RNA structure landscape of *S. cerevisiae* introns

Ramya Rangan[1], Rui Huang[2], Oarteze Hunter[3], Phillip Pham[2], Manuel Ares, Jr.[3], Rhiju Das[1, 2, 4]

Affiliations:
[1]Biophysics Program, Stanford University, Stanford, CA, USA
[2]Department of Biochemistry, Stanford University, Stanford, CA, USA
[3]RNA Center and Department of Molecular, Cell & Developmental Biology, University of California, Santa Cruz, Santa Cruz CA, USA 95064
[4]Howard Hughes Medical Institute, Stanford University

Abstract

Pre-mRNA secondary structures are hypothesized to play widespread roles in regulating RNA processing pathways, but these structures have been difficult to visualize *in vivo*. Here, we characterize *S. cerevisiae* pre-mRNA structures through transcriptome-wide dimethyl sulfate (DMS) probing, enriching for low-abundance pre-mRNA through splicing inhibition. These data enable evaluation of structures from phylogenetic and mutational studies as well as identification of new structures within the majority of probed introns (102 of 161). We find widespread formation of "zipper stems" between the 5' splice site and branch point, "downstream stems" between the branch point and the 3' splice site, and previously uncharacterized long stems that distinguish pre-mRNA from spliced mRNA. Multi-dimensional chemical mapping reveals that intron structures can form *in vitro* without the presence of binding partners, and structure ensemble prediction suggests that such structures appear in introns across the *Saccharomyces* genus. We develop a high-throughput functional assay to characterize variants of RNA structure (VARS-seq) and we apply the method on 135 sets of stems across 7 introns, finding that some structured elements can increase spliced mRNA levels despite being distal from canonical splice sites. Unexpectedly, other structures, including zipper stems, can increase retained intron levels. This transcriptome-wide inference of intron RNA structures suggests new ideas and model systems for understanding how pre-mRNA folding influences gene expression.

Introduction

Introns are widely prevalent features of eukaryotic genomes. Many genes contain long stretches of these non-coding RNA sequences, which are ultimately excised from mRNA precursors through RNA splicing. In the splicing reaction, a large and dynamic macromolecular machine, the spliceosome, precisely recognizes and positions three key intronic sequences termed the 5' splice site, the branch point, and the 3' splice site, carrying out the two catalytic steps required for removing introns (Fig. 1A).[1] Despite their prevalence, the functional roles for many introns remain underexplored. In some cases, intron sequences beyond splice sites regulate gene expression by controlling splicing rates and promoting alternative splicing.[2,3] In addition, introns can contain functional non-coding RNAs, alter pre-mRNA decay rates, and facilitate the evolution of new genes.[4-10]

Intron RNA sequences are complex macromolecules that occupy an ensemble of secondary structures including single-stranded and double-stranded regions. Intron secondary structure can regulate splice site selection and splicing efficiency, and structure in nascent RNA can function in numerous other nuclear processes such as RNA editing and RNA end-processing.[11,12] *S. cerevisiae* provides a useful model system for studying the role of pre-mRNA secondary structures, with the catalytic steps of splicing, the spliceosomal machinery, processes of RNA modification, and RNA decay pathways highly conserved across eukaryotes from *S. cerevisiae* to humans.[1,13,14] Studies in *S. cerevisiae* have revealed that intron stems called "zipper stems" can link the 5' splice site to the branch point,[15-18] and hairpins can lower the effective distance between the branch point and 3' splice site.[19] These structural motifs are hypothesized to enable efficient splicing by co-localizing key splicing sequences. In addition, pre-mRNA structures can

bind to protein products translated from their corresponding mRNA to regulate autogenous gene expression control at the level of splicing.[7,20-22]

These case studies highlight the potential variety of functions for pre-mRNA secondary structures, but the extent to which these findings extend across the transcriptome is unknown. We are missing a broad experimental survey of structures in introns across the transcriptome of any organism, and functional data that test the role for native intron structures is limited. Though scans for covariation across intron sequence alignments have been able to identify potentially functional structures,[23] signals for covariation are limited by high sequence divergence in intronic regions, and functional structures from some prior case studies have not been identified by these scans.[20-22,24] Moreover, while prior experiments have tested individual sequence variants for some intron structures,[3,25] deep mutagenesis aiming to rationally interrogate intron structures has been limited.[26] At this stage, the structural landscape and functional roles for intron secondary structures across any transcriptome are largely unknown.

In vivo chemical probing approaches provide an avenue for obtaining necessary structural data on RNA across the transcriptome.[27] However, low abundance introns sequences typically escape detection by these methods which have high sequencing coverage requirements. Here, we use splicing inhibition to enrich for unspliced pre-mRNA in transcriptome-wide structure probing experiments, assessing the presence of structural motifs in *S. cerevisiae* introns *in vivo* and identifying patterns in accessibility and folding that distinguish introns from coding regions. We combine this structural information with phylogenetic analysis across a broad set of yeast genomes, and we develop a strategy for high-throughput analysis of variants of RNA structure (VARS-seq) to evaluate levels of unspliced and spliced RNA for 135 sets of stems in 7 introns. Our combined structural, functional, and evolutionary analysis provides an atlas of pre-mRNA structural elements that serves as a foundation for understanding the previously hidden roles that introns play in gene regulation and genome evolution.

Results

Transcriptome-wide structure probing with splicing inhibition in *S. cerevisiae*. Structure probing experiments like DMS-MaPseq[27] can evaluate the formation of RNA secondary structures across the transcriptome. However, since splicing proceeds rapidly in yeast,[28] the coverage of pre-mRNA is limited in existing DMS-MaPseq datasets for *S. cerevisiae*, preventing the analysis of introns. Recently, *S. cerevisiae* strains with mutations in U2 snRNP component HSH155 (SF3B1 in human) have been developed that sensitize these yeast to splicing inhibition by Pladeinolide B (pladB).[29] Typically, splicing proceeds in two catalytic steps, with sequences in the intron termed the 5' splice site and the branch point interacting in the first catalytic step, and with the 5' splice site and another sequence, the 3' splice site, participating in the second catalytic step (Fig. 1A).[1] By using a yeast strain carrying mutations in HSH155 HEAT repeat domains 15-16 to match sequences from human SF3B1, we enable pladB to bind the yeast U2 snRNP and stall the assembling pre-spliceosome (A-complex) before the first catalytic step of splicing has occurred (Fig. 1A). For our probing experiments, we chemically inhibit splicing with 5 μM pladB for 1 hour prior to DMS treatment, library preparation, and sequencing (Fig. 1B).

As a confirmation of this strategy, treatment with pladB led to an accumulation of pre-mRNA in constructs tested by RT-PCR, increasing the proportion of unspliced mRNA for RPL36B and MATa1 (Fig. 1C). The sequencing data obtained from DMS-MaPseq additionally demonstrated increased intron retention for many genes, with read coverage in introns increasing from negligible levels prior to pladB treatment to levels approaching that of coding regions (Fig. 1D). Most intron-containing genes showed increased intron retention upon pladB treatment, with a median ratio of retained intron fraction (RI fraction) in pladB versus control of 10.0 (Fig. 1E). Before pladB treatment, 143 out of 288 annotated introns had an RI fraction less than 0.05, and only 28 introns had an RI fraction over 0.5. However, after pladB treatment, only 11 introns had an RI fraction less than 0.05, and 180 introns had an RI fraction over 0.5. Longer introns (over 200 nucleotides) and introns in ribosomal protein genes (RPGs) exhibited increased RI fractions upon pladB treatment compared to other introns (Fig. S1).

To assess the quality of our DMS-MaPseq reactivity data, we computed per-base mutational frequency distributions and evaluated reactivity profiles for control RNAs with known structure. As expected, DMS primarily modified A and C residues, leading to mutational frequencies of 2.7% and 2.3% respectively, and G and U residues exhibited mutation frequency rates closer to background (0.3 - 0.4%) (Fig. S2A). The DMS reactivity data for ribosomal RNA (rRNA) aligned closely with prior experiments, with mutational frequency values for the 18S and 25S rRNA highly correlated with values from an earlier DMS-MaPseq study (Pearson correlation's coefficient r=0.91, Fig. S2B).[27] Moreover, DMS reactivity values permitted successful classification of accessible (solvent accessible and not Watson-Crick-Franklin base-paired) from inaccessible (Watson-Crick-Franklin base-paired) rRNA residues, with an AUC of 0.91 (Fig. S2B). Finally, the reactivity profiles generated for stem-loops in the HAC1 and ASH1 mRNAs align with the well characterized secondary structures for these regions,[27] with base-paired residues less reactive than positions in loops (Fig. S2C).

The global increase in the proportion of unspliced mRNA allows for more robust reactivity calculations for intron regions from DMS-MaPseq. The reactivity values for 54.8% of introns exhibit R-value correlations of over 0.5 between replicate experiments, and introns with higher sequencing coverage have better replicability between experiments (Fig. S2D). Furthermore, with high expression levels, most of the introns in RPGs have sufficient coverage for reactivity-guided structure prediction. We proceeded with combining DMS data from both replicates and focused further analyses on the 161 introns meeting the coverage threshold for having a between-replicate reactivity correlation of r>0.5.

Assessing structure prediction from DMS-MaPseq after splicing inhibition. For each of the 161 introns with sufficient coverage, we predicted secondary structures guided by DMS data using RNAstructure[30,31], and we additionally assigned confidence estimates for helices in the secondary structure by performing nonparametric bootstrapping on reactivity values.[32] Helix confidence estimates from bootstrapping have empirically been useful for improving the quality of stem predictions.[32] To assess the impact of helix confidence estimate cutoffs on stem predictions, we additionally made DMS-guided structure predictions for a range of positive control secondary structures represented in our DMS-MaPseq dataset, including rRNAs, snRNAs, tRNAs, and mRNA segments. Helix confidence estimate thresholds improved the positive predictive value (PPV) and F1 score for stem predictions (Fig. S3A, Fig. S3C). When using reactivity data to make DMS-guided structure predictions) with a helix confidence estimate

threshold of 70% and minimum stem length cutoff of 5 base-pairs, predicted stems had a PPV of 82.3% (Table S1). DMS-guided structure prediction with helix confidence estimation was able to recover stems with this accuracy even for larger control RNAs like the U1 snRNA (Fig. S3D-E, Table S1). We therefore designated stems with over 70% helix confidence estimates as high confidence stems, aiming to prioritize PPV while maintaining reasonable sensitivity (66.2%, Fig. S3B). With this approach, we identified 3824 high confidence stems across the 161 introns. The majority of introns (102 of 161) had at least one such high confidence stem. We explored additional structure prediction approaches to identify potential pseudoknots and alternative conformations in introns using the DMS data, finding that no introns included high confidence predicted pseudoknots, and that all introns tested for multiple conformations were best explained by a single structure (see Supplemental Text).

We were concerned that the use of PladB to permit detection of intron structure might result in alterations in intron folding. To test this, we examined the three introns that gave sufficient sequencing data to meet our coverage thresholds in the DMS-MaPseq experiments carried out with and without pladB treatment. For these three introns, reactivity profiles are similar between conditions with and without pladB treatment, with highly reactive positions often shared between these conditions (Fig. 1F). The reactivity values at A and C residues gave an R-value between these conditions of 0.76, 0.93, and 0.84 for the introns in RPL26B, RPL28, and RPS13 respectively. Additionally, the secondary structures predicted for these three introns using DMS agreed well between conditions with and without pladB treatment, with all high confidence stems (helix confidence estimate over 70%) shared between conditions (Fig. S4). These comparisons suggested that structural profiles adopted by introns after pladB treatment would be informative for evaluating and discovering intron structures in untreated *S. cerevisiae*.

Support for proposed functional structures in *S. cerevisiae* introns. Secondary structures have been proposed for some introns in *S. cerevisiae* through prior functional experiments,[7,15,16,19-22,24] covariation scans that pinpoint functional base-pairs,[23] or *de novo* prediction methods based on conservation in sequence alignments.[33] With our structure data across introns, we were able to evaluate the support for these proposed structures *in vivo*. We note that our DMS-guided stem predictions include some false positives and false negatives (16.4% false positive rate and 33.8% false negative rate on controls, Fig. S3A-B). Additionally, in cases where DMS data do not lead to confident stem prediction, the proposed structure may still be a functional conformation that represents a minor portion of the intron's structural ensemble or appears only in a particular cellular state.

Using our DMS data, we compared the quality of structures proposed from prior functional experiments, computational structure predictions, and scans for covariation. We first focused on seven introns that included regulatory structures identified in functional experiments: introns in RPS17B,[15,16] RPS23B,[19] RPL32,[20] RPS9A,[7] RPS14B,[21,22] RPL18A[24] and RPS22B.[24] Most structures in this class received medium or high support from our DMS data (Fig. 2A-E) with only two exceptions (Fig. S5A-B), in general validating structures found in these experiments that assessed mutants and compensatory mutants (Fig. 2G, see Supplemental Text for details). In contrast, structures identified through computational predictions in Hooks, et al. (2016)[33] using CMfinder,[34] RNAz,[35] and Evofold[36] showed low to medium support from DMS data (Fig. S6A-G), suggesting that these approaches less reliably identify structures that form *in vivo* compared to functional experiments (Fig. 2G, see Supplemental Text for details.) Finally, we evaluated

structures detected by R-scape, which identifies pairs of residues with significant covariation compared to phylogenetic sequence backgrounds. In particular, we evaluated DMS support for covarying residues identified by R-scape in Gao, et al. (2021),[23] and we additionally evaluated 11 introns with significantly covarying residues from our own R-scape scan that used less stringent cutoffs (Fig. S7, see Methods). As a positive control for functional secondary structures in introns, we noted that seven of the eight *S. cerevisiae* introns that encode snoRNAs[37] included significantly covarying residues, suggesting that the remaining introns with covariation may also encode functional structural elements. Since most intron structures with covarying residues were supported by DMS data (Fig. S8, Fig. 2D-F) with one exception (Fig. S5C), DMS data largely validated covariation both in snoRNAs as well as other intron structures (see Supplemental Text for details). The prevalence of DMS-validated covarying residues suggests that covariation from R-scape[38] can reliably identify structures that form *in vivo* (Fig. 2G). Thus, overall, we found that compared to structures predicted by CMfinder,[34] RNAz,[35] and Evofold[36], structures that were identified through functional experiments and R-scape[38] covariation were more consistently supported by DMS data.

New structures found by probing *S. cerevisiae* pre-mRNA. Having established above that our DMS-guided structure predictions are coherent with previously identified intron structures, we inspected the 161 introns passing coverage thresholds for the presence of new intron structures and enriched structural features. First, these DMS data support the widespread formation of zipper stems, i.e., stems that reduce the distance between the 5' splice site and branch point. The potential for forming zipper stems has been noted in various introns,[17] and a zipper stem in RPS17B has been shown to be essential for efficient splicing.[15,16] To identify zipper stems across introns, we sought to precisely define the positional constraints on zipper stem formation, modeling intronic stems with various linker lengths in the context of the A-complex spliceosome using Rosetta[39] (Fig. S9A-B). Based on this modeling, we defined zipper stems as the longest stem with between 42 and 85 total nucleotides linking the stem, the 5' splice site, and branch point sequences. We find that high confidence zipper stems (including at least 6 base-pairs with at least 70% helix confidence estimate) are predicted for 42 of 161 introns with sufficient read coverage (Fig. 3A). Reactivity values align with the formation of these zipper stems, with stem residues having low reactivity values while neighboring loop residues are highly reactive (Fig. 3A).

We next noted the presence of stems between the branch point and 3' splice site across introns, which we term "downstream stems". It has been proposed that these stems serve to decrease the effective distance between the branch point and 3' splice site to facilitate splicing.[19] We find that downstream stems (at least 6 base-pairs in the stem with at least 70% helix confidence estimate) are present in 30 of 161 introns (Fig. 3B). Again, the reactivity profiles for these stems, for instance in the case of RPL40B, RPS14A, and RPS23B, show higher reactivity in loop or junction residues than in base-paired positions, supporting the *in vivo* formation of these structures (Fig. 2E, Fig. 3B).

Beyond previously proposed structures and structural patterns, DMS reactivity for some introns suggest new elaborate extended secondary structures with long stems and multiway junctions. For instance, in the case of RPL28, in addition to a zipper stem connecting the 5' splice site to the branch point and a downstream stem between the branch point and 3' splice site, the pre-mRNA includes a three-way junction and long stems of over 10 base-pairs through the length of

the intron (Fig. 3C). We speculate that these secondary structures allow internal stems to extend beyond the spliceosome, potentially poised to interact with other factors. In fact, when the 3D structure of the RPL28 intron is modeled in the context of the A-complex spliceosome, the introns' stems extrude from the core spliceosome, potentially enabling binding partners to avoid steric clashes with the splicing machinery (Fig. 3D). Intron structures can extend beyond the spliceosome even for shorter introns like RPL36B (238 nucleotides), facilitated by stems through the length of the intron (Fig. S9C).

Introns are more structured than coding regions and unspliced decoys. Our DMS-guided structure predictions reveal that intron secondary structures have distinct properties compared to the secondary structures adopted by coding regions of the mRNA. As found in mammalian compartment-specific structure probing,[40] intron regions have higher Gini coefficients, a measure quantifying the extent to which reactivity values diverge from an even distribution (Fig. 3E). Therefore, introns include more non-random secondary structure elements compared to coding regions. Additionally, intron structures predicted using DMS data extend further from their sequence endpoints than coding regions, which we measure as the normalized maximum extrusion from ends (MEE) (Fig. 3F). To compute the extrusion from ends metric, we measure the distance from positions in the sequence to the sequence endpoints (splice sites for an intron sequence) in a graph-based representation of the secondary structure, with higher values indicating that the structure extends further from the sequence endpoints. Introns additionally include longer stems with helix confidence estimates of at least 90%, with some high confidence stems extended to over 20 base-pairs (Fig. 3G). Finally, intron stems with at least 6 base-pairs in the predicted minimum free energy structure tend to have higher helix confidence estimates on average, indicating that they more often included reactivity values that agreed well with the modeled secondary structure (Fig. 3H). Together, these metrics demonstrate that compared to coding regions, introns in *S. cerevisiae* adopt extended secondary structures with non-random local structures and long stems supported by DMS-guided structural analysis.

The secondary structure patterns enriched in introns may distinguish spliced introns from other transcribed sequences that do not splice efficiently. To explore this possibility, we assembled a set of unspliced decoy introns in *S. cerevisiae* that included consensus 5' splice site, branch point and 3' splice site sequences in positions that matched canonical introns' length distributions (Fig. S10A). When comparing DMS-guided structure predictions for authentic intron sequences to those for control sequences shifted to nearby genomic regions, authentic introns exhibited more stable (lower ΔG) zipper stems, more stable downstream stems, longer stems, higher maximum extrusion from ends, and higher secondary structure distances from the 5' splice site to the branch point (Fig. S10B). In contrast to spliced introns, the DMS-guided secondary structures of unspliced decoys were not enriched for these features (Fig. S10B).

Comparing *in vivo* intron RNA folding with *in vitro* folding. Having identified extended structures and long stems in *S. cerevisiae* introns, we next sought to understand whether these structures would be present in introns folded *in vitro*, when potential protein binding partners and other components of the nuclear environment are missing. We initially assayed intron structures *in vitro* by refolding RNA extracted from yeast and probing accessible residues with DMS. We analyzed *in vitro* reactivity profiles for 81 introns meeting the threshold for 0.5 average replicate reactivity correlation (Fig. S11A) and found that in this case, introns were largely unfolded (Fig. S11B-C, Fig. S12, see Supplemental Text for details). We hypothesized that introns could be

unfolded in *in vitro* refolded total RNA due to intermolecular interactions in the complex RNA mixture.

