**Molecular Cell, Volume *57***

# Supplemental Information

# Global Analysis of the RNA-Protein Interaction and RNA Secondary Structure Landscapes of the *Arabidopsis* Nucleus

Sager J. Gosai, Shawn W. Foley, Dongxue Wang, Ian M. Silverman, Nur Selamoglu, Andrew D.L. Nelson, Mark A. Beilstein, Fevzi Daldal, Roger B. Deal, and Brian D. Gregory

## SUPPLEMENTAL FIGURES

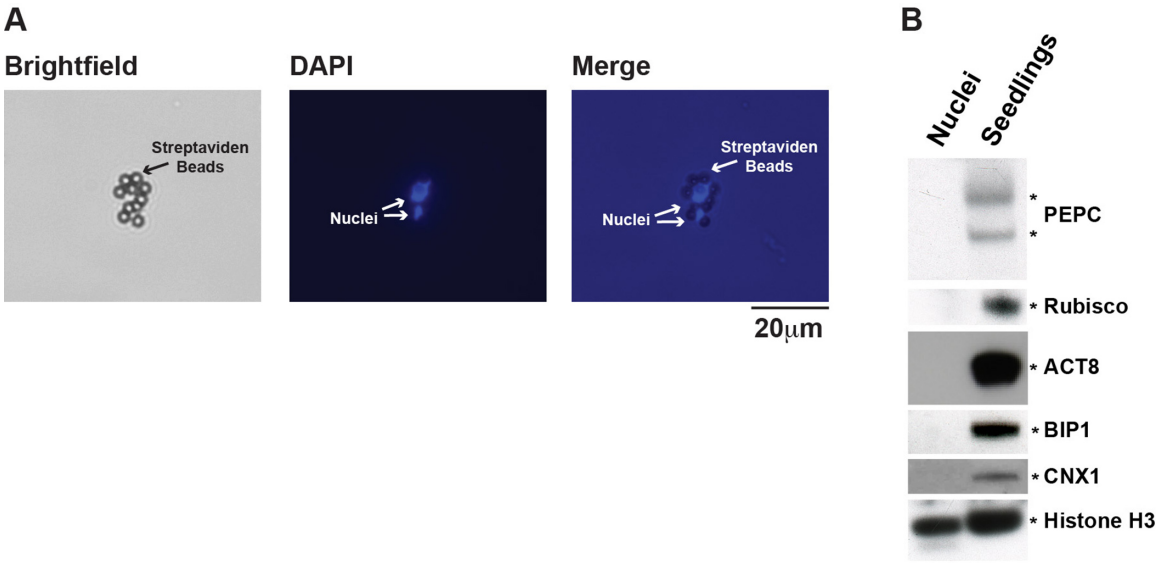



**Figure S1, Related to Figures 1-7: INTACT purified nuclei are free of cytoplasmic, ER, and chloroplastic contamination**

(A) Microscopy imaging of DAPI stained nuclei during the INTACT purification process. The images show that only the DAPI stained nuclei are bound to the streptavidin beads. (B) Western blot of lysates from INTACT purified nuclei and 10-day-old seedlings for the chloroplastic PEPC and RUBISCO, the mostly cytoplasmic ACT8, the endoplasmic reticulum (ER)-localized BIP1 and CNX1, as well as the nuclear histone H3 proteins.

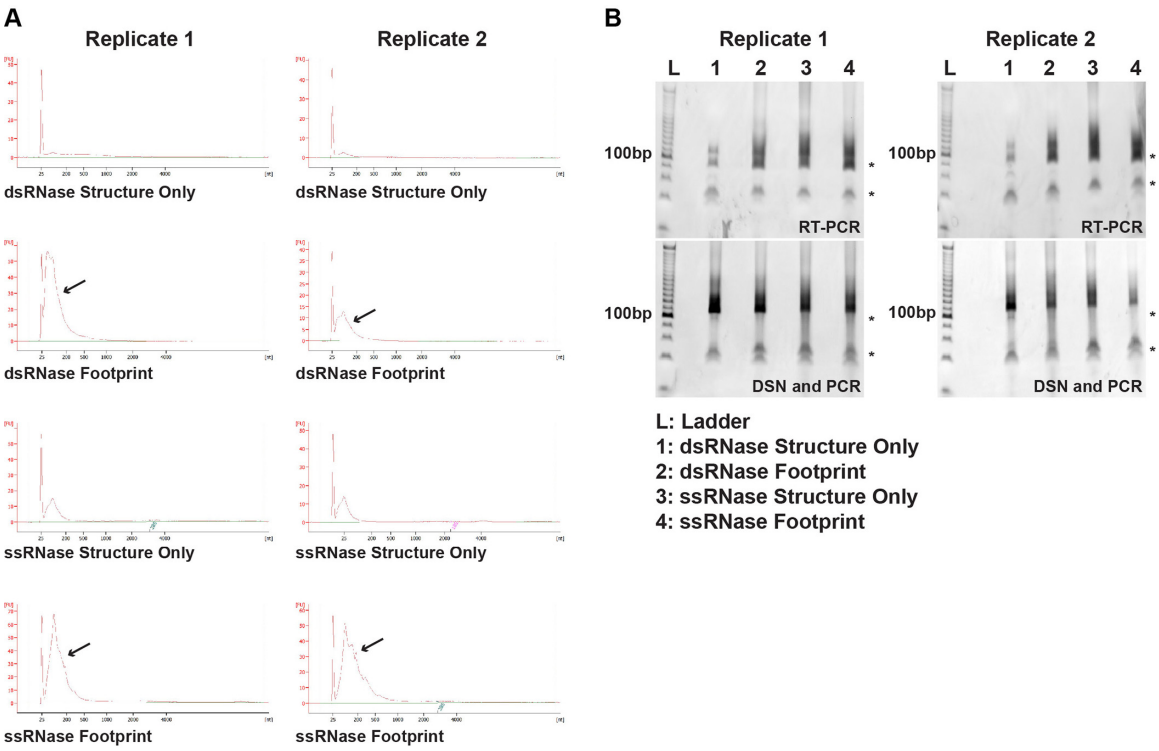



**Figure S2, Related to Figures 1-6: The PIP-seq libraries passed all three quality control checkpoints during library preparation**

(A) Profiles from a BioAnalyzer run of the digested RNA for each of the eight PIP-seq libraries. These profiles show the expected sizes and quality for these libraries when compared to profiles from previous PIP-seq experiments. The arrows point to the larger fragments found specifically in the footprinting samples that likely represent the protein protected sites (PPSs). (B) Two size selection gels run after the initial RT-PCR or after DSN treatment and PCR. These gels show that the libraries are still of the expected high quality, and have been shifted to the expected sizes after adapter ligations. The top and bottom asterisks (*) to the right of each gel image denote adapter-adapter products and unused primers, respectively. These contaminants were avoided during the gel purification process, ensuring the high quality of our sequenced libraries.

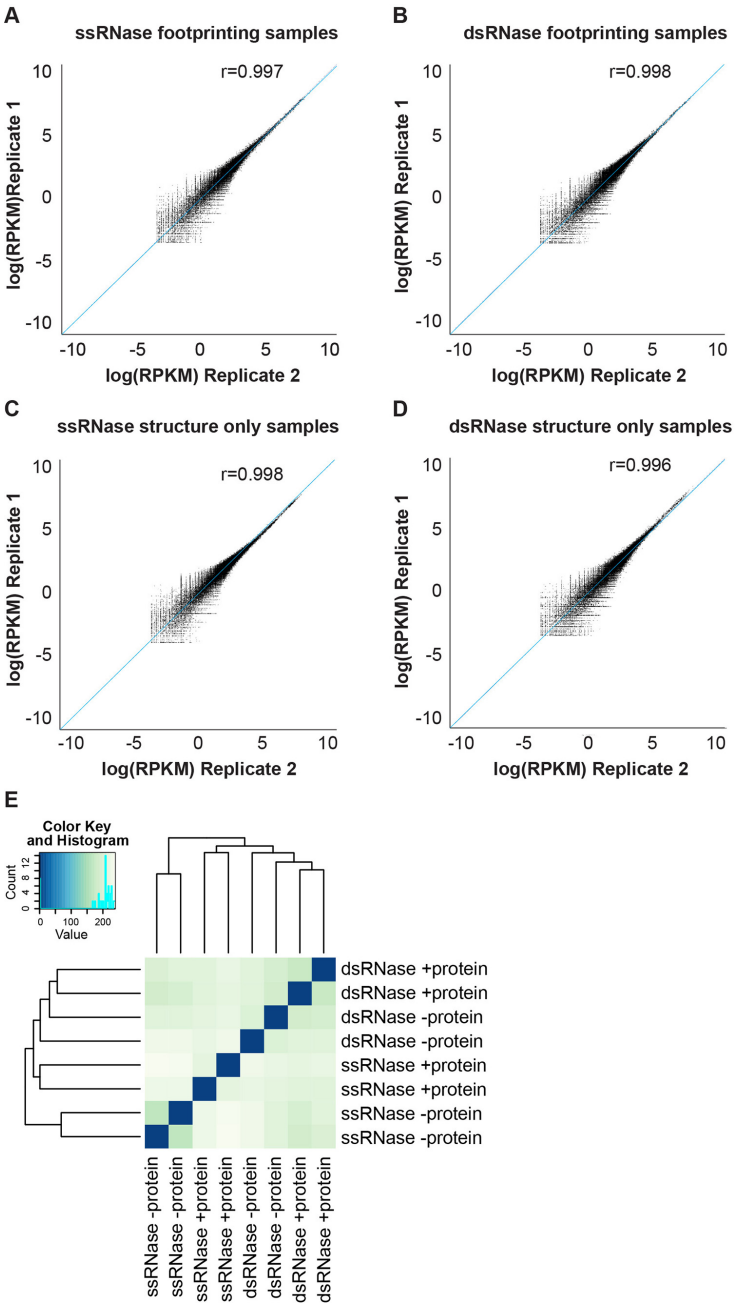



**Figure S3, Related to Figures 1-6: PIP-seq is a highly reproducible method**

(A-B) Correlation in read counts in a 50 nt sliding window between both ssRNase (A) and dsRNase (B) footprinting replicates. (C-D) Correlation in read counts in a 50 nt sliding window between both ssRNase (C) and dsRNase (D) structure only replicates. (E) Principle component analysis of 500 nt bins between each of the eight libraries. All replicate pairs cluster together, as do both RNases demonstrating the high reproducibility of these PIP-seq libraries.

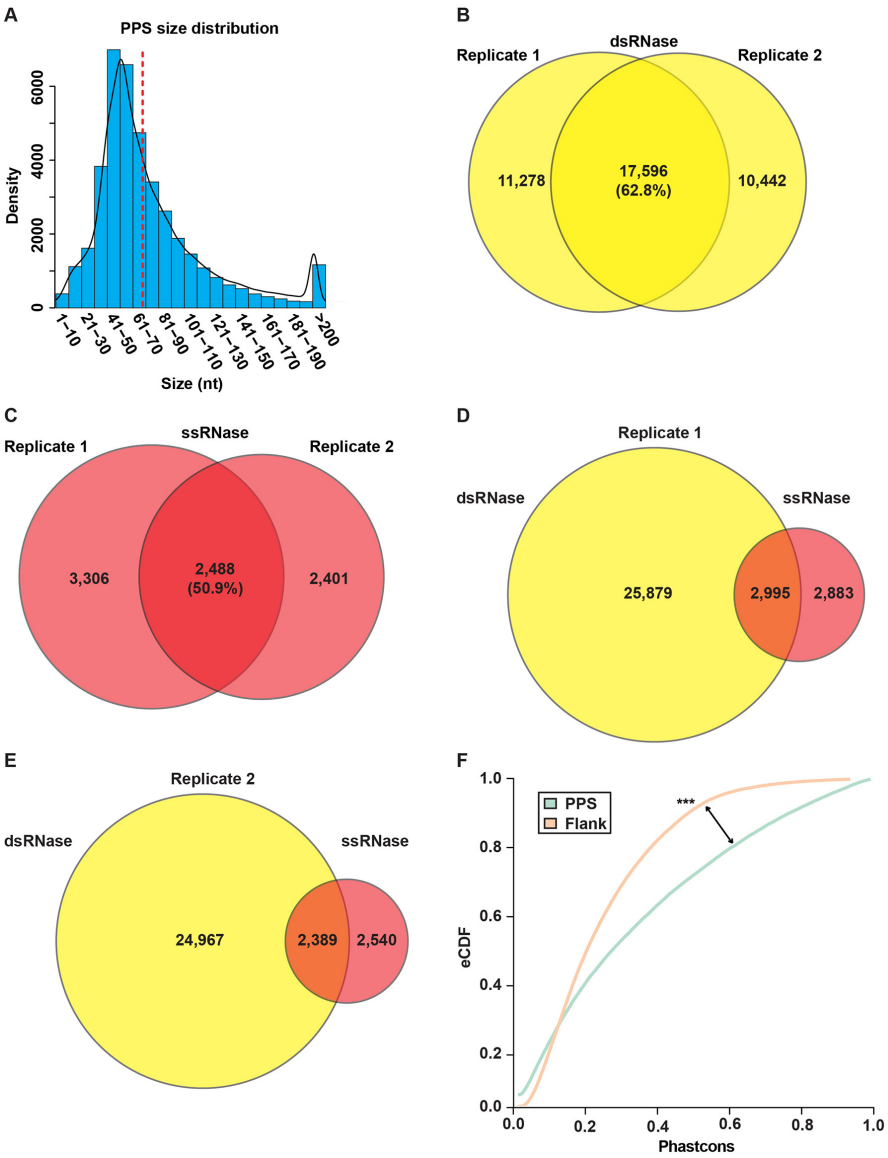



**Figure S4, Related to Figure 2: Further characterization of Arabidopsis nuclear PPSs**

(A) Distribution of sizes (nt) for the total set of 40,131 distinct PPSs. Dashed red line represents the median PPS size (~68 nt). (B-C) Overlap in PPS calls between dsRNase- (B) and ssRNase-treated (C) PIP-seq replicates. (D-E) Overlap in PPS calls between the dsRNase- (yellow circle) and ssRNase-treated (red circle) samples for replicate 1 (D) and replicate 2 (E). (F) Cumulative distribution of average PhastCons scores in PPSs (green line) versus similarly sized flanking regions (orange line). *** denotes p-value < $1x10^{-10}$, Kolmogorov-Smirnov test.

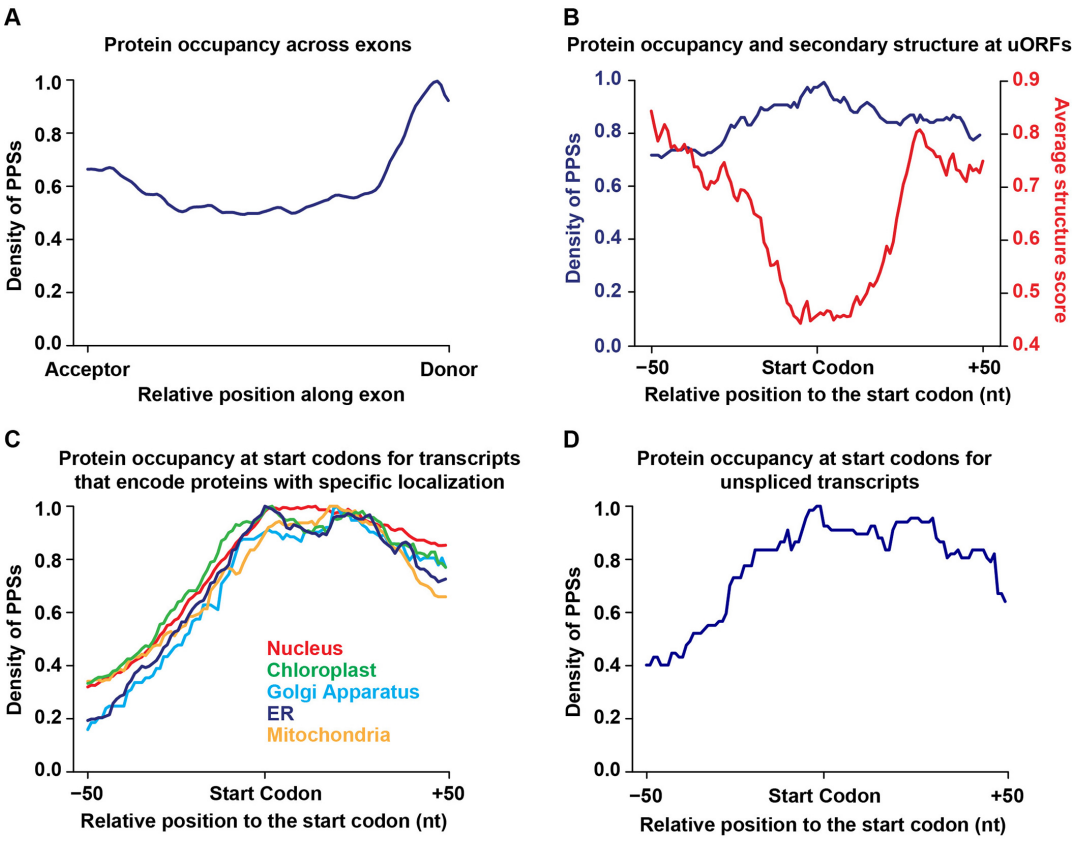



**Figure S5, Related to Figures 2-3: Protein occupancy at Arabidopsis constitutive exons, protein binding and secondary structure landscapes at upstream open reading frames (uORFs), as well as protein occupancy at start codons of transcripts encoding specifically localized proteins or that are unspliced**.

(A) PPS density across Arabidopsis constitutive exons (excluding exons containing star and stop codons) (B) PPS density and structure score profiles for highly confident upstream open reading frames (uORFs) (von Arnim et al., 2014). Average PPS density (blue) and structure score (red) at each position +/- 50 nt at uORF start codons. (C) Average PPS density at each position +/- 50 nt at canonical start codons for transcripts encoding proteins that are localized to specific cellular compartments (as specified by colored line and label) based on TAIR10 annotation. These transcripts show similar profiles. (D) Average PPS density at each position +/- 50 nt at canonical start codons for transcripts that are unspliced and likely nuclear localized in our RNA sequencing experiments.

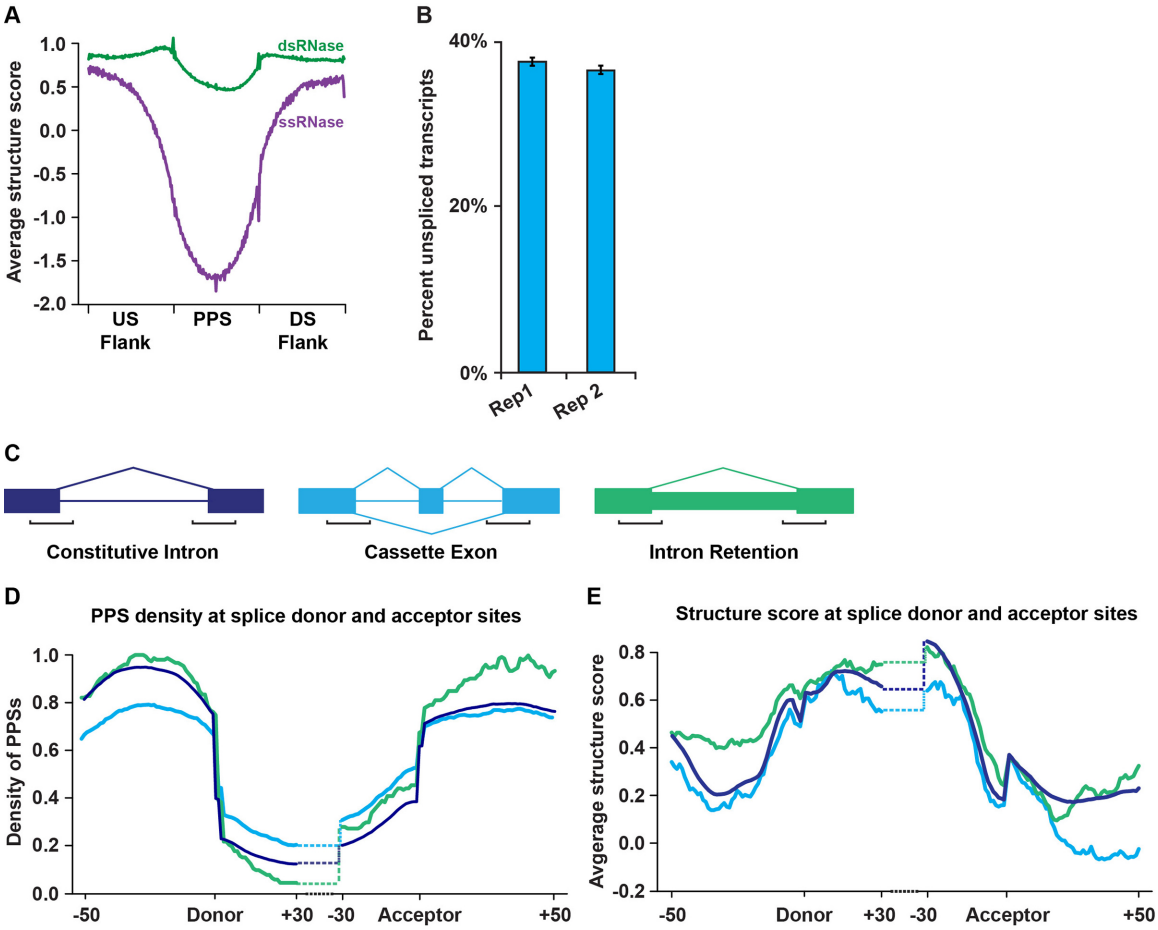



**Figure S6, Related to Figures 3-5: Secondary structure and protein binding landscapes at protein interaction sites and isolated alternative splicing events**.

(A) The average structure score of exonic PPSs from the dsRNase (green) or ssRNase (purple) treated libraries, and equal sized flanking regions, for 100 equal sized bins. The average structure score for the dsRNase treated PPSs is significantly (p-value < $2.2 \times 10^{-16}$, Wilcoxon test) greater than ssRNase treated PPSs. (B) The mean percentage of exon/intron junction mapping reads per transcript from two replicates (as indicated) of total RNA sequencing for congruently purified nuclei. Error bars represent standard error of the mean (SEM). (C) Diagram of constitutive introns (blue), cassette exons (turquoise), and intron retention events (green). Large boxes represent exons, lines represent constitutive introns, and small boxes represent alternatively spliced introns, with

brackets indicating regions graphed in D and E for reference. (D) Average PPS density at each position -50 to +30 nt at the donor splice site, and -30 to +50 at the acceptor splice site. Line colors correspond to examples shown in C. CEs show significantly (p-values < 0.001, Wilcoxon test) higher PPS density across both intronic and exonic sequences at the acceptor splice site (-30 to +40). IR events demonstrate significantly (p-values < $6.0 \times 10^{-6}$, Wilcoxon test) higher PPS density across all interrogated regions. (E) Structure score profiles for constitutive and isolated alternative splicing events in Arabidopsis covering the same regions as D. Line colors correspond to examples shown in C. IR events displayed significantly (p-value < $6.5 \times 10^{-3}$, Wilcoxon test) increased structure upstream of the donor splice site (-45 to -1). Conversely, CEs did not demonstrate any significant differences in secondary structure as compared to constitutive introns.

