT half-life (labeled gray headers show half-life ranges in min; square brackets: inclusive, parentheses: exclusive). Note that the correlation between VCS dependence and increased abundance only holds for fast-decaying RNAs.

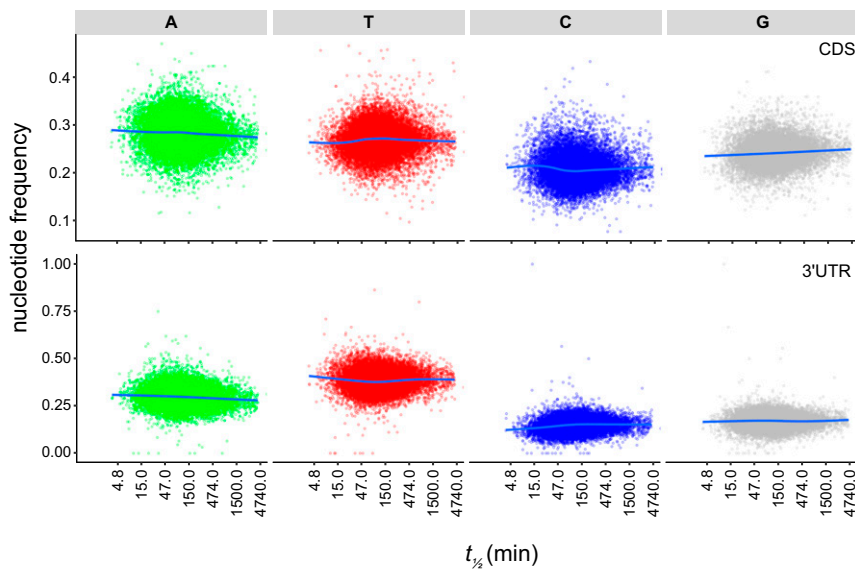


Fig. 56. Nucleotide frequency in coding sequence and 3' UTR in relationship to RNA half-life. (*Upper*) We found no consistent relationship between the frequency of A, T, C, and G in the CDS and RNA half-life. (*Lower*) Nucleotide frequency in the 3' UTR also did not vary with half-life, although we did find that the 3' UTR was relatively depleted in C and G, and had elevated levels of T.

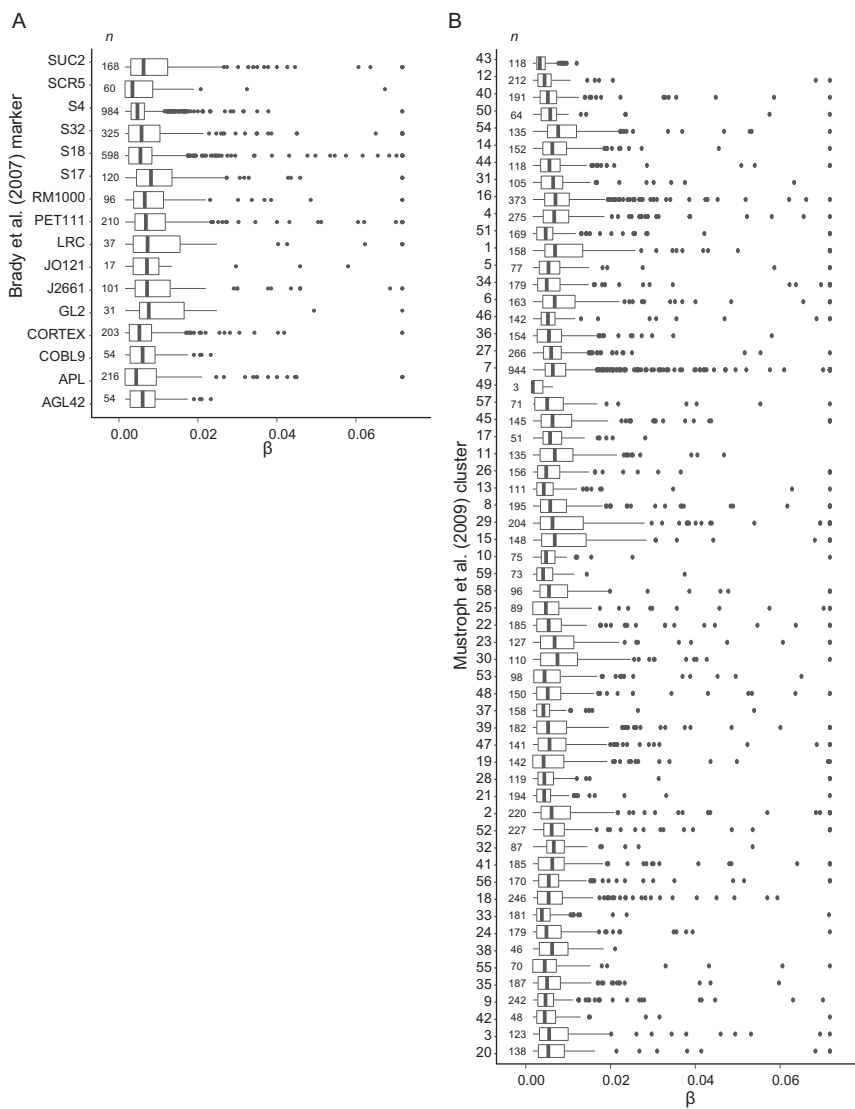


Fig. S7. Analysis of β values of genes expressed in specific cell types. To determine whether β (the decay of decay rate) arose because cordycepin poorly penetrated some tissues, we compared the distributions of β in cell-specific gene sets (1, 2). The mean β and range of β values were similar for cell type populations on the exterior and interior of the seedling, suggesting that differential cordycepin penetration was unlikely to have influenced estimated decay rates or decay-of-decay rates.

1. Brady SM, et al. (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* 318:801–806.

2. Mustrup A, et al. (2009) Profiling transcriptomes of discrete cell populations resolves altered cellular priorities during hypoxia in *Arabidopsis*. *Proc Natl Acad Sci USA* 106:18843–18848.

Other Supporting Information Files

[Dataset S1 \(XLSX\)](#)