To better assess the folding of individual intron RNAs *in vitro* and confirm intramolecular base pairing, we DMS probed isolated introns transcribed *in vitro* and additionally made use of mutate-and-map readout through next-generation sequencing (M2-seq[41]), which can identify base-pairing residues in addition to providing average per-residue accessibility data. We chose five introns from DMS-MaPseq that included zipper stems and had lengths short enough to map completely with *in vitro* M2-seq. In the case of the introns in QCR9 and RPL36B, Z-score plots from *in vitro* M2-seq included off-diagonal signals that are indicative of the presence of stems, and high base-pairing probabilities support the formation of these stems as these introns' primary *in vitro* structure (Fig. 4A-B and Fig. S13A-B, see Supplemental Text for details). Secondary structures predicted from M2-seq agree with the stems observed *in vivo*, with most high confidence stems shared between these structures (Fig. 4C-D, Fig. S13C-D). For the introns in RPS11A, RPL37A, and RPS7B, though M2-seq Z-scores did not include visually apparent off-diagonal signals, helix confidence estimates computed using M2-seq data showed support for the stems observed *in vivo* (Fig. S14). Although these structures are disrupted during *in vitro* refolding of total RNA when introns are exposed to flanking sequences and other RNA molecules (Fig. S15), our *in vitro* M2-seq results suggest that intron sequences have the capacity to form structures found *in vivo*, even when outside the context of the nucleus and without protein binding partners.

Structural landscape for *S. cerevisiae* introns. To understand the distribution of structural patterns in introns across the yeast transcriptome, we clustered introns based on their structural features. For each intron with sufficient DMS-MaPseq coverage, we assembled a full set of features using probing data: zipper stem and downstream stem free energy, the maximum extrusion from ends, the length of the longest stem, the average helix confidence estimate from bootstrapping for stems in the intron, the maximum Gini coefficient window, and the accessibility (average DMS reactivity) across the 5' splice site, branch point, and 3' splice site. These secondary structure features and DMS-guided secondary structures for *S. cerevisiae* introns are tabulated in Table S2.

The clustered heatmap (Fig. 5) depicts the global distribution of secondary structure features across all introns with sufficient sequencing coverage. The first class of introns includes 12 that have both a zipper stem and downstream stem, with high average helix confidence estimates (Class 1, blue in Fig. 5). Class 2 includes 30 introns with zipper stems, and within this class, introns with more stable zipper stems clustered together (orange Fig. 5). The next class includes 18 introns with downstream stems; this class included lower longest stem lengths compared to introns with zipper stems (Class 3, yellow in Fig. 5). In *S. cerevisiae*, intron lengths are bimodal, with a set of shorter introns (shorter than 200 nucleotides, with most around 80 nucleotides) and a set of longer introns (with most between 400 and 500 nucleotides).[42] Our structure probing data had sufficient coverage for 45 short introns and 116 long introns. Of the long introns, a sizable fraction include either a zipper stem or downstream stem in Class 1-3, with 51.7% of long introns including at least one of these motifs. Class 4 includes structured short introns with at least one high confidence stem (green in Fig. 5). With only 5 introns in this class, a minority of short introns show signal for high-confidence structure. The fifth class includes 13 structured long introns that do not meet the criteria for zipper stems or downstream stems, and yet include

structures with high helix confidence, long stems, and windows with high Gini coefficients (Class 5, pink in Fig. 5). In contrast, the next class includes long introns which in some cases do not include high Gini coefficients or do not include long stems, indicating that introns in this class tend to be less structured than Class 5 (Class 6, brown in Fig. 5). Class 7 includes most short introns, which tend to be depleted of stem structures, without zipper stems, downstream stems, high helix confidence estimates, or high Gini coefficient windows (purple in Fig. 5). Finally, Class 8 includes a small set of long introns depleted of structure, including only short stems with low helix confidence (red in Fig. 5).

A high-throughput assay to test the function of intron stems. To understand the roles that intron structures play in influencing gene expression, we developed a high-throughput assay for evaluating variants of RNA structure (VARS-seq). VARS-seq includes a variant design pipeline for specifically interrogating structures of interest, a functional assay with a sequencing readout, and an analysis strategy for measuring effects. While VARS-seq's variant design approach can be used for various functional readouts, in our case, we measured spliced and unspliced mRNA levels for intron structure variants, expecting that these structures may influence splicing or pre-mRNA decay rates. Structures that slow pre-mRNA decay or splicing could lead to an accumulation of unspliced mRNA, whereas structures that increase splicing rates could in turn increase spliced mRNA levels (Fig. 6A). We would for instance expect that zipper stems would contribute to higher levels of spliced RNA, since they are hypothesized to improve splicing efficiency by co-localizing the 5' splice site and branch point. We chose 7 structured introns to assay with VARS-seq, including introns in Class 1 and Class 2 with zipper stems (RPL7A, QCR9, RPL36B, and RPL28), covarying base-pairs (RPS9A in Class 5, RPS9B in Class 5, and RPL7A in Class 2), and long stems that distinguished pre-mRNA from coding RNA.

For each intron, we designed variants that systematically mutated secondary structure elements and rescued them with compensatory mutations where possible (Table S3). More specifically, for each target set of stems of interest, we chose variants that were computationally predicted to disrupt base-pairing within the stem set while maintaining base-pairing elsewhere in the structure (Fig. 6B, Fig. S16A-B). Where technically feasible within the length limits of library construction (see Methods), we also included rescue sequences that were computationally predicted to restore the native secondary structure, with compensatory mutations for each primary mutation (Fig. 6B, Fig. S16A-B). We ultimately designed 4-8 distinct intron variants for each target set of stems and loops, with around 200 variant sequences for each intron (Table S3). To obtain reliable measurements of unspliced and spliced mRNA levels, we integrated these intron variant libraries into the yeast genome[43] in their native gene context (Fig. 6C, Fig. S16C), and we installed unique random barcodes upstream of intron variants to help match spliced RNAs to their original pre-mRNA variant (Fig. S16C-D, Fig. S17). With multiple barcodes and variant sequences assigned to each stem set of interest, we were able to observe even subtle effects on gene expression (see Supplemental Text for details). Ultimately, using our sequencing data, we computed two key readouts for each barcode: the retained intron (RI) fraction (fraction of total RNA that is unspliced) and the normalized mRNA level (spliced mRNA levels normalized by transformation frequencies from genomic DNA sequencing).

Across the 7 tested introns, we found that structure-disrupting mutations often altered pre-mRNA and mRNA levels compared with wildtype sequences (Fig. S18). We detected statistically significant effects on RI or normalized mRNA in many of the tested stem sets (52 of 136) and

loops (4 in 15). Some stem disruptions significantly decreased RI levels (Fig. 6D), while others significantly increased them (Fig. 6E). For many introns, independently designed mutations in overlapping sets of stems produced similar effects in our assay, providing a useful cross-check for our approach while also suggesting larger functional domains (Fig. S18A-D). For instance, for all four stem sets that included nucleotides 45-48 in the RPL28 intron, independently designed variants mutating these residues significantly decreased RI (Fig. 6D, Fig. S18A). Since mutations in junction residues 36-42 produced a similar effect (Fig. 6D), we hypothesize that positions 36-48 in the wildtype RPL28 intron help to reduce splicing efficiency or slow pre-mRNA decay. Similarly, all mutations to residues 73-96 of the RPL28 intron show increased RI (Fig. 6E), suggesting that the wildtype nucleotides promote splicing or pre-mRNA decay. Our secondary structures from DMS-MaPseq demonstrate that in this case and others, mutations can influence gene expression even at positions distant from splice sites in introns' secondary structures (Fig. 6E, Fig. S18). Additionally, we find that zipper stem variants in the RPL28 intron (Fig. 6D) along with introns across out dataset (i.e. in RPL7A, QCR9, and RPL36B) lower retained intron levels (Fig. S18-19). Counter to the simple model that intron zipper stems enable efficient splicing by co-localizing splice sites, these data suggest that these stems may inhibit splicing or pre-mRNA decay.

In cases where compensatory mutations were accessible within our library's length constraints, our stem disruption and rescue variants pinpoint functional intronic structural elements. For instance, in the case of RPL36B, we identified structures that facilitate splicing, such that sequence variants in this region reduced normalized mRNA levels and rescue variants restored higher mRNA levels (Fig. 6F). In contrast, in the case of RPS9A, we identified a functional stem where stem disruptions reduced RI and stem rescue restored RI (Stem 191-195, Fig. 6G). An analogous stem-loop is present in the RPS9B intron with similar effects on RI levels (Stems 165-183, Fig. S18B). These stem-loops in RPS9A and RPS9B likely influence gene expression by inhibiting splicing or pre-mRNA decay, aligning with the prediction for covariation in these stems (Fig. 2H-I) and corroborating a prior study[3] for an RPS9A intron variant in this region. In these cases with disruption and rescue variants, VARS-seq is able to identify intron structures where base-pairing is critical for influencing gene expression. To enable further dissection of effects of intron stem variants, we summarize all variant and rescue comparisons in heatmaps in Fig. S19 and we include per-barcode spliced and unspliced RNA counts in Table S4.

Computational prediction of enriched structural patterns across the *Saccharomyces* genus. Our transcriptome-wide structure mapping provides evidence for widespread structure in *S. cerevisiae* introns, and with VARS-seq, we find that intron structures can impact gene expression regulation. To extend our observations to other species in the *Saccharomyces* genus, we turned to computational structure prediction. We first evaluated whether *de novo* secondary structure prediction could recapitulate structural features observed through DMS-guided structure analysis, comparing structural features between introns and length-matched controls (Fig. 7A, Fig. S20). For *de novo* structure prediction, as a first approach, we predicted minimum free energy structures,[30] and as a second approach, we generated secondary structure ensembles through stochastic sampling of suboptimal structures,[44] which has previously enabled the study of structural patterns in introns.[16,19] With either of these approaches, structural features from *de novo* secondary structure prediction were largely consistent with patterns from DMS-guided structures for *S. cerevisiae* (Fig. 7B, Fig. S21, see Supplemental Text for details).

Having confirmed coherence between DMS-guided and *de novo* prediction methods for *S. cerevisiae* introns, we made *de novo* structure predictions across species in the *Saccharomyces* genus, focusing on the 20 species in this genus with curated intron alignments from Hooks et al. (2014).[45] Numerous secondary structure patterns remain enriched in introns across the *Saccharomyces* genus. In particular, when compared to shuffled sequence controls, introns across the genus include more stable zipper stems and downstream stems (Fig. 7C), along with higher maximum extrusion from ends and shorter distances between the 5' splice site and branch point (Fig. S22A). Furthermore, many of these features remain enriched when comparing each species' introns to phylogenetic controls, which include random sequences constructed to match the mutation and indel frequency between an intron and its homologous *S. cerevisiae* intron (Fig. S22B).

Secondary structure patterns are maintained across *Saccharomyces* species despite extreme sequence-level divergence between introns, suggesting that these structures have conserved functions. Complete intron deletions between these species are common, with many introns having orthologs in only a subset of the *Saccharomyces* species (Fig. S22C), and most intron regions have low sequence conservation between species, with most intron sequences 50-60% conserved between these species (Fig. S22D). Some key functional intervals in introns are more conserved across the *Saccharomyces* species, with higher sequence conservation (72.3% to 83.6%) across the seven snoRNAs that are included in these intron sequence alignments. In contrast, zipper stem regions in *S. cerevisiae* introns diverge significantly between *Saccharomyces* species, with most zipper stems only 20-30% conserved. Strikingly, although most species in the genus include zipper stems in 20-40% of their introns, this structural motif appears in disparate introns across species, with many zipper stems only present in a small number of orthologous introns (Fig. 7D). Therefore, in some introns, zipper stems may have evolved separately across species, pointing to a potential functional role for these stems. It is possible that functional secondary structures in these regions have been challenging to find by covariation due to high intron sequence divergence in closely related fungal genomes.

Discussion

RNA structures play critical roles in regulating a wide array of nuclear processes including transcription, RNA modification and editing, and splicing. Intronic RNA structures are poised to participate in these processes, for instance by altering splicing kinetics, by changing RNA decay rates, or by interacting with other nuclear factors and performing orthogonal regulatory functions. Here, we sought to understand the structural landscape of introns in *S. cerevisiae*, a model system for eukaryotic splicing. First, we evaluated the presence of structural patterns through transcriptome-wide DMS probing while enriching for unspliced RNA, revealing extended secondary structures in introns that distinguished these regions from coding mRNA. These structural data allow for clustering introns into eight classes, with most falling into classes that contain introns with zipper stems or downstream stems (present in 51.7% of the long introns), long introns with intermediate structure, and unstructured short introns. In Fig. S23 and Fig. S24, we display secondary structures and reactivity profiles for all introns from DMS-MaPseq, grouped into these classes. With a high-throughput structure-function assay, VARS-seq, we evaluated the roles for these structures through deep mutagenesis of 7 introns, finding structural elements that influence spliced and unspliced mRNA levels. Finally, through

computational structure prediction, we identified signals for structure in introns across the *Saccharomyces* genus.

Structure probing experiments have enabled the measurement of RNA structure transcriptome-wide, but these experiments are often unable to provide information for low abundance transcripts including many unspliced pre-mRNA molecules. While low abundance introns have been probed previously by target-specific enrichment[27,46] or nuclear RNA enrichment,[40,47] here we broaden the capture of these targets by using global splicing inhibition. It is possible that splicing inhibition could alter the structures of pre-mRNA compared to untreated cells, as accumulating pre-mRNA are expected to outnumber spliceosome components[48] and may instead be unbound or interacting with the nuclear pore complex,[49] nuclear exosome,[50] or nonsense mediated decay machinery.[51] Nevertheless, for the cases with sufficient sequencing coverage from experiments with untreated cells, we found similar structure profiles with and without splicing inhibition. Then, with DMS data for introns transcriptome-wide, we discerned structural patterns that distinguish introns from coding regions. For instance, with higher Gini coefficients and longer stems, stable structures were enriched in introns compared to coding regions. While it is possible that active translation may unwind structures in coding regions, our computational secondary structure predictions suggest that intron sequences have the capacity to form more stable structures regardless of their translation status. It is for instance possible that introns' intramolecular structures are necessary to avoid spurious interactions between intron RNA and other nucleic acids in the crowded nuclear environment, preventing gene expression dysregulation. Additionally, while there is evidence across species for depletion of double stranded RNA in coding RNA to avoid activating cytosolic antiviral cellular responses,[52] this selection would not be present for introns in the nucleus. Finally, without the evolutionary constraints of adhering to a viable coding sequence, introns may be free to form extended structures that can play roles in regulating a host of nuclear processes.

With structural motifs enriched in *S. cerevisiae* introns compared to coding regions and random control sequences, we hypothesized that these introns could harbor functional structures. While scans for covariation have provided evidence for functional structures in a handful of *S. cerevisiae* introns (RPS9A,[7] RPS9B, RPS13,[23] RPL7B,[23] and RPL7A), this analysis may miss functional structures due to the limited power of existing intron sequence alignments,[53] with high sequence variation and large-scale deletions often present between aligned intron sequences. By assaying intron variants and rescue sequences in high-throughput with VARS-seq, we identified additional functional structures in introns. While prior intron variant libraries have been designed to assess the effects of sequence motifs on splicing patterns,[54,55] libraries assessing the role for structured elements in introns have been limited. For interrogating the role of structure, it was critical to include both variant and rescue sequences where possible and to test a library of multiple sequences and barcodes rather than single sequence variants due to variation between constructs. With VARS-seq, we found that some intron structures increase retained intron levels and others increase spliced mRNA levels. In fact, some introns included domains that could influence gene expression despite being distal from splice sites in the intron's secondary structure. These examples challenge the notion that introns are largely nonfunctional junk DNA sequences, adding to the known complexity of potential functional roles for introns.[9,10]

Together, our structural, computational, and functional experiments point to the formation of structural motifs across introns that impact gene expression. One pattern that emerged was the

formation of zipper stems co-localizing the 5' splice site and the branch point. Introns included highly stable zipper stem structures *in vivo*, and these structures were additionally supported by multi-dimensional chemical mapping (M2-seq) *in vitro*. These zipper stems could enable efficient splicing by reducing the physical distance between the 5' splice site and branch point[56] or by enhancing specific interaction with the spliceosome, with intron helical density seen interacting directly with the E and pre-A spliceosome complexes.[25,57] Surprisingly, from VARS-seq we observe that disrupting intron zipper stems actually decreases retained intron levels, seemingly contradicting prior observations that zipper stems enable efficient splicing in the RPS17B intron.[15] However, zipper stems could both enable efficient splicing while slowing pre-mRNA decay, explaining how zipper stem disruption would lower retained intron levels from VARS-seq along with prior experiments.[15,26,58]

To interrogate the mechanisms by which these and other structured elements regulate gene expression, we will need customized experiments for individual introns. Structures located close to splice sites could have functional roles in recruiting the spliceosome, while structures extruded away from splice sites could play a role in recruiting other nuclear factors. Intron secondary structures can regulate splicing by sequestering alternative splice sites,[59,60] occluding exonic splicing enhancers,[61] co-localizing selected splice sites,[62] facilitating co-transcriptional splicing,[63] or mediating protein interactions that influence splicing patterns.[64] Additionally, these intronic structures could have regulatory functions in pathways orthogonal to splicing, much like the snoRNAs encoded in *S. cerevisiae* introns.[37] In fact, RNA secondary structures in introns associate with RNA-binding proteins involved in varied processes including transcription, tRNA and rRNA processing, ribosome biogenesis and assembly, and metabolic processes.[65,66] Furthermore, secondary structures in *S. cerevisiae* introns are poised to regulate gene expression in auto-regulatory circuits as seen previously in the cases of RPS14B[21] or RPS9A[7], with structural elements within a gene's introns binding its protein products and thereby downregulating subsequent splicing and gene expression. Finally, structures can play a functional role in regulating adaptation to starvation by influencing the accumulation of linearized introns under nutrient depletion,[5,6] and it will be interesting to probe structures in saturated-growth conditions or other stress conditions.

Our work identifies a set of intron structures with properties distinct from those in coding regions, support for *in vivo* formation from DMS-MaPseq, signals in other related yeast species, and in some cases, covarying residues. Furthermore, we identify structured intervals of introns that modulate gene expression, impacting levels of spliced mRNA and retained introns in cells. These functional experiments provide candidates for further mechanistic characterization and provide a glimpse into the broad regulatory potential for intron sequences beyond splice sites. The widespread presence of structured elements in *S. cerevisiae* introns raises the possibility that similar motifs and stable secondary structures play a role in introns in higher-order eukaryotes, perhaps forming regulatory elements in human pre-mRNA.

Online Methods

Strains, media, and growth conditions for DMS probing. The strain OHY001 was constructed from the background strain JRY8012,[67] which includes three deletions of ABC transporter genes (*prd5::kan*^{r}, *snq2::kan*^{r}, *yor1::kan*^{r}) to reduce drug efflux. OHY001 was generated via CRISPR

editing of JRY8012 to mutate portions of HSH155 HEAT repeat domains 15-16 to match the sequence found in human SF3B1 (see sequence in Table S5).

Strains were grown at 30 ^{o}C on YPD plates and in YPD liquid medium. Single colonies of OHY001 were used to inoculate overnight cultures, diluted to OD600 0.1, and grown to OD600 0.5-0.6. Biological replicates were obtained from distinct single colonies.