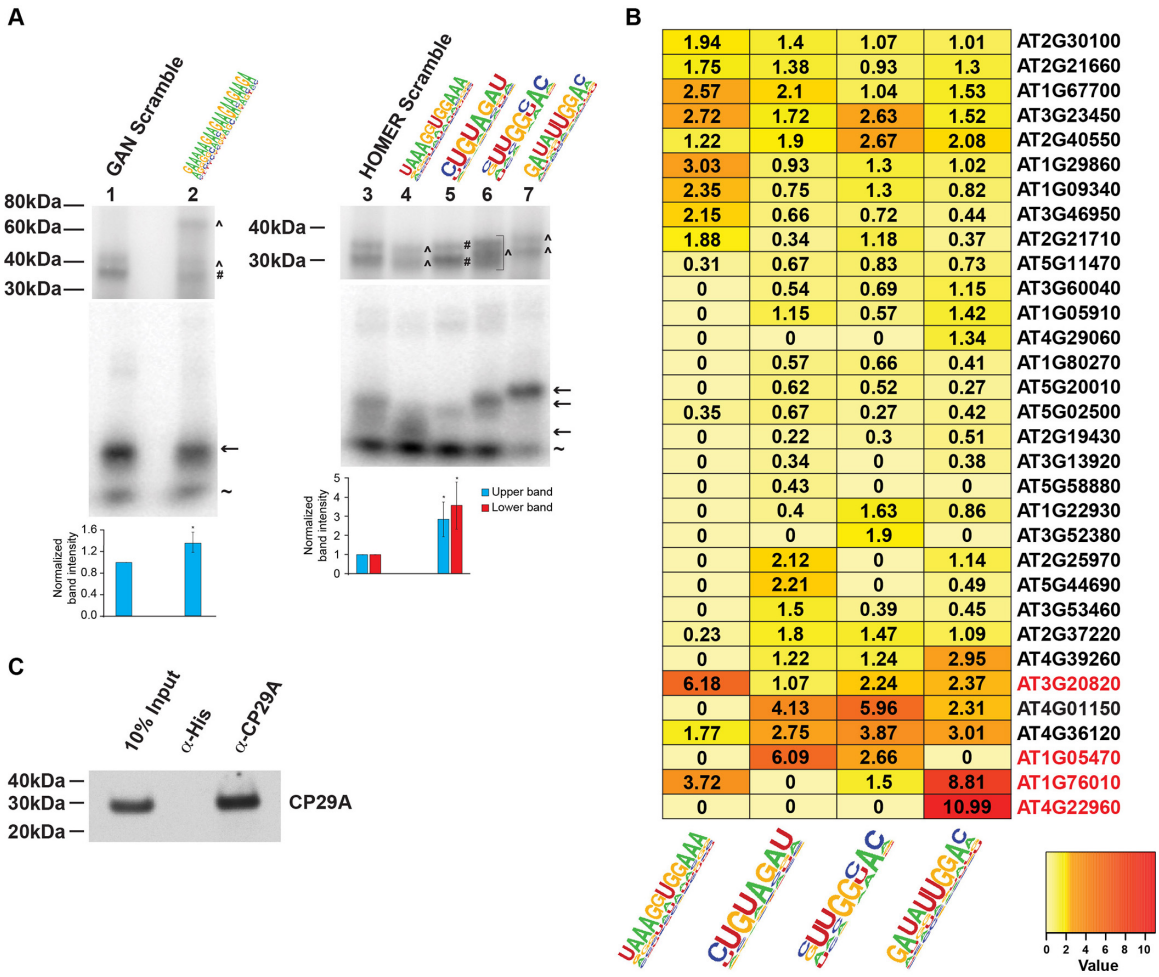



**Figure S7, Related to Figure 7: Identification of putative RBPs using synthetic RNA motifs**

(A) UV-crosslinking analysis for the indicated RBP-interacting motifs compared to non-specific controls using Arabidopsis 4-week-old leaf lysate. Three biological replicates were performed, and a representative gel is shown. For bands that are present in both the motif and scrambled control lanes, the intensity of the band was quantified and normalized to the unbound probe, and is graphed below the chart as fold change relative to scrambled control. ^ denotes bands that are present in a motif lane, but are absent from the scrambled controls. # denotes a band that is present in both the motif lane and the scrambled control, and therefore was quantified in the graphs below the respective lane. The arrows denote unbound probes. ~ denotes the unincorporated radiolabeled ATP. * denotes $p < 0.05$, Fisher's t-test. Error bars represent standard deviation, n=3.

(B) Enrichment of peptides from the indicated Arabidopsis proteins as compared to negative control pulldown samples. The number of peptide spectrum matches (PSM) for each sample was taken and the percentage of the total PSM for each identified protein was calculated. The fold change relative to the average of the empty bead and scramble bait negative controls is shown. Proteins denoted in red are candidate RBPs that passed a 6 fold enrichment threshold. (C) Western blot with an $\alpha$-CP29A monoclonal antibody on 10% input and eluents from RNA immunoprecipitations (RIPs) performed with $\alpha$-CP29A and $\alpha$-His monoclonal antibodies.

**SUPPLEMENTAL TABLES**

1. **Supplemental Table S1, Related to Figures 1-6:** The pertinent information for the 40,131 distinct PPSs identified in the Arabidopsis nuclear transcriptome.
2. **Supplemental Table S2, Related to Figures 5-7:** The pertinent information for the 41 identified protein-bound motifs.
3. **Supplemental Table S3, Related to Figures 7 and S7:** The pertinent information for all oligonucleotide probes and RT-qPCR primers that were used in this study.
4. **Supplemental Table S4, Related to Figure 7:** The number of peptides identified by LC-MS/MS in each RNA-affinity chromatography experiment, which is presented graphically in Figures 7A-B and S7B.

## SUPPLEMENTAL EXPERIMENTAL PROCEDURES

### Plant Materials

The purified nuclei used in this study were extracted from 10-day-old seedlings of *UBQ10*:NTF*/ACT2p*:BirA Columbia-0 (Col-0) ecotype of *Arabidopsis thaliana* using the INTACT methodology. Additionally, the lysates for all western blots were from these same 10-day-old seedlings. The lysates used for RNA immunoprecipitation (RIP) RT-qPCR and motif-interacting protein analyses were from whole leaves extracted from four-week-old Col-0 plants. All plants were grown at 20°C, in a 16 h light/8 h dark cycle.

### Cross-linking and INTACT purification

Immediately before nuclei purification, 10-day-old seedlings of *UBQ10*:NTF*/ACT2p*:BirA were crosslinked in nuclear purification buffer (20 mM MOPS (pH = 7), 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA) plus 1% (vol/vol) formaldehyde under vacuum for 10 minutes, followed by a five minute quench with 125 mM Glycine under vacuum for an additional five minutes. Crosslinked seedlings then underwent INTACT purification as previously described (Deal and Henikoff, 2010).

### Total RNA sequencing library preparation

10-day-old seedlings of UBQ10:NTF/*ATC2p*:BirA underwent the INTACT purification as previously described (Deal and Henikoff, 2010). The resulting nuclei were lysed and the RNA was isolated using the Qiagen miRNeasy RNA isolation kit following the included protocol (Qiagen; Valencia, CA, USA). Finally, the purified RNA was used as the substrate for strand-specific total RNA sequencing library preparation as previously described (Elliott et al., 2013), with the exception that no poly(A) purification was performed but was replaced by DSN treatments as previously described (Silverman et al., 2014). The resulting libraries were sequenced on an Illumina HiSeq2000 using the standard protocol for 50 base pair single read sequencing.

**PIP-seq library preparation**

~Two million INTACT purified nuclei were lysed in 850 μl RIP buffer (25 mM Tris-HCl, pH = 7.4; 150 mM KCl, 5 mM EDTA, pH = 7.5; 0.5% NP40; 10 μM DTT; 1 tablet protease inhbitors and 0.5 μl/ml RNaseOUT (Life Technologies; Carlsbad, CA, USA)) by manual grinding. The resulting cell lysate was treated with RNase-free DNase (Qiagen; Valencia, CA, USA). The lysates were then split and treated with either 100 U/ml of a single-stranded RNase (ssRNase) (RNaseONE (Promega; Madison, WI, USA)) with 200 μg/ml BSA in 1X RNaseONE buffer for 1 hour at room temperature (RT), or 2.5 U/ml of a double-stranded RNase (dsRNase) (RNaseV1 (Ambion; Austin, TX, USA)) in 1X RNA structure buffer for 1 hour at 37°C as previously described (Silverman et al., 2014). See Figure 1A for a schematic representation of library preparation. Proteins were then denatured and digested by treatment with 1% SDS and 0.1 mg/ml Proteinase K (Roche; Basel, Switzerland) for 15 minutes at RT. Proteinase digestion was followed by a 2 hour incubation at 65°C to reverse the RNA-protein cross-links.

To determine whether nuclease resistant regions in RNAs are due to protein binding or specific secondary structures, we also determined the digestion patterns of ds- and ssRNases immediately following protein digestion. To do this, we performed the identical treatments as described above except that the cross-linked nuclear lysates were treated with 1% SDS and 0.1 mg/ml Proteinase K (Roche; Basel, Switzerland) and ethanol precipitated prior to being treated with the two RNases. In this way, the SDS and Proteinase K solubilized and digested the proteins allowing us to deduce PPSs within all detectable RNAs in the cells of interest (see Figure 1A for schematic).

The digested RNA was then isolated using the Qiagen miRNeasy RNA isolation kit following the included protocol (Qiagen; Valencia, CA, USA). To ensure that only high quality RNA samples were used for PIP-seq library preparation, the purified RNA was run on a Eukaryotic Total RNA Pico Series II chip (5067-1513; Agilent Technologies; Wilmington, DE, USA) using a

BioAnalyzer 2100 system. Finally, the purified RNA was used as the substrate for strand-specific sequencing library preparation as previously described (Silverman et al., 2014). All of the RNase footprinting libraries (a total of 4 for each replicate: ss- and dsRNase treatments, footprint and structure only) were sequenced on an Illumina HiSeq2000 using the standard protocol for 50 base pair single read sequencing.

**Read processing and alignment**

PIP-seq reads were first trimmed to remove 3' sequencing adapters using cutadapt (version 1.2.1 with parameters -e 0.06 -O 6 -m 14). The resulting trimmed sequences were collapsed to unique reads and aligned to the TAIR10 Arabidopsis genome sequence using Tophat (version 2.0.10 with parameters --library-type fr-secondstrand --read-mismatches 2 --read-edit-dist 2 --max-multihits 10 --b2-very-sensitive --transcriptome-max-hits 10 --no-coverage-search --no-novel-juncs). PCR duplicates were collapsed to single reads for all subsequent analyses.

**Estimating unspliced transcripts**

All reads from the total RNA-sequencing data that mapped to all detectable first TAIR10 annotated constitutively spliced introns were collected, removing reads that were entirely within the intron. We quantified the number of reads that had mapped through the exon/intron boundary (unspliced) compared to those that contained the exon/exon boundary (spliced). We then determined the fraction of junction mapping reads that were unspliced for each gene.

**Identification of PPSs**

PPSs were identified using a modified version of the CSAR software package (Muiño et al., 2011). Specifically, read coverage values were calculated for each base position in the genome and a Poisson test was used to compute an enrichment score for footprint versus structure only libraries. PPSs were then

called with a false discovery rate of 5% as previously described (Silverman et al., 2014).

**Functional analysis of PPSs**

PPS annotation was done 'greedily' using the TAIR10 genome annotations, such that all functional annotations that overlapped with a given PPS were counted equally. Conservation was assessed by comparing both PhastCons scores and the number of SNPs, within PPSs relative to equally sized flanking regions. PhastCons scores for PPSs compared to same sized flanking regions were calculated as previously described (Li et al., 2012; Silverman et al., 2014).

To perform the SNP occurrence analysis we first identified SNPs located in transcriptionally active region (TARs), defined as intervals at least 15 nt long with greater than 20 reads of coverage, while allowing for a gap of 10 nt with less coverage, as calculated using an aggregate list of alignments from both replicates of the PIP-seq libraries. Ten permutations of random shuffling of TARs were then performed to generate the control set with similar numbers and fragment sizes to our list of PPSs. We then quantified the number of non-redundant, substitution SNP sites cataloged by the 1001 Genomes Project (Cao et al., 2011) within the total list of PPSs and the 10 shuffled intervals, which were statistically compared to one another using a $\chi^2$-test.

***lincRNA conservation analysis***

*Brassica rapa* lincRNAs were identified from a list of 3,450 intergenic transcripts generated previously (Tong et al., 2013), then further filtered by removal of transcripts with an open reading frame >100 codons. A total of 1908 *B. rapa* lincRNAs were then used as the dataset in BLAST analyses with Arabidopsis lincRNAs using an E-value of $10^{-10}$.

**Calculating the structure score statistic**

For every base of detectable transcripts, we calculated the dsRNA-seq and ssRNA-seq coverages from the structure only samples, then calculated the structure score as described previously (Li et al., 2012). Briefly, when given the dsRNA-seq and ssRNA-seq coverages ($n_{ds}, n_{ss}$) of a given base *i*, the structure score is determined as:

$$S_i = \mathrm{glog}(ds_i) - \mathrm{glog}(ss_i) = \log_2\left(ds_i + \sqrt{1+ds_i^2}\right) - \log_2\left(ss_i + \sqrt{1+ss_i^2}\right)$$

$$ds_i = n_{ds}\frac{\max(L_{ds}, L_{ss})}{L_{ds}}, \quad ss_i = n_{ss}\frac{\max(L_{ds}, L_{ss})}{L_{ss}}$$

where $S_i$ is the structure score, $ds_i$ and $ss_i$ are the normalized read coverages, and $L_{ds}, L_{ss}$ are the total covered length by mapped dsRNA-seq and ssRNA-seq reads, respectively. The total coverage length was used as the normalization constant instead of the total number of mapped reads used previously, because we believe it is a more reasonable assumption for the transcriptome to have comparable levels of paired/unpaired regions. It is of note that we used a generalized log ratio (glog) instead of normal log-odds because it can tolerate 0 values (positions with no dsRNA or ssRNA read coverage) as well as being asymptotically equivalent to the standard log ratio when the coverage values are large. Only sense-mapping reads were used, as we are entirely concerned with the intra-molecular interactions contributing to the self-folding secondary structure.

**Structure score profile analysis of mRNAs**

The structure score for every base of each detected transcript was first calculated using all mapped and spliced reads. In addition to the minimum dsRNA-seq plus ssRNA-seq read coverage requirement discussed above, we only considered mRNAs with intact CDS regions, ≥ 45 nt 5'UTRs, ≥ 140 nt 3'UTRs and a minimum coverage of 100 reads across the entire transcript. For the profiles near CDS boundaries, structure scores for up/downstream of the

CDS start or end sites were extracted, aligned for each detectable mRNA and averaged to produce the profiles.

**PPS profile across constitutive exons**

All constitutive exons that did not contain a start or stop codon were taken and subdivided into one hundred equal sized bins. PPS density was then calculated and graphed across the bins as previously described (Silverman et al., 2014).

**Secondary structure and PPS density at upstream Open Reading Frames (uORFs)**

Annotated Arabidopsis uORFs of high confidence (defined as a purine at the -3 position and a glycine at the +4 position) were extracted from a previously annotated dataset (von Arnim et al., 2014). We then calculated average structure score (see above) and PPS density (average number of PPS covered bases) for uORFs with >10 mapped reads in the regions 50 bp up- or downstream of uORF start codons.

**PPS profiles across canonical start codons for transcripts localized to specific cellular compartments**

Transcripts were subdivided based on their TAIR10 annotated cellular component gene ontology (mitochondria: 0005739, chloroplast: 0009507, ER: 0005829, Golgi apparatus: 0005794, nucleus: 0005634). PPS density was then calculated and graphed for 50 nt up- and downstream of the start codon as previously described (Silverman et al., 2014).

**PPS profiles across canonical start codons for unspliced transcripts**

Unspliced genes were defined as transcripts in our total RNA-seq libraries with high coverage (above the median) in which junction-spanning reads at the first constitutive intron only crossed the exon/intron boundary, and not the exon/exon junction. PPS density was then calculated and graphed at 50 nt up-

and downstream of the start codon as previously described (Silverman et al., 2014).

**Structure profile at dsRNase- and ssRNase-identified PPSs**

All exonic PPSs and equal sized flanking regions were taken and subdivided into one hundred equal sized bins. The calculated structure scores (see above) were averaged for each bin, and the resulting profiles were graphed.

***Analysis of alternatively spliced exons and introns***

In order to identify specific subsets of alternative splicing events, we took all TAIR10 annotated mRNA transcripts and used the ASTALAVISTA suite (parameters -t asta –i) to identify every annotated alternative splicing event (Foissac and Sammeth, 2007; Sammeth et al., 2008). We then used the ASTALAVISTA code assigned to each event to identify single cassette exons or intron retention sites (0,1^2- or 0,1-2^, respectively). Additionally, we extracted all cassette exon and intron retention events, regardless of adjacent exons, using the list of alternative events and corresponding ASTALAVISTA codes previously described in Arabidopsis (Marquez et al., 2012). Taking these annotated events, we then identified the splice donor and acceptor sites of the nearest constitutive introns for our analysis (e.g. if exons 4, 5, and 6 are alternatively spliced together we looked at the donor and acceptor sites at exons 3 and 7, respectively). PPS and structure score profiles were then calculated (see above) for regions where the donor exon was ≥ 50 nt, acceptor exon was ≥ 50 nt, and intron was ≥ 60 nt and at least 5 reads mapped to the intron. Thus, these profiles can cover the fifty exonic and thirty intronic nucleotides flanking the splice donor and acceptor sites. P-values were calculated by non-pairwise Wilcoxon tests.

***Analysis of alternative polyadenylation sites***

We extracted the cleavage and polyadenylation sites previously identified by direct RNA sequencing (Sherstnev et al., 2012) and filtered out sites that were located outside of TAIR10 annotated 3’UTRs. A second filtering step was

performed to remove alternative polyadenylation (APA) sites within 60 nt of one another, preventing any overlap between analyzed flanking regions. PPS density and structure score profiles were then calculated (see above) for 30 nt flanking each side of these cleavage and polyadenylation sites. P-values were calculated by non-pairwise Wilcoxon tests.

**RBP bound sequence motif identification and profiling secondary structure at these sites**

MEME (Bailey et al., 2009) and HOMER (Heinz et al., 2010) were used to identify enriched RBP interaction motifs with parameters -p 8 -dna -nmotifs 100 -maxw 20 -evt 0.01 -maxsize 100000000, and -rna -size given -p 2 respectively. Motifs from Figures 5A-E were mapped to the genome using HOMER (Heinz et al.) to identify every occurrence of the motifs in nuclear mRNAs. We then identified protein bound and unbound occurrences using our mapped PPSs. Average structure scores for each position were calculated as described above.

**Motif and co-occurrence analysis**

Motif co-occurrence was defined at the transcript level, and k-means clustering of the resultant weighted adjacency matrix was used to identify clusters of co-occuring motifs. We set k=3 based on manual inspection of clusters on a multidimensional scaling (MDS) plot of the adjacency matrix. Gene Ontology (GO) analysis on the lists of transcripts that contained at least three protein bound occurrences of the motifs in each cluster was performed using agriGO (Du et al., 2010).

**UV Cross-linking analysis of motifs**

Synthetic RNA oligonucleotides (Table S3) were radiolabeled in a T4 polynucleotide kinase (PNK) reaction (New England Biolabs; Cambridge, MA, USA) using 500 μCi of $\gamma$-$^{32}$P ATP following the manufacturer's recommendation, followed by phenol-chloroform extraction and precipitation. Each RNA probe was diluted to equal counts per minute (cpm), and was added to separate 10.2 μL

binding reactions comprising 0.2 mM Tris (pH = 7.5), 0.02 mM EDTA, 40 mM KCl, 1.3% polyvinyl alcohol, 25 ng/μl tRNA, 3 mM $MgCl_2$, 1 mM ATP, 50 mM creatine phosphate, and 2.8 μg/μl Arabidopsis leaf lysate in RIP buffer (25 mM Tris-HCl, pH = 7.4; 150 mM KCl, 5 mM EDTA, pH = 7.5; 0.5% NP40; 10 μM DTT; 1 tablet/10ml protease inhibitors (Roche; Basel, Switzerland)) and incubated at 30°C for 20 minutes. The binding reaction was then subjected to UV cross-linking for 20 minutes using a 254 nm UV lamp (Mineralight Lamp Model R-52G (UVP; Upland, CA, USA)). RNA bound proteins were denatured in 1X SDS sample buffer and 1 mM β-mercaptoethanol and boiled for 5 minutes. Samples were separated on NuPAGE 3-8% Tris-Acetate gel (Life Technologies; Carlsbad, CA, USA) at 120V for 1 h. The gel was then fixed in a 10% methanol and 10% acetic acid solution for five minutes, and dried for 90 minutes. Phosphorimaging was used to visualize protein-bound and unbound RNA probes. This assay was replicated three times, and densitometry was used to quantify the bands that were present in both the motif and scramble probe lanes. The intensity of these bands was normalized to the intensity of the unbound probes from the corresponding lane, and the normalized intensity of the band in the scramble lane was set to one for comparison.

**Identification of proteins that interact with motifs identified in PPSs**

We used five of the most enriched motifs that we identified within PPS sequences (Figure 5 and Supplemental Table S2) as baits to isolate interacting RBPs by RNA-affinity chromatography. Specifically, RNA baits (covalently-linked to agarose beads) containing the identified motif of interest (IDT; Coralville, IA, USA) were incubated in a binding reaction (3.2 mM $MgCl_2$, 20 mM creatine phosphate, 1 mM ATP, 1.3% polyvinyl alcohol, 25 ng of yeast tRNA, 70 mM KCl, 10 mM Tris (pH = 7.5), 0.1 mM EDTA) with 56 μg of 4-week-old Arabidopsis whole leaf lysate at RT for 30 minutes. Beads were washed four times with GFB-100 (20 mM TE, 100 mM KCl) plus 4 mM $MgCl_2$ and once with 20 mM Tris-HCl (pH = 7.4). The RNA-bound proteins were then directly trypsinized on the beads.