Splicing inhibition and DMS treatment. We carried out splicing inhibition and DMS treatment for two biological replicates, splicing inhibition only for a no-modification control, and DMS treatment only for the control without splicing inhibition. We grew 15 mL of culture to OD600 ~0.5 for replicate 1, 60 mL for replicate 2, 15 mL for the no-modification control, 20 mL for the condition without pladB treatment, 20 mL for *in vitro* replicate 1, and 20 mL for *in vitro* replicate 2. We treated cultures with 5 μM Pladeinolide B (pladB, Cayman Chemicals) using 50 μL of 1 mM pladB for every 10 mL of culture, and we incubated cultures at 30 ^{o}C for 1 hour with shaking. The condition without splicing inhibition was treated with an equal volume of DMSO.

For *in vivo* DMS modification, we treated cultures with 3% DMS or an equivalent volume of H_2O for the no-modification control. Treated cells were incubated with occasional stirring in a water bath at 30 °C for 5 minutes, and the reaction was quenched by adding 20 mL stop solution (30% 2-mercaptoethanol, 50% isoamyl alcohol) for every 10 mL of culture. The culture was mixed multiple times by inversion and transferred quickly to ice. Cultures were spun down for 3 minutes at 1500g at 4 ^{o}C, and washed first with 5 mL wash solution (30% 2-mercaptoethanol) for every 10 mL of culture. A second wash was performed with 3 mL YPD per 10 mL culture. RNA was extracted using the YeaStar RNA Kit (Zymo Research), using 7.5 μL of Zymolase for every 2.5 mL starting cell culture and shortening the Zymolase incubation to 15 minutes at 30 ^{o}C to reduce RNA fragmentation.

For *in vitro* DMS modification, we followed a protocol similar to the one used in Rouskin, et al. (2014).[68] We first obtained RNA from cultures after splicing inhibition, pelleting cells by spinning at 1500g for 3 minutes, resuspending in YPD, and extracting RNA with the YeaStar RNA Kit (Zymo Research) using 1 column for every 5 mL starting culture. We then re-folded RNA *in vitro*, first denaturing 200 μg of RNA at 95 ^{o}C for 2 minutes, cooling RNA on ice for 2 minutes, and then folding RNA at 30 ^{o}C for 30 minutes in 10 mM Tris HCl pH 8.0, 100 mM NaCl, and 6 mM $MgCl_2$. We treated RNA with 3% DMS at 30 ^{o}C for 5 minutes, and we quenched with 25% 2-mercaptoethanol to stop the reaction. RNA was purified by ethanol precipitation and eluted in 20 μL RNAse-free H_2O.

RT-PCR for verifying splicing inhibition. As initial verification of splicing inhibition by pladB, we used RT-PCR to compare unmodified RNA extracted after 1 hour of either 5 μM pladB or DMSO treatment. We first DNase treated the +pladB and -pladB samples by mixing 1 μL of TURBO Dnase (Thermo Fisher), 20 ng of RNA, and Rnase-free H_2O for a 20 μL reaction volume. We incubated the reactions at 37 ^{o}C for 1 hour, added 2 μL of 50 mM EDTA, incubated the reactions at 65 ^{o}C for 10 minutes, and placed them on ice. We then carried out reverse transcription with the iScript Reverse Transcription Supermix (Bio-Rad), using 10 μL of each RNA sample for a reaction that included the reverse transcriptase, and the remaining 10 μL for a control without this enzyme. 4 μL of the 5X RT Supermix was mixed with 10 μL of RNA

sample and water to a reaction volume of 20 μL, and the reaction was incubated in a thermocycler for 5 minutes at 25 ^{o}C, 40 minutes at 46 ^{o}C, and 1 minute at 95 ^{o}C. The samples were purified with Oligo Clean and Concentrator columns (Zymo Research) and eluted in 15 μL. PCR was then performed using 2 μL of cDNA template, 25 μL of the 2X NEBNext Ultra II Q5 Master Mix (NEB), 18 μL of H_2O, and 2.5 μL of 10 μM primers (RR063 and RR064 for RPL36B, and RR067 and RR070 for MATa1; Table S5). The reactions were denatured 98 ^{o}C for 30 seconds, followed by 30 cycles of 98 ^{o}C for 10 seconds, 56 ^{o}C for 30 seconds, and 72 ^{o}C for 30 seconds, and ending with a final extension period at 72 ^{o}C for 5 minutes. PCR products were visualized on a 2% agarose gel with ethidium bromide.

DMS-MaPseq sequencing library preparation. For preparation of DMS-treated RNA for sequencing, we first depleted the extracted RNA of rRNA using Rnase H to deplete rRNA with complementary oligos. We concentrated extracted RNA for each condition using RNA Clean and Concentrator-5 columns (Zymo Research), yielding 47.4 μg RNA for *in vivo* replicate 1, 200 μg RNA for *in vivo* replicate 2, 47.2 μg RNA for the no-modification condition, 53.2 μg RNA for the condition without splicing inhibition, 58.0 μg RNA for *in vitro* replicate 1, and 46.1 μg RNA for *in vitro* replicate 2. We pooled 108 50-mer oligos (RR-rRNAdep-1-108; Table S5) that tiled the 5S, 5.8S, 18S, and 25S rRNA in *S. cerevisiae*, and we concentrated these oligos with Oligo Clean and Concentrator column. Up to 40 μg of total RNA was included in each rRNA depletion reaction, with multiple reactions as needed for each sample. First, a 15 μL annealing reaction was prepared with total RNA, an equal mass of rRNA depletion oligos, and 3 μL of 5X hybridization buffer (500 mM Tris-HCl pH 7.5 and 1 M NaCl). For the annealing reaction, the reaction mix was heated to 95 ^{o}C for 2 minutes and the temperature was ramped down to 45 ^{o}C by 0.1 ^{o}C/sec with a thermocycler. 7.5 μL of Hybridase thermostable Rnase H (Lucigen) with 2.5 μL of 10X digestion buffer (500 mM Tris-HCl pH 7.5, 1M NaCl, 100 mM $MgCl_2$) was preheated to 45 ^{o}C. We combed the annealing mix and the Rnase H mix and incubated at 45 ^{o}C for 30 minutes. For each reaction, we then used an RNA Clean and Concentrator-5 column with the size-selection protocol to exclude RNA below a size cutoff of 200 nucleotides, cleaning up the shorter oligos used for rRNA depletion and removing small RNA from the total RNA sample. We then Dnase treated each reaction by combining each sample with 10 μL of TURBO Dnase in a 167 μL reaction. We purified the reaction using an RNA Clean and Concentrator-5 column, eluting in 9 μL of Rnase-free H_2O. After rRNA depletion and size-exclusion of small RNA, we obtained 1.76 μg RNA for *in vivo* replicate 1, 4.58 μg RNA for *in vivo* replicate 2, 1.64 μg RNA for the no-modification control, 1.57 μg RNA for the condition without splicing inhibition, 2.74 μg RNA for *in vitro* replicate 1, and 1.81 μg RNA for *in vitro* replicate 2. We split the RNA from replicate 2 into 4 samples with 1.15 μg RNA each to scale up the enzyme amounts used for the following library preparation steps.

We fragmented each reaction by combining 1 μL of 10X RNA Fragmentation Reagent (Ambion) with 9 μL of RNA sample, incubating for 8 minutes at 70 ^{o}C. We added 1 μL of the Stop Solution (Ambion), mixed thoroughly, and placed immediately on ice. Each reaction was cleaned with an RNA Clean and Concentrator-5 column, eluting in 11 μL of Rnase-free H_2O. We removed the 3' phosphate groups left by fragmentation by adding 1.5 μL rSAP (NEB), 1.5 μL 10X Cutsmart buffer (NEB), and 1 μL SUPERase inhibitor (Thermo Fisher), and incubating the 15 μL reaction at 37 ^{o}C for 1 hour. The reaction was stopped at 65 ^{o}C for 4 minutes.

We next ligated a universal cloning linker to the RNA to serve as a handle for reverse transcription. To prepare linker for this reaction, we first phosphorylated 1 nmol of the DNA universal cloning linker with a 3' amino blocking group (oligo RR118; Table S5) by adding 3 μL of T4 PNK (NEB), 15 μL of 10X T4 PNK buffer (NEB), and 15 μL of 10 mM ATP in a 150 μL reaction; incubating at 37 ^{o}C for 30 minutes; and inactivating at 65 ^{o}C for 20 minutes. We then purified the reaction with an Oligo Clean and Concentrator column. Next, we adenylated the linker by adding 20 μL of Mth RNA Ligase (NEB), 20 μL of 5' DNA adenylation reaction buffer (NEB), and 20 μL of 1 mM ATP to the sample in a 200 μL reaction. We incubated the reaction at 65 ^{o}C for 1 hour and inactivated the reaction at 85 ^{o}C for 4 minutes. We then purified the reaction with an Oligo Clean and Concentrator column and measured the final adenylated linker concentration. We added adenylated linker in 2-fold molar excess to each RNA sample, obtaining the molarity of the RNA samples by estimating the library to contain RNA of average length 150. 3 μL of T4 RNA ligase 2 truncated KQ (NEB), 3.5 μL 10X NEBuffer 2 (NEB), 1.75 μL 100 mM DTT, and 10 μL 50% PEG-8000 were added to adenylated linker and RNA sample to make a 35 μL reaction. The reaction was incubated at 25 ^{o}C for 2 hours in a thermocycler. RNA ligated to the DNA linker was purified using an RNA Clean and Concentrator-5 column, eluting in 15 μL of Rnase-free H_2O. The excess DNA linker was then degraded by adding 1 μL of 5' Deadenylase (NEB), 1 μL of RecJf (NEB), 1 μL of SUPERase inhibitor, and 2 μL of NEBuffer 2 in a 20 μL reaction. The mixture was incubated at 30 ^{o}C for 1 hour and purified with an RNA Clean and Concentrator-5 column.

We then proceeded to reverse transcription (RT) with mutational readthrough and denaturing PAGE gel size selection of the resulting cDNA library. For the RT primer, we used oligo RR114 (Table S5) which included a sequence complementary to the universal cloning linker, a 5' phosphate modification that would allow for circularization after RT, a 10-nucleotide randomized UMI sequence, and sequences complementary to Illumina sequencing primers that are separated by a spacer to allow for PCR amplification of the final library. We added 2 μL of 5X TGIRT buffer (250 mM Tris-HCl pH 8.0, 375 mM KCl, 15 mM $MgCl_2$), 0.5 μL of 2 μM RT primer, and 6 μL of the RNA sample. The reaction was denatured at 80 ^{o}C for 2 minutes and left at room temperature for 5 minutes. 0.5 μL TGIRT enzyme (InGex), 0.5 μL SUPERase inhibitor, 0.5 μL 100 mM DTT, and 1 μL 10 mM dNTPs were added to the reaction. The reaction was incubated at 57 ^{o}C for 1.5 hours in a thermocycler for reverse transcription. RNA was then degraded by adding 5 μL of 0.4 M NaOH for 3 minutes at 90 ^{o}C, and the reactions were neutralized by adding 5 μL of an acid quench mix (from a stock solution of 2 mL of 5 M NaCl, 2 mL of 2 M HCl, and 3 mL of 3 M NaOAc). The reaction was purified with an Oligo Clean and Concentrator column, eluting in 7.5 μL of Rnase-free H_2O. cDNA was then purified with a denaturing PAGE (dPAGE) gel to remove excess RT primer. dPAGE gels were cast with 10% 29:1 bis-acrylamide, 7 M urea, and 1X TBE using 150 μL of 10% APS and 15 μL of TEMED. After pre-running the gel at 17 W for 1 hour, we loaded the gel with samples that had been denatured at 90 ^{o}C for 3 minutes in 50% formamide and a 1X TBE loading buffer with xylene cyanol and bromophenol blue markers. Gels were run for 40 minutes at 17W and stained for 20 minutes with 5 μL of 10,000X Sybr Gold (Invitrogen) in 50 mL of 1X TBE. We cut cDNA in the 200 to 400 nucleotide size range and eluted cDNA using the ZR small-RNA PAGE Recovery Kit (Zymo Research), following the manufacturer's protocol except for the changes noted as follows. We centrifuged the macerated gel slice for 15 minutes at max speed for the first column step, centrifuged at 5000 g for 5 minutes for the second column, and eluted the final sample in 7.5 μL of Rnase-free H_2O warmed to 65 ^{o}C.

The purified and size-selected cDNA was circularized and then amplified for sequencing. For circularization, 0.5 μL of CircLigase ssDNA Ligase (Lucigen), 2 μL of 10X CircLigase buffer (Lucigen), 1 μL of 1 mM ATP, and 1 μL of 50 mM $MnCl_2$ were added to the purified cDNA, and the reaction volume was adjusted to 20 μL. The reaction was left to incubate overnight at 60 ^{o}C in a thermocycler with a heated lid, and then the reaction was stopped at 80 ^{o}C for 10 minutes. The circularized cDNA was purified with an Oligo Clean and Concentrator column and eluted in 7.5 μL of Rnase-free H_2O. Samples from replicate 2 that were processed in parallel from RNA fragmentation to circularization were combined in this column purification step. Residual RT primer was removed by degrading linear DNA. First, we added 1 μL of RecJf and 1 μL of NEBuffer2 to the sample and adjusted with H_2O to make a 20 μL reaction, incubating at 37 ^{o}C for 10 minutes. We then added 2 μL each of ExoCIP A and ExoCIP B (NEB) and incubated for another 10 minutes at 37 ^{o}C, followed by 1 minute of heat inactivation at 80 ^{o}C. The sample was purified with an Oligo Clean and Concentrator column, with elution volume adjusted to 10 μL.

We carried out 9 cycles of indexing PCR to add i5 and i7 index sequences to the sequencing library, making a 50 μL reaction with 25 μL of the 2X NEBNext Ultra II Q5 Master Mix (NEB), 10 μL of cDNA, and 2.5 μL of 10 μM primers including different index sequences for each sample (i5_4 and i7_4 for replicate 1, i5_3 and i7_3 for replicate 2, i5_5 and i7_5 for the no-modification control, and i5_2 and i7_2 for the -pladB control; Table S5). The reactions were denatured at 98 ^{o}C for 30 seconds, followed by 9 cycles of denaturing at 98 ^{o}C for 10 seconds, annealing at 70 ^{o}C for 30 seconds, and extending 72 ^{o}C for 30 seconds, and ending with a final extension period at 72 ^{o}C for 5 minutes. The reaction mix was purified with the DNA Clean and Concentrator-5 kit (Zymo Research), with elution in 10 μL of H_2O. To determine the number of amplification cycles required for obtaining sufficient library concentration for sequencing, we quantified the library size through qPCR using the P5 and P7 primer sequences (Table S5), using the iTaq Universal SYBR Green Supermix (Bio-Rad). Based on this qPCR quantification, we carried out 2 more cycles of PCR for *in vivo* replicates 1 and 2 and the no-modification control, 5 more cycles for the -pladB control, 9 more cycles for *in vitro* replicate 1, and 3 more cycles for *in vitro* replicate 2. For this final PCR, we again used 25 μL of the Q5 Master Mix and 2.5 μL of 10 μM primers (P5 and P7 primers), using the same cycling parameters as the first indexing PCR. We purified the final library with two rounds of bead purification with size selection to remove remaining excess RT primer. We used RNACleanXP beads (Beckman Coulter), mixing 42.5 μL beads with the 50 μL PCR reaction (0.85 reaction volume ratio for size selection), shaking for 10 minutes and then separating beads from supernatant using a magnetic post for 10 minutes. We then washed the beads twice with 200 μL 70% EtOH, dried the beads, and eluted in H_2O after shaking the beads with H_2O for 10 minutes. For the first round of bead-based size selection we eluted in 50 μL of H_2O, and for the second round, we eluted in 20 μL of H_2O. The sequencing libraries were quantified with a Qubit high sensitivity dsDNA Assay Kit (Invitrogen) and Bioanalyzer HS DNA assay, and sequenced with a 10% PhiX library spike-in. The dsDNA libraries for *in vivo* replicate 1, the no-modification control, and the -pladB control were sequenced across three Illumina HiSeq lanes with paired-end reads of length 150. The dsDNA library for *in vivo* replicate 2 and the *in vitro* replicates were sequenced with NovaSeq S4 partial lanes, also with paired-end reads of length 150.

DMS-MaPseq sequencing data analysis. We obtained 557 million reads for *in vivo* replicate 1, 1.30 billion reads for *in vivo* replicate 2, 375 million reads for the no-modification control, 169 million reads for the control without splicing inhibition, 335 million reads for *in vitro* replicate 1,

and 174 million reads for *in vitro* replicate 2. We used UMI-tools[69] to extract UMI tags from reads by matching to the expected pattern from our RT primers, and we used cutadapt to trim low quality reads (Q-score cutoff 20) and remove adapter sequences including indexing primer sequences and the cloning linker. We then aligned sequencing reads to sequence sets of interest, including rRNA sequences, introns, complete pre-mRNA ORFs for genes containing introns, coding mRNA sequences for genes containing introns, decoy intron sequences (see below), and sequences for structured controls (the ASH1 and HAC1 mRNA sequences). *S. cerevisiae* intron annotations including genomic coordinates and branch point positions were obtained from Talkish, et al. 2019.[70] Coding ORF annotations were obtained from the Saccharomyces Genome Database for the S288C reference genome, and coding sequences corresponding to introns were identified for all cases except the two introns in snoRNAs (SNR17A and SNR17B). Paired-end alignment was performed with Bowtie2[71] using the alignment parameters used in ShapeMapper 2[72]: --local –sensitive-local –maxins=800 –ignore-quals –no-unal –mp 3,1 –rdg 5,1 –rfg 5,1 –dpad 30. To obtain mutational frequencies and normalized reactivities for both replicates combined, we additionally used alignment files merged with samtools[73]. Alignments were sorted and indexed using samtools. Reads with matching mapped positions and UMI tags were then deduplicated using the UMI-tools[69] dedup function with default parameters for paired-end reads.

Mutational frequencies, coverage values, and normalized reactivities were obtained by processing alignment files using RNAframework[74] executables. For each library and reference sequence set, we first ran rf-count with the flag -m to compute mutation counts, using all other default parameters. We obtained coverage statistics for each sequence with rf-rctools stats, and we obtained per-position mutation counts and coverage statistics with rf-rctools view. We obtained reactivity values with rf-norm, using Rouskin scoring with 90% Winsorizing for normalization, reactive bases A and C, and dynamic resizing of the normalization window to account for A/C frequencies (flags: -sm 2 -nm 2 -ow -rb AC -dw).

DMS-MaPseq data quality assessment. For each construct, per-base coverage was computed as the number of paired-end reads mapped to the construct obtained from rf-rctools stats multiplied by the total read length and divided by the length of the construct. To compare the retained intron fraction across all introns before and after pladB treatment, these coverage values were obtained for all introns and all coding regions for genes containing introns in the -pladB and +pladB conditions. The retained intron fraction was the ratio of these values. We additionally found the Pearson correlation between intron reactivities between replicate 1 and replicate 2 as it related to the average coverage between these replicates. A linear fit for log coverage versus replicate correlation for each intron indicated that the replicate correlation was best approximated as 0.2 log(coverage) – 1.03 for the *in vivo* DMS-MaPseq replicates. Based on this relationship, to reach 0.5 Pearson correlation between replicates, we used the coverage cutoff of 1971 averaged between replicate 1 and replicate 2, or 3942 when combining data from the two replicates. We chose a coverage cutoff of 3160 for the *in vitro* dataset, as this provided an average of 0.5 Pearson correlation between replicates when considering introns above this coverage cutoff. We used annotations for the ribosomal protein-coding genes from Hooks, et al. 2014.[45] We excluded eight snoRNA-containing introns from further analysis,[37] as the majority of these constructs' reactivities represented the excised snoRNA structure rather than the complete intron structure.