**MS-ready sample preparation**

Multiple independent samples for the selected motifs and their corresponding controls were used to average out experimental variability, optimize detection limits, and improve signal to noise ratio for robust specific identification. MS sample preparations and analyses were performed as described previously (Onder et al., 2008; Onder et al., 2006). Briefly, RNA-bound proteins were treated directly on the beads with 100 mM $NH_4HCO_3$ containing ~6 ng/μl of MS-grade trypsin (Promega; Madison, WI, USA) and incubated at 37ºC for 12-18 hrs. These samples were extracted first with 1%HCOOH/2%$CH_3CN$, and several times with 50% $CH_3CN$; combined peptide extracts were vacuum dried and desalted using a ZipTip procedure before resuspending in ~5-10 μL LC buffer A (0.1% HCOOH (v/v) in 5:95 $CH_3CN$:$H_2O$) for MS analysis.

**Mass Spectrometry Analyses**

Tryptic peptide extracts were analyzed using nLC-MS/MS (Dionex/LCPackings Ultimate nano-LC coupled to a Thermo LCQ Deca XP+ ion trap mass spectrometer) in duplicate. 1 μl of the peptide sample (in LC buffer A, 0.1% HCOOH (v/v) in 5:95 $CH_3CN$:$H_2O$) was first loaded onto a μ-Precolumn (PepMapTM C18, LC-Packings), washed for 4 min at a flow rate of 25 μl/min with LC buffer A, then transferred onto an analytical C18-nanocapillary HPLC column (PepMapAcclaim100). Peptides were eluted at 280 nl/min flow rate with a 120 minute gradient of LC buffers A and B (0.1% (v/v) formic acid in 80:20 acetonitrile:water) ranging from 5%-95% B. A fused silica emitter tip with 8 μm aperture (FS360-75-8-N-5-C12; New Objective) mounted to a Thermo nanospray ionization (NSI) source at 1.8 kV was used for positive ionization of peptides. Mass spectra were collected using Thermo Xcalibur 2.0 software. The top 3 principal ions from each MS scan were trapped and fragmented during the chromatographic gradient, using dynamic exclusion to maximize detection of ions (range 200-2000 m/z). The trapped ions were subjected to collision-induced dissociation (CID) with He, and ~4000 spectra (MS/MS) were collected to cover the entire chromatography elution profile.

**Spectral Data Analyses and Protein ID**

Experimentally collected MS/MS tandem data were searched against the Arabidopsis Proteome Database (NCBI, latest version) using Thermo Proteome Discoverer 1.4 software. The search was restricted to full trypsin digestion with a maximum of 3 missed cleavages and potential modifications for methionine (oxidation) and cysteine (carbamidomethylation); other parameters were standard for LCQ Deca XP+ instrumentation. Peptide filters were set to standard Xcorr vs charge state values; X corr = (1.5, 2.0, 2.25, 2.5) for charges (+1,+2,+3,+4), respectively. Spectral assignments were manually scrutinized to validate the reliability of the protein identifications. Mass spectral data are summarized in Supplemental Table 4. Raw mass spectral data for key peptides can be found at http://gregorylab.bio.upenn.edu/PIPSeq_AtTotalNuc.

***RIP-RT-qPCR***

RNA immunoprecipitaion (RIP) was performed on frozen four-week-old Col-0 leaves as described previously (Kupsch et al., 2012). To begin, the frozen leaves were manually ground and homogenized before crosslinking in nuclear purification buffer (20 mM MOPS (pH = 7), 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA) plus 1% (vol/vol) formaldehyde, rotating at RT for 10 minutes. One molar Glycine (Sigma-Aldrich; St. Louis, MO, USA) was added to a final concentration of 125 mM before an additional five minutes of rotation. The homogenized leaves were then washed twice with PBS followed by lysis and resuspension in RIP buffer (150 mM NaCl, 20 mM Tris (pH=8.6), 1 mM EDTA, 5 mM $MgCl_2$, 0.5% NP40, 1 tablet/10 ml protease inhibitor (Roche; Basel, Switzerland), 0.5 μl/ml RNaseOUT RNaseOUT (Life Technologies; Carlsbad, CA, USA). This lysate was then subjected to 30 min of sonication and centrifugation to remove any remaining precipitate. Eight microliters of α-CP29A (Kupsch et al., 2012) or α-His antibodies (MA1-21315; Thermo Scientific; Waltham, MA, USA) were added to 400 μl aliquots of lysate and incubated while rotating at 4°C. Protein A beads (Life Technologies; Carlsbad, CA, USA) were washed with RIP

buffer and added to the reaction for an additional one hour of rotation at 4°C, followed by four washes with RIP buffer. Immunoprecipitated RNA was then isolated using the miRNeasy mini kit (Qiagen; Valencia, CA, USA) and target specific reverse primers (Table S3) were used for cDNA synthesis using SuperScript II Reverse Transcriptase (Life Technologies; Carlsbad, CA, USA) following the manufacturers protocol. mRNA standards were amplified from Arabidopsis cDNA using the Phusion 2X High Fidelity PCR Master Mix (New England Biolabs; Ipswich, MA, USA) and used to create standard curves of each target during quantitative PCR performed as previously described (Younis et al., 2013).

***Western blotting***

Western blots using lysates from INTACT purified nuclei or 10-day-old seedlings were performed using $\alpha$-ACT8 (1:5,000), $\alpha$-PEPC (1:5,000; 200-4163S; Rockland; Boyertown, PA, USA), $\alpha$-RUBISCO (1:5,000; ab62391; Abcam; Cambridge, MA, USA), ), $\alpha$-BIP1 (1:200; sc-33757; Santa Cruz Biotechnology; Dallas, TX, USA), $\alpha$-CNX1 (1:2,500; AS12 2365; Agrisera; Vännäs, Sweden), $\alpha$-H3 (1:1,000; ab1791; Abcam; Cambridge, MA, USA), or $\alpha$-CP29A (1:5,000) antibodies were performed as previously described (Kupsch et al., 2012).

**SUPPLEMENTAL REFERENCES**

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acid Research *37*, W202-W208.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C.*, et al.* (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics *43*, 956-963.

Deal, R.B., and Henikoff, S. (2010). A Simple Method for Gene Expression and Chromatin Profiling of Individual Cell Types within a Tissue. Developmental Cell *18*, 1030-1040.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Z., S. (2010). agriGO: a GO analysis toolkit for the agricultural community Nucleic Acid Research *38*, W64-W70.

Elliott, R., Li, F., Dragomir, I., Chua, M.M.W., Gregory, B.D., and Weiss, S.R. (2013). Analysis of the Host Transcriptome from Demyelinating Spinal Cord of Murine Coronavirus-Infected Mice. PLoS ONE *8*, e75346.

Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acid Research *35*, W297-W299.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell *38*, 576-589.

Kupsch, C., Ruwe, H., Gusewski, S., Tillich, M., Small, I., and Schmitz-Linneweber, C. (2012). Arabidopsis Chloroplast RNA Binding Proteins CP31A and CP29A Associate with Large Transcript Pools and Confer Cold Stress Tolerance by Influencing Multiple Chloroplast RNA Processing Steps. The Plant Cell *24*, 4266-4280.

Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. (2012). Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. The Plant Cell *24*, 4346-4359.

Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Research *22*, 1184-1195.
Muiño, J.M., Kaufmann, K., van Ham, R.C.H.J., Angenent, G.C., and Krajewski, P. (2011). ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. Plant Methods *7*.
Onder, O., Turkarslan, S., Sun, D., and Daldal, F. (2008). Overproduction or absence of the periplasmic protease DegP severely compromises bacterial growth in the absence of the dithiol: disulfide oxidoreductase DsbA. Mol Cell Proteomics *7*, 875-890.
Onder, O., Yoon, H., Naumann, B., Hippler, M., Dancis, A., and Daldal, F. (2006). Modifications of the lipoamide-containing mitochondrial subproteome in a yeast mutant defective in cysteine desulfurase. Mol Cell Proteomics *5*, 1426-1436.
Sammeth, M., Foissac, S., and Guigo, R. (2008). A General Definition and Nomenclature for Alternative Splicing Events. PLoS Computational Biology *4*, e1000147.
Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., Milos, P.M., Barton, G.J., and Simpson, G.G. (2012). Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. Nature Structural and Molecular Biology *19*, 845-852.
Silverman, I.M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J.L., and Gregory, B.D. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biology *15*, R3.
Tong, C., Wang, X., Yu, J., Wu, J., Li, W., Huang, J., Dong, C., Hua, W., and Liu, S. (2013). Comprehensive analysis of RNA-seq data reveals
the complexity of the transcriptome in
Brassica rapa. BMC Genomics *14*, 689.
von Arnim, A.G., Jia, Q., and Vaughn, J.N. (2014). Regulation of plant translation by upstream open reading frames. Plant Science *214*, 1-12.

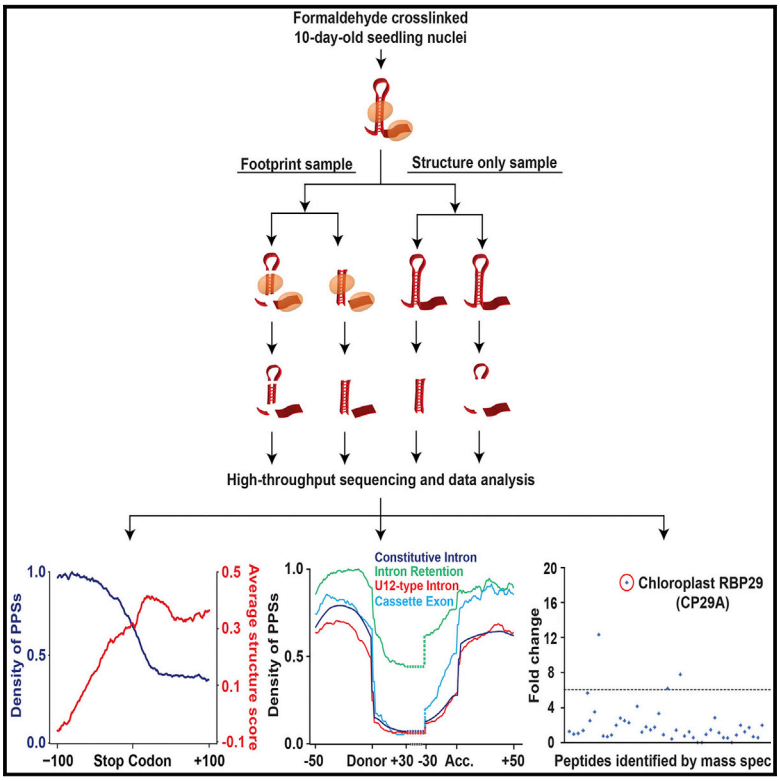
Younis, I., Dittmar, K., Wang, W., Foley, S.W., Berg, M.G., Hu, K.Y., Wei, Z., Wan, L., and Dreyfuss, G. (2013). Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. eLife *2*, e00780.

Resource

Molecular Cell

# Global Analysis of the RNA-Protein Interaction and RNA Secondary Structure Landscapes of the *Arabidopsis* Nucleus

## Graphical Abstract




## Highlights

- Patterns of RNA secondary structure and RBP binding are anticorrelated
- Alternative splice sites have distinct RBP binding and secondary structure profiles
- Groups of *Arabidopsis* RBP-bound motifs co-occur on functionally related mRNAs
- The chloroplast RBP CP29A also interacts with nuclear mRNAs

## Authors

Sager J. Gosai, Shawn W. Foley, ..., Roger B. Deal, Brian D. Gregory

## Correspondence

bdgregor@sas.upenn.edu

## In Brief

RNA secondary structure and RNA-protein interactions regulate all aspects of a transcript's life cycle. Using a ribonuclease-mediated protein-footprinting approach, Gosai et al. provide a simultaneous genome-wide view of RNA-protein interaction sites and RNA secondary structure in the *Arabidopsis* nucleus.

## Accession Numbers

GSE58974

Gosai et al., 2015, Molecular Cell *57*, 376–388
January 22, 2015 
http://dx.doi.org/10.1016/j.molcel.2014.12.004

CellPress

Molecular Cell
Resource

# Global Analysis of the RNA-Protein Interaction and RNA Secondary Structure Landscapes of the *Arabidopsis* Nucleus

Sager J. Gosai,[1,5] Shawn W. Foley,[1,2,5] Dongxue Wang,[3] Ian M. Silverman,[1,2] Nur Selamoglu,[1] Andrew D.L. Nelson,[4] Mark A. Beilstein,[4] Fevzi Daldal,[1] Roger B. Deal,[3] and Brian D. Gregory[1,2,*]
[1]Department of Biology
[2]Cell and Molecular Biology Graduate Group
University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Biology, Emory University, Atlanta, GA 30322, USA
[4]School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
[5]Co-first author
*Correspondence: bdgregor@sas.upenn.edu
http://dx.doi.org/10.1016/j.molcel.2014.12.004

## SUMMARY

**Posttranscriptional regulation in eukaryotes requires *cis*- and *trans*-acting features and factors including RNA secondary structure and RNA-binding proteins (RBPs). However, a comprehensive view of the structural and RBP interaction landscape of nuclear RNAs has yet to be compiled for any organism. Here, we use our ribonuclease-mediated structure and RBP-binding site mapping approaches to globally profile these features in *Arabidopsis* seedling nuclei in vivo. We reveal anticorrelated patterns of secondary structure and RBP binding throughout nuclear mRNAs that demarcate sites of alternative splicing and polyadenylation. We also uncover a collection of protein-bound sequence motifs, and identify their structural contexts, co-occurrences in transcripts encoding functionally related proteins, and interactions with putative RBPs. Finally, using these motifs, we find that the chloroplast RBP CP29A also interacts with nuclear mRNAs. In total, we provide a simultaneous view of the RNA secondary structure and RBP interaction landscapes in a eukaryotic nucleus.**

## INTRODUCTION

RNA molecules are bound throughout their life cycle by dynamic complexes of proteins that regulate their splicing, polyadenylation, nuclear export, localization, translation, and degradation (Bailey et al., 2009). These RNA-binding proteins (RBPs) interact with their targets in a sequence- and secondary structure-specific manner (Cruz and Westhof, 2009). Therefore, both the bound RBPs and secondary structure are key regulatory features of these molecules (Ding et al., 2014; Li et al., 2012a, 2012b). For instance, recent studies have linked secondary structure of mRNA to translation efficiency, stability, splicing regulation, and polyadenylation (Ding et al., 2014; Li et al., 2012a, 2012b; Zheng et al., 2010).

Due to the importance of RNA secondary structure in eukaryotic posttranscriptional processing and regulation, several high-throughput approaches have been developed to globally profile single- and double-stranded RNAs (ssRNAs and dsRNAs, respectively) (Rouskin et al., 2014; Zheng et al., 2010). For example, ss- and dsRNA-seq employ single- and double-stranded RNases (ssRNases and dsRNases, respectively) to provide direct evidence for both single- and double-stranded regions of the transcriptome (Li et al., 2012a, 2012b; Zheng et al., 2010). Alternatively, dimethylsulfate sequencing (DMS-seq) is a technique where samples are treated with DMS, which specifically modifies unpaired adenines (As) and cytosines (Cs) resulting in the termination of reverse transcriptase products, providing evidence for unpaired As and Cs in RNAs (Ding et al., 2014; Rouskin et al., 2014). However, recent studies have demonstrated that DMS modification is obstructed at RBP-binding sites (Talkish et al., 2014), making protein-bound regions indistinguishable from truly structured regions of RNAs.

Most studies of RBP-RNA interactions identify the binding partners of a single protein of interest. This is often accomplished by crosslinking and immunoprecipitation (CLIP) (Ule et al., 2003), in which RNA-protein interactions are crosslinked via UV irradiation followed by immunoprecipitation of a protein of interest. Recently, two methods have reported development of unbiased approaches to study RNA-RBP binding (Baltz et al., 2012; Silverman et al., 2014). Protein interaction profile sequencing (PIP-seq) crosslinks RNA-protein interactions via formaldehyde and subsequently digests ssRNA and dsRNA using structure-specific RNases before high-throughput sequencing, providing a global view of both RNA secondary structure and RBP-bound RNA sequences across the transcriptome (Silverman et al., 2014). Additionally, global photoactivatable ribonucleoside CLIP (gPAR-CLIP) utilizes the incorporation of a synthetic nucleotide into RNAs to identify RNA-protein crosslinking events after exposure to long-wave UV radiation (Baltz et al., 2012). To date,

there have been no global studies of either RBP binding or RNA secondary structure performed in the nucleus of any organism.

All aspects of posttranscriptional mRNA maturation are tightly controlled by RNA-protein interactions acting to positively or negatively regulate recruitment of catalytic molecular machines. For instance, splicing is performed by one of two large complexes, the U2- or U12-type spliceosomes, which identify and excise ~170,000 or ~1,800 introns in *Arabidopsis*, respectively (Marquez et al., 2012). In addition to being regulated by multiple spliceosomes, pre-mRNA transcripts can undergo alternative splicing (AS), resulting in mature mRNAs of different sequences (Wahl et al., 2009). In *Arabidopsis*, over 60% of introns are alternatively spliced, with failure to excise an intron (intron retention [IR]) or exclusion of an exon (exon skipping/cassette exon [CE]) in specific isoforms comprising > 64% of these events (Marquez et al., 2012). Additionally, more than 70% of *Arabidopsis* pre-mRNAs can undergo alternative polyadenylation (APA), resulting in transcript isoforms that differ in their 3′ termini (Hunt et al., 2012; Wu et al., 2011). Previous studies have shown that perturbing RNA secondary structure at alternatively spliced exons can result in decreased RBP recruitment and a shift in spliceoform abundance (Raker et al., 2009). Thus, both AS and APA are important regulatory processes driven by large collections of RBPs and their interactions with specific RNA sequences and structures.

The interplay between RBPs that bind functionally related genes has become a topic of great interest. Recent studies have attempted to identify posttranscriptional operons (Tenenbaum et al., 2011), transcripts with the same gene ontology that are bound by similar populations of RBPs. Thus, the binding of these RBPs would allow coregulation of genes encoding functionally related proteins. Evidence for posttranscriptional operons has been seen in human cells (Silverman et al., 2014); however, this analysis has yet to be performed in *Arabidopsis*.

Here, we simultaneously profile the global landscapes of RBP binding and RNA secondary structure in nuclei of 10-day-old *Arabidopsis* seedlings using our PIP-seq and structure-mapping approaches. In total, this study produces an unbiased view of RBP binding and RNA secondary structure for a nuclear transcriptome, providing a rich resource for future hypothesis generation and testing.

## RESULTS AND DISCUSSION

### PIP-seq on Purified *Arabidopsis* Seedling Nuclei

To probe the RNA-RBP interaction site and RNA secondary structure landscapes of the *Arabidopsis* nucleus, we performed PIP-seq (Silverman et al., 2014) on total nuclei from 10-day-old seedlings. The nuclei were crosslinked with formaldehyde prior to purification via the isolation of nuclei in tagged cell types (INTACT) approach (Deal and Henikoff, 2010). We confirmed nuclei purity by direct imaging (Figure S1A available online), revealing only DAPI-stained nuclei bound to the streptavidin-coated beads. Additionally, we found an enrichment of the nuclear histone H3 protein and undetectable levels of the mostly cytoplasmic ACT8 (Kandasamy et al., 1999), the ER-localized BIP1 and CNX1, and chloroplastic RUBISCO and PEPC proteins in our INTACT-purified nuclei preparations (Figure S1B), confirming that there is no chloroplastic, ER, or cytoplasmic contamination. We used ~2 million of these highly pure nuclei for each of two PIP-seq replicates, which were split into footprinting and structure-only samples (four total libraries per replicate) (Figure 1A). Our structure-only samples provide in vivo structure data, and additionally serve as a background to our footprinting samples accounting for regions that are insensitive to the structure-specific RNases.

Footprint samples were directly treated with either an ss- or dsRNase (see Experimental Procedures). In contrast, the structure-only samples first had proteins denatured in SDS and degraded with Proteinase K prior to RNase digestion. Denaturation of RBPs before RNase treatment will make protein-bound sequences in the footprinting sample accessible to RNases in these reactions. Thus, RBP-bound sequences were enriched in footprinting relative to structure-only samples (Figure 1B). Additionally, analysis of the structure-only samples as previously described (Li et al., 2012a) allowed us to determine the native (protein-bound) RNA base-pairing probabilities for the *Arabidopsis* nuclear transcriptome (example shown in Figures 1C–1E).

The resulting high-quality PIP-seq libraries (Figures S2A and S2B) were sequenced and provided ~24–38 million raw reads per library. To determine reproducibility, we used a 50 nucleotide (nt) sliding window to define the correlation of nonredundant sequence read abundance between biological replicates of footprinting and structure-only libraries. We observed a high correlation in read counts between all footprinting and structure-only libraries (Pearson correlation > 0.81) (Figures S3A–S3D). Similarly, principle component analysis of read coverage in 500 nt bins revealed that replicates of each library type clustered together (Figure S3E), further indicating the high quality and reproducibility of our PIP-seq libraries.

### The RNA-Protein Interaction Landscape of the *Arabidopsis* Nucleus

To identify protein-protected sites (PPSs), we used a Poisson distribution model to identify enriched regions in the footprinting compared to the structure-only libraries at a false-discovery rate of 5% as previously described (Silverman et al., 2014) (Figure 1B). We identified 61,632 total PPSs in our experiments, 64.7% of which overlap between the two replicates (Figure 2A). Consolidation of all PPSs yields 40,131 distinct sites (Table S1) with an average size of 68 nt (Figure S4A). This reproducibility is much higher than many CLIP-seq experiments, which often produce < 35% overlap between replicates (Lebedeva et al., 2011). The majority of PPSs were identified by the dsRNase (~30,000 PPSs) as compared to the ssRNase (~10,000 PPSs) (Figures S4B and S4C) treatment, with ~50% of the sites uncovered by the ssRNase overlapping those from the dsRNase libraries (Figures S4D and S4E).