Per-base mutational frequencies were computed starting with the per-position mutation counts and coverage values from rf-rctools view. For each nucleotide identity, we found the average mutational frequency for each intron by averaging mutation rates across all positions in the intron with this nucleotide identity. We then found the average mutation frequencies for each nucleotide across all introns meeting the coverage threshold of 3942.

To obtain the correlation between Zubradt, et al. (2017)[27] rRNA mutational frequencies and our data, we obtained this paper's DMS-MaPseq reads for *S. cerevisiae* with TGIRT reverse transcriptase (replicate 1 accession number SRX1959209, run number SRR3929621). We aligned these reads to the 18S and 25S yeast rRNA sequences with Tophat v2.1.0[75] using alignment parameters stated in Zubradt, et al. (2017)[27]: -N 5 –read-gap-length 7 –read-edit-dist 7 –max-insertion-length 5 –max-deletion-length 5 -g 3. We obtained mutational frequencies with rf-count as described above. We compared mutational frequency values for all A/C positions where per-position coverage values met our coverage threshold, finding the Pearson correlation for these values between our data and those in Zubradt, et al. (2017)[27].

We evaluated whether reactivity values could distinguish surface accessible, unpaired residues from base-paired residues across the 18S and 25S rRNA. To identify surface accessible residues in rRNA, we followed a similar protocol to Rouskin, et al. (2014)[68]. We used the high-resolution *S. cerevisiae* ribosome structure with PDB ID 4V88,[76] finding solvent accessibility for rRNA residues in the complete context of the ribosome. We computed solvent-accessible surface area (SASA) values for the rRNAs' N1 atoms on A residues and N3 atoms on C residues in PyMOL, approximating DMS as a sphere with solvent_radius 3 and with dot_solvent and dot_density parameters set to 1. We additionally used this ribosome structure with DSSR[77] to find the secondary structure of the 18S and 25S rRNA sequences. We found the ROC curve for distinguishing unpaired and solvent-accessible rRNA residues from Watson-Crick base-paired positions, using a SASA cutoff of 2 to determine solvent accessibility.

We assessed whether reactivity profiles for known stems in the HAC1 and ASH1 mRNA sequences accurately reflected their secondary structures. Secondary structures for these stems were obtained from Zubradt, et al. (2017)[27].

DMS-guided structure prediction and validation. For introns meeting coverage cutoffs as discussed above, we performed secondary structure prediction by RNAstructure[30] guided by DMS reactivity with 1000 bootstrapping iterations without pseudoknot predictions, using the package Biers[41] with default parameters for RNAstructure. Minimum free energy structures and base-pair confidence matrices from bootstrapping were obtained for each construct. Structures were visualized with VARNA, using Biers to display DMS reactivity values and per-helix helix confidence scores from bootstrapping.

To assess DMS-guided structure prediction and the helix confidence scores, we performed structure prediction for a set of controls with known secondary structures. These included rRNAs, snRNAs, tRNAs, and mRNA segments, all of which were included in our transcriptome-wide DMS-MaPseq dataset at high coverage. The ground truth secondary structures for the 5S, 5.8S, and 18S rRNA was obtained from a high-resolution X-ray crystallography structure of the eukaryotic ribosome (PDB ID: 4V88)[76], and DSSR[77] was used to obtain the secondary structure from these 3D models. The ground truth U5 snRNA secondary structure was obtained from

Nguyen, et. Al. (2016)[78], and the U1 snRNA secondary structure was obtained from Li, et. Al. (2017)[79]. Four tRNA ground truth structures were obtained from Rfam-derived secondary structures.[80] Finally, secondary structures for mRNA segments in HAC1, ASH1, RPS28B, and SFT2 were obtained from Zubradt, et al. (2017)[27]. We did not include RNAs with known pseudoknots in the control set (for instance, the Rnase P RNA[81]), as our structure prediction approach would not be able to recover pseudoknots. Additionally, we excluded control cases where the RNA was expected to take on multiple distinct structural conformations, such as the U2 snRNA.[82] For DMS-guided structure prediction, 1000 bootstrapping iterations were performed for all cases except the 18S rRNA, for which we performed 100 bootstrapping iterations. Predicted stems were called using varying helix confidence estimate cutoffs. As a point of comparison, structure predictions were also made using ViennaRNA 2.0[44] to generate predictions unguided by DMS data. All stems in control structures and predictions were evaluated for true positive, false positive, and false negative stem predictions. Positive stem predictions were defined as cases in which a stem in the predicted structure included at least five base-pairs and the helix confidence estimate threshold was met. False positive stem predictions were those where less than 50% of the predicted stem's base-pairs were included in the ground truth structure. False negative stem predictions were stems of length at least five in the native structure that either had fewer than 50% of their base-pairs in the predicted structure, or did not meet the helix confidence estimate threshold in the predicted structure.

We additionally explored other methods for structure prediction using DMS data. First, using the Arnie package (https://github.com/DasLab/arnie), we ran ShapeKnots[83] (which allows for pseudoknot prediction) on all 161 introns with sufficient coverage, guiding predictions with DMS data in 100 bootstrapping iterations with default parameters. As a control, we ran structure predictions with ShapeKnots for the RNase P RNA with 100 bootstrapping iterations, comparing to the secondary structure obtained by using DSSR on the high-resolution cryo-EM structure of the yeast RNase P (PDB ID: 6AGB)[81]. We additionally made predictions using DREEM[84] for the following eight intron regions that included predicted structures and exhibited high coverage from DMS-MaPseq, with intervals numbered starting at the beginning of each intron: RPL28 75-300, RPL7A 120-340 (first intron), ECM33 35-270, RPL26B 300-400, RPS13 140-212, RPL25 148-255, RPS9B 147-266, and RPL30 25-165. We ran DREEM using treating sequencing reads as single-stranded, and used default parameters otherwise.

Two-dimensional chemical mapping (M2-seq). We carried out two-dimensional chemical probing to assess the formation of base pairs in secondary structures predicted from DMS-MaPseq. DNA sequences for the introns in QCR9, RPL36B, RPS11A, RPL37A, and RPS7B were obtained as gene fragments from Twist Biosciences (Table S5). Each gene fragment included the full intron sequence, a T7 promoter sequence, reference hairpins for normalizing structure probing data, and sequences complementary to universal RT and PCR primers used for library preparation. To generate a pool of DNA variants through error-prone PCR, we first assembled a reaction mix with the following for each intron: 10 µL of 2 ng/µL template, 10 µL of 100 mM Tris pH 8.3, 2.5 µL of 2 M KCl, 3.5 µL of 200 mM $MgCl_2$, 4 µL of 25 mM dTTP, 4 µL of 25 mM dCTP, 4 µL of 5 mM dATP, 4 µL of 5 mM dGTP, 2 µL of 25 mM $MnCl_2$, 1 µL of Taq polymerase (Thermo Fisher), 2 µL of 100 mM primers (RR1 and RR107; Table S5), and 51 µL of H_2O. We ran the reaction on a thermocycler for an initial 98 ^oC denaturation, and then 24 cycles of 94 ^oC denaturation for 30 seconds, 64 ^oC annealing for 60 seconds, and 72 ^oC elongation for 3 minutes, followed by a final elongation at 72 ^oC for 10 minutes. Construct sizes

were verified on a 2% agarose gel, and samples were purified as described above using RNACleanXP beads (Beckman Coulter) with a 1.8 Ampure bead to sample volume ratio. We then proceeded with *in vitro* transcription of these RNA fragments by assembling the following reaction mix: 5 μL of 1X T7 RNA polymerase (NEB), 2 μL of 1 M DTT, 6 μL of 25 mM NTPs, 5 μL of 40% PEG-8000, 5 μL of T7 transcription buffer (NEB), template DNA, and Rnase-free H_2O added to a final reaction volume of 50 μL. Reactions were incubated at 37 ^oC overnight. Samples were Dnase treated by adding 1 μL of TURBO Dnase (Thermo Fisher) to the reaction and incubating at 37 ^oC for 30 minutes. RNA was purified with RNACleanXP beads using a 70:30 ratio of beads to 40% PEG-8000 as the beads mixture, and using a 1.8 bead mixture to sample volume ratio.

We proceeded with structure probing, reverse transcription, and library preparation for these RNA pools. We prepared 3 μL of 12.5 pmol RNA for each intron for DMS treatment and a no-modification control. We denatured the RNA by unfolding at 95 ^oC for 2 minutes and left RNA on ice for 1 minute. RNA was folded by adding 5 μL of 5x folding buffer (1.5 M sodium cacodylate pH 7.0. 50 mM $MgCl_2$) and 14.5 μL of Rnase-free H_2O, and incubating for 30 minutes at 37 ^oC. We modified RNA by adding 2.5 μL of 15% DMS (DMS condition) or 100% EtOH (no-modification control), and modifying at 37 ^oC for 6 minutes. The reaction was quenched by adding 25 μL of 2-mercaptoethanol and purified by Ampure bead purification as described above, eluting in 7 μL of Rnase-free H_2O. We reverse-transcribed the modified RNA with mutational read-through using TGIRT. We assembled a TGIRT reverse transcription master-mix by combining 2.4 μL 5X TGIRT buffer (see above), 1.2 μL of 10 mM dNTPs, 0.6 μL of 100 mM DTT, 0.5 μL TGIRT enzyme (InGex), and 1.82 μL of H_2O. For reverse transcription, we used FAM-labeled primers that included a distinct index sequence for each construct and condition. We mixed 0.93 μL of RT primer (RTB primers in Table S5), 4.6 μL of RNA, and 6.52 μL of the reverse transcription master mix. We incubated the reaction mix at room temperature for 5 minutes and then at 57 ^oC for 3 hours. The reaction was stopped by adding 5 μL of 0.4 M NaOH and incubating at 90 ^oC for 3 minutes to degrade RNA, and the solution was neutralized by adding 2.2 μL of an acid quench mix (see above). cDNA was purified using RNACleanXP beads and eluted in 10 μL H_2O, using the protocol for RNA bead purification as described above. The cDNA library was then amplified, adding Illumina adapters for sequencing. We mixed 5 μL of 5X Phusion HF buffer (NEB), 0.5 μL of 10 mM dNTPs, 0.5 μL of Phusion high-fidelity DNA polymerase (NEB), 2.5 μL of cDNA, and 1 μL of 100 μM primers (MaP forward and reverse primers; Table S5) in a reaction volume of 25 μL. The PCR reaction used the following thermocycler conditions: denaturation at 98 ^oC for 30 seconds, 20 cycles of 98 ^oC for 10 seconds, 65 ^oC for 30 seconds, and 72 ^oC for 30 seconds, and a final elongation at 72 ^oC for 10 minutes. Amplicon sizes were verified on a 2% agarose gel with ethidium bromide staining. We collected library pools for the sequencing run, combining 0.57 fmol for each DMS treatment case and 0.18 fmol for each no-modification control, and adding a 25% PhiX spike-in. The resulting libraries were sequenced in two partial sequencing runs using Illumina MiSeq v3 600-cycle reagent kits, providing paired-end reads of length 300.

M2-seq sequencing reads were processed using the M2seq package.[41] First, barcodes for the different introns and conditions were demultiplexed. We obtained at least 500,000 reads for each intron's DMS treatment condition. The ShapeMapper 1.2[72] software was used to align reads to each reference sequence using Bowtie2 and compute mutation rates. The output from ShapeMapper was processed by the simple_to_rdat.py script to obtain 2D reactivity values and

raw counts in an RDAT file. These RDAT files were obtained for both DMS and no-modification conditions, and processed with the Biers package[41] to generate Z-score plots, base-pairing probability matrices from bootstrapping, and predicted secondary structures. These secondary structure predictions were made using RNAstructure's Fold executable[30] with 500 bootstrapping iterations, guided predictions with both 1D and 2D DMS reactivity data. Structures were visualized using VARNA, using Biers to display reactivity profiles and helix confidence estimates from bootstrapping for all stems.

Assessing proposed functional structures with DMS data. Using our DMS data, we assessed proposed structures from prior functional experiments and analysis of sequence alignments. For each evaluated structure, we displayed the proposed structure with reactivity values and helix confidence estimates using Biers.[85] Structure from Hooks, et al. 2016[33] were obtained by request. For structures with covarying residues, DMS-guided structure predictions were displayed when they agreed with the proposed covarying base-pairs (RPS9A, RPS9B, and RPL7A), and structures from CaCoFold[86] were displayed otherwise (RPS13).

To expand the set of covarying residues identified across introns, we used the protocol and software developed by Gao, et al. 2020[23] with thresholds as follows. We obtained all complete and chromosome-level assemblies of genomes in the Ascomycota phylum from NCBI GenBank[87] (accession date: August 17, 2021), yielding 107 genome sequences. For each *S. cerevisiae* intron, we generated multiple-sequence alignments using the scripts from Gao, et al. 2020[23] in flanked mode with a noncoding threshold of 50. We then used R-scape[38] to predict covarying residues with a 0.05 E-value threshold, and we generated structures with CaCoFold.[86] We identified all introns for which at least 1 significantly covarying pair was identified between two non-consecutive residues in a stem that contained at least 3 base-pairs in the CaCoFold structure. We evaluated the presence of these covarying residues in minimum free-energy secondary structures predicted from ViennaRNA 2.0[44]. The set of snoRNA-containing introns was obtained from the *S. cerevisiae* snoRNA database[37], and snoRNA gene coordinates were obtained from the Saccharomyces Genome Database.[88]

Zipper stem identification and stability calculation. We identified the positional constraints for forming a stem between the 5' splice site and branch point sequence (termed "zipper stems") by modeling 3D structures for introns in the context of the spliceosome using the Rosetta protocol FARFAR2.[39] As test constructs, we used variants of the intron in RPS17B since a zipper stem in this sequence has been characterized.[16] To identify the minimum linker length for which formation of a zipper stem is sterically compatible with an intron binding to the spliceosome, we built models for variants of the RPS17B intron with a range of linker lengths between the 5' splice site, zipper stem, and branch point using the cryo-EM structure of the *S. cerevisiae* A-complex spliceosome (PDB ID: 6G90).[89] The tested variants included 36, 40, 45, 50, and 56 total nucleotides of linker sequence between the 5' splice site, zipper stem strands, and the branch point sequence.

For each modeled RPS17B intron variant, we replaced the nucleotides proximal to the 5' splice site and branch point that were resolved in the A-complex spliceosome structure with the sequence from RPS17B, inserting the corresponding sequences from the RPS17B intron using Rosetta's rna_thread application. We docked the RPS17B zipper stem into the binding site for an intron stem in the U1 snRNP, as determined from the E state spliceosome structure.[25] For each

variant 1250 structures were sampled using Rosetta's rna_denovo application. The A-complex spliceosome structure, docked zipper stem, and rethreaded RPS17B intron residues were treated as rigid bodies through these modeling runs, and remaining linker residues and other positions in RPS17B were sampled freely. To accelerate structure sampling in the large context of the spliceosome, we generated models using a simplified score function that only included three terms: rna_vdw at weight 10.0 (the RNA van der Waals score term), rnp_vdw at weight 50.0 (the RNA-protein van der Waals score term), and linear_chainbreak at weight 10.0 (a score term penalizing chain breaks). We collected the linear chain break score for each sampled structure, a score term which penalizes breaks in covalently attached neighboring residues that can occur when linker lengths are too short to stretch between docked nucleotides.

Based on the linear chain break scores for the tested series of linker lengths, to designate an intron stem as a zipper stem we required at least 42 total nucleotides between the 5' splice site and the zipper stem's 5' end, and between the zipper stem's 3' end and the branch point. Additionally, we required that these stems begin at least 10 nucleotides from the 5' splice site and end at least 20 nucleotides from the branch point, as positions more proximal to the splice sites were resolved in the A-complex spliceosome structure and would not be able to form base-pairs. Finally, we required that zipper stems are separated from the 5' splice site and branch point at most 85 total linker nucleotides, so that they serve to co-localize the 5' splice site and branch point. For a given secondary structure, we first identified the longest zipper stem in these windows allowing for at most 5 bulge nucleotides, and we computed the stability of this stem by finding the dG of folding in RNAcofold[90] as implemented in Vienna 2.0[44]. In cases where no stem was present that satisfied these positional constraints in the secondary structure, we gave the intron a best zipper stem dG of 0 kcal/mol.

DMS reactivity and structure analysis for introns. We made reactivity-guided secondary structure predictions to identify structural features of *S. cerevisiae* introns, making predictions for all introns and coding regions meeting the coverage threshold. Secondary structure predictions were made separately for introns and coding regions using *in vivo* DMS-MaPseq data and *in vitro* refolded DMS probing data. We restricted our coding region structure predictions to genes containing introns. We counted the number of introns that included zipper stems and stems between the branch point and 3' splice site, using the following criteria. Our zipper stem definition was based on modeling of zipper stems in the context of the spliceosome as described above. We additionally identified "downstream stems" between the branch point and 3' splice site with length at least 6 base-pairs and bootstrapping support at least 70%.

To calculate secondary structure properties for each intron and coding region with sufficient coverage, we represented the secondary structure for each sequence as a fine-grained and coarse-grained graph. Fine-grained graphs included a node for every base-pair and single-stranded position, with edges between consecutive positions or base-pairs. Coarse-grained graphs included a single node for each stem, junction, loop, single-stranded 5' end, and single-stranded 3' end. Fine-grained graphs were used to compute the maximum extrusion from ends metric, using the 23rnie2323 package[91] to find the length of the shortest path from each position in the sequence to either end of the sequence. The maximum value for this shortest path across all nodes in the network yielded the maximum extrusion from ends for the sequence, corresponding to the secondary structure distance between the sequence ends and the position furthest from these ends. The maximum extrusion from ends values were normalized by the length of each

construct and compared between intron and coding regions with the Mann-Whitney U rank test. Coarse-grained graphs were used to identify the longest stem with at most 10 total loop nucleotides and at least 90% bootstrapping probability. Additionally, coarse-grained graphs were used to compute average helix confidence estimates for all stems of length at least 6 in introns and coding regions meeting coverage thresholds. The average helix confidence estimates and longest stems were compared between intron and coding sequences with nonparametric Mann-Whitney U rank tests.

Gini coefficients for comparing intron and coding reactivity values were calculated based on mutational frequency values for intron and coding regions. We obtained mutational frequencies across intron and coding sequences in all windows of 20 nucleotides, scanning in intervals of 10 nucleotides. For *in vivo* samples, we required that every position in each included window had per-position coverage at least 6456 as computed from rf-rctools view, corresponding to a 0.6 Pearson correlation between replicates for normalized reactivities. To compare *in vivo* to *in vitro* reactivity windows, we required that all positions in each window had at least 3160 coverage (the coverage cutoff used for the *in vitro* reactivity profiles). Gini coefficients across all passing intron and coding regions were compared with a nonparametric Mann-Whitney U rank test.

Modeling of intron stems in spliceosome structure. We modeled the RPL28 and RPL36B introns in the context of the A-complex spliceosome (PDB ID: 6G90)[89], using a protocol similar to the RPS17B zipper stem modeling as described above. Intron nucleotides resolved in 6G90 were replaced by the corresponding sequences from the intron in RPL28 and in RPL36B with the rna_thread application.[92] The secondary structure used for modeling was obtained from M2-seq guided RNAstructure prediction, and helices were supplied as rigid bodies for modeling. 4564 models were sampled using the Rosetta application rna_denovo.[39] To accelerate structure sampling, we ran modeling with only rna_vdw, rnp_vdw, and linear_chainbreak score terms as described previously. The top-scoring models were visualized in PyMOL using the RiboVis package (https://ribokit.gihub.io/RiboVis/).