Given the high reproducibility between our PIP-seq replicates (Figures 2A, S3, and S4), we focused on the complete set of 40,131 distinct PPSs for all subsequent analyses. To estimate the functional relevance of these nuclear PPSs, we compared flowering plant PhastCons conservation scores (Li et al., 2012b) for PPSs versus same-sized flanking regions. We found that PPS sequences were significantly (p values < $1\times10^{-200}$, Kolmogorov-Smirnov test) more evolutionarily

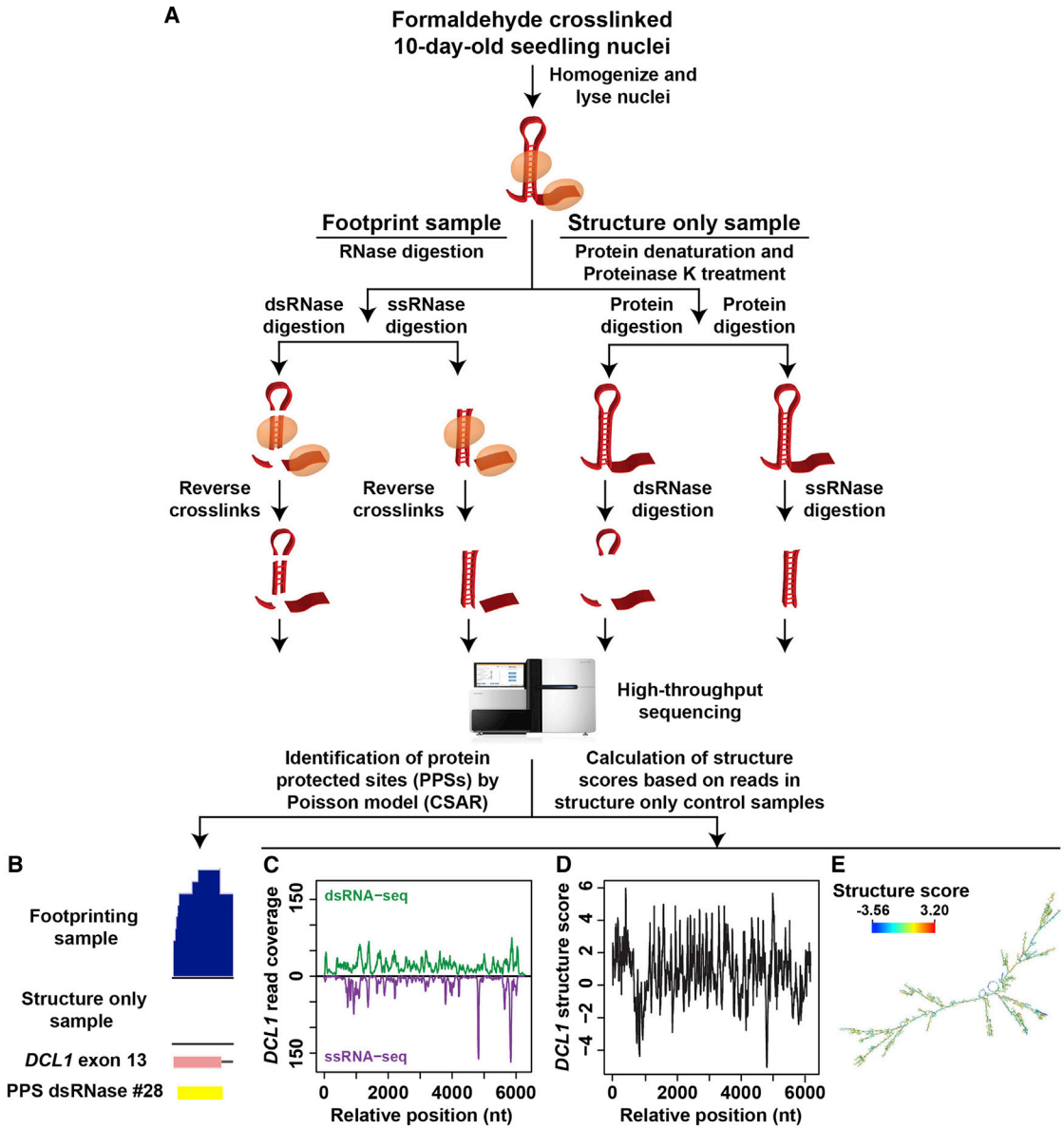



**Figure 1. Overview of PIP-seq in *Arabidopsis* Nuclei**

(A) The PIP-seq approach in the *Arabidopsis* nucleus. Nuclei were purified from 10-day-old *Arabidopsis* seedlings that were crosslinked using a 1% formaldehyde solution. Nuclei were lysed and separated into footprinting and structure-only samples. Four total sequencing libraries were then prepared for each replicate experiment as previously described (Silverman et al., 2014).

(B) An example of PPS identification (dsRNase #28) in exon 13 of *DCL1*.

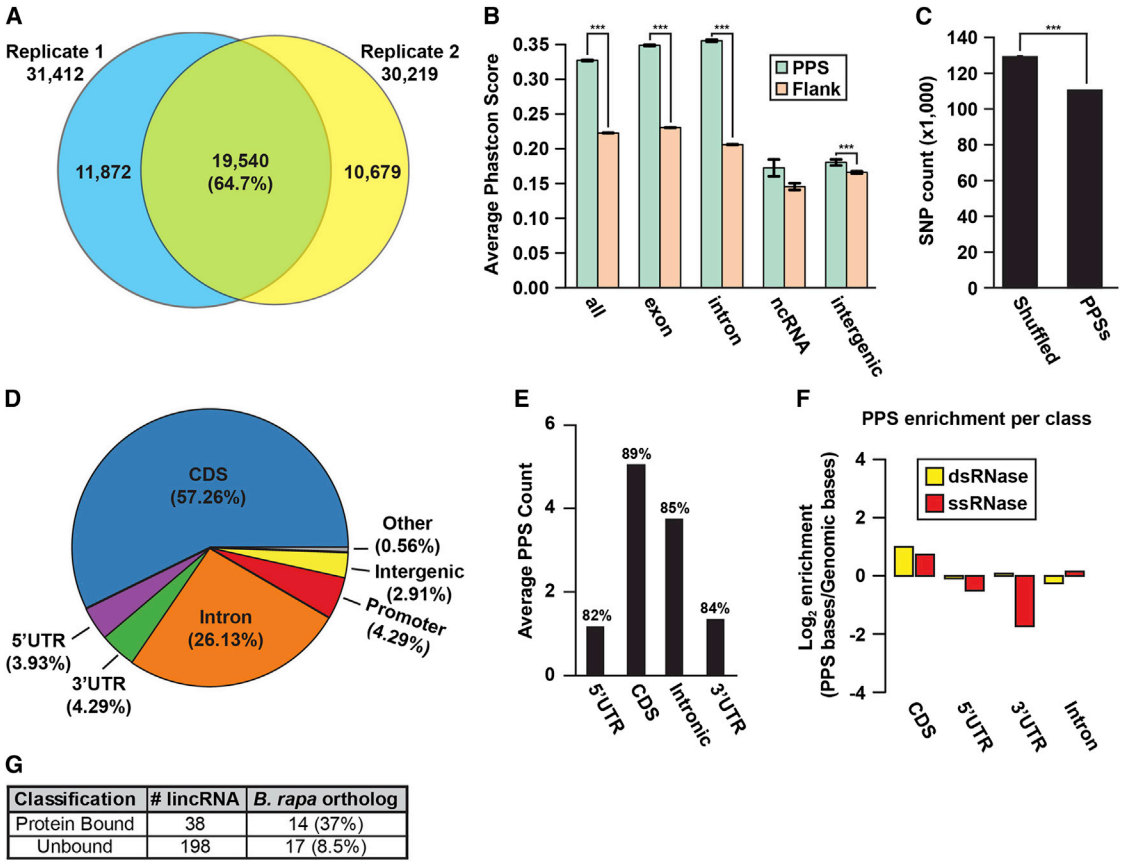
(C) Read coverage across the *DCL1* transcript for the ds- (top, green line) and ssRNA-seq (bottom, purple line) structure-only samples.

(D) Structure scores for the *DCL1* transcript based on read coverage seen in (C).

(E) mRNA secondary structure model for *DCL1* determined using our methodology. See also Figures S1–S3.

conserved than flanking regions (Figures 2B and S4F). Importantly, this was true for PPS sequences in both exonic and intronic portions of the nuclear collection of mature and pre-mRNA transcripts (nuclear mRNAs), but not for ncRNAs (Figure 2B). These results support the notion that nuclear mRNA sequences are constrained by their ability to interact with RBPs, while decreased PPS conservation within ncRNAs is consistent with their low conservation rates across plant species (Liu et al., 2012).

We also reasoned that functional RBP-interacting sequences would contain less nucleotide diversity across closely related strains when compared to an equal number of same-sized regions randomly selected from detected transcripts. To address this, we used data from the 1001 Genomes Project,




| Classification | # lincRNA | *B. rapa* ortholog |
|---|---|---|
| Protein Bound | 38 | 14 (37%) |
| Unbound | 198 | 17 (8.5%) |

**Figure 2. Characterization of *Arabidopsis* Nuclear PPSs**

(A) Overlap between PPSs identified from two replicate nuclear PIP-seq experiments.

(B) Comparison of average PhastCons scores between PPSs (green bars) and equal-sized flanking regions (orange bars) for various genomic regions. *** denotes p value < $1\times10^{-10}$, Kolmogorov-Smirnov test. Error bars, ± SEM.

(C) Analysis of the total number of SNPs identified by the 1001 Genomes Project (Cao et al., 2011) in PPSs compared to a shuffled background control. *** denotes p value < $1\times10^{-10}$, $\chi^2$ test. Error bars, ± SD.

(D) Absolute distribution of PPSs throughout various RNA species and transcript regions.

(E) Average PPS count per pre-mRNA transcript region. Percentages indicate the fraction of annotated RNAs that contain sequencing information for that region.

(F) Genomic enrichment of PPS density, measured as $\log_2$ enrichment of the fraction of PPS base coverage normalized to the fraction of genomic bases covered by indicated nuclear mRNA regions for the dsRNase (yellow bars) and ssRNase (red bars) libraries.

(G) Breakdown of bound compared to unbound nuclear lincRNAs that are conserved between *Arabidopsis thaliana* and *Brassica rapa*. See also Figure S4 and Table S1.

which has cataloged naturally occurring single-nucleotide polymorphisms (SNPs) between 80 strains of *Arabidopsis thaliana* (Cao et al., 2011). We found a significant (p value < $2.2\times10^{-16}$, $\chi^2$ test) decrease in nucleotide diversity within PPSs compared to shuffled regions (Figure 2C). Therefore, *Arabidopsis* PPSs resist the effects of random genetic drift occurring in the numerous populations across the globe, indicating their functional significance.

A classification of all distinct PPSs revealed the majority of these sites were located in nuclear mRNAs, with the largest fractions occupying the coding sequence (CDS) (57.3%) and introns (26.1%) (Figure 2D). Closer examination of PPSs broken down by genic features (e.g., 5′ and 3′ UTR, CDS, and intron) revealed that detected *Arabidopsis* nuclear mRNAs contained multiple binding events in both the CDS (~5 total/gene) and introns (~4 total/gene), while the 5′ and 3′ UTRs averaged only a single interaction per expressed transcript (Figure 2E).

We then tested the enrichment of PPSs in specific nuclear mRNA regions (e.g., 3′ and 5′ UTRs) normalized to the number of bases annotated as these features in the TAIR10 *Arabidopsis* genome. We found that PPSs identified by both RNases were enriched in CDSs, while being underrepresented in 5′ UTRs (Figure 2F). Interestingly, both introns and 3′ UTRs show opposite enrichment trends for ds- and ssRNases, suggesting that PPSs preferentially occur in more highly or lowly structured regions, respectively. When interrogating the enrichment of PPSs in the CDSs of nuclear mRNAs we found that the intron flanking ends of exons tend to be more protein bound than their middle segments (Figure S5A). This binding suggests that we can detect a high level of splicing factor/machinery binding through nuclear

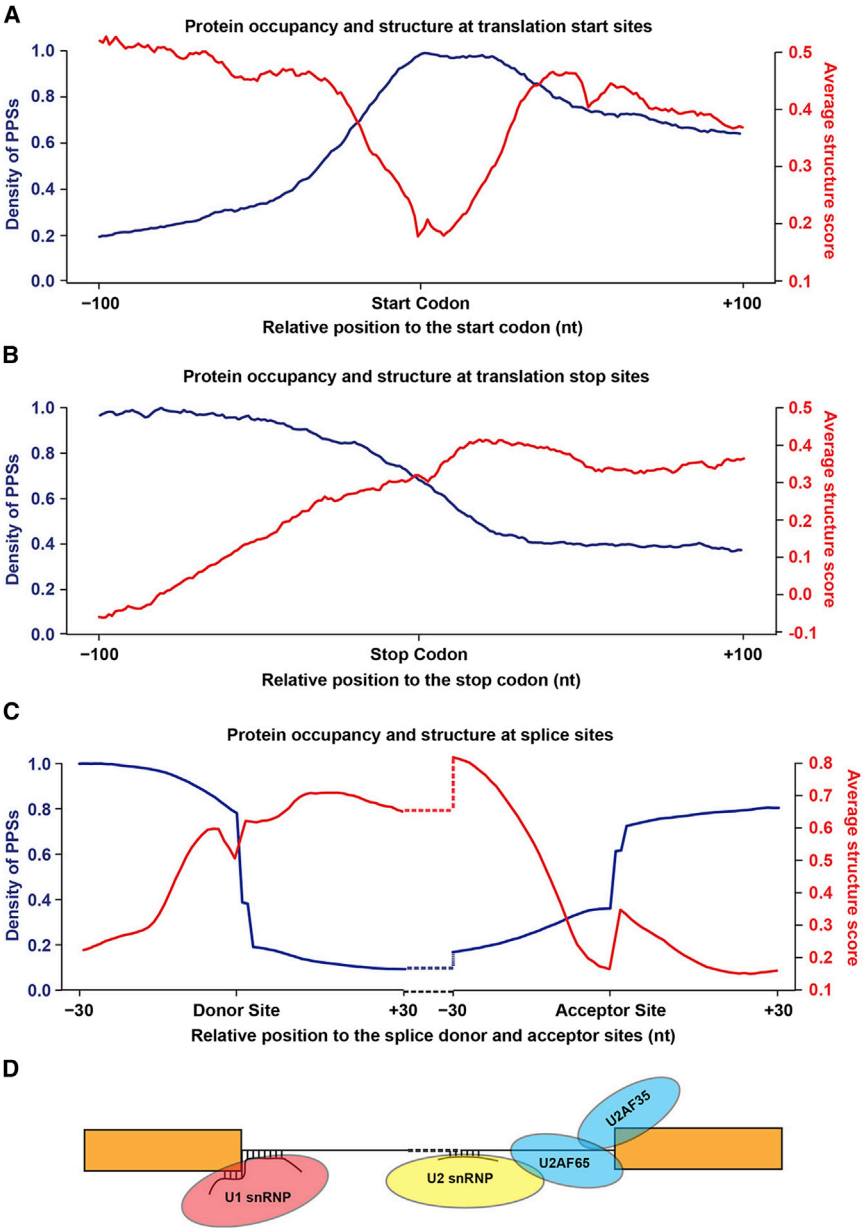



**Figure 3. Patterns of Protein Occupancy and Secondary Structure in Specific Nuclear mRNA Regions**

(A and B) PPS density and structure score profiles for nuclear mRNAs based on our PIP-seq experiments. Average PPS density (blue lines) and structure scores (red lines) at each position ± 100 nt from canonical (A) start and (B) stop codons for *Arabidopsis* nuclear mRNAs.

(C) PPS density and structure score profiles for exon/intron boundaries of nuclear mRNAs. Average PPS density (blue lines) and structure scores (red lines) at each position ± 30 nt from splice donor and acceptor sites.

(D) Model depicting the canonical protein and RNA interactions of the U2-type spliceosome at the splice donor and acceptor sites depicted in (C). See also Figure S5.

PIP-seq as described below. In total, our results reveal that the CDSs of mRNAs are enriched for RBP binding in the *Arabidopsis* nucleus.

Although PPSs in ncRNAs were not conserved, this category consists of many RNA subgroups, thus conserved classes might be obscured. Long intergenic noncoding RNAs (lincRNAs) are a recently discovered class of ncRNAs that are necessary for vertebrate development (Cech and Steitz, 2014; Sauvageau et al., 2013), but are not well characterized in plants (Hacisuleyman et al., 2014; Liu et al., 2012). We examined the relationship between our PIP-seq data and a set of ~2,700 curated lincRNAs in *Arabidopsis* (Liu et al., 2012) to identify nuclear protein-bound RNAs. We detected 236 lincRNAs in our nuclear sequencing data, 38 of which contained one to four PPSs (Figure 2G). We found that these protein-bound lincRNAs were significantly (p value < $4.5\times10^{-30}$, $\chi^2$ test) more conserved within the related crop species *Brassica rapa* (37%, 14 total) as compared to unbound nuclear lincRNAs (8.5%, 17 total) (Figure 2G). The combination of nuclear protein binding and conservation in *B. rapa* suggests that RBP-bound nuclear lincRNAs have important functions in plant systems.

## Patterns of RNA Secondary Structure and RBP Binding Are Anticorrelated

To interrogate the landscape of RBP binding and RNA secondary structure in specific regions of nuclear mRNAs, we calculated the structure scores and PPS densities and examined the average profiles for all detectable transcripts. The structure score is a generalized log ratio of dsRNA-seq to ssRNA-seq reads at each nucleotide position, with positive and negative scores indicating ds- and ssRNA, respectively (see Supplemental Experimental Procedures). To examine the relationship between PPS density and structure score, we focused on the boundaries between the UTRs and CDS of nuclear mRNAs. We observed the highest PPS density in the CDS with decreased occupancy within the 5′ and 3′ UTRs (Figures 3A and 3B), consistent with the gross PPS localization and enrichment analysis (Figures 2D–2F). Interestingly, we observed significantly (p value < $8.2\times10^{-32}$, Wilcoxon test) higher levels of protein binding directly over the start codon (Figure 3A) relative to flanking regions. Similarly, we examined the start codons at high-confidence upstream open reading frames (uORFs) (von Arnim et al., 2014) and found a significant (p value < 0.01, Wilcoxon test) increase in PPS density

over uORF start codons relative to the upstream flanking region (Figure S5B). Similar increases in PPS density over the start and stop codon were speculated to be due to ribosome binding (Baltz et al., 2012; Silverman et al., 2014). However, the nuclear preparations used in this study are free of the cellular compartments containing functional ribosomes (cytoplasm and ER) (Figure S1B), and RBP-binding profiles for transcripts that are not translated in the rough ER (Figure S5C) or are unspliced and likely localized in the nucleus (Figure S5D) demonstrate very similar protein-binding profiles. Taken together, these results suggest that one or more nuclear RBPs occupy this region.

In contrast to RBP occupancy, we found that secondary structure was higher in both UTRs compared to the CDS at the regions analyzed, with a significant (p values $< 0.05$, Wilcoxon test) dip directly over uORF and canonical start codons, as well as upstream of the stop codon, as observed previously (Ding et al., 2014; Li et al., 2012b) (Figures 3A, 3B, and S5B). Thus, these structural characteristics at the start and stop codons seem to be a consistent feature of both *Arabidopsis* nuclear and mature mRNAs. Interestingly, our analyses revealed that secondary structure and PPS density are anticorrelated to one another. Specifically, we looked at both PPS density and structure score simultaneously, and found a significant (p value $< 2.2\times10^{-16}$, asymptotic *t* approximation) anticorrelation (Spearman's rho $< -0.82$) between these metrics at both canonical start and stop codons. Although the correlation is milder (likely due to fewer instances), there is a significant (p value $< 3.6\times10^{-9}$, asymptotic *t* approximation) negative correlation (Spearman's rho $< -0.55$) for uORF start codons as well.

It is worth noting that although the majority of PPSs were identified in the dsRNase-treated samples, this does not necessitate that the interacting RBPs are binding dsRNA. In support of this hypothesis, we found that more highly structured regions generally surrounded PPSs, with a lower average structure score directly over the RBP-bound sequence (Figure S6A). Although the dsRNase-identified PPSs have a significantly (p value $< 2.2\times10^{-16}$, Wilcoxon test) higher average structure score than those uncovered by the ssRNase (Figure S6A), the dip in structure score directly over these regions suggests that they can be ds- and/or ssRNAs. Taken together, these results suggest that many *Arabidopsis* RBPs bind ssRNA flanked by structured regions.

It should also be noted that the higher overall structure of the UTRs compared to the CDS is opposite to what has been observed previously both in vivo and in vitro when profiling total (mostly mature cytoplasmic) RNA in *Arabidopsis* (Ding et al., 2014; Li et al., 2012b). Together, these results suggest that the structural landscape of the nucleus is distinct from that of the cytoplasm. These differences in secondary structure in specific cellular locales will need to be further investigated.

As we were probing the nuclear transcriptome, we next examined the PPS density and structure scores across all TAIR10 annotated splice donor and acceptor sites (Figure 3C). We first determined that the RNA population consisted of a high percentage of unspliced pre-mRNA. Specifically, we found that ~40% of reads mapping to the first and last constitutively spliced intron junctions cross the exon-intron boundary in total RNA sequencing data sets from congruently purified nuclei (see Supplemental Experimental Procedures), suggesting comparable levels of spliced and unspliced transcripts in our data sets (Figure S6B). Despite the large percentage of detectable unspliced transcripts (pre-mRNAs), exonic and intronic regions cannot be directly compared due to slightly lower read coverage in introns. Therefore, we first compared 30 nt regions up- or downstream of acceptor and donor intron sites, respectively, and found that the 3′ end of introns had significantly (p value $< 1\times10^{-30}$, Wilcoxon test) higher protein binding relative to the 5′ end. These results are consistent with the U2 auxiliary factors (U2AFs) occupying the acceptor splice site (Wahl et al., 2009). Intriguingly, there were distinct patterns of secondary structure at both the splice donor and acceptor sites (Figure 3C). Upstream of the donor site, we observed a dramatic decrease in secondary structure from nt −3 to −1, corresponding to the U1 snRNA binding site (−3 to +8) (Chiou et al., 2013). This dip in secondary structure mirrors what we have seen over the translation start codon (Figure 3A), revealing that this region is more accessible to intermolecular RNA pairing than flanking sequences, perhaps facilitating binding of the U1 snRNA. Additionally, we found a drop in secondary structure immediately upstream of the splice acceptor site, suggesting an increased accessibility to U2AFs and other splicing factors in this region (Wahl et al., 2009) (Figure 3D).

We again observed opposing patterns of secondary structure and PPS density at all regions examined in these analyses (Figure 3C). Specifically, we found that this anticorrelation (Spearman's rho $< -0.93$) between PPS density and RNA secondary structure was significant (p value $< 2.2\times10^{-16}$, asymptotic *t* approximation) at regions flanking the acceptor sites, as well as the upstream exonic sequence at donor sites. The proximal intronic region at donor sites had a milder (Spearman's rho $< -0.38$), but still significant (p value $< 0.05$, asymptotic *t* approximation) anticorrelation between structure score and PPS density, which may be due to the intermolecular base pairing between the U1 snRNA and the intron (Figure 3D) that occurs at 8 of the 30 nt probed. In total, our findings reveal that RBP binding and RNA secondary structure are anticorrelated features in the *Arabidopsis* nuclear transcriptome.

## Distinct RNA Secondary Structure and RBP-Binding Profiles Demarcate AS and Polyadenylation Sites

The specific patterns of RBP binding and RNA secondary structure at exon/intron boundaries suggest that these features may also have distinct distributions at sites of AS. Therefore, we compared the profiles for these two features at several types of alternatively spliced exons. To do this, we used ASTALAVISTA (Foissac and Sammeth, 2007) to annotate AS events in the TAIR10 transcript assembly, and isolated all examples of CE and IR. We also focused on TAIR10 introns that have been previously described as U12-type splice sites (Marquez et al., 2012). We compared average PPS density and structure score for 50 nt in the exonic region and 30 nt in the intronic sequence at both the splice donor and acceptor sites for these splicing events (Figure 4A). We found that IR events have significantly (p values $< 4.3\times10^{-7}$, Wilcoxon test) higher PPS density in the 40 nt upstream (−40 to −1) of the splice donor, while CE and U12-type introns do not significantly (p value $> 0.05$, Wilcoxon test) differ from constitutive introns. This trend for

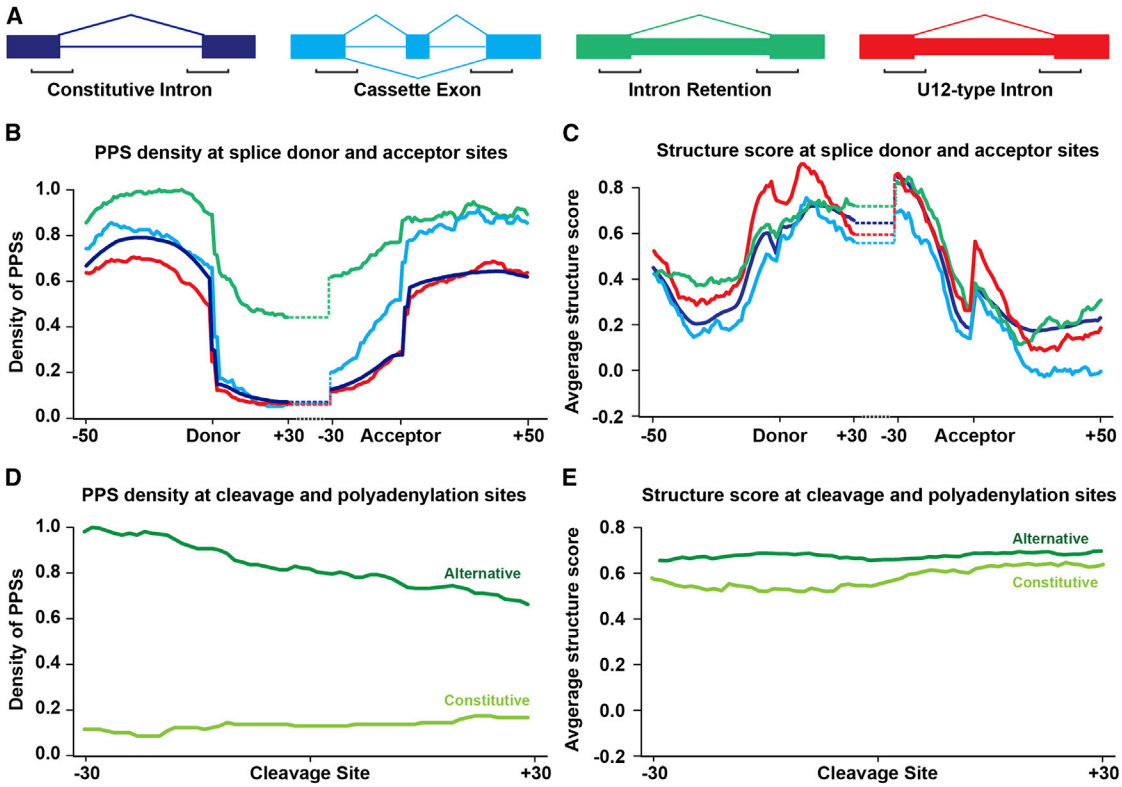



**Figure 4. Protein Occupancy and Secondary Structure Landscapes at Alternative Splicing and Polyadenylation Sites**

(A) Diagram of constitutive introns (blue), cassette exons (turquoise), intron retention events (green), and U12-type introns (red). Large boxes represent exons, lines represent constitutive introns, and small boxes represent alternatively spliced sequences, with the black brackets indicating the regions graphed in (B) and (C) for reference.

(B) PPS density profiles for constitutive and alternative splicing events in *Arabidopsis*. Average PPS density at each position −50 to +30 nt at the donor splice site, and −30 to +50 nt at the acceptor splice site. Line colors correspond to examples shown in (A).

(C) Structure score profiles for constitutive and alternative splicing events in *Arabidopsis* covering the same regions as (B). Line colors correspond to examples shown in (A).

(D) PPS density profiles for constitutive and alternative poly(A) sites of nuclear mRNAs. Average PPS density at each position ± 30 nt from constitutive (light-green line) and alternative (dark-green line) cleavage and polyadenylation sites.

(E) Average structure score profiles for constitutive and APA sites covering the same regions as (D). See also Figure S6.

increased PPS density continues in IR events 30 nt into the intron at splice donor sites, with these events showing ~4.5-fold higher protein binding than constitutive introns (p value < $1.9\times10^{-44}$, Wilcoxon test) (Figure 4B). The increased binding within these introns is consistent with the presence of intronic splicing silencers, *cis* elements that recruit proteins to inhibit spliceosome assembly (Chen and Manley, 2009). We observed increased PPS density at the splicing acceptor for both CE and IR sites in the downstream exon (p values < $6.7\times10^{-6}$, Wilcoxon test) and in the 30 nt of intron directly upstream of this splice site (p values < 0.001, Wilcoxon test) (Figure 4B). This can likely be explained by recruitment of RBPs through a combination of both positive and negative *cis* regulatory elements, such as exonic splicing silencers to induce exon skipping, and intronic splicing enhancers to increase inclusion, working additively to regulate each exon in a cell type-specific manner (Chen and Manley, 2009). These same trends are observed when specifically examining CE and IR events with adjacent constitutive exons (Figures S6C and S6D). In total, these results reveal that IR and CE events can be differentiated from one another based on the patterns of protein binding density just up- and downstream of both splice sites.

We next probed the structural profiles for each of these subsets of introns across splice sites (Figure 4C). The most striking feature we observed was the dramatic difference in overall profile shape between U12-type introns and constitutive introns upstream of the donor splice site (−16 to −1). We found a significantly (p value < 0.01, Wilcoxon test) higher structure score for these introns in this region, which have a PPS profile that is indistinguishable from constitutive introns. This structural profile likely influences the identity of the proteins binding this region (Cech and Steitz, 2014), resulting in distinct RBP populations at each type of intron. Additionally, IR events are also significantly (p value < $4.5\times10^{-3}$, Wilcoxon test) more structured 40 nt upstream of the donor splice site (−40 to −1). Specifically, these profiles reveal highly structured regions that are associated with increased binding levels of regulatory proteins. Thus, in both U12-type and IR events, the increased structure in specific

regions likely limits the accessibility of binding sites to specific RBPs allowing for a tighter control over the splicing machinery. Interestingly, CEs are the only subset of events that are consistently less structured than constitutive introns. This trend is only statistically significant (p value < 0.05, Wilcoxon test) upstream of the acceptor site (−30 to −1), but the analysis is limited by a low number of annotated events (< 700) (Figure 4C). Constitutive exon-flanked CE and IR events exhibit similar patterns (Figure S6E). In total, these results reveal that each of these three subtypes of AS has a distinct combination of PPS and structural profiles, supporting the idea that both structure and protein occupancy are required for their proper regulation.

Addition of the poly(A) tail (polyadenylation) during eukaryotic mRNA maturation is also highly regulated. Therefore, we calculated average PPS density and structure score 30 nt up- and downstream of expressed transcripts with constitutive or APA sites (Sherstnev et al., 2012). We found that APA events were on average 3.7-fold (p value < $4.8\times10^{-16}$, Wilcoxon test) more protein bound up- and downstream of the cleavage site as compared to constitutive events (Figure 4D). Interestingly, there is no significant (p value > 0.05, Wilcoxon test) difference in structure scores between the alternative and constitutive sites (Figure 4E), revealing that this differential protein binding is independent of secondary structure. These results indicate that APA sites do not exhibit altered secondary structure compared to constitutive sites; however, the increased protein binding could be used to differentiate these two types of events from one another.

## The Structural Landscape of Protein-Bound RNA Motifs

To identify RBP-bound motifs, we employed the motif finding algorithms MEME (Bailey et al., 2009) and HOMER (Heinz et al., 2010) on PPSs partitioned by specific region (e.g., CDS) or on the entire collection, respectively. We identified one GAN repeat motif by MEME that was common to both the CDS and 5′ UTR (Figure 5A), while HOMER identified 40 octamers that were significantly (p values < $10^{-7}$) enriched in our PPSs (Table S2), of which we further characterized four of the most significantly enriched (p values < $1.0\times10^{-67}$) (Figures 5B–5E).

We identified the percentage of PPS-bound and -unbound motif occurrences in specific regions of nuclear mRNAs normalized by their overall length in the genome (Figures 5F–5J). Comparing the localization of bound and unbound motif instances revealed stark differences. We saw an overall enrichment of bound sites within the CDS and 5′ UTR. Conversely, the unbound HOMER motif instances were generally more prevalent in introns (Figures 5G–5J), while the 5′ UTR is overrepresented in the unbound GAN repeat occurrences (Figure 5F). In total, these results indicate that within the nucleus RBP binding is enriched within 5′ UTR and CDS instances of specific sequence motifs.

To define the structural context at these five sequence motifs, we calculated average structure scores at the core motif and 50 nt flanking regions for bound and unbound instances. We observed that the five motifs have low structure scores, but are flanked by more structured regions (Figures 5K–5O). As mentioned above, the high levels of this conformation within the nuclear transcriptome may explain increased PPS identification by the dsRNase (Figures S4B, S4C, and S6A). Interestingly, protein-bound instances of all five motifs and their flanking sequences are significantly (p values < $7.3\times10^{-12}$, Wilcoxon test) less structured relative to unbound instances of these sequences (Figures 5K–5O). In total, these findings support the observations that PPSs occur preferentially at less-structured regions of transcripts. Whether this is a cause or consequence of protein binding to these sequence elements will need to be further investigated.

## Evidence of Posttranscriptional Operons in the *Arabidopsis* Nuclear Transcriptome

RBP-interacting motifs often co-occur in functionally related genes in human cells (Silverman et al., 2014), but it is not known if this happens in the *Arabidopsis* nuclear transcriptome. To address this, we interrogated the interactions between protein-bound motifs discovered by our PIP-seq approach. Thus, we identified all bound instances of each identified motif (Table S2) in target RNAs using the HOMER suite (Bailey et al., 2009) on the total set of nuclear PPSs. We then quantified co-occurrences of each pair of these protein-bound motifs within all nuclear mRNAs. We used k-means clustering of the resultant weighted adjacency matrix and identified three clusters of motifs that co-occur on highly similar sets of target transcripts (Figure 6A). Interestingly, Clusters 1 and 2 have only five and four motifs, respectively, while Cluster 3 consisted of the remaining 32 motifs, although no transcripts contained more than four of these co-occurring PPS-bound motifs. The number of transcripts containing at least three bound motifs within each cluster varied greatly, with Clusters 2 and 3 having 188 and 204 transcripts, respectively, while Cluster 1 had the most co-occurring bound motifs with 5,887. These findings indicate that many *Arabidopsis* transcripts contain numerous RBP-interacting motifs.

We used agriGO (Du et al., 2010) to interrogate overrepresented biological processes for these collections of RNAs with co-occurring RBP-bound motifs (Figure 6A). We found that the most highly overrepresented functional terms were related to distinct processes, including cell death/apoptosis and postembryonic and organ development (Cluster 1); response to desiccation, abscisic acid, and cold (Cluster 2); as well as stress response, posttranslational modification, and mRNA processing (Cluster 3) (Figure 6B). The identification of groups of functionally related transcripts bound by the same collection of RBPs during their nuclear life cycle supports the idea of posttranscriptional operons (Keene and Tenenbaum, 2002; Tenenbaum et al., 2011) functioning in the *Arabidopsis* nucleus.

## CP29A Localizes to the *Arabidopsis* Nucleus

After identifying enriched motifs within our PPS list we used these motifs to identify putative *Arabidopsis* RBPs. To begin, we confirmed that these sequences interact in vitro with specific RBPs using a UV crosslinking assay with radiolabeled RNA probes (from Figures 5A–5E; Table S3) or a scrambled control sequence. We found that each sequence motif interacted with one or more distinct RBPs (Figure S7A). We then used these same probes in RNA-affinity chromatography followed by mass spectrometry analysis. Using this approach with four significant HOMER motifs (Figures 5B–5E), we identified 25 proteins

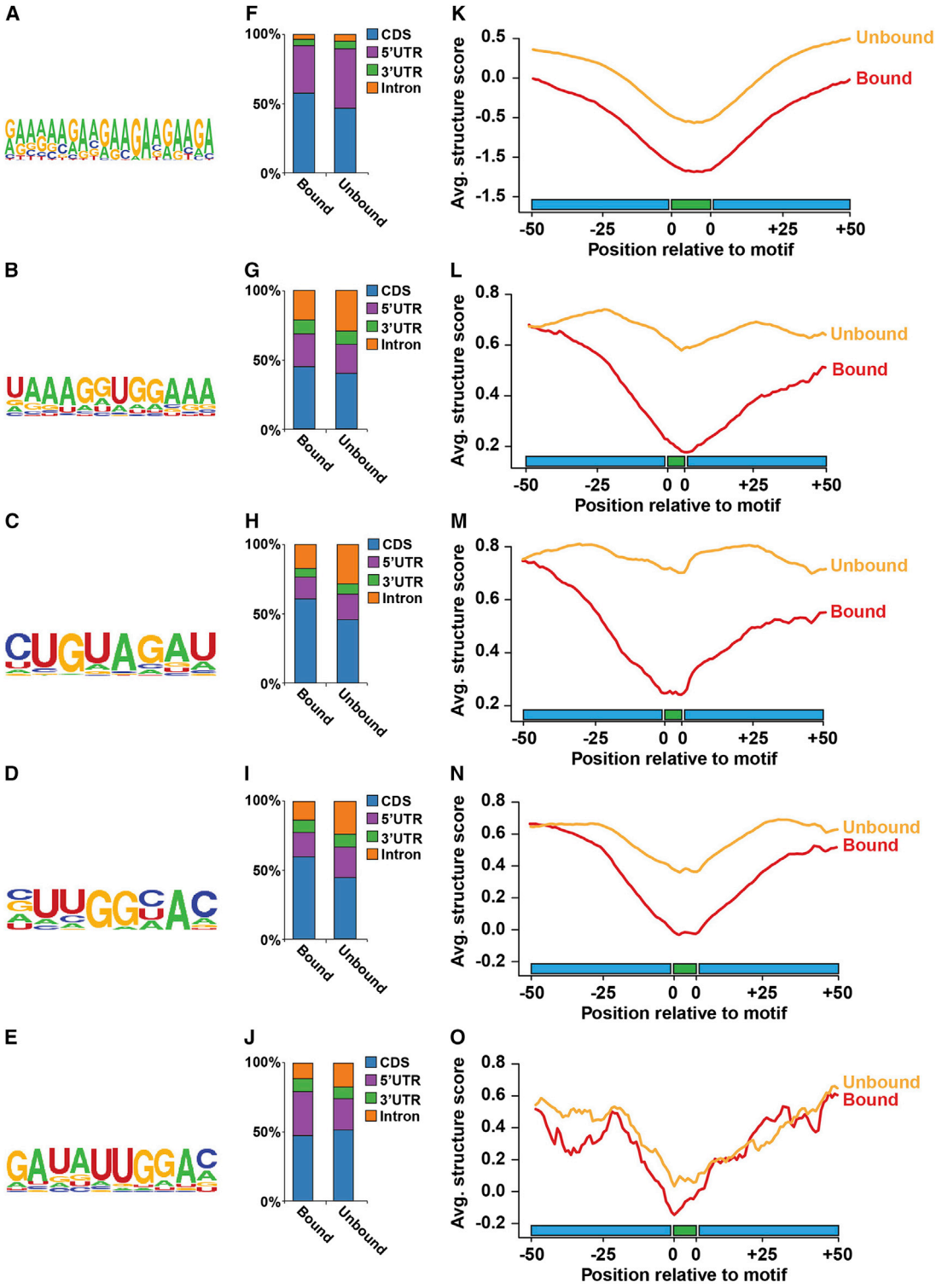


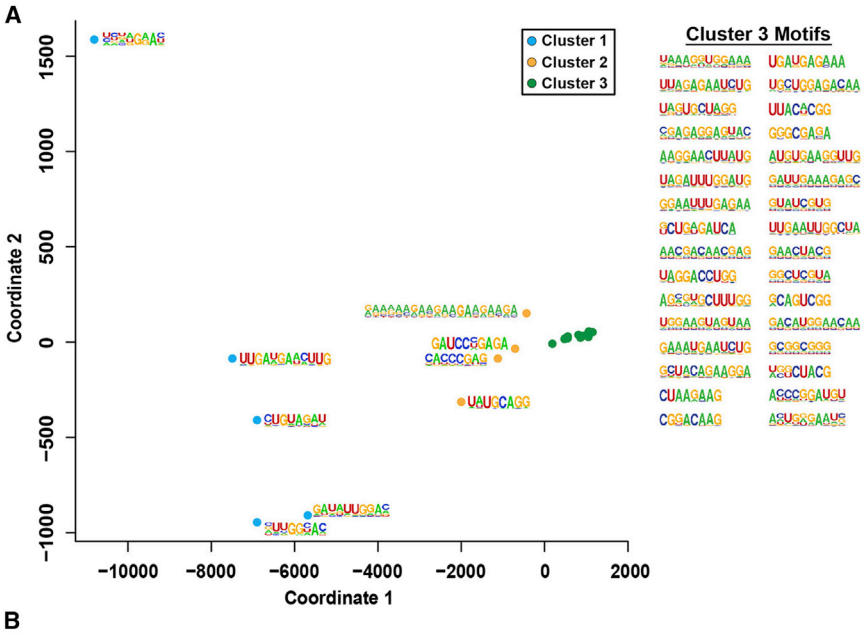

**Figure 5. The Landscape of Protein-Bound RNA Motifs**

(A–E) Overrepresented sequence motifs identified by MEME (A) or HOMER (B–E) analysis of PPS sequences.

(F–J) The relative distribution of protein-bound and unbound motifs from (A) to (E) throughout specific pre-mRNA regions, including the CDS (blue), 5′ UTR (purple), 3′ UTR (green), and intron (orange).

(K–O) Structure score profiles at RBP-binding motifs from (A) to (E). Average structure score at each position ± 50 nt up- and downstream of bound (red lines) and unbound (orange lines) motif occurrences from (A) to (E).

A




B

| Cluster # | GO Annotation | P-value |
|---|---|---|
| Cluster 1 | Post-embryonic development | $2.5 \times 10^{-64}$ |
| | Organ development | $4.2 \times 10^{-17}$ |
| | Cell death/apoptosis | $1.2 \times 10^{-13}$ |
| Cluster 2 | Response to desiccation | $2.8 \times 10^{-10}$ |
| | Response to abscisic acid | $1.1 \times 10^{-5}$ |
| | Response to cold | $2.2 \times 10^{-5}$ |
| Cluster 3 | Stress response | $2.8 \times 10^{-8}$ |
| | Post-translational modification | $1.1 \times 10^{-5}$ |
| | mRNA processing | $1.8 \times 10^{-5}$ |

**Figure 6. Clusters of Motifs Are Present in Functionally Related Genes**

(A) Multidimensional scaling analysis of RBP-bound motif co-occurrence in *Arabidopsis* transcripts. The motifs used for this analysis were identified by HOMER- and MEME-based analyses of PPS sequences. Sequences for all of the motifs used in this analysis can be seen in the figure and found in Table S2. Colored dots indicate cluster membership as defined by k-means clustering (k = 3).

(B) The most significantly enriched biological processes (and corresponding p value) for target transcripts of the specified clusters of motifs identified in (A) where three or more of the motifs are protein-bound and co-occurring. See also Table S2.

with peptides that were enriched over our negative controls, with four proteins that passed a threshold of > 6-fold enrichment for interaction with at least one sequence (Figures S7B and 7A; Table S4). Interestingly, CVP2 as well as the LRR family and DUF544-containing proteins do not have canonical RNA-binding domains (RBDs). This is similar to recent findings in human RBP identification (Baltz et al., 2012; Castello et al., 2012), suggesting that these proteins interact with their target motifs via noncanonical RBDs or an RBP partner.

The GAN repeat motif is of particular interest because it has been linked to splicing regulation in *Physcomitrella patens* (Wu et al., 2014). The UV crosslinking assay indicated that numerous proteins were capable of binding this motif, with several 25–40 kDa proteins significantly (p value < 0.05, Fisher's t test) enriched over the negative control (Figure S7A). However, from mass spectrometry analysis of interacting proteins only four passed a threshold of 6-fold enrichment over negative controls, with the strongest candidate RBPs being CP29A (> 18-fold enrichment) (Figure 7B). This protein has previously been identified as an RBP that functions in the chloroplast (Ye et al., 1991), but nuclear localization had not been demonstrated. We used an *Arabidopsis* CP29A monoclonal antibody (Kupsch et al., 2012) to perform western blots on lysates from INTACT-purified nuclei and 10-day-old seedlings. Although at low levels, we could reproducibly detect CP29A in the *Arabidopsis* nucleus (Figure 7C), in contrast to other chloroplastic proteins (Figure S1B), showing that a subset of CP29A is localized in the nucleus.

To confirm that CP29A could interact with both nuclear and chloroplast transcripts containing the predicted GAN repeat motif in vivo we performed RNA immunoprecipitation (RIP). We took lysates from formaldehyde-treated leaves and incubated them with either a monoclonal α-CP29A or α-His antibody (negative control) (Figure S7C) followed by RT-qPCR for three nuclear transcripts and two chloroplast RNAs as positive controls. All three nuclear and one chloroplast (*ATCG00490*) transcript contain the GAN repeat motif. We found that all five transcripts were significantly (all p values < 0.05) enriched > 1.5-fold in the α-CP29A compared to the α-His control RIP samples, as opposed to the *ACTIN* negative control (Figure 7D). Taken together, these results indicate that CP29A localizes to both the chloroplast and nucleus, and interacts with a subset of GAN repeat motif-containing transcripts in *Arabidopsis*, suggesting a new functionality for this plant RBP.