Comparison of introns with decoys. To determine whether secondary structure features distinguish authentic introns from unspliced genomic sequences that match splicing motifs, we assembled a set of decoy introns from the genome. To find these decoy introns, we first computed the position weight matrix (PWM) for the 5' splice site (6 nucleotides with consensus GUAUGU), branch point (8 nucleotides with consensus UACUAACN), and 3' splice site (3 nucleotides with consensus YAG) for all canonical annotated introns in *S. cerevisae*, finding the log of the fraction of sequences that included each of the 4 nucleotides for every position in the sequence motif. We then obtained three length distributions for annotated *S. cerevisiae* introns: the complete intron length, the distance between the 5' splice site and branch point, and the distance between the branch point and 3' splice site. Gene annotations were obtained from the sacCer3 UCSC genome assembly. To find candidate decoy intron windows, we scanned across all genes for intervals containing 5' splice site, branch point, and 3' splice site sequences matching the PWMs and three length distributions for introns, using PWM score and length cutoffs that captured at least 95% of canonical introns. We filtered these sequences to exclude annotated introns. The resulting sequences represented "decoy introns", transcribed gene intervals that matched splice site motifs and intron length distributions and yet were not spliced. For each authentic intron and decoy sequence, we assembled length-matched control sequences that were shifted in the genome by 500 nucleotides outside the intron sequence.

DMS-guided structure predictions were made for each intron, decoy, and matched control sequence. Sequences were only considered if both the construct and its shifted genomic control had sufficient coverage from DMS-MaPseq for structure prediction. For each secondary structure, we computed zipper stem dG values, downstream stem dG values, maximum extrusion from ends, and longest stem lengths as described above. Additionally, we computed the distance between the 5' splice site and branch point in the intron's secondary structure by computing the length of the shortest path in secondary structure graph between the 5' splice site and branch point. We then determined whether authentic introns and decoy intron sequences enriched for these structural features more than shifted genomic controls. We compared the resulting values between these sets using the non-parametric Wilcoxon ranked sum test. We displayed the difference between intron and control feature values in violin plots after normalizing each metric between 0 and 1 and computing the feature value percent change.

Visualizing the structure landscape for *S. cerevisiae* introns. We assembled secondary structure features for all introns passing DMS-MaPseq coverage thresholds and used these features to classify *S. cerevisiae* introns. For each intron, we included the following metrics which were calculated as described previously: the total sequence length, the presence of a zipper stem (binary 0 or 1), zipper stem free energy, the presence of a downstream stem (binary 0 or 1) between the branch point and 3' splice site, downstream stem free energy, maximum extrusion from ends, the longest stem length, the average stem helix confidence estimate, the maximum Gini coefficient window, and average DMS accessibility for the 5' splice site, branch point, and 3' splice site. Hierarchical clustering was performed using scipy's hierarchy module[93] with the Ward linkage criteria and optimal leaf ordering, generating 9 clusters, and seaborn's clustermap was used to generate a dendrogram and heatmap visualization. To create more interpretable classes, starting from the hierarchical clustering classes, we combined two clusters to generate Class 2 (zipper stems), reordered dendrogram classes to place Class 3 (downstream stems) after Class 2, and moved two short introns from Class 8 (unstructured long introns) to Class 7 (unstructured short introns). Additionally, tSNE classification was performed with all structural features using sklearn, with 2 embedded space dimensions, a perplexity of 40, and 300 optimization iterations.

Designing intron structure variants for structure-function assay. We chose 7 structured introns to mutate systematically in our variant library by identifying secondary structures with zipper stems, covarying residues, and long stems. Intron library sequences were constrained to 300 total nucleotides including fixed primer binding sites and extensions into the 5'UTR for inserting barcodes, leaving around 200-250 nucleotides from the 5' splice site in introns as the variable region for each construct. For each intron, we designated stem and loop sets to mutate within these length constraints, with each stem set containing one or more continuous stems in the secondary structure, and each loop set containing a single junction or loop in the secondary structure.

For each stem set and loop set, we generated candidate variants and rescue sequences where available, and we scored these candidates to find a final set of designs based on desired secondary structure properties. For each stem set, to generate candidate library sequences, we generated 10000 randomized nucleotide sequences and 10000 shuffled sequences mutating the 5' strand of the stems. Additionally, if the 3' end of the stem set fell within the length constraints of the library, we generated rescue sequences for each of these mutation candidates, installing

compensatory mutations in the 3' strand. For each loop set, we generated 20000 random variants as discussed above for all nucleotides in the loop accessible within the library. For each variant and rescue sequence, we computed the minimum free energy structure with Contrafold[94] after placing the intron variable region in the full intron context.

We then assessed whether the variant disrupted any targeted stem structures while leaving the remainder of the secondary structure unchanged. More specifically, for each variant, we accumulated any junction or stem residues in the stem or loop set, and we computed two penalty scores: a variable region penalty and a constant region penalty. The variable region penalty was the sum of the paired residues in the variable region along with the number of base-pairs present in this region in the variant. The constant region penalty focused on nucleotides outside the variable region, computing the number of base-pairs in this region of the native secondary structure disrupted by the variant, and adding half of the number of base-pairs gained in the variant. If the variant also had a rescue sequence available, we computed the rescue penalty as the number of base-pairs lost or gained between the native secondary structure and the rescue sequence. The final total score for the variant sequence was the sum of the variable region penalty, constant region penalty, and the rescue penalty. We ranked variants by this total score and chose the top 4-8 unique variants per region. We visually checked the base-pair probability matrices for wildtype, variant, and rescue sequences to ensure disruption and rescue of stems in the desired sequence interval. Additionally, we ensured that variants did not introduce BglII and XcmI cut sites, as these restriction enzymes would be used subsequently for library integration. Our variants did not mutate the 5' splice site, branch point, or 3' splice site sequences. For each variant set, we additionally included 4-8 copies of the wildtype sequence to ensure sufficient coverage of the native intron in our experiment. The final chosen variant sequences along with the variable region penalty, constant region penalty, and total penalty scores are in Table S3.

Designing gene context for intron variants. Introns were cloned into their native gene context to better mimic the effects of structures in native *S. cerevisisae* pre-mRNA on gene expression. For each construct, we included the promoter TDH3, the 5' UTR beginning at the transcription start site, the full gene including the intron sequence, and the ADH1 terminator (Fig. S16C). The TDH3 promoter was chosen for yielding high expression levels,[43] and the ADH1 terminator was chosen for yielding high mRNA half life.[95] For each gene, we used the YeasTSS database to find the most prominent transcription start site upstream of the gene, using this position as the start of the 5'UTR in our constructs.[96] When cloning intron libraries into this gene context, we included a short 8-nucleotide fixed sequence in the 5'UTR to distinguish the inserted gene from the endogenous gene copy (AGCGGACG for QCR9, RPL28, RPS9A and RPL36B; AGAAGACG for RPS14B; AGAAGAGC for RPL7A; and AACTGCCC for RPS9B). Additionally, we included a random 12-nucleotide (12N) barcode ahead of the intron libraries in the 5'UTR that we aimed to use to identify pre-mRNA variants even after introns are spliced out. We ran simulations by randomly sampling 12N barcodes to check that for our expected number of E. coli and yeast library clones (10,000-50,000), we would not have substantial overlap between barcodes for separate clones, with most barcodes expected to have at least edit distance two from all other barcodes (Fig. S16D). We therefore expected that these random barcodes would allow us to unambiguously assign spliced reads to intron variants.

Growth media for structure variant library assay. We used the following growth media and plates for our structure-function assay:

- YPD media: YPD powder (Difco; Yeast extract 10g/L, Peptone 20g/L, Dextrose 20g/L)
- YPD plates: Difco YPD powder 50g/L, Difco agar powder 20g/L
- SD-LEU plates: Premix from Takara Bio
- SD-HIS plates: Premix from Takara Bio

Constructing yeast background strain for structure variant library integration. We aimed to integrate our variant library into the yeast genome using homologous combination and CRISPR/Cas9-induced double stranded breaks to increase efficiency. We chose the ARS416 locus as our integration site, as it was previously identified as a highly efficient integration site when using Cas9-facilitated genomic integration.[97] To ensure integration of our library in the expected locus, we first constructed the background strain PLH001 that inserted a partial LEU2 selection marker at this locus (starting from the second codon of the coding sequence). We then designed our genomic inserts for our structure variant library to include the LEU2 promoter and start codon, such that genomic integration of the library at the correct locus was required to complete the LEU2 selection marker and allow for selection with SD-LEU.

The strain PLH001 was constructed from background strain BY4741 (Mata, his3Δ1, leu2Δ0, met15Δ0, ura3Δ0) to serve as a landing pad for genomic integration of the structure variant library. A partial LEU2 selection marker and a complete HIS3 selection marker were integrated into the ARS416 gene locus immediately downstream of the previously tested CRISPR/Cas9 cut site.[97] To clone the construct for yeast genomic integration, LEU2 and HIS3 were obtained through PCR from the pHLUM plasmid,[98] adding homology arms for genomic integration into ARS416. These constructs were cloned into the pUC19 backbone[99] using NEBuilder Hifi DNA Assembly (NEB) (Fig. S16C, insert sequence RRLH in Table S5). We obtained a linear insert for transformation with PCR and integrated into the ARS416 locus with heat shock transformation of BY4741 using SD-HIS selection.[100]

Constructing plasmid backbones for structure variant library. We first cloned a plasmid backbone for each intron variant containing the intron's gene context (pRR1-pRR7), and we then cloned in our intron libraries into this background with a high efficiency transformation protocol. For each intron, the background gene consisted of the TDH3 promoter, the 5'UTR, the gene and intron excluding the variant library, the ADH1 terminator, and homology arms for insertion into the ARS416 locus (homologous to ARS416 on the 5' end and part of the inserted LEU2 selection marker on the 3' end) (Fig. S16C). All PCR reactions in the following section followed the following protocol unless otherwise specified. We used 25 μL of the 2X NEBNext Ultra II Q5 Master Mix (NEB) with 2.5 μL of 10 μM primers using the following thermocycler conditions: denaturation at 98 ^{o}C for 30 seconds; 30 cycles of 98 ^{o}C for 10 seconds, 30 seconds at the primers' annealing temperature, and 72 ^{o}C for 30 seconds; and a final elongation at 72 ^{o}C for 10 minutes.

We used the following gene fragments from IDT to build gene background constructs: 1. Gene blocks with the homology arm and TDH3 promoter; 2. A gene block with the ADH1 terminator, LEU2 promoter, LEU2 start codon, and LEU2 homology arm; and 3. Gene blocks with the target gene and constant regions in the intron (gblockRR4-12; Table S5). Gene blocks were amplified with PCR using primers RR222-RR233, adding homology sequences used for construct assembly (Table S5). pUC19 was linearized with PCR and treated with DpnI (NEB) to remove

intact plasmid template. Gene blocks and the background plasmid were cloned using a 4-part NEBuilder Hifi DNA Assembly (NEB) reaction.

Cloning structure variant library with high efficiency transformation. Intron library sequences for the 7 intron sub-libraries were obtained as a single oPool from IDT, and the 7 sub-libraries were then amplified by PCR. For each intron, we carried out 8 parallel PCR reactions with limited PCR cycling to reduce bias. We used the following primer pairs: RR273, RR249 for RPS14B; RR274, RR251 for RPS9B; RR275, RR253 for RPS9A; RR276, RR255 for QCR9; RR277, RR257 for RPL36B; RR279, RR261 for RPL28; and RR280, RR263 for RPL7A (Table S5). Forward primers for these PCR reactions included 12 nucleotide random barcodes and 5' homology to the corresponding plasmid backbone (pRR1-pRR7), and reverse primers included 3' homology to the plasmid backbone. We used the Q5 PCR amplification protocol from the previous section for each of the 8 reactions, but in this case with 30 ng template, 1 minute extension time, and 14 amplification cycles. The 8 PCR reactions for library inserts were concentrated together with the QIAQuick PCR Purification Kit (Qiagen) and gel purified with a 1% agarose gel and MinElute PCR Purification Kit (Qiagen). Plasmid backbones pRR1-pRR7 were linearized via PCR primers RR223 and RR239-RR245 using the Q5 protocol from the previous section with 3.5 minutes extension time, and then purified from 0.8% agarose gels with the MinElute PCR Purification kit, and treated with DpnI (NEB). Intron library inserts were assembled into these linearized backbones used NEBuilder HiFi DNA Assembly (NEB) with 100 ng backbone and 3X molar excess of library inserts.

For library cloning, we transformed NEB Stable Competent *E. coli* cells (NEB) with NEBuilder assembly reactions for each of our 7 sub-libraries using the high efficiency transformation protocol from the manufacturer. We used 2.5 μL ligation product and 32 μL competent cells per reaction. For each intron, we plated 100 μL of cells on each of 3 LB plates with carbenicillin. We obtained 6000-46000 clones per sub-library (30-230X library coverage), and we found correct designed variants in 50-100% of the clones from Sanger sequencing of 8-16 clones per sub-library. Plasmid DNA was extracted by scraping cells and then using the ZymoPURE II Plasmid Midiprep Kit (Zymo Research).

Yeast genomic integration for structure variant library assay. The plasmid library was linearized by restriction digest and transformed into PLH001 for genomic integration. For the first digestion, we combined 3 μL of XcmI (NEB) with 10 μL 10X R2.1 buffer (NEB), 5 μg of each plasmid sub-library, and H_2O to 100 μL. Reactions were left at 37 ^{o}C for 60 minutes, 65 ^{o}C for 20 minutes, and then on ice for 5 minutes. Then for the second digestion, we added 2.0 μL of 2.5 M NaCl, 4.16 μL of Tris HCl pH 7.9, and 1.5 μL of the BglII enzyme, followed by an incubation at 37 ^{o}C for 60 minutes. To increase efficiency of integration, the linearized plasmid library was transformed with Cas9 to encourage genomic double-stranded breaks at the integration site. We obtained 50 μg of plasmid p426[97] with the ZymoPUREII Plasmid Maxiprep Kit (Zymo Research) and linearized with AhdI digest.

We used heat-shock transformation in PLH001 with the linearized Cas9 plasmid and library insert for genomic integration with the LiAc/ssDNA/PEG transformation protocol.[100] For each sub-library transformation, we used 5 μg linearized plasmid library and 5 μg linearized p426 plasmid to transform 25 mL of OD600 0.5 cells. We plated transformed cells on SD-LEU selection plates, with 4 plates per sub-library. As negative controls, we transformed DNA from

one sub-library into BY4741 (without the partial LEU2 landing pad in the genome), and we transformed p426 alone without a library insert into PLH001. We saw no colony growth for either of these negative control conditions on SD-LEU selection plates. We scraped colonies from all library transformation plates and stored in 15% filtered glycerol at -80 ^{o}C.

Dilution plates from the sub-library transformations were used for counting colonies and verifying genomic integration. We counted 9000-35000 colonies for each sub-library, yielding 45-175X library coverage for each intron. To verify genomic integration, we extracted genomic DNA from 4-8 colonies per sub-library with the YeaStar Genomic DNA Kit (Zymo Research) and sequenced PCR amplicons with Sanger sequencing, finding that 57-100% of clones were successfully integrated.

Genomic DNA extraction and sequencing for structure variant library assay. We sequenced genomic DNA to assign barcodes to each designed variant sequence and to obtain transformation frequencies. For genomic DNA extraction, we started overnight cultures from 100 μL of yeast sub-library glycerol stocks. Cultures were diluted and grown to OD600 0.5-0.6 the following day. We used the YeaStar Genomic DNA Kit (Zymo Research) for DNA extraction.

To sequence genomic DNA, we first carried out an indexing PCR with limited cycles to amplify target intron library regions and add i5/i7 indices for Illumina sequencing. We used the primers RR281-RR294 to amplify our sub-libraries and add Illumina adapters (Table S5). These primers amplified the target sub-libraries including barcodes and complete variant sequences, with forward primers binding to the fixed 8-nucleotide sequences that distinguished the inserted intron library from the endogenous gene copy. Additionally, these primers included unique i5 and i7 index sequences for demultiplexing sub-libraries. For each indexing PCR, we used a 50 μL PCR reactions using Q5 (NEB) with 50 ng genomic DNA template and 7 amplification cycles. We purified PCR reactions with the DNA-5 Clean & Concentrator-5 kit (Zymo Research). Next, we ran qPCR reactions with two replicates per sub-library to determine the number of cycles of additional PCR required for sequencing. We then amplified our indexing PCR sample with P5/P7 primers, again using the Q5 protocol described earlier, now with 11-14 cycles and annealing at 66 ^{o}C. Samples were purified using RNACleanXP beads (Beckman Coulter) with a 0.8 Ampure bead to sample volume ratio for size selection, aiming to select amplicons and remove any primers or primer dimers from the sample. The sequencing libraries were quantified with a Qubit high sensitivity dsDNA Assay Kit (Invitrogen) and the libraries were sequenced using an Illumina MiSeq v3 600-cycle kit, providing paired-end reads of length 300 each.

RNA extraction and targeted RNA sequencing for structure variant library assay. We aimed to sequence RNA to obtain levels of spliced and unspliced RNA. For RNA extraction, we again started overnight cultures from 100 μL of yeast sub-library glycerol stocks. For RNA extraction, when cultures reached OD600 0.5-0.6 the following day, we used 5 mL culture per sub-library in two columns from the YeaStar RNA Kit (Zymo Research), with 2.5 μL of Zymolase for every 2.5 mL starting cell culture.

For RNA sequencing, many library preparation steps mirrored our DMS-MaPseq library preparation described earlier. We pooled RNA extracted from each sub-library, combining 50 μg total RNA from each sub-library to make 350 μg total RNA. As described previously, we first concentrated extracted RNA with the RNA Clean and Concentrator-5 kit (Zymo Research), and

then we depleted rRNA from our total RNA with RNAseH, using 50-mer oligos to tile rRNA sequences. We then carried out size selection with RNA Clean and Concentrator-5 kit (Zymo Research) to remove the smaller rRNA depletion oligos and small RNAs from the total RNA sample. We DNAse treated each reaction with 10 μL TURBO Dnase and a 30 minute incubation at 37 ^{o}C, and we purified reactions with the RNA Clean and Concentrator-5 kit (Zymo Research). After rRNA depletion, size selection, and DNAse treatment, we were left with 1.56 μg RNA.

To assess the removal of contaminating genomic DNA from our total RNA, we carried out a reverse transcription reaction with and without reverse transcriptase and checked for amplification of control RNA regions with PCR. For reverse transcription, we used the iScript Reverse Transcription Supermix (Bio-Rad), with 50 ng of our DNAse-treated RNA in samples with and without the reverse transcriptase. We purified samples with Oligo Clean and Concentrator columns (Zymo Research), eluting in 15 μL H_2O. We then carried out PCR with the MATa1 and RPL36B primers as described in the section "RT-PCR for verifying splicing inhibition". We visualized products on a 2% E-Gel EX Agarose Gel (Thermo Fisher).