## Conclusion

Here, we characterized the global landscapes of RNA secondary structure and RBP occupancy of the *Arabidopsis* nuclear transcriptome (Figure 1). We demonstrated that these data are highly reproducible, and that the identified protein-binding sites are significantly more conserved than their flanking sequences (Figure 2). Additionally, we calculated the structure score for nuclear RNAs that passed filtering criteria (see Supplemental Experimental Procedures), creating a comprehensive database of in vivo RNA secondary structure for the *Arabidopsis* nucleus (Figures 1C–1E). Together, these data sets provide a vast resource of RBP binding and secondary structure information

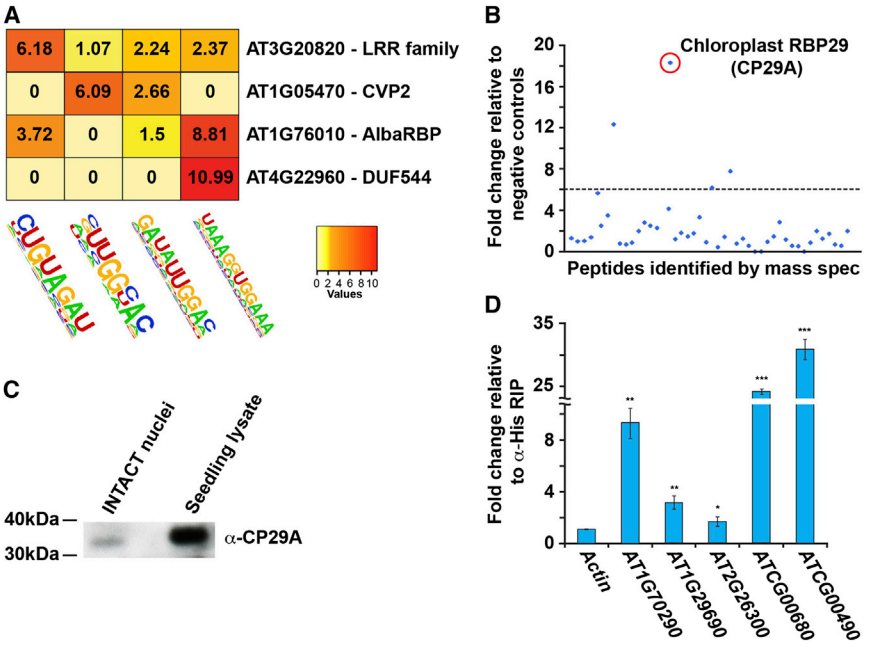



**Figure 7. Identification of *Arabidopsis* RNA-Interacting Proteins**

(A and B) Identification of proteins that interact with specific overrepresented sequence motifs.

(A) The fold enrichment over negative control of peptides from each designated protein identified by mass spectrometry analysis of eluates after RNA-affinity chromatography with each specified motif.

(B) The fold enrichment of peptides from proteins identified by mass spectrometry analysis after RNA-affinity chromatography with the GAN repeat motif (Figure 5A). The top candidate identified by this analysis, CP29A, is annotated and denoted with a red circle. Dotted line indicates cutoff of 6-fold enrichment.

(C) Western blot analysis of INTACT-purified nuclei and *Arabidopsis* 10-day-old seedling lysates using an antibody to CP29A.

(D) RT-qPCR analysis of three nuclear GAN motif-containing genes (*AT1G70290*, *AT1G29690*, and *AT2G26300*), two positive control chloroplast transcripts (*ATCG00680* and *ATCG00490* [also with motif]), and an *ACTIN* negative control following RIP with an α-CP29A or α-His antibody. The data is presented as the fold change in the α-CP29A relative to α-His RIP samples. Error bars, ± SD. *, **, and *** indicate p value < 0.05, < 0.001, and < $1\times10^{-10}$, respectively, Fisher's t test. See also Figure S7 and Table S4.

for the *Arabidopsis* nuclear transcriptome that can inform future experiments focused on understanding posttranscriptional regulation.

Using the data generated here, we searched for patterns of global RBP binding and RNA secondary structure. The most striking association that we identified was a distinct anticorrelation between RNA secondary structure and RBP occupancy within the RNA regions that were examined (Figures 3, S5B, 4, and 5). This pattern was present when focusing on uORF and canonical translation start codons (Figures 3A and S5B), stop codons (Figure 3B), exon/intron junctions (Figure 3C), and specific RBP-binding motifs (Figures 5K–5O). Furthermore, we found that the RBP-interacting motifs identified by our study tend to be less structured when protein bound (Figures 5K–5O). Although we cannot discern causality, our findings reveal that in general RBPs bind to unstructured sequence elements in target transcripts resulting in the overall opposing patterns of these features in the *Arabidopsis* nucleus.

When initially examining these data we questioned whether the structure score was artificially lowered in regions of high PPS density by occlusion of the RNase through the incomplete digestion of bound RBPs. However, if this were true these regions would not be called PPSs in our initial analyses because their read levels would be artificially raised in the structure-only libraries. Furthermore, we find that the presence of PPSs is actually associated with more negative structure scores (Figures 5K and S6A). Thus, our results are likely true biological observations of decreased structure at RBP-binding sites, not an artifact of the PIP-seq methodology.

We also examined subsets of annotated alternative exons and identified unique profiles of PPS density and secondary structure in constitutive, CE, IR, and U12-type introns (Figures 4B and 4C). These profiles suggest that gross protein binding can regulate AS, while secondary structure can influence the population of proteins that occupies each region. Although it is known that RBP binding in the exon or intron can regulate AS (Chen and Manley, 2009; Simpson et al., 2010), our observations demonstrate that protein occupancy levels in regions near the splice site can differentiate subsets of alternative exons. Our observations have provided the resources for identifying these populations of proteins and specific structural features in these alternative events.

Finally, we uncovered motifs that were enriched within our PPSs and identified co-occurrences of RBP-bound instances of these sequences in functionally related transcripts (Figure 6). These findings are similar to previous observations in human cells (Silverman et al., 2014), and support a model in which RNA transcripts encoding proteins with related functions also share a set of interacting RBPs through underlying sequence motifs allowing their coregulation. Taken together, our findings suggest that both plants and humans use different groups of RBPs to allow specific sets of proteins, especially those functioning in development, stress responses, and apoptosis, to be precisely coregulated in an operon-like fashion.

## EXPERIMENTAL PROCEDURES

### Supplemental Experimental Procedures

Further details on the experimental procedures, high-throughput sequencing, and processing, mapping, and analysis of PIP-seq data are provided in the Supplemental Experimental Procedures.

### INTACT-Purified Nuclei

Seedlings of *UBQ10*:NTF/*ACT2p*:BirA *Arabidopsis thaliana* ecotype Col-0 were grown for 10 days (20°C, 16 hr light/8 hr dark) before RNA-protein

interactions were crosslinked in a 1% formaldehyde solution under a vacuum and subsequently quenched with 125 mM glycine. INTACT purification was then performed as previously described (Deal and Henikoff, 2010). This same ecotype of *Arabidopsis* was used for all analyses in this study.

### PIP-seq and PPS Analysis

We used 2 million purified nuclei for each PIP-seq replicate, which was performed as previously described (Silverman et al., 2014). Read processing and alignment, PPS identification, and all other PPS analyses were done as previously described (Silverman et al., 2014).

## ACCESSION NUMBERS

The raw and processed data for PIP-seq and total RNA sequencing from our analyses have been deposited into the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE58974. All of our data (i.e., files of all identified PPSs, complete list of nuclear mRNA structure scores, etc.) can also be accessed at http://gregorylab.bio.upenn.edu/PIPseq_AtTotalNuc.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2014.12.004.

## AUTHOR CONTRIBUTIONS

S.J.G., S.W.F., R.B.D., and B.D.G. conceived the study and designed the experiments. S.J.G., S.W.F., D.W., N.S., A.D.L.N., M.A.B., F.D., and B.D.G. performed the experiments. S.J.G., S.W.F., I.M.S., and B.D.G. analyzed the data. S.W.F., S.J.G., I.M.S., R.B.D., and B.D.G. wrote the paper with assistance from all authors. The authors have read and approved the manuscript for publication.

## ACKNOWLEDGMENTS

The authors thank Dr. Christian Schmitz-Linneweber for providing the α-CP29A antibody, Alsu Ibragimova for optimizing the RNA-affinity chromatography, Jennifer Nemhauser for providing the *UBQ10p*:NTF/*ACT2p*:BirA transgenic line, and members of the B.D.G. lab for helpful discussions. This work was funded by NSF grant MCB-1243947 to B.D.G. and NIGMS 5T32GM008216-26 to I.M.S.

Received: July 23, 2014
Revised: September 15, 2014
Accepted: November 25, 2014
Published: December 31, 2014

## REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37* (Web Server issue), W202–W208.

Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell *46*, 674–690.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. *43*, 956–963.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393–1406.

Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution—trashing old rules to forge new ones. Cell *157*, 77–94.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat. Rev. Mol. Cell Biol. *10*, 741–754.

Chiou, N.T., Shankarling, G., and Lynch, K.W. (2013). hnRNP L and hnRNP A1 induce extended U1 snRNA interactions with an exon to repress spliceosome assembly. Mol. Cell *49*, 972–982.

Cruz, J.A., and Westhof, E. (2009). The dynamic landscapes of RNA architecture. Cell *136*, 604–609.

Deal, R.B., and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev. Cell *18*, 1030–1040.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696–700.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acid Res. *38*, W64–W70.

Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Res. *35* (Web Server issue), W297–W299.

Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R., et al. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat. Struct. Mol. Biol. *21*, 198–206.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hunt, A.G., Xing, D., and Li, Q.Q. (2012). Plant polyadenylation factors: conservation and variety in the polyadenylation complex in plants. BMC Genomics *13*, 641.

Kandasamy, M.K., McKinney, E.C., and Meagher, R.B. (1999). The late pollen-specific actins in angiosperms. Plant J. *18*, 681–691.

Keene, J.D., and Tenenbaum, S.A. (2002). Eukaryotic mRNPs may represent posttranscriptional operons. Mol. Cell *9*, 1161–1167.

Kupsch, C., Ruwe, H., Gusewski, S., Tillich, M., Small, I., and Schmitz-Linneweber, C. (2012). *Arabidopsis* chloroplast RNA binding proteins CP31A and CP29A associate with large transcript pools and confer cold stress tolerance by influencing multiple chloroplast RNA processing steps. Plant Cell *24*, 4266–4280.

Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M., and Rajewsky, N. (2011). Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Mol. Cell *43*, 340–352.

Li, F., Zheng, Q., Ryvkin, P., Dragomir, I., Desai, Y., Aiyer, S., Valladares, O., Yang, J., Bambina, S., Sabin, L.R., et al. (2012a). Global analysis of RNA secondary structure in two metazoans. Cell Rep. *1*, 69–82.

Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. (2012b). Regulatory impact of RNA secondary structure across the *Arabidopsis* transcriptome. Plant Cell *24*, 4346–4359.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. Plant Cell *24*, 4333–4345.

Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res. *22*, 1184–1195.

Raker, V.A., Mironov, A.A., Gelfand, M.S., and Pervouchine, D.D. (2009). Modulation of alternative splicing by long-range RNA structures in *Drosophila*. Nucleic Acids Res. *37*, 4533–4544.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.

Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., Milos, P.M., Barton, G.J., and Simpson, G.G. (2012). Direct sequencing of *Arabidopsis thaliana* RNA reveals patterns of cleavage and polyadenylation. Nat. Struct. Mol. Biol. *19*, 845–852.

Silverman, I.M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J.L., and Gregory, B.D. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol. *15*, R3.

Simpson, C.G., Manthri, S., Raczynska, K.D., Kalyna, M., Lewandowska, D., Kusenda, B., Maronova, M., Szweykowska-Kulinska, Z., Jarmolowski, A., Barta, A., and Brown, J.W. (2010). Regulation of plant gene expression by alternative splicing. Biochem. Soc. Trans. *38*, 667–671.

Talkish, J., May, G., Lin, Y., Woolford, J.L., Jr., and McManus, C.J. (2014). Mod-seq: high-throughput sequencing for chemical probing of RNA structure. RNA *20*, 713–720.

Tenenbaum, S.A., Christiansen, J., and Nielsen, H. (2011). The post-transcriptional operon. Methods Mol. Biol. *703*, 237–245.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

von Arnim, A.G., Jia, Q., and Vaughn, J.N. (2014). Regulation of plant translation by upstream open reading frames. Plant Sci. *214*, 1–12.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701–718.

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q.Q., and Hunt, A.G. (2011). Genome-wide landscape of polyadenylation in *Arabidopsis* provides evidence for extensive alternative polyadenylation. Proc. Natl. Acad. Sci. USA *108*, 12533–12538.

Wu, H.P., Su, Y.S., Chen, H.C., Chen, Y.R., Wu, C.C., Lin, W.D., and Tu, S.L. (2014). Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in *Physcomitrella patens*. Genome Biol. *15*, R10.

Ye, L.H., Li, Y.Q., Fukami-Kobayashi, K., Go, M., Konishi, T., Watanabe, A., and Sugiura, M. (1991). Diversity of a ribonucleoprotein family in tobacco chloroplasts: two new chloroplast ribonucleoproteins and a phylogenetic tree of ten chloroplast RNA-binding domains. Nucleic Acids Res. *19*, 6485–6490.

Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.S., and Gregory, B.D. (2010). Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. PLoS Genet. *6*, e1001141.

**Molecular Cell, Volume *57***

# Supplemental Information

# Global Analysis of the RNA-Protein Interaction and RNA Secondary Structure Landscapes of the *Arabidopsis* Nucleus

Sager J. Gosai, Shawn W. Foley, Dongxue Wang, Ian M. Silverman, Nur Selamoglu, Andrew D.L. Nelson, Mark A. Beilstein, Fevzi Daldal, Roger B. Deal, and Brian D. Gregory

## SUPPLEMENTAL FIGURES

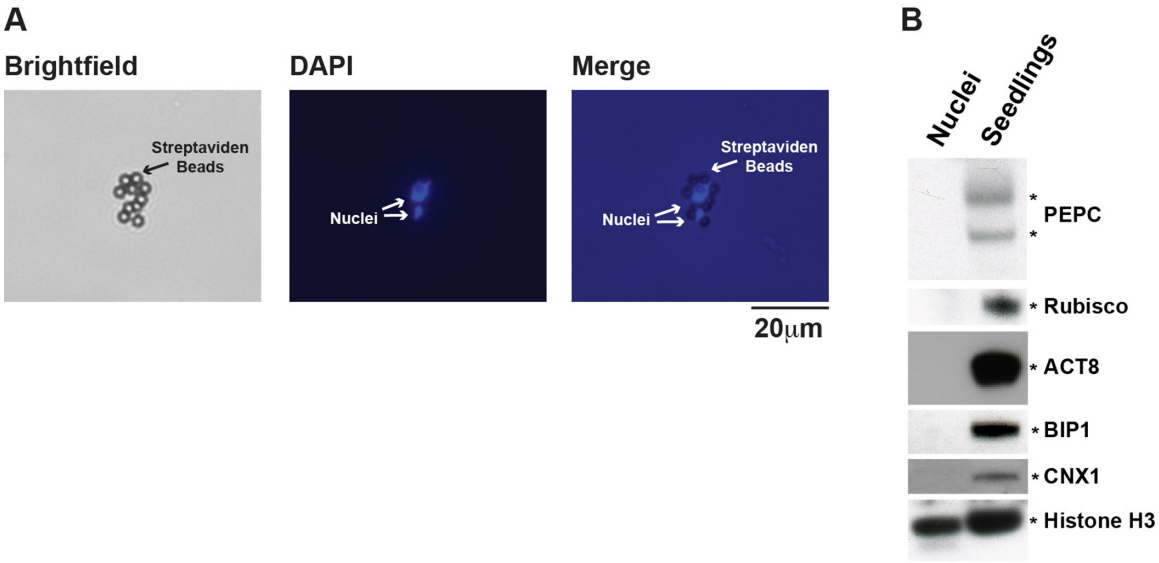



**Figure S1, Related to Figures 1-7: INTACT purified nuclei are free of cytoplasmic, ER, and chloroplastic contamination**

(A) Microscopy imaging of DAPI stained nuclei during the INTACT purification process. The images show that only the DAPI stained nuclei are bound to the streptavidin beads. (B) Western blot of lysates from INTACT purified nuclei and 10-day-old seedlings for the chloroplastic PEPC and RUBISCO, the mostly cytoplasmic ACT8, the endoplasmic reticulum (ER)-localized BIP1 and CNX1, as well as the nuclear histone H3 proteins.

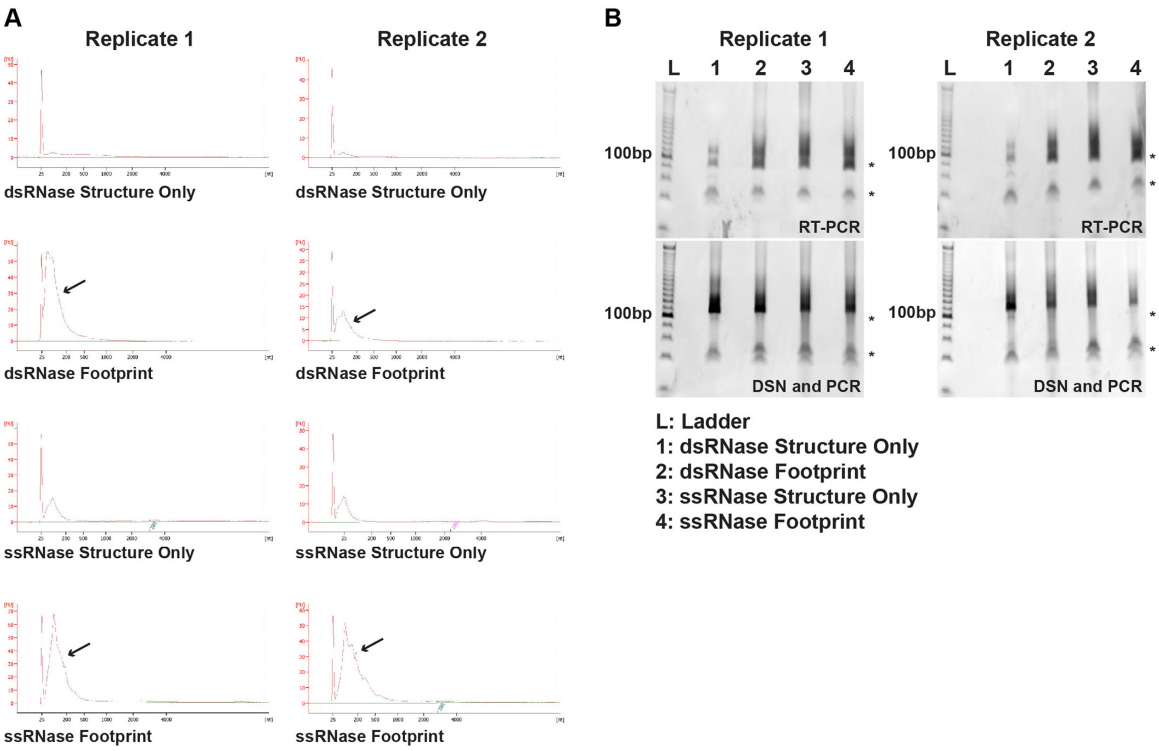



**Figure S2, Related to Figures 1-6: The PIP-seq libraries passed all three quality control checkpoints during library preparation**

(A) Profiles from a BioAnalyzer run of the digested RNA for each of the eight PIP-seq libraries. These profiles show the expected sizes and quality for these libraries when compared to profiles from previous PIP-seq experiments. The arrows point to the larger fragments found specifically in the footprinting samples that likely represent the protein protected sites (PPSs). (B) Two size selection gels run after the initial RT-PCR or after DSN treatment and PCR. These gels show that the libraries are still of the expected high quality, and have been shifted to the expected sizes after adapter ligations. The top and bottom asterisks (*) to the right of each gel image denote adapter-adapter products and unused primers, respectively. These contaminants were avoided during the gel purification process, ensuring the high quality of our sequenced libraries.

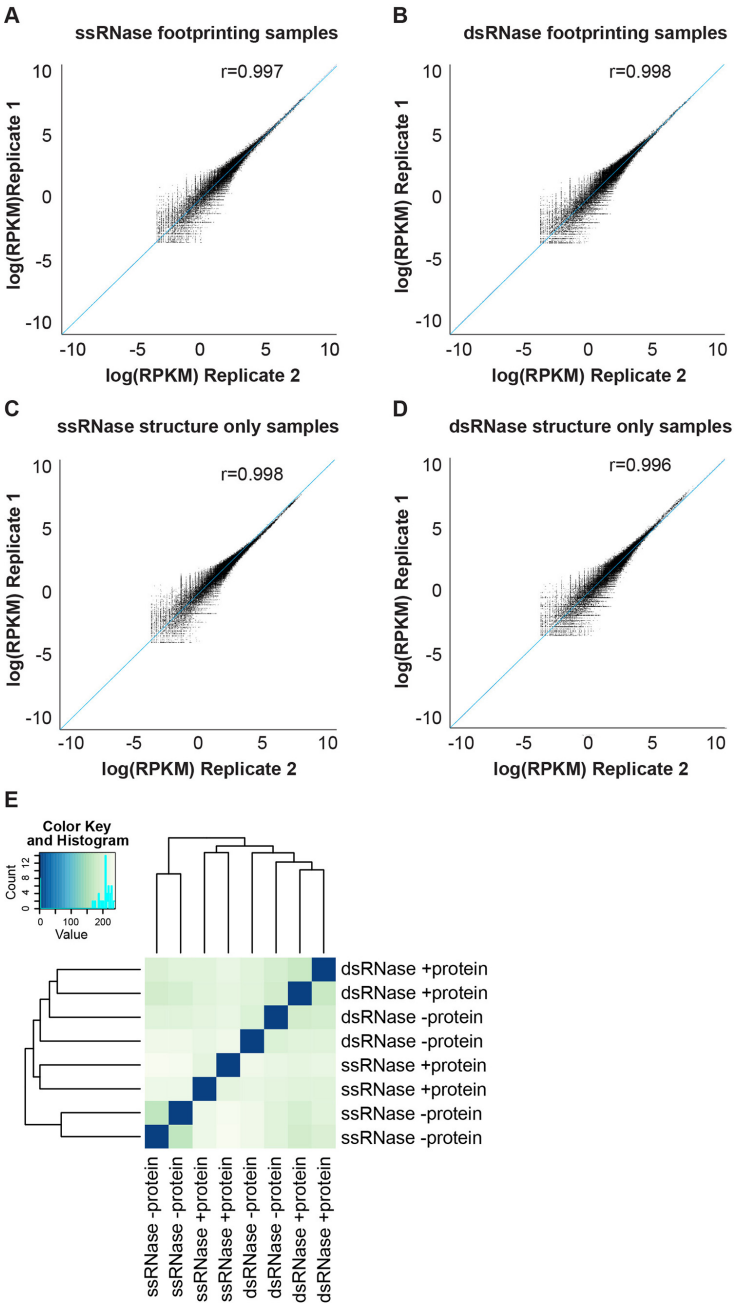



**Figure S3, Related to Figures 1-6: PIP-seq is a highly reproducible method**

(A-B) Correlation in read counts in a 50 nt sliding window between both ssRNase (A) and dsRNase (B) footprinting replicates. (C-D) Correlation in read counts in a 50 nt sliding window between both ssRNase (C) and dsRNase (D) structure only replicates. (E) Principle component analysis of 500 nt bins between each of the eight libraries. All replicate pairs cluster together, as do both RNases demonstrating the high reproducibility of these PIP-seq libraries.

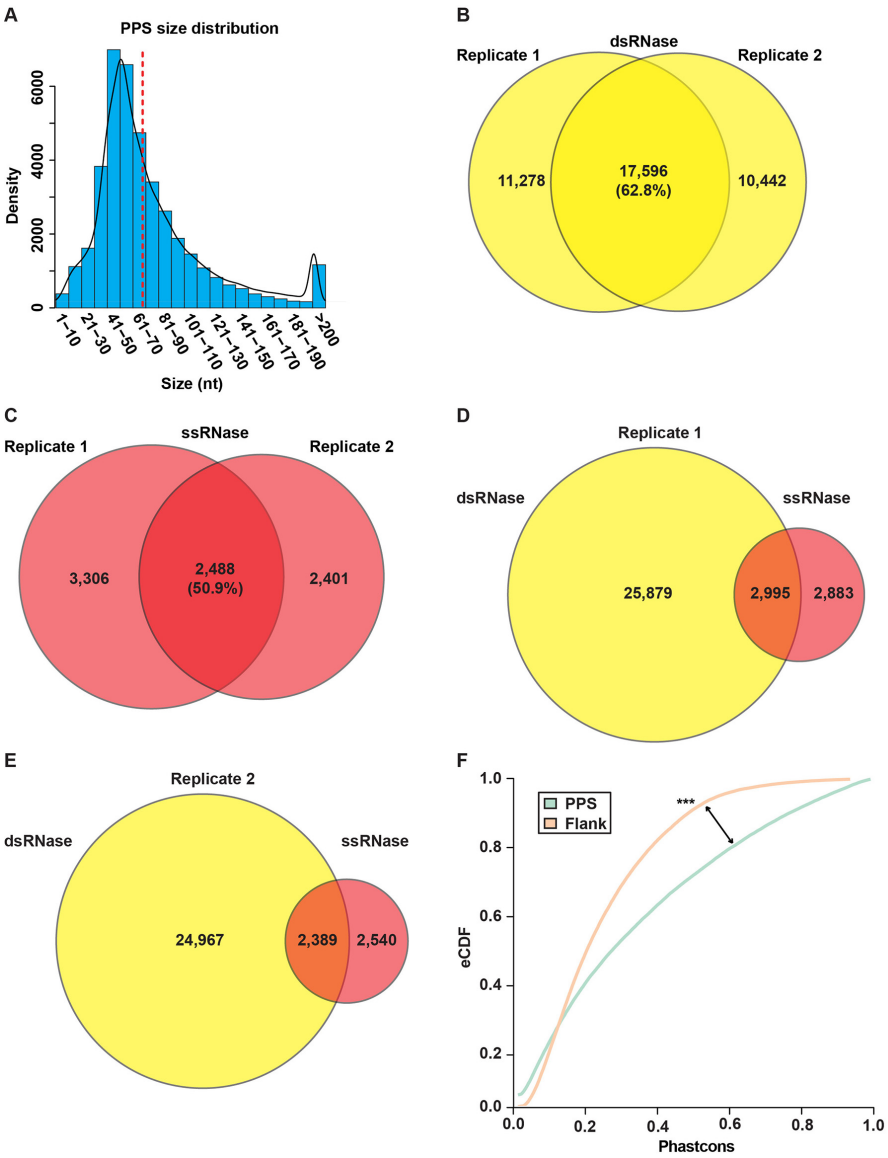



**Figure S4, Related to Figure 2: Further characterization of Arabidopsis nuclear PPSs**

(A) Distribution of sizes (nt) for the total set of 40,131 distinct PPSs. Dashed red line represents the median PPS size (~68 nt). (B-C) Overlap in PPS calls between dsRNase- (B) and ssRNase-treated (C) PIP-seq replicates. (D-E) Overlap in PPS calls between the dsRNase- (yellow circle) and ssRNase-treated (red circle) samples for replicate 1 (D) and replicate 2 (E). (F) Cumulative distribution of average PhastCons scores in PPSs (green line) versus similarly sized flanking regions (orange line). *** denotes p-value < $1x10^{-10}$, Kolmogorov-Smirnov test.

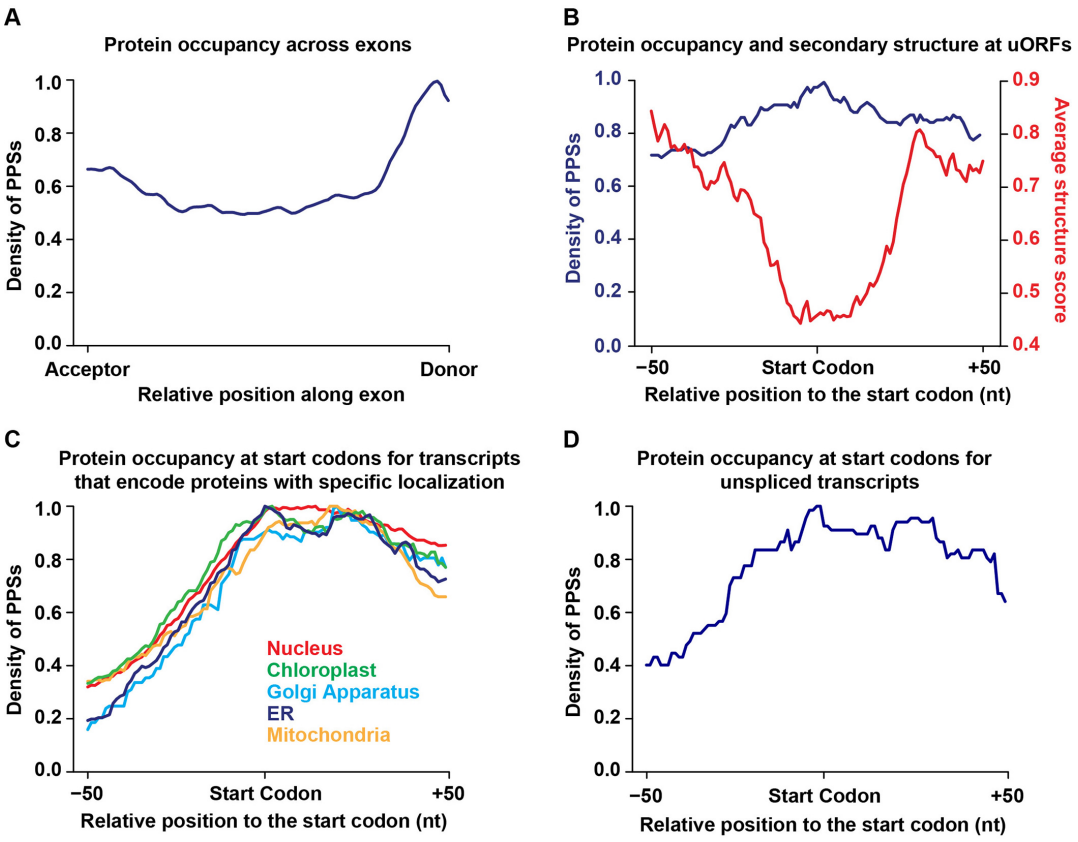



**Figure S5, Related to Figures 2-3: Protein occupancy at Arabidopsis constitutive exons, protein binding and secondary structure landscapes at upstream open reading frames (uORFs), as well as protein occupancy at start codons of transcripts encoding specifically localized proteins or that are unspliced**.

(A) PPS density across Arabidopsis constitutive exons (excluding exons containing star and stop codons) (B) PPS density and structure score profiles for highly confident upstream open reading frames (uORFs) (von Arnim et al., 2014). Average PPS density (blue) and structure score (red) at each position +/- 50 nt at uORF start codons. (C) Average PPS density at each position +/- 50 nt at canonical start codons for transcripts encoding proteins that are localized to specific cellular compartments (as specified by colored line and label) based on TAIR10 annotation. These transcripts show similar profiles. (D) Average PPS density at each position +/- 50 nt at canonical start codons for transcripts that are unspliced and likely nuclear localized in our RNA sequencing experiments.

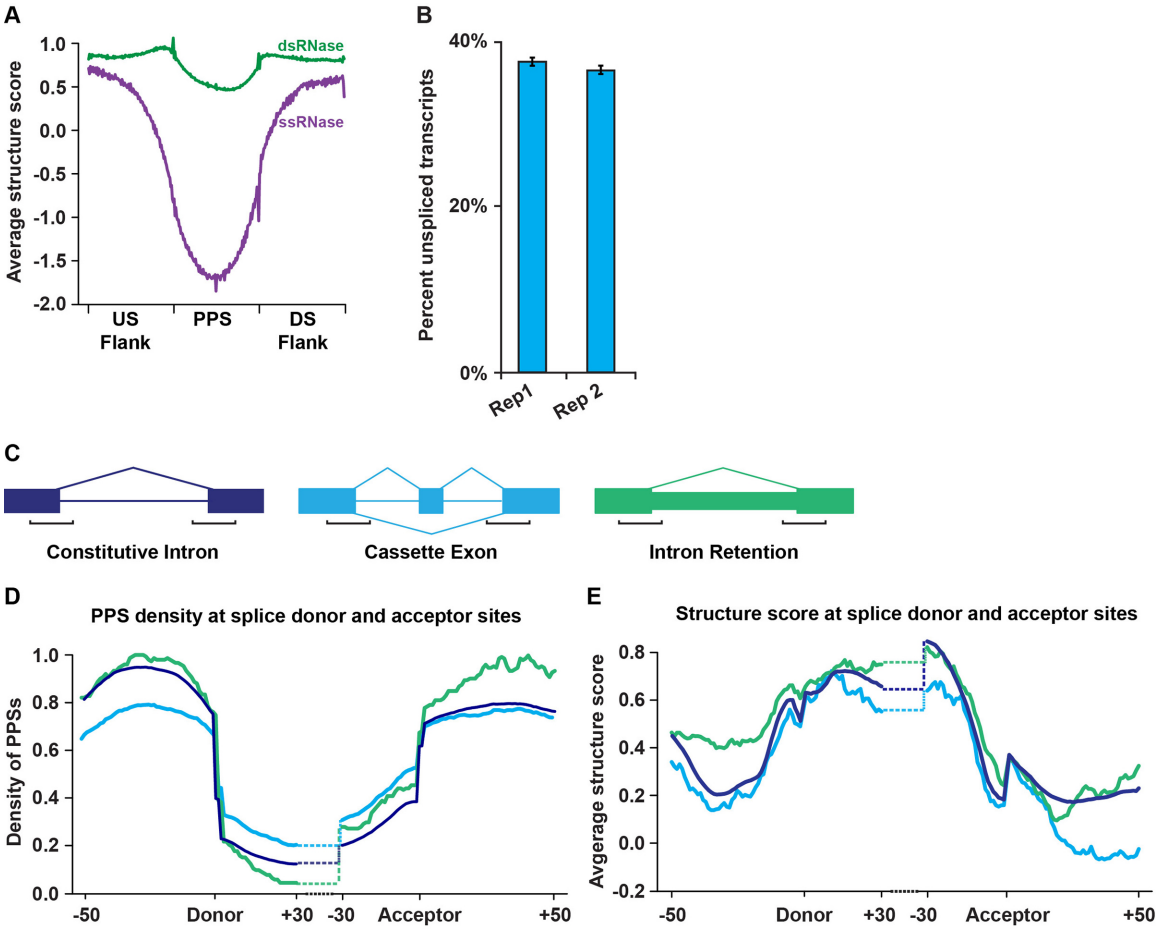



**Figure S6, Related to Figures 3-5: Secondary structure and protein binding landscapes at protein interaction sites and isolated alternative splicing events**.

(A) The average structure score of exonic PPSs from the dsRNase (green) or ssRNase (purple) treated libraries, and equal sized flanking regions, for 100 equal sized bins. The average structure score for the dsRNase treated PPSs is significantly (p-value < $2.2 \times 10^{-16}$, Wilcoxon test) greater than ssRNase treated PPSs. (B) The mean percentage of exon/intron junction mapping reads per transcript from two replicates (as indicated) of total RNA sequencing for congruently purified nuclei. Error bars represent standard error of the mean (SEM). (C) Diagram of constitutive introns (blue), cassette exons (turquoise), and intron retention events (green). Large boxes represent exons, lines represent constitutive introns, and small boxes represent alternatively spliced introns, with

brackets indicating regions graphed in D and E for reference. (D) Average PPS density at each position -50 to +30 nt at the donor splice site, and -30 to +50 at the acceptor splice site. Line colors correspond to examples shown in C. CEs show significantly (p-values < 0.001, Wilcoxon test) higher PPS density across both intronic and exonic sequences at the acceptor splice site (-30 to +40). IR events demonstrate significantly (p-values < $6.0 \times 10^{-6}$, Wilcoxon test) higher PPS density across all interrogated regions. (E) Structure score profiles for constitutive and isolated alternative splicing events in Arabidopsis covering the same regions as D. Line colors correspond to examples shown in C. IR events displayed significantly (p-value < $6.5 \times 10^{-3}$, Wilcoxon test) increased structure upstream of the donor splice site (-45 to -1). Conversely, CEs did not demonstrate any significant differences in secondary structure as compared to constitutive introns.

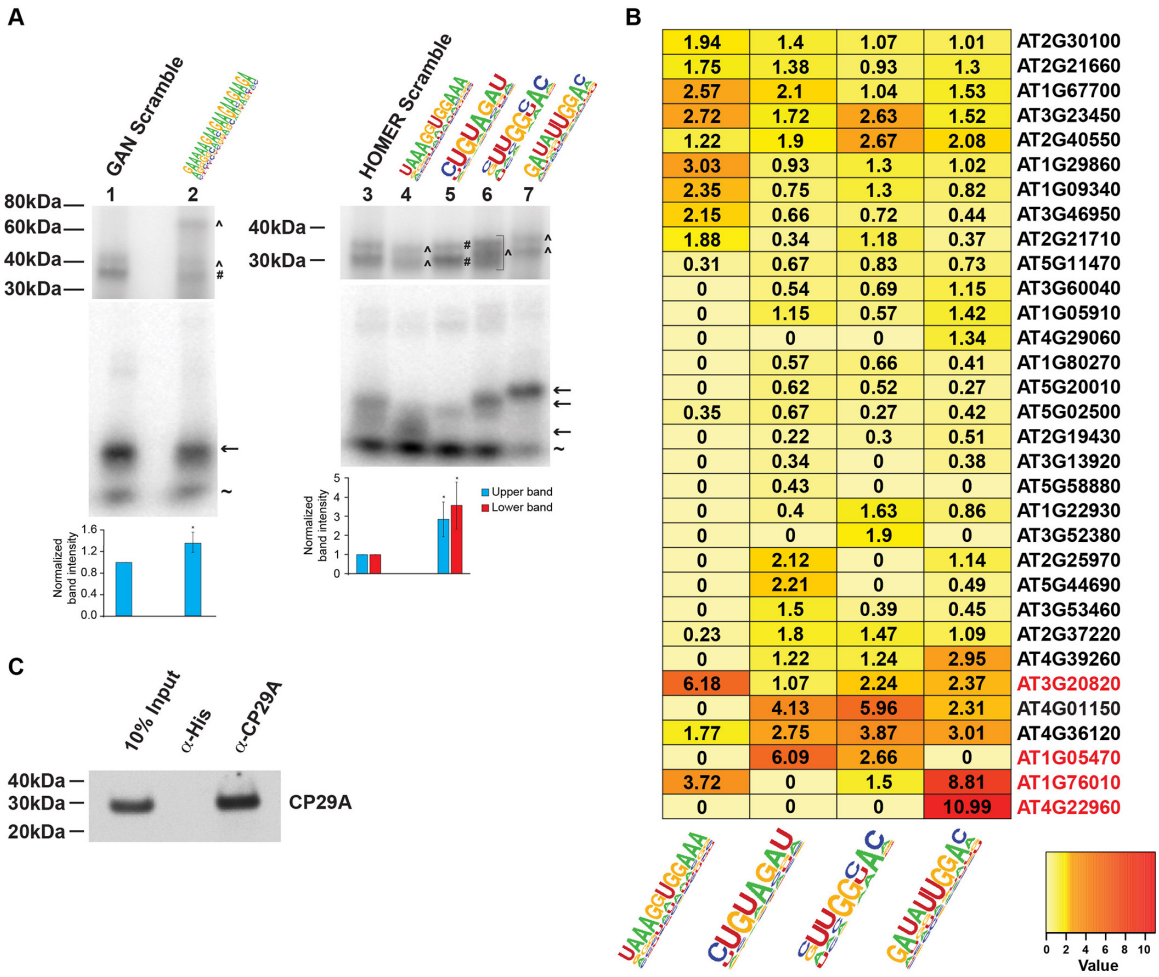



**Figure S7, Related to Figure 7: Identification of putative RBPs using synthetic RNA motifs**

(A) UV-crosslinking analysis for the indicated RBP-interacting motifs compared to non-specific controls using Arabidopsis 4-week-old leaf lysate. Three biological replicates were performed, and a representative gel is shown. For bands that are present in both the motif and scrambled control lanes, the intensity of the band was quantified and normalized to the unbound probe, and is graphed below the chart as fold change relative to scrambled control. ^ denotes bands that are present in a motif lane, but are absent from the scrambled controls. # denotes a band that is present in both the motif lane and the scrambled control, and therefore was quantified in the graphs below the respective lane. The arrows denote unbound probes. ~ denotes the unincorporated radiolabeled ATP. * denotes $p < 0.05$, Fisher's t-test. Error bars represent standard deviation, n=3.

(B) Enrichment of peptides from the indicated Arabidopsis proteins as compared to negative control pulldown samples. The number of peptide spectrum matches (PSM) for each sample was taken and the percentage of the total PSM for each identified protein was calculated. The fold change relative to the average of the empty bead and scramble bait negative controls is shown. Proteins denoted in red are candidate RBPs that passed a 6 fold enrichment threshold. (C) Western blot with an $\alpha$-CP29A monoclonal antibody on 10% input and eluents from RNA immunoprecipitations (RIPs) performed with $\alpha$-CP29A and $\alpha$-His monoclonal antibodies.

**SUPPLEMENTAL TABLES**

1. **Supplemental Table S1, Related to Figures 1-6:** The pertinent information for the 40,131 distinct PPSs identified in the Arabidopsis nuclear transcriptome.
2. **Supplemental Table S2, Related to Figures 5-7:** The pertinent information for the 41 identified protein-bound motifs.
3. **Supplemental Table S3, Related to Figures 7 and S7:** The pertinent information for all oligonucleotide probes and RT-qPCR primers that were used in this study.
4. **Supplemental Table S4, Related to Figure 7:** The number of peptides identified by LC-MS/MS in each RNA-affinity chromatography experiment, which is presented graphically in Figures 7A-B and S7B.

## SUPPLEMENTAL EXPERIMENTAL PROCEDURES

### Plant Materials

The purified nuclei used in this study were extracted from 10-day-old seedlings of *UBQ10*:NTF*/ACT2p*:BirA Columbia-0 (Col-0) ecotype of *Arabidopsis thaliana* using the INTACT methodology. Additionally, the lysates for all western blots were from these same 10-day-old seedlings. The lysates used for RNA immunoprecipitation (RIP) RT-qPCR and motif-interacting protein analyses were from whole leaves extracted from four-week-old Col-0 plants. All plants were grown at 20°C, in a 16 h light/8 h dark cycle.

### Cross-linking and INTACT purification

Immediately before nuclei purification, 10-day-old seedlings of *UBQ10*:NTF*/ACT2p*:BirA were crosslinked in nuclear purification buffer (20 mM MOPS (pH = 7), 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA) plus 1% (vol/vol) formaldehyde under vacuum for 10 minutes, followed by a five minute quench with 125 mM Glycine under vacuum for an additional five minutes. Crosslinked seedlings then underwent INTACT purification as previously described (Deal and Henikoff, 2010).

### Total RNA sequencing library preparation

10-day-old seedlings of UBQ10:NTF/*ATC2p*:BirA underwent the INTACT purification as previously described (Deal and Henikoff, 2010). The resulting nuclei were lysed and the RNA was isolated using the Qiagen miRNeasy RNA isolation kit following the included protocol (Qiagen; Valencia, CA, USA). Finally, the purified RNA was used as the substrate for strand-specific total RNA sequencing library preparation as previously described (Elliott et al., 2013), with the exception that no poly(A) purification was performed but was replaced by DSN treatments as previously described (Silverman et al., 2014). The resulting libraries were sequenced on an Illumina HiSeq2000 using the standard protocol for 50 base pair single read sequencing.

**PIP-seq library preparation**

~Two million INTACT purified nuclei were lysed in 850 μl RIP buffer (25 mM Tris-HCl, pH = 7.4; 150 mM KCl, 5 mM EDTA, pH = 7.5; 0.5% NP40; 10 μM DTT; 1 tablet protease inhbitors and 0.5 μl/ml RNaseOUT (Life Technologies; Carlsbad, CA, USA)) by manual grinding. The resulting cell lysate was treated with RNase-free DNase (Qiagen; Valencia, CA, USA). The lysates were then split and treated with either 100 U/ml of a single-stranded RNase (ssRNase) (RNaseONE (Promega; Madison, WI, USA)) with 200 μg/ml BSA in 1X RNaseONE buffer for 1 hour at room temperature (RT), or 2.5 U/ml of a double-stranded RNase (dsRNase) (RNaseV1 (Ambion; Austin, TX, USA)) in 1X RNA structure buffer for 1 hour at 37°C as previously described (Silverman et al., 2014). See Figure 1A for a schematic representation of library preparation. Proteins were then denatured and digested by treatment with 1% SDS and 0.1 mg/ml Proteinase K (Roche; Basel, Switzerland) for 15 minutes at RT. Proteinase digestion was followed by a 2 hour incubation at 65°C to reverse the RNA-protein cross-links.