We next generated cDNA that would be used for targeted PCR amplification and sequencing from our RNA sample. We chose to generate cDNA after fragmenting RNA and ligating on a universal cloning adapter as a handle for reverse transcription. Compared to library preparation approaches that would reverse transcribe RNA with targeted primers downstream of the intron, we expected that our approach would avoid large length differences between unspliced and spliced amplicons, which could lead to biases in reverse transcription and PCR amplification. Our approach mirrors the library preparation strategy described previously for DMS-MaPseq. Briefly, first, we fragmented our RNA and purified with the RNA Clean and Concentrator-5 kit (Zymo Research). We then removed 3' phosphate groups from fragmentation with rSAP (NEB) treatment followed by rSAP inactivation. We prepared adenylated linker from RR118 (Table S5), and we then ligated this linker to our RNA sample with a T4 RNA ligase 2 truncated KQ (NEB) reaction, adding adenylated linker in 2X molar excess of the RNA sample. Excess linker was degraded with 5' Deadenylase (NEB) and RecJf (NEB) exonuclease. The sample was cleaned with the RNA Clean and Concentrator-5 kit to prepare for reverse transcription. For reverse transcription, we used TGIRT enzyme (InGex) with the protocol described previously, using the reverse transcription primer RR301 which includes a 10 nucleotide UMI and Illumina adapters (Table S5). We degraded RNA and then purified cDNA with an Oligo Clean and Concentrator column (Zymo Research).

Next, we amplified cDNA with targeted PCR primers for each sub-library to generate final sequencing samples. We first added i5/i7 indices for Illumina sequencing along with Illumina adapters in a short indexing PCR for each sample. For these PCR reactions, we used the following primer pairs: RR281, RR302; RR283, RR303; RR285, RR304; RR287, RR305; RR289, RR306; RR291, RR307; and RR293, RR308 (Table S4). Our forward PCR primers were complimentary to cDNA just 5' of our barcode sequences, and they included the 8-nucleotide fixed sequence that distinguished the inserted library of interest from the endogenous gene. These primers allowed for the amplification of cDNA including barcode sequences and the splice site junction. Additionally, these primers included unique i5 and i7 index sequences, allowing libraries to be demultiplexed after sequencing. For this indexing PCR, we used the Q5 reaction protocol described above, in this case with 7 amplification cycles. To obtain sufficient

material for sequencing, we then carried out 24-28 additional PCR cycles with P5/P7 primers (Table S5), using Q5 with the same cycling parameters as the indexing PCR reaction. Finally, to obtain dsDNA for sequencing, we carried out two rounds of size-selection and purification with RNACleanXP beads (Beckman Coulter), using a 0.7X Ampure bead to sample volume ratio and eluting in 15 μL H_2O. The sequencing libraries were quantified with a Qubit high sensitivity dsDNA Assay Kit (Invitrogen) and sequenced using a partial Illumina NovaSeq S4 lane, providing around 700 million paired-end reads of length 150.

Genomic DNA sequencing data analysis for structure variant library. Genomic DNA sequencing provided links from barcodes to intron variants, a critical step for quantifying spliced reads for intron variants since the barcodes remain in spliced mRNA while variant sequences are spliced out. Additionally, genomic DNA sequencing provides quantification for the number of transformants for each variant, providing a normalization factor for mRNA levels. To prepare genomic DNA sequencing reads for analysis, we first used UMI-tools[69] to extract barcodes by matching primer binding sites immediately upstream of the barcode. We then error corrected paired-end reads using bbmerge,[101] removing Illumina adapter sequences and ensuring consistency between paired-end reads.

We used the following pipeline to assemble consensus intron variant sequences corresponding to each barcode, making use of functions within the fgbio toolkit (fulcrumgenomics.github.io/fgbio/). First, paired-end reads were aligned to introns using bwa[102] with default settings. The resulting alignments were sorted using fgbio's SortBam directive and mate-pairs were annotated with fgbio's SetMateInformation. With reads aligned to the reference introns, we then grouped reads with the same barcode using fgbio's GroupReadsByUmi, choosing the adjacency method to combine barcode sets. Consensus reads were called using fgbio's CallMolecularConsensusReads with mostly default settings except the flags -m 20 -M 5, which required that only bases with Q score above 20 were included, and that at least 5 reads covered each consensus sequence. We then quality filtered these sequences using fgbio's FilterConsensusReads with the flags -M 5 -q 20 -N 20 -E 0.01, again using the same quality and read depth scores and additionally requiring that consensus bases have quality at least 20 and reads have average error rates lower than 0.01. With samtools,[73] we sorted the resulting alignment file and retrieved paired-end consensus reads. We then merged paired-end reads using bbmerge[101] with default settings, obtaining consensus length distributions depicted in Fig. S17A. Finally, to ensure that barcodes are mapped primarily to single consensus reads, we kept the primary consensus sequence for each barcode and required that at most 5% of reads map to secondary consensus sequences. This filter removed 0.4-4.7% of barcodes from further analysis across all sub-libraries.

From our consensus sequences, we obtained barcode mappings for designed variants along with transformation frequencies. We collated barcodes that mapped to designed variant sequences, finding at least 10 unique barcodes for many of the designed variants in each sub-library (Fig. S17B). This indicates that each designed variant was transformed into at least 10 *E. coli* and *S. cerevisiae* colonies, with some variants having much higher transformation frequencies. We additionally found that many colonies had consensus sequence variants that were mutations of wildtype sequences or designed variants. From these sequences, we retrieved all barcodes corresponding to QCR9 intron variants with branchpoint mutations, providing a control sample

for our RNA sequencing analysis. To obtain transformation frequencies, we found total read coverage for each barcode using the output from fgbio's FilterConsensusReads.

RNA sequencing data analysis for structure variant library. Our targeted RNA-sequencing data provided normalized mRNA levels and retained intron fractions for each intron variant. To process RNA-sequencing data, we first extracted barcodes (one for each transformant) and UMI's (one for each cDNA molecule) from our sequencing reads with UMI-tools.[69] We then removed Illumina adapters and the ligated linker for reverse transcription from the reads using cutadapt, additionally trimming bases with Q-score lower than 20. We then aligned reads to spliced isoform reference sequences to filter out reads from our RNA-sequencing run that did not correspond to the desired target sequence. For our reference sequence annotations, we collated all possible isoforms capturing spliced and unspliced transcripts (4 isoforms for two-intron RPL7A and 2 for all other genes), and we generated gff annotation files with gmap.[103] We used TopHat2[75] to align sequencing reads to these reference isoform annotations, using the following flags to enable potentially distant alignments between variants and the reference sequence: --no-novel-juncs -T –b2-mp 3,1 –b2-rdg 5,1 –b2-rfg 5,1 –segment-length 20 –segment-mismatches 3 –read-gap-length 7 –read-edit-dist 50 -m 1 –max-insertion-length 19 –max-deletion-length 19. We generated fasta files from the resulting alignment files with samtools.[73]

We classified each aligned sequence into one of three categories: unspliced, spliced at the expected junction, or other. We required an exact agreement with 14 nucleotides representing the unspliced junction (for unspliced sequences) or spliced junction (for spliced sequences). For each barcode, we computed the number of spliced and unspliced reads, deduplicating UMI's at this stage to reduce the effects of PCR amplification biases on our quantification. We then computed two key metrics for each barcode: the retained intron fraction and the normalized mRNA level. The retained intron fraction was computed as the ratio between unspliced read counts and total read counts (sum of spliced and unspliced reads). The normalized mRNA levels were computed as the ratio between the spliced read count and the transformation frequency obtained from genomic DNA sequencing.

For each intron sub-library, we used the barcode-variant correspondence from genomic DNA sequencing to collect the retained intron fractions and normalized mRNA levels for all the barcodes corresponding to each variant sequence. For all the stem and loop sets assessed in each intron sub-library, we compared RI fractions and normalized mRNA levels between the following sets of sequences: the wildtype sequence versus stem disruption variants, and the stem disruption variants versus the rescue sequences. Comparisons were made using permutation tests (scipy[93]) for the difference in mean statistic with 100,000 resamples of data labels per comparison.

***De novo* computational structure prediction and structure metrics.** We predicted secondary structure ensembles for *S. cerevisiae* introns and control sequence sets with *de novo* structure prediction. For direct comparison between DMS-guided RNAstructure prediction and *de novo* RNAstructure prediction, we computed features in both cases only for the 161 introns with sufficient coverage from DMS-MaPseq; for other cases, we made predictions for the complete set of introns filtered to exclude sequences smaller than 50 nucleotides or larger than 600 nucleotides, leaving 288 of 297 introns. Three control sequence sets were assembled, each with length-matched control sequences for every intron. To construct the first set, the "shifted

control", the genome coordinates for each intron were shifted 500 nucleotides downstream of the intron's 5' start site and a sequence of the same length as the intron was taken from these shifted coordinates. We constructed the second control set, the "sequence-matched control", by replacing the shifted control's sequences at the positions of splice sites in the matched intron, substituting the 5' splice site, branch point, and 3' splice site sequences from the intron at these positions. This set was constructed to control for structural differences caused primarily by sequence preferences at these splice sites. The final set, the "shuffled control", was obtained by shuffling each intron's sequence randomly.

Secondary structure ensemble predictions for each set of sequences were obtained by either making minimum free energy predictions or by sampling suboptimal structures using computational secondary structure prediction packages. For suboptimal structure sampling, 1000 structures were sampled for each sequence using the Arnie package (https://github.com/DasLab/arrnie) to call structure prediction executables from Vienna 2.0[44] and RNAstructure[30] with default parameters. We additionally made structure predictions for intron sequences with surrounding exon context, extending each intron to include 50 bases upstream and downstream of the intron. For each intron and matched control, we computed average values across these structure ensembles for the zipper stem free energy, downstream stem free energy, length of the longest stem, maximum extrusion from ends, and distance between the 5' splice site and branchpoint. As described previously, we compared the intron and control values using the non-parametric Wilcoxon ranked sum test. When we compared feature values using a t-test for related samples, we identified the same comparisons as significant in all cases.

Structure prediction for *Saccharomyces* genus. We obtained multiple sequence alignments for introns in 20 species in the *Saccharomyces* genus from Hooks, et al. 2014.[45] We computed statistics on the alignments for each intron, finding its number of orthologs and the average percent sequence conservation. To count the number of orthologs for each intron, we found all the species with non-empty sequences aligned to the *S. cerevisiae* intron. We additionally counted the number of zipper stem orthologs by finding all the non-empty regions aligned to the longest zipper stem in each *S. cerevisiae* intron. Finally, we computed the average percent sequence conservation across each complete intron and across the intron's zipper stem if one was present.

We next evaluated secondary structure feature enrichment for the introns across the *Saccharomyces* genus. We tested each intron in these 20 species against matched control sets, with each intron having a matching shuffled sequence control. We additionally compared the secondary structure ensembles of all the introns in each of these species' genomes to phylogenetic controls, assessing whether they preserved structural features more than expected compared to sequences altered at similar levels from the *S. cerevisiae* intron homolog. To construct phylogenetic control sequence sets, we measured mutation and indel rates between each intron and its homologous *S. cerevisiae* intron, and we created an intron's matched control sequence by inserting mutations and indels randomly into the homolog *S. cerevisiae* intron at this measured average frequency. As described earlier for the analysis of *S. cerevisiae* introns, for each of these comparisons, we predicted secondary structure ensembles for the intron set and control set, computed secondary structure features, and compared between sequence sets with a Wilcoxon ranked sum test.

Reagent and resource sharing. Yeast strains are available upon request from the authors. Requests for resources and reagents should be directed to the corresponding author, Rhiju Das (rhiju@stanford.edu).

Code availability

Code for assessing DMS-MaPseq reactivity values, evaluating control RNA structures, and analyzing DMS-derived secondary structures is included here: https://github.com/ramyarangan/DMS_intron_analysis. Code for computational secondary structure prediction, secondary structure feature calculation, statistical comparisons, and evolutionary analyses are included at https://github.com/ramyarangan/pre-mRNA_Secstruct. Code for designing VARS-seq sequences and analyzing gDNA and RNA sequencing data from VARS-seq is included in https://github.com/ramyarangan/VARSseq.

Data availability

All sequencing data are available at the Gene Expression Omnibus (GEO) accession number GSE209857. DMS-derived secondary structures are included in Table S2. Processed DMS reactivity data, control RNA secondary structures, and base-pairing probabilities from structure prediction with bootstrapping are included in https://github.com/ramyarangan/DMS_intron_analysis. VARS-seq intron variant sequences along with spliced and unspliced read counts are included in Tables S3-S4.

Acknowledgments

We thank K. B. Hooks and D. Delneri for providing secondary structure data upon request. We thank A. Hoskins, J. Puglisi, T. Lan, and S. Rouskin for illuminating discussions on this project, and we thank V. V. Topkar and I. Zheludev for advice on experimental protocols. We thank Stanford University for providing computational resources and support for the Sherlock 2.0 cluster that contributed to these results. This work was supported by the National Science Foundation Graduate Research Fellowship Program grant no. 1650114 (R.R.); the Gerald J. Lieberman Fellowship (R.R.); the National Institutes of Health grant R35GM145266 (M.A.), a Stanford Bio-X Interdisciplinary Initiative Award (R.D.); the National Institutes of Health grant R35GM122579 (R.D.); and the Howard Hughes Medical Institute (HHMI) Investigator Program (R.D.). This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

Author information

R.R. and R.D. conceived the project. R.R. designed and performed experiments and carried out computational analyses. O.H. and M.A. developed the *S. cerevisiae* strain OHY001 used in the

project. R.H. conducted *in vitro* DMS probing experiments. P.P. tested alternate structure prediction approaches. R.R., M.A., and R.D. interpreted the results. R.R. wrote the initial draft, and R.R., M.A., and R.D. wrote the final manuscript with input from all authors.

References

1. Wilkinson, M.E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu Rev Biochem* **89**, 359-388 (2020).
2. Wang, Z. & Burge, C.B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802-13 (2008).
3. Petibon, C., Parenteau, J., Catala, M. & Elela, S.A. Introns regulate the production of ribosomal proteins by modulating splicing of duplicated ribosomal protein genes. *Nucleic Acids Res* **44**, 3878-91 (2016).
4. Lukacisin, M., Espinosa-Cantu, A. & Bollenbach, T. Intron-mediated induction of phenotypic heterogeneity. *Nature* **605**, 113-118 (2022).
5. Morgan, J.T., Fink, G.R. & Bartel, D.P. Excised linear introns regulate growth in yeast. *Nature* **565**, 606-611 (2019).
6. Parenteau, J. et al. Introns are mediators of cell response to starvation. *Nature* **565**, 612-617 (2019).
7. Plocik, A.M. & Guthrie, C. Diverse forms of RPS9 splicing are part of an evolving autoregulatory circuit. *PLoS Genet* **8**, e1002620 (2012).
8. Vincenti, S., De Chiara, V., Bozzoni, I. & Presutti, C. The position of yeast snoRNA-coding regions within host introns is essential for their biosynthesis and for efficient splicing of the host pre-mRNA. *RNA* **13**, 138-50 (2007).
9. Jo, B.S. & Choi, S.S. Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform* **13**, 112-8 (2015).
10. Chorev, M. & Carmel, L. The function of introns. *Front Genet* **3**, 55 (2012).
11. Warf, M.B. & Berglund, J.A. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* **35**, 169-78 (2010).
12. Scharfen, L. & Neugebauer, K.M. Transcription regulation through nascent RNA folding. *J Mol Biol*, 166975 (2021).
13. Andjus, S., Morillon, A. & Wery, M. From Yeast to Mammals, the Nonsense-Mediated mRNA Decay as a Master Regulator of Long Non-Coding RNAs Functional Trajectory. *Noncoding RNA* **7**(2021).
14. Roundtree, I.A., Evans, M.E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187-1200 (2017).
15. Charpentier, B. & Rosbash, M. Intramolecular structure in yeast introns aids early steps of in vitro spliceosome assembly. *RNA* (1996).
16. Rogic, S. et al. Correlation between the secondary structure of pre-mRNA introns and the efficiency of splicing in Saccharomyces cerevisiae. *BMC Genomics* **9**, 355 (2008).
17. Parker, R. & Patterson, B. *Architecture of Fungal Introns: Implications for Spliceosome Assembly*, (1987).
18. Howe, K.J. & Ares, M., Jr. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc Natl Acad Sci U S A* **94**, 12467-72 (1997).

19. Meyer, M., Plass, M., Perez-Valle, J., Eyras, E. & Vilardell, J. Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol Cell* **43**, 1033-9 (2011).
20. Vilardell, J. & Warner, J.R. Regulation of splicing at an intermediate step in the formation of the spliceosome. *Genes Dev* **8**, 211-20 (1994).
21. Li, Z., Paulovich, A.G. & Woolford, J.L., Jr. Feedback inhibition of the yeast ribosomal protein gene CRY2 is mediated by the nucleotide sequence and secondary structure of CRY2 pre-mRNA. *Mol Cell Biol* **15**, 6454-64 (1995).
22. Fewell, S.W. & Woolford, J.L., Jr. Ribosomal protein S14 of Saccharomyces cerevisiae regulates its expression by binding to RPS14B pre-mRNA and to 18S rRNA. *Mol Cell Biol* **19**, 826-34 (1999).
23. Gao, W., Jones, T.A. & Rivas, E. Discovery of 17 conserved structural RNAs in fungi. *Nucleic Acids Res* **49**, 6128-6143 (2021).
24. Danin-Kreiselman, M., Lee, C.Y. & Chanfreau, G. RNAse III-Mediated Degradation of Unspliced Pre-mRNAs and Lariat Introns. *Molecular Cell* **11**, 1279-1289 (2003).
25. Li, X. et al. A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375-380 (2019).
26. Libri, D., Stutz, F., McCarthy, T. & Rosbash, M. RNA structural patterns and splicing: molecular basis for an RNA-based enhancer. *RNA* **1**, 425-36 (1995).
27. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods* **14**, 75-82 (2017).
28. Wallace, E.W.J. & Beggs, J.D. Extremely fast and incredibly close: cotranscriptional splicing in budding yeast. *RNA* **23**, 601-610 (2017).
29. Hansen, S.R., Nikolai, B.J., Spreacker, P.J., Carrocci, T.J. & Hoskins, A.A. Chemical Inhibition of Pre-mRNA Splicing in Living Saccharomyces cerevisiae. *Cell Chem Biol* **26**, 443-448 e3 (2019).
30. Reuter, J.S. & Mathews, D.H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
31. Cordero, P., Kladwang, W., VanLang, C.C. & Das, R. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* **51**, 7037-9 (2012).
32. Kladwang, W., VanLang, C.C., Cordero, P. & Das, R. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* **50**, 8049-56 (2011).
33. Hooks, K.B., Naseeb, S., Parker, S., Griffiths-Jones, S. & Delneri, D. Novel Intronic RNA Structures Contribute to Maintenance of Phenotype in Saccharomyces cerevisiae. *Genetics* **203**, 1469-81 (2016).
34. Yao, Z., Weinberg, Z. & Ruzzo, W.L. CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445-52 (2006).
35. Gruber, A.R., Neubock, R., Hofacker, I.L. & Washietl, S. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res* **35**, W335-8 (2007).
36. Pedersen, J.S. et al. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**, e33 (2006).
37. Piekna-Przybylska, D., Decatur, W.A. & Fournier, M.J. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA* **13**, 305-12 (2007).

38. Rivas, E., Clements, J. & Eddy, S.R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods* **14**, 45-48 (2017).
39. Watkins, A.M., Rangan, R. & Das, R. FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure* **28**, 963-976 e6 (2020).
40. Sun, L. et al. RNA structure maps across mammalian cellular compartments. *Nat Struct Mol Biol* **26**, 322-330 (2019).
41. Cheng, C.Y., Kladwang, W., Yesselman, J.D. & Das, R. RNA structure inference through chemical mapping after accidental or intentional mutations. *Proc Natl Acad Sci U S A* **114**, 9876-9881 (2017).
42. Neuveglise, C., Marck, C. & Gaillardin, C. The intronome of budding yeasts. *C R Biol* **334**, 662-70 (2011).
43. Lee, M.E., DeLoache, W.C., Cervantes, B. & Dueber, J.E. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synth Biol* **4**, 975-86 (2015).
44. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
45. Hooks, K.B., Delneri, D. & Griffiths-Jones, S. Intron Evolution in Saccharomycetaceae. *Genome Biology and Evolution* **6**, 2543-2556 (2014).
46. Bubenik, J.L. et al. RNA structure probing to characterize RNA-protein interactions on a low abundance pre-mRNA in living cells. *RNA* (2020).
47. Liu, Z. et al. In vivo nuclear RNA structurome reveals RNA-structure regulation of mRNA processing in plants. *Genome Biol* **22**, 11 (2021).
48. Ares, M., Jr., Grate, L. & Pauling, M.H. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5**, 1138-9 (1999).
49. Galy, V. et al. Nuclear Retention of Unspliced mRNAs in Yeast Is Mediated by Perinuclear Mlp1. *Cell* **116**, 63-73 (2004).
50. Bousquet-Antonelli, C., Presutti, C. & Tollervey, D. Identification of a Regulated Pathway for Nuclear Pre-mRNA Turnover. *Cell* **102**, 765-775 (2000).
51. Sayani, S., Janis, M., Lee, C.Y., Toesca, I. & Chanfreau, G.F. Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. *Mol Cell* **31**, 360-70 (2008).
52. Barak, M. et al. Purifying selection of long dsRNA is the first line of defense against false activation of innate immunity. *Genome Biol* **21**, 26 (2020).
53. Rivas, E., Clements, J. & Eddy, S.R. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics* **36**, 3072-3076 (2020).
54. Schirman, D., Yakhini, Z., Pilpel, Y. & Dahan, O. A broad analysis of splicing regulation in yeast using a large library of synthetic introns. *PLoS Genet* **17**, e1009805 (2021).
55. Rosenberg, A.B., Patwardhan, R.P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698-711 (2015).
56. Abelson, J. et al. Conformational dynamics of single pre-mRNA molecules during in vitro splicing. *Nat Struct Mol Biol* **17**, 504-12 (2010).
57. Zhang, Z. et al. Structural insights into how Prp5 proofreads the pre-mRNA branch site. *Nature* **596**, 296-300 (2021).
58. Goguel, V. & Rosbash, M. Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. *Cell* **72**, 893-901 (1993).
59. Tan, J. et al. A Disease-Causing Intronic Point Mutation C19G Alters Tau Exon 10 Splicing via RNA Secondary Structure Rearrangement. *Biochemistry* (2019).

60. Shilo, A., Tosto, F.A., Rausch, J.W., Le Grice, S.F.J. & Misteli, T. Interplay of primary sequence, position and secondary RNA structure determines alternative splicing of LMNA in a pre-mature aging syndrome. *Nucleic Acids Res* **47**, 5922-5935 (2019).
61. Saha, K. et al. Structural disruption of exonic stem-loops immediately upstream of the intron regulates mammalian splicing. *Nucleic Acids Res* **48**, 6294-6309 (2020).
62. Graveley, B.R. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **123**, 65-73 (2005).
63. Saldi, T., Riemondy, K., Erickson, B. & Bentley, D.L. Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing. *Mol Cell* (2021).
64. Fish, L. et al. A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements. *Science* **372**(2021).
65. Casas-Vila, N., Sayols, S., Perez-Martinez, L., Scheibe, M. & Butter, F. The RNA fold interactome of evolutionary conserved RNA structures in S. cerevisiae. *Nat Commun* **11**, 2789 (2020).
66. Gosai, S.J. et al. Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the Arabidopsis nucleus. *Mol Cell* **57**, 376-88 (2015).
67. Jeong, H., Herskowitz, I., DKroetz, D.L. & Rine, J. Function-Altering SNPs in the Human Multidrug Transporter Gene ABCB1 Identified Using a Saccharomyces-Based Assay. (2007).
68. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J.S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701-5 (2014).
69. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499 (2017).
70. Talkish, J. et al. Rapidly evolving protointrons in Saccharomyces genomes revealed by a hungry spliceosome. *PLoS Genet* **15**, e1008249 (2019).
71. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
72. Busan, S. & Weeks, K.M. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA* **24**, 143-148 (2018).
73. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
74. Incarnato, D., Morandi, E., Simon, L.M. & Oliviero, S. RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic Acids Res* **46**, e97 (2018).
75. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
76. Ben-Shem, A. et al. The structure of the eukaryotic ribosome at 3.0 A resolution. *Science* **334**, 1524-9 (2011).
77. Lu, X.J. & Olson, W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **31**, 5108-21 (2003).
78. Nguyen, T.H.D. et al. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 A resolution. *Nature* **530**, 298-302 (2016).

79. Li, X. et al. CryoEM structure of Saccharomyces cerevisiae U1 snRNP offers insight into alternative splicing. *Nat Commun* **8**, 1035 (2017).
80. Kalvari, I. et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**, D192-D200 (2021).
81. Lan, P. et al. Structural insight into precursor tRNA processing by yeast ribonuclease P. *Science* **362**(2018).
82. Rodgers, M.L. et al. Conformational dynamics of stem II of the U2 snRNA. *RNA* **22**, 225-36 (2016).
83. Hajdin, C.E. et al. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci U S A* **110**, 5498-503 (2013).
84. Tomezsko, P.J. et al. Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature* **582**, 438-442 (2020).
85. Tian, S., Cordero, P., Kladwang, W. & Das, R. High-throughput mutate-map-rescue evaluates SHAPE-directed RNA structure and uncovers excited states. *RNA* **20**, 1815-26 (2014).
86. Rivas, E. RNA structure prediction using positive and negative evolutionary information. *PLoS Comput Biol* **16**, e1008387 (2020).
87. Benson, D.A. et al. GenBank. *Nucleic Acids Res* **41**, D36-42 (2013).
88. Cherry, J.M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**, D700-5 (2012).
89. Plaschka, C., Lin, P.C., Charenton, C. & Nagai, K. Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature* **559**, 419-422 (2018).
90. Bernhart, S.H. et al. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol* **1**, 3 (2006).
91. Hagberg, A., Schult, D. & Swart, P. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science conference (SciPy 2008)*, 11-15 (2008).
92. Cheng, C.Y., Chou, F.C. & Das, R. Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol* **553**, 35-64 (2015).
93. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261-272 (2020).
94. Do, C.B., Woods, D.A. & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90-8 (2006).
95. Curran, K.A., Karim, A.S., Gupta, A. & Alper, H.S. Use of expression-enhancing terminators in Saccharomyces cerevisiae to increase mRNA half-life and improve gene expression control for metabolic engineering applications. *Metab Eng* **19**, 88-97 (2013).
96. McMillan, J., Lu, Z., Rodriguez, J.S., Ahn, T.H. & Lin, Z. YeasTSS: an integrative web database of yeast transcription start sites. *Database (Oxford)* **2019**(2019).
97. Reider Apel, A. et al. A Cas9-based toolkit to program gene expression in Saccharomyces cerevisiae. *Nucleic Acids Res* **45**, 496-508 (2017).
98. Mulleder, M. et al. A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nat Biotechnol* **30**, 1176-8 (2012).
99. Norrander, J., Kempe, T. & Messing, J. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* **26**, 101-6 (1983).
100. Gietz, R.D. & Woods, R.A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods Enzymol* **350**, 87-96 (2002).

101. Bushnell, B., Rood, J. & Singer, E. BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
102. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
103. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).

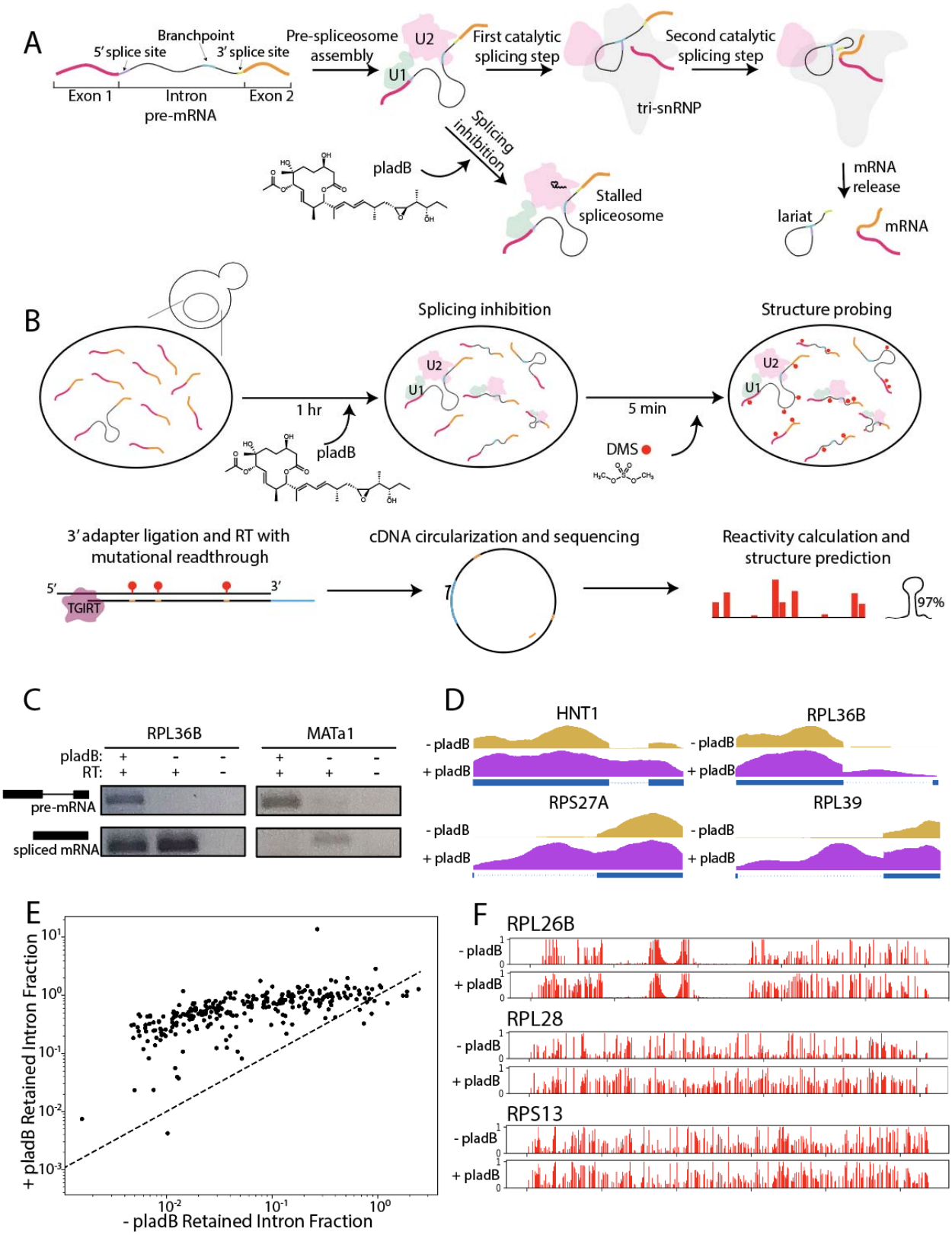
A
5′ splice site
Branchpoint
3′ splice site
Exon 1
Intron
Exon 2
pre-mRNA
Pre-spliceosome assembly
U2
U1
First catalytic splicing step
tri-snRNP
Second catalytic splicing step
Splicing inhibition
pladB
Stalled spliceosome
mRNA release
lariat
mRNA
B
Splicing inhibition
Structure probing
1 hr
5 min
pladB
DMS
3′ adapter ligation and RT with mutational readthrough
cDNA circularization and sequencing
Reactivity calculation and structure prediction
TGIRT
97%
C
RPL36B
MATa1
pladB:
RT:
pre-mRNA
spliced mRNA
D
HNT1
RPL36B
RPS27A
RPL39
- pladB
+ pladB
E
+ pladB Retained Intron Fraction
- pladB Retained Intron Fraction
F
RPL26B
RPL28
RPS13
- pladB
+ pladB

Figure 1: **Splicing inhibition by pladB allows for accumulation of pre-mRNA.** A) Schematic of RNA splicing. B) Schematic for splicing inhibition followed by DMS-MaPseq experiment. C) Accumulation of pre-mRNA shown for RPL36B and MATa1 by RT-PCR. D) Read coverage across intron-containing pre-mRNA with (purple) and without (orange) pladB treatment. E) Retained intron fraction with and without pladB treatment. Points are plotted on a log scale, and equal retained fraction from both conditions is indicated by a dotted line. F) Comparison of reactivity values for three introns with and without pladB treatment.

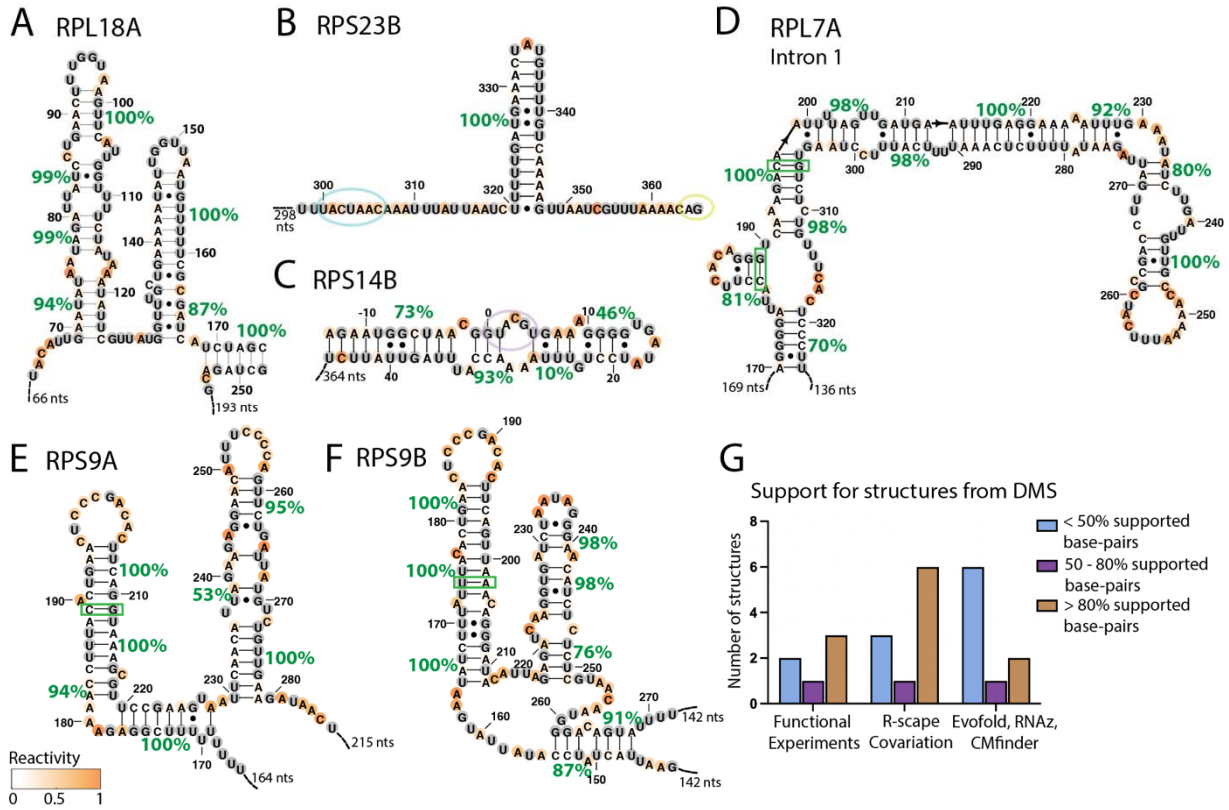

Figure 2: Support from DMS reactivity for *in vivo* formation of control structures and proposed functional structures. Helix confidence estimates and covariation for intron structures reported in A) RPL18A[20], B) RPS23B[15], C) RPS14B[17], D) the first intron in RPL7A, E) RPS9A, and F) RPS9B[7]. Secondary structures are colored by DMS reactivity and helix confidence estimates are depicted as green percentages. 5' splice site, branchpoint, and 3' splice site sequences are circled in purple, blue, and yellow respectively. Covarying base pairs in RPS9A, RPS9B, and RPL7A are indicated as green boxes. G) Summary of the percent of supported base-pairs in structures proposed in prior functional studies, R-scape scans for covariation in multiple sequence alignments (MSAs), and in other approaches using sequence alignments to pinpoint structures (Evofold, RNAz, and CMfinder). A base-pair is supported if included in a stem with over 70% helix confidence estimate, and base-pair support statistics are computed based on all base-pairs in proposed structures (functional experiments, Evofold, RNAz, CMfinder) or significantly covarying base-pairs (R-scape covariation).

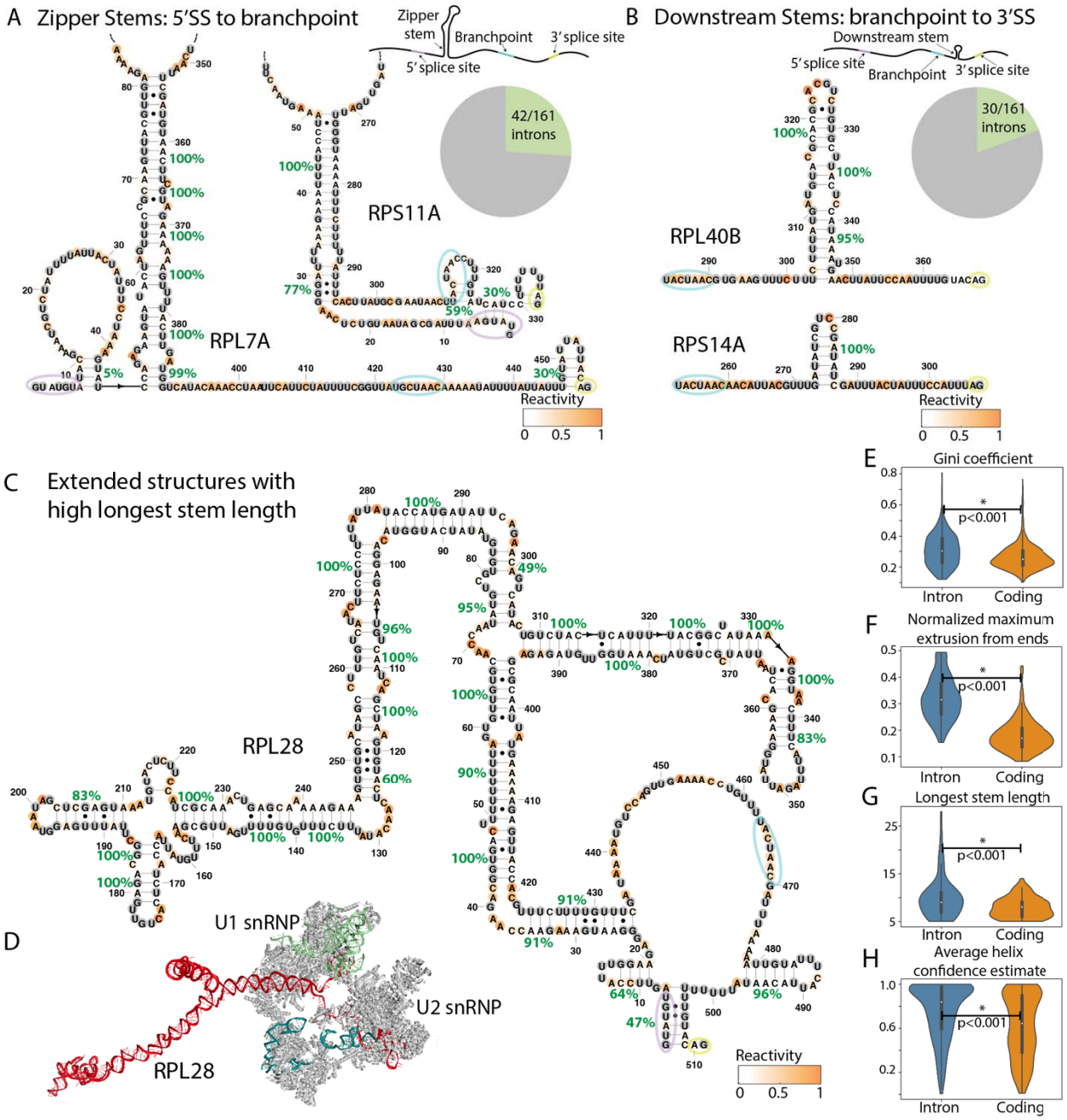

Figure 3: Structural insights from DMS probing of *S. cerevisiae* introns. A) Reactivity support for zipper stems in RPL7A and RPS11A. B) Reactivity support for stems connecting the branch point and 3' splice site in RPL40B and RPS14A. C) Secondary structure for RPL28 predicted by RNAstructure guided with DMS reactivity. 5' splice site, branchpoint, and 3' splice site sequences are circled in purple, blue, and yellow respectively. D) Top-scoring 3D model for the RPL28 intron in the context of the A complex spliceosome (PDB ID: 6G90)[70], modeled using the secondary structure derived from DMS-MaPseq. E)-H) Comparisons of secondary structure features between introns and coding regions were computed using Wilcoxon ranked sum tests to compare classes.