To determine whether nuclease resistant regions in RNAs are due to protein binding or specific secondary structures, we also determined the digestion patterns of ds- and ssRNases immediately following protein digestion. To do this, we performed the identical treatments as described above except that the cross-linked nuclear lysates were treated with 1% SDS and 0.1 mg/ml Proteinase K (Roche; Basel, Switzerland) and ethanol precipitated prior to being treated with the two RNases. In this way, the SDS and Proteinase K solubilized and digested the proteins allowing us to deduce PPSs within all detectable RNAs in the cells of interest (see Figure 1A for schematic).

The digested RNA was then isolated using the Qiagen miRNeasy RNA isolation kit following the included protocol (Qiagen; Valencia, CA, USA). To ensure that only high quality RNA samples were used for PIP-seq library preparation, the purified RNA was run on a Eukaryotic Total RNA Pico Series II chip (5067-1513; Agilent Technologies; Wilmington, DE, USA) using a

BioAnalyzer 2100 system. Finally, the purified RNA was used as the substrate for strand-specific sequencing library preparation as previously described (Silverman et al., 2014). All of the RNase footprinting libraries (a total of 4 for each replicate: ss- and dsRNase treatments, footprint and structure only) were sequenced on an Illumina HiSeq2000 using the standard protocol for 50 base pair single read sequencing.

**Read processing and alignment**

PIP-seq reads were first trimmed to remove 3' sequencing adapters using cutadapt (version 1.2.1 with parameters -e 0.06 -O 6 -m 14). The resulting trimmed sequences were collapsed to unique reads and aligned to the TAIR10 Arabidopsis genome sequence using Tophat (version 2.0.10 with parameters --library-type fr-secondstrand --read-mismatches 2 --read-edit-dist 2 --max-multihits 10 --b2-very-sensitive --transcriptome-max-hits 10 --no-coverage-search --no-novel-juncs). PCR duplicates were collapsed to single reads for all subsequent analyses.

**Estimating unspliced transcripts**

All reads from the total RNA-sequencing data that mapped to all detectable first TAIR10 annotated constitutively spliced introns were collected, removing reads that were entirely within the intron. We quantified the number of reads that had mapped through the exon/intron boundary (unspliced) compared to those that contained the exon/exon boundary (spliced). We then determined the fraction of junction mapping reads that were unspliced for each gene.

**Identification of PPSs**

PPSs were identified using a modified version of the CSAR software package (Muiño et al., 2011). Specifically, read coverage values were calculated for each base position in the genome and a Poisson test was used to compute an enrichment score for footprint versus structure only libraries. PPSs were then

called with a false discovery rate of 5% as previously described (Silverman et al., 2014).

**Functional analysis of PPSs**

PPS annotation was done 'greedily' using the TAIR10 genome annotations, such that all functional annotations that overlapped with a given PPS were counted equally. Conservation was assessed by comparing both PhastCons scores and the number of SNPs, within PPSs relative to equally sized flanking regions. PhastCons scores for PPSs compared to same sized flanking regions were calculated as previously described (Li et al., 2012; Silverman et al., 2014).

To perform the SNP occurrence analysis we first identified SNPs located in transcriptionally active region (TARs), defined as intervals at least 15 nt long with greater than 20 reads of coverage, while allowing for a gap of 10 nt with less coverage, as calculated using an aggregate list of alignments from both replicates of the PIP-seq libraries. Ten permutations of random shuffling of TARs were then performed to generate the control set with similar numbers and fragment sizes to our list of PPSs. We then quantified the number of non-redundant, substitution SNP sites cataloged by the 1001 Genomes Project (Cao et al., 2011) within the total list of PPSs and the 10 shuffled intervals, which were statistically compared to one another using a $\chi^2$-test.

***lincRNA conservation analysis***

*Brassica rapa* lincRNAs were identified from a list of 3,450 intergenic transcripts generated previously (Tong et al., 2013), then further filtered by removal of transcripts with an open reading frame >100 codons. A total of 1908 *B. rapa* lincRNAs were then used as the dataset in BLAST analyses with Arabidopsis lincRNAs using an E-value of $10^{-10}$.

**Calculating the structure score statistic**

For every base of detectable transcripts, we calculated the dsRNA-seq and ssRNA-seq coverages from the structure only samples, then calculated the structure score as described previously (Li et al., 2012). Briefly, when given the dsRNA-seq and ssRNA-seq coverages ($n_{ds}, n_{ss}$) of a given base *i*, the structure score is determined as:

$$S_i = \mathrm{glog}(ds_i) - \mathrm{glog}(ss_i) = \log_2\left(ds_i + \sqrt{1+ds_i^2}\right) - \log_2\left(ss_i + \sqrt{1+ss_i^2}\right)$$

$$ds_i = n_{ds}\frac{\max(L_{ds}, L_{ss})}{L_{ds}}, \quad ss_i = n_{ss}\frac{\max(L_{ds}, L_{ss})}{L_{ss}}$$

where $S_i$ is the structure score, $ds_i$ and $ss_i$ are the normalized read coverages, and $L_{ds}, L_{ss}$ are the total covered length by mapped dsRNA-seq and ssRNA-seq reads, respectively. The total coverage length was used as the normalization constant instead of the total number of mapped reads used previously, because we believe it is a more reasonable assumption for the transcriptome to have comparable levels of paired/unpaired regions. It is of note that we used a generalized log ratio (glog) instead of normal log-odds because it can tolerate 0 values (positions with no dsRNA or ssRNA read coverage) as well as being asymptotically equivalent to the standard log ratio when the coverage values are large. Only sense-mapping reads were used, as we are entirely concerned with the intra-molecular interactions contributing to the self-folding secondary structure.

**Structure score profile analysis of mRNAs**

The structure score for every base of each detected transcript was first calculated using all mapped and spliced reads. In addition to the minimum dsRNA-seq plus ssRNA-seq read coverage requirement discussed above, we only considered mRNAs with intact CDS regions, ≥ 45 nt 5'UTRs, ≥ 140 nt 3'UTRs and a minimum coverage of 100 reads across the entire transcript. For the profiles near CDS boundaries, structure scores for up/downstream of the

CDS start or end sites were extracted, aligned for each detectable mRNA and averaged to produce the profiles.

**PPS profile across constitutive exons**

All constitutive exons that did not contain a start or stop codon were taken and subdivided into one hundred equal sized bins. PPS density was then calculated and graphed across the bins as previously described (Silverman et al., 2014).

**Secondary structure and PPS density at upstream Open Reading Frames (uORFs)**

Annotated Arabidopsis uORFs of high confidence (defined as a purine at the -3 position and a glycine at the +4 position) were extracted from a previously annotated dataset (von Arnim et al., 2014). We then calculated average structure score (see above) and PPS density (average number of PPS covered bases) for uORFs with >10 mapped reads in the regions 50 bp up- or downstream of uORF start codons.

**PPS profiles across canonical start codons for transcripts localized to specific cellular compartments**

Transcripts were subdivided based on their TAIR10 annotated cellular component gene ontology (mitochondria: 0005739, chloroplast: 0009507, ER: 0005829, Golgi apparatus: 0005794, nucleus: 0005634). PPS density was then calculated and graphed for 50 nt up- and downstream of the start codon as previously described (Silverman et al., 2014).

**PPS profiles across canonical start codons for unspliced transcripts**

Unspliced genes were defined as transcripts in our total RNA-seq libraries with high coverage (above the median) in which junction-spanning reads at the first constitutive intron only crossed the exon/intron boundary, and not the exon/exon junction. PPS density was then calculated and graphed at 50 nt up-

and downstream of the start codon as previously described (Silverman et al., 2014).

**Structure profile at dsRNase- and ssRNase-identified PPSs**

All exonic PPSs and equal sized flanking regions were taken and subdivided into one hundred equal sized bins. The calculated structure scores (see above) were averaged for each bin, and the resulting profiles were graphed.

***Analysis of alternatively spliced exons and introns***

In order to identify specific subsets of alternative splicing events, we took all TAIR10 annotated mRNA transcripts and used the ASTALAVISTA suite (parameters -t asta –i) to identify every annotated alternative splicing event (Foissac and Sammeth, 2007; Sammeth et al., 2008). We then used the ASTALAVISTA code assigned to each event to identify single cassette exons or intron retention sites (0,1^2- or 0,1-2^, respectively). Additionally, we extracted all cassette exon and intron retention events, regardless of adjacent exons, using the list of alternative events and corresponding ASTALAVISTA codes previously described in Arabidopsis (Marquez et al., 2012). Taking these annotated events, we then identified the splice donor and acceptor sites of the nearest constitutive introns for our analysis (e.g. if exons 4, 5, and 6 are alternatively spliced together we looked at the donor and acceptor sites at exons 3 and 7, respectively). PPS and structure score profiles were then calculated (see above) for regions where the donor exon was ≥ 50 nt, acceptor exon was ≥ 50 nt, and intron was ≥ 60 nt and at least 5 reads mapped to the intron. Thus, these profiles can cover the fifty exonic and thirty intronic nucleotides flanking the splice donor and acceptor sites. P-values were calculated by non-pairwise Wilcoxon tests.

***Analysis of alternative polyadenylation sites***

We extracted the cleavage and polyadenylation sites previously identified by direct RNA sequencing (Sherstnev et al., 2012) and filtered out sites that were located outside of TAIR10 annotated 3’UTRs. A second filtering step was

performed to remove alternative polyadenylation (APA) sites within 60 nt of one another, preventing any overlap between analyzed flanking regions. PPS density and structure score profiles were then calculated (see above) for 30 nt flanking each side of these cleavage and polyadenylation sites. P-values were calculated by non-pairwise Wilcoxon tests.

**RBP bound sequence motif identification and profiling secondary structure at these sites**

MEME (Bailey et al., 2009) and HOMER (Heinz et al., 2010) were used to identify enriched RBP interaction motifs with parameters -p 8 -dna -nmotifs 100 -maxw 20 -evt 0.01 -maxsize 100000000, and -rna -size given -p 2 respectively. Motifs from Figures 5A-E were mapped to the genome using HOMER (Heinz et al.) to identify every occurrence of the motifs in nuclear mRNAs. We then identified protein bound and unbound occurrences using our mapped PPSs. Average structure scores for each position were calculated as described above.

**Motif and co-occurrence analysis**

Motif co-occurrence was defined at the transcript level, and k-means clustering of the resultant weighted adjacency matrix was used to identify clusters of co-occuring motifs. We set k=3 based on manual inspection of clusters on a multidimensional scaling (MDS) plot of the adjacency matrix. Gene Ontology (GO) analysis on the lists of transcripts that contained at least three protein bound occurrences of the motifs in each cluster was performed using agriGO (Du et al., 2010).

**UV Cross-linking analysis of motifs**

Synthetic RNA oligonucleotides (Table S3) were radiolabeled in a T4 polynucleotide kinase (PNK) reaction (New England Biolabs; Cambridge, MA, USA) using 500 μCi of $\gamma$-$^{32}$P ATP following the manufacturer's recommendation, followed by phenol-chloroform extraction and precipitation. Each RNA probe was diluted to equal counts per minute (cpm), and was added to separate 10.2 μL

binding reactions comprising 0.2 mM Tris (pH = 7.5), 0.02 mM EDTA, 40 mM KCl, 1.3% polyvinyl alcohol, 25 ng/μl tRNA, 3 mM $MgCl_2$, 1 mM ATP, 50 mM creatine phosphate, and 2.8 μg/μl Arabidopsis leaf lysate in RIP buffer (25 mM Tris-HCl, pH = 7.4; 150 mM KCl, 5 mM EDTA, pH = 7.5; 0.5% NP40; 10 μM DTT; 1 tablet/10ml protease inhibitors (Roche; Basel, Switzerland)) and incubated at 30°C for 20 minutes. The binding reaction was then subjected to UV cross-linking for 20 minutes using a 254 nm UV lamp (Mineralight Lamp Model R-52G (UVP; Upland, CA, USA)). RNA bound proteins were denatured in 1X SDS sample buffer and 1 mM β-mercaptoethanol and boiled for 5 minutes. Samples were separated on NuPAGE 3-8% Tris-Acetate gel (Life Technologies; Carlsbad, CA, USA) at 120V for 1 h. The gel was then fixed in a 10% methanol and 10% acetic acid solution for five minutes, and dried for 90 minutes. Phosphorimaging was used to visualize protein-bound and unbound RNA probes. This assay was replicated three times, and densitometry was used to quantify the bands that were present in both the motif and scramble probe lanes. The intensity of these bands was normalized to the intensity of the unbound probes from the corresponding lane, and the normalized intensity of the band in the scramble lane was set to one for comparison.

**Identification of proteins that interact with motifs identified in PPSs**

We used five of the most enriched motifs that we identified within PPS sequences (Figure 5 and Supplemental Table S2) as baits to isolate interacting RBPs by RNA-affinity chromatography. Specifically, RNA baits (covalently-linked to agarose beads) containing the identified motif of interest (IDT; Coralville, IA, USA) were incubated in a binding reaction (3.2 mM $MgCl_2$, 20 mM creatine phosphate, 1 mM ATP, 1.3% polyvinyl alcohol, 25 ng of yeast tRNA, 70 mM KCl, 10 mM Tris (pH = 7.5), 0.1 mM EDTA) with 56 μg of 4-week-old Arabidopsis whole leaf lysate at RT for 30 minutes. Beads were washed four times with GFB-100 (20 mM TE, 100 mM KCl) plus 4 mM $MgCl_2$ and once with 20 mM Tris-HCl (pH = 7.4). The RNA-bound proteins were then directly trypsinized on the beads.

**MS-ready sample preparation**

Multiple independent samples for the selected motifs and their corresponding controls were used to average out experimental variability, optimize detection limits, and improve signal to noise ratio for robust specific identification. MS sample preparations and analyses were performed as described previously (Onder et al., 2008; Onder et al., 2006). Briefly, RNA-bound proteins were treated directly on the beads with 100 mM $NH_4HCO_3$ containing ~6 ng/μl of MS-grade trypsin (Promega; Madison, WI, USA) and incubated at 37ºC for 12-18 hrs. These samples were extracted first with 1%HCOOH/2%$CH_3CN$, and several times with 50% $CH_3CN$; combined peptide extracts were vacuum dried and desalted using a ZipTip procedure before resuspending in ~5-10 μL LC buffer A (0.1% HCOOH (v/v) in 5:95 $CH_3CN$:$H_2O$) for MS analysis.

**Mass Spectrometry Analyses**

Tryptic peptide extracts were analyzed using nLC-MS/MS (Dionex/LCPackings Ultimate nano-LC coupled to a Thermo LCQ Deca XP+ ion trap mass spectrometer) in duplicate. 1 μl of the peptide sample (in LC buffer A, 0.1% HCOOH (v/v) in 5:95 $CH_3CN$:$H_2O$) was first loaded onto a μ-Precolumn (PepMapTM C18, LC-Packings), washed for 4 min at a flow rate of 25 μl/min with LC buffer A, then transferred onto an analytical C18-nanocapillary HPLC column (PepMapAcclaim100). Peptides were eluted at 280 nl/min flow rate with a 120 minute gradient of LC buffers A and B (0.1% (v/v) formic acid in 80:20 acetonitrile:water) ranging from 5%-95% B. A fused silica emitter tip with 8 μm aperture (FS360-75-8-N-5-C12; New Objective) mounted to a Thermo nanospray ionization (NSI) source at 1.8 kV was used for positive ionization of peptides. Mass spectra were collected using Thermo Xcalibur 2.0 software. The top 3 principal ions from each MS scan were trapped and fragmented during the chromatographic gradient, using dynamic exclusion to maximize detection of ions (range 200-2000 m/z). The trapped ions were subjected to collision-induced dissociation (CID) with He, and ~4000 spectra (MS/MS) were collected to cover the entire chromatography elution profile.

**Spectral Data Analyses and Protein ID**

Experimentally collected MS/MS tandem data were searched against the Arabidopsis Proteome Database (NCBI, latest version) using Thermo Proteome Discoverer 1.4 software. The search was restricted to full trypsin digestion with a maximum of 3 missed cleavages and potential modifications for methionine (oxidation) and cysteine (carbamidomethylation); other parameters were standard for LCQ Deca XP+ instrumentation. Peptide filters were set to standard Xcorr vs charge state values; X corr = (1.5, 2.0, 2.25, 2.5) for charges (+1,+2,+3,+4), respectively. Spectral assignments were manually scrutinized to validate the reliability of the protein identifications. Mass spectral data are summarized in Supplemental Table 4. Raw mass spectral data for key peptides can be found at http://gregorylab.bio.upenn.edu/PIPSeq_AtTotalNuc.

***RIP-RT-qPCR***

RNA immunoprecipitaion (RIP) was performed on frozen four-week-old Col-0 leaves as described previously (Kupsch et al., 2012). To begin, the frozen leaves were manually ground and homogenized before crosslinking in nuclear purification buffer (20 mM MOPS (pH = 7), 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA) plus 1% (vol/vol) formaldehyde, rotating at RT for 10 minutes. One molar Glycine (Sigma-Aldrich; St. Louis, MO, USA) was added to a final concentration of 125 mM before an additional five minutes of rotation. The homogenized leaves were then washed twice with PBS followed by lysis and resuspension in RIP buffer (150 mM NaCl, 20 mM Tris (pH=8.6), 1 mM EDTA, 5 mM $MgCl_2$, 0.5% NP40, 1 tablet/10 ml protease inhibitor (Roche; Basel, Switzerland), 0.5 μl/ml RNaseOUT RNaseOUT (Life Technologies; Carlsbad, CA, USA). This lysate was then subjected to 30 min of sonication and centrifugation to remove any remaining precipitate. Eight microliters of α-CP29A (Kupsch et al., 2012) or α-His antibodies (MA1-21315; Thermo Scientific; Waltham, MA, USA) were added to 400 μl aliquots of lysate and incubated while rotating at 4°C. Protein A beads (Life Technologies; Carlsbad, CA, USA) were washed with RIP

buffer and added to the reaction for an additional one hour of rotation at 4°C, followed by four washes with RIP buffer. Immunoprecipitated RNA was then isolated using the miRNeasy mini kit (Qiagen; Valencia, CA, USA) and target specific reverse primers (Table S3) were used for cDNA synthesis using SuperScript II Reverse Transcriptase (Life Technologies; Carlsbad, CA, USA) following the manufacturers protocol. mRNA standards were amplified from Arabidopsis cDNA using the Phusion 2X High Fidelity PCR Master Mix (New England Biolabs; Ipswich, MA, USA) and used to create standard curves of each target during quantitative PCR performed as previously described (Younis et al., 2013).

***Western blotting***

Western blots using lysates from INTACT purified nuclei or 10-day-old seedlings were performed using $\alpha$-ACT8 (1:5,000), $\alpha$-PEPC (1:5,000; 200-4163S; Rockland; Boyertown, PA, USA), $\alpha$-RUBISCO (1:5,000; ab62391; Abcam; Cambridge, MA, USA), ), $\alpha$-BIP1 (1:200; sc-33757; Santa Cruz Biotechnology; Dallas, TX, USA), $\alpha$-CNX1 (1:2,500; AS12 2365; Agrisera; Vännäs, Sweden), $\alpha$-H3 (1:1,000; ab1791; Abcam; Cambridge, MA, USA), or $\alpha$-CP29A (1:5,000) antibodies were performed as previously described (Kupsch et al., 2012).

**SUPPLEMENTAL REFERENCES**

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acid Research *37*, W202-W208.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C.*, et al.* (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics *43*, 956-963.

Deal, R.B., and Henikoff, S. (2010). A Simple Method for Gene Expression and Chromatin Profiling of Individual Cell Types within a Tissue. Developmental Cell *18*, 1030-1040.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Z., S. (2010). agriGO: a GO analysis toolkit for the agricultural community Nucleic Acid Research *38*, W64-W70.

Elliott, R., Li, F., Dragomir, I., Chua, M.M.W., Gregory, B.D., and Weiss, S.R. (2013). Analysis of the Host Transcriptome from Demyelinating Spinal Cord of Murine Coronavirus-Infected Mice. PLoS ONE *8*, e75346.

Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acid Research *35*, W297-W299.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell *38*, 576-589.

Kupsch, C., Ruwe, H., Gusewski, S., Tillich, M., Small, I., and Schmitz-Linneweber, C. (2012). Arabidopsis Chloroplast RNA Binding Proteins CP31A and CP29A Associate with Large Transcript Pools and Confer Cold Stress Tolerance by Influencing Multiple Chloroplast RNA Processing Steps. The Plant Cell *24*, 4266-4280.

Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. (2012). Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. The Plant Cell *24*, 4346-4359.

Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Research *22*, 1184-1195.
Muiño, J.M., Kaufmann, K., van Ham, R.C.H.J., Angenent, G.C., and Krajewski, P. (2011). ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. Plant Methods *7*.
Onder, O., Turkarslan, S., Sun, D., and Daldal, F. (2008). Overproduction or absence of the periplasmic protease DegP severely compromises bacterial growth in the absence of the dithiol: disulfide oxidoreductase DsbA. Mol Cell Proteomics *7*, 875-890.
Onder, O., Yoon, H., Naumann, B., Hippler, M., Dancis, A., and Daldal, F. (2006). Modifications of the lipoamide-containing mitochondrial subproteome in a yeast mutant defective in cysteine desulfurase. Mol Cell Proteomics *5*, 1426-1436.
Sammeth, M., Foissac, S., and Guigo, R. (2008). A General Definition and Nomenclature for Alternative Splicing Events. PLoS Computational Biology *4*, e1000147.
Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., Milos, P.M., Barton, G.J., and Simpson, G.G. (2012). Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. Nature Structural and Molecular Biology *19*, 845-852.
Silverman, I.M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J.L., and Gregory, B.D. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biology *15*, R3.
Tong, C., Wang, X., Yu, J., Wu, J., Li, W., Huang, J., Dong, C., Hua, W., and Liu, S. (2013). Comprehensive analysis of RNA-seq data reveals
the complexity of the transcriptome in
Brassica rapa. BMC Genomics *14*, 689.
von Arnim, A.G., Jia, Q., and Vaughn, J.N. (2014). Regulation of plant translation by upstream open reading frames. Plant Science *214*, 1-12.

Younis, I., Dittmar, K., Wang, W., Foley, S.W., Berg, M.G., Hu, K.Y., Wei, Z., Wan, L., and Dreyfuss, G. (2013). Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. eLife *2*, e00780.