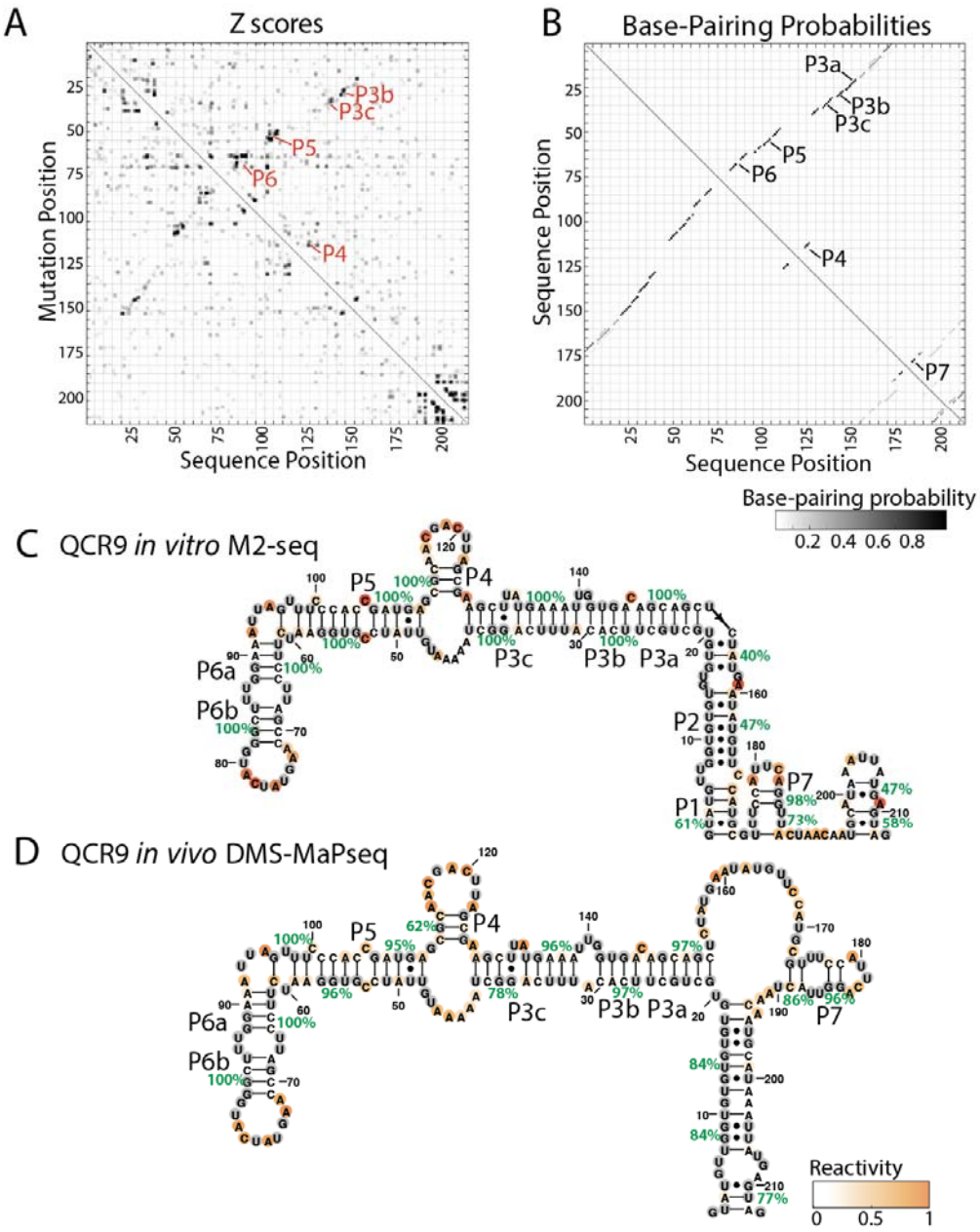

Figure 4. Comparing *in vivo* and *in vitro* folding of intron RNA structures. A) *In vitro* M2-seq Z scores for the intron in QCR9, with peaks representing helices annotated in red. B) *In vitro* chemical reactivity base-pairing probabilities for QCR9 using 1D and 2D chemical reactivity from M2-seq, with peaks representing helices annotated in black. Secondary structure predictions guided by 1D and 2D DMS probing data for the intron in QCR9 C) from *in vitro* M2-seq, and D) from *in vivo* DMS-MaPseq.

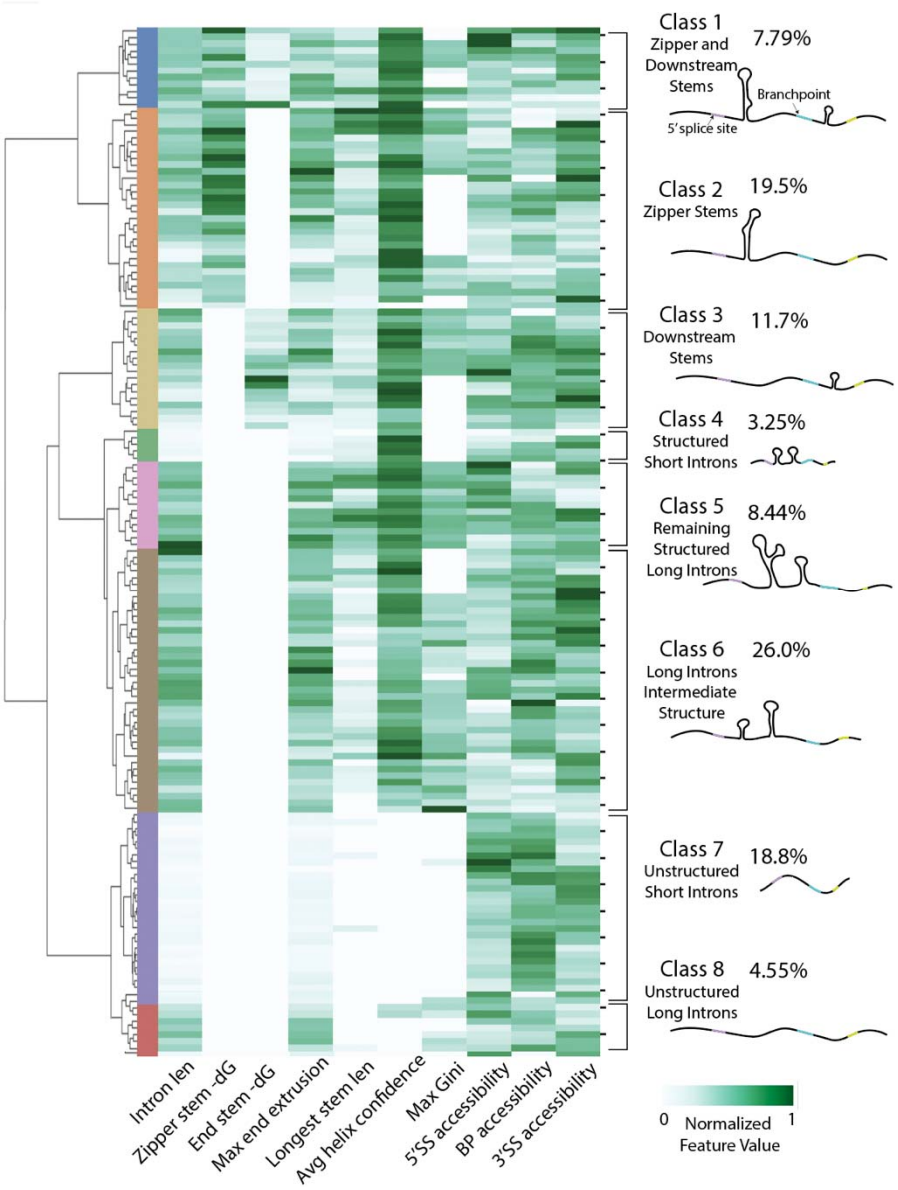

Figure 5: **Structural landscape for *S. cerevisiae* introns.** Heatmap and dendrogram summarizing intron structural classes, with hierarchical clustering based on secondary structure features. In addition to the features displayed on the heatmap, flags indicating whether zipper stems and downstream stems were present were included as features for hierarchical clustering.

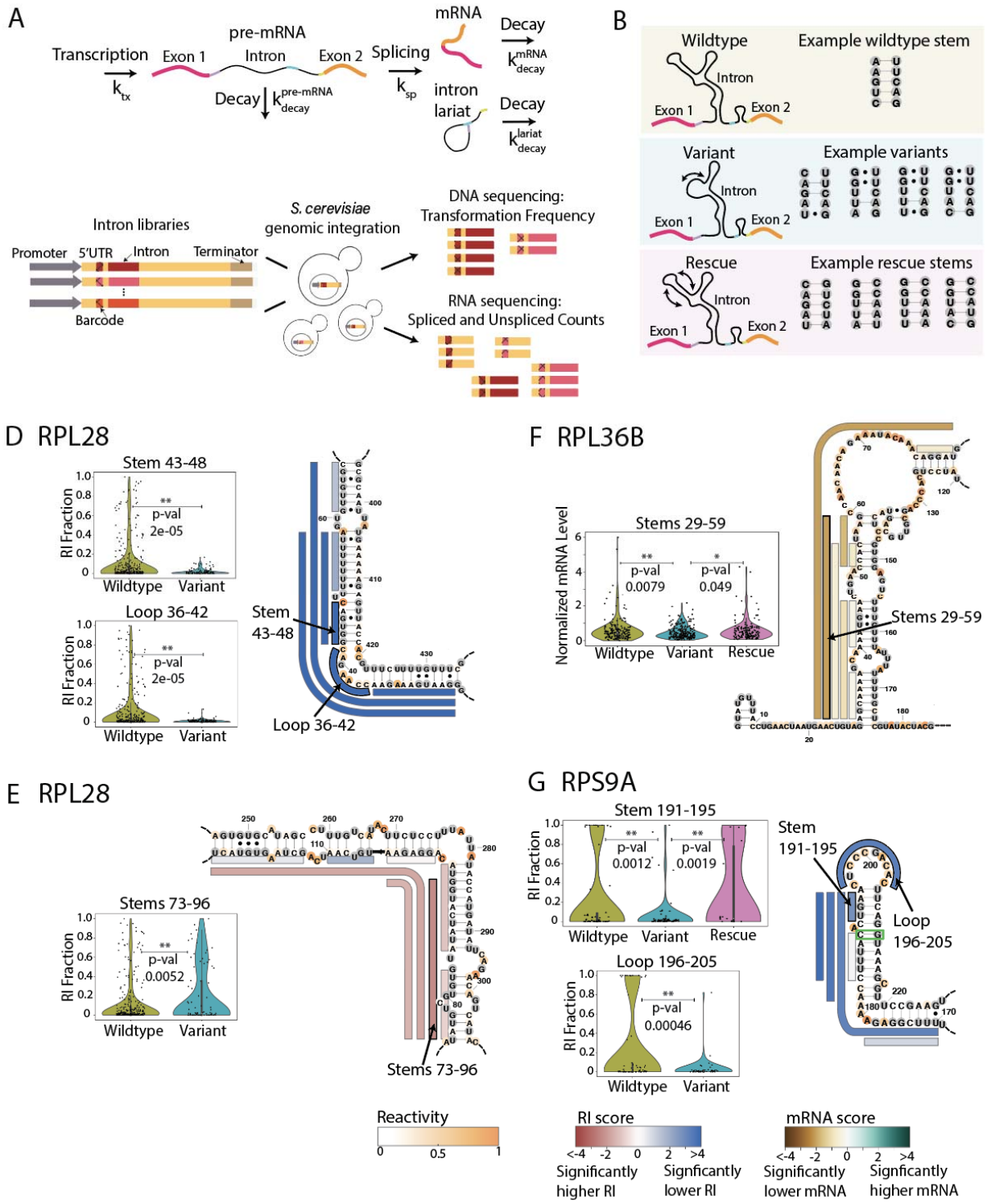

Figure 6: High-throughput structure-function assay for evaluating intron stem variants. A) Schematic for an example library design for one intron stem, with variants disrupting the stem and rescue sequences restoring base-pairing. B) Experiment overview for structure-function experiment. C)-F) The effects of structure variants on retained intron fraction for two regions of the RPL28 intron in C)-D) and for stems in RPS9A in F), and effects on normalized mRNA

levels for RPL36B in F). All p-values are computed using permutation tests for the difference in mean statistic. For a given stem set or loop, violin plots depict data for the wildtype sequence and all variant sequences, with black points for each unique barcode. Data for rescue sequences are shown when included in the intron library. p-values are indicated for comparisons between wildtype and variant sequence sets, and between variant and rescue sequence sets. Secondary structures are colored by reactivity data. Bars alongside the secondary structure indicate stem and loop disruption sets, with each bar representing a set of variant sequences mutating nucleotides across the full extent of the bar. These bars are colored by the retained intron (RI) score (C-D, F) and mRNA scores (E) for the corresponding stem or loop disruption set. The RI score is computed as the negative log(p-value) when comparing RI values between wildtype and variant sequences, and the sign is used to indicate the effect direction, with positive values (shown as blue) for lower variant RI compared to wildtype, and negative values (shown as red) for higher variant RI compared to wildtype. Similarly, the mRNA score is computed as the negative log(p-value) when comparing normalized mRNA levels between wildtype and variant sequences, with positive (green) values indicating higher variant mRNA levels, and negative (brown) values indicating lower variant mRNA levels. The green box in F) indicates a significantly covarying residue.

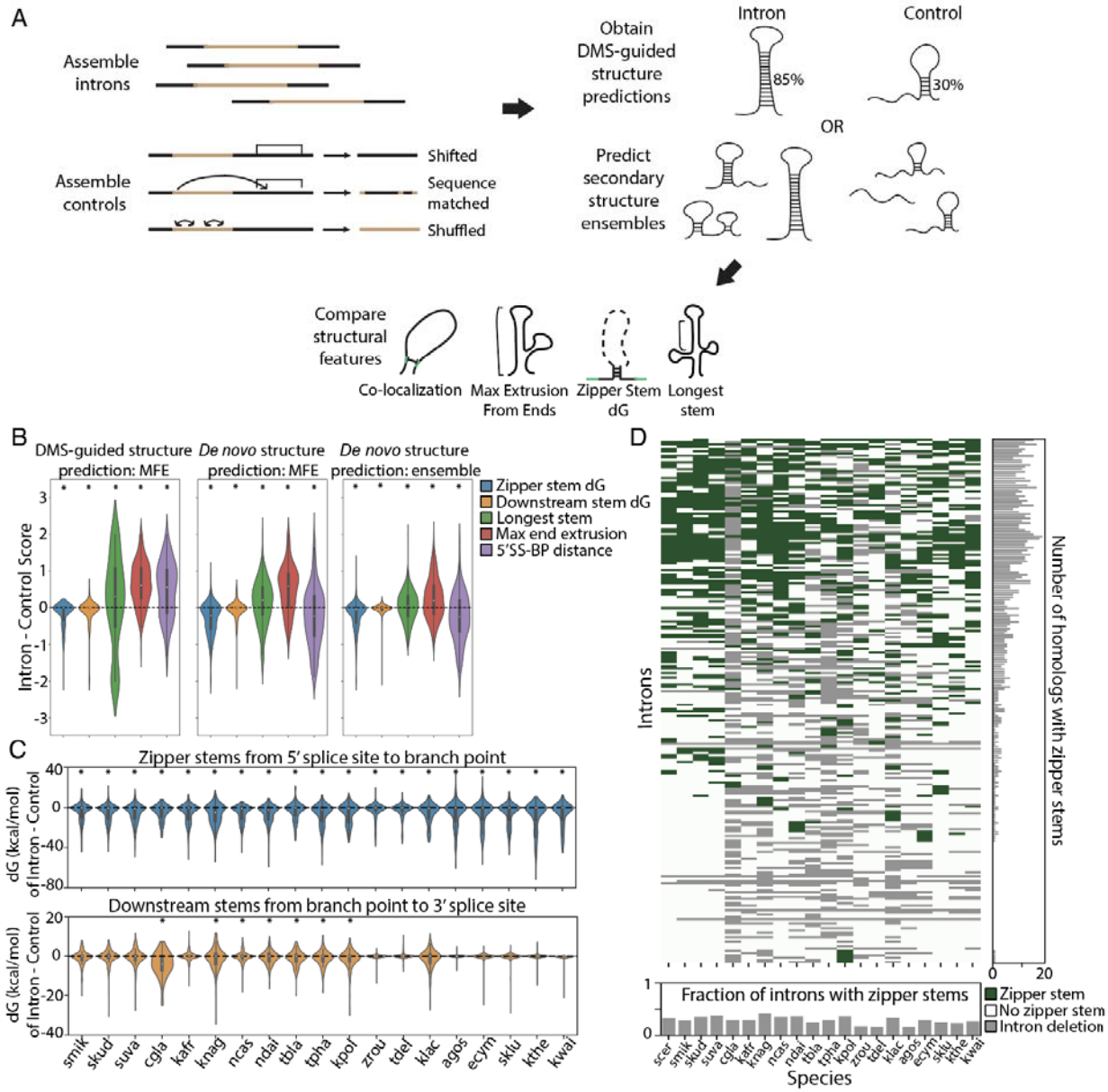

Figure 7: *De novo* secondary structure feature prediction for *S. cerevisiae* and the *Saccharomyces* genus. A) Workflow for comparisons between introns' secondary structure ensembles and those of control sequences. B) Comparison of secondary structure feature enrichment between introns and control sequences for DMS-guided structure prediction (with folding engine RNAstructure, comparing introns to shifted genomic controls), *de novo* MFE structure prediction (with folding engine RNAstructure, comparing introns to shifted genomic controls), and *de novo* ensemble structure prediction (with folding engine Vienna 2.0, comparing introns to shuffled sequence controls). *p-value < 0.01 by Wilcoxon ranked-sum test. C) Difference in zipper stem (top) and downstream stem (bottom) dG between introns and shuffled sequence controls for introns in the *Saccharomyces* genus, using Vienna 2.0 ensemble predictions. *p-value < 0.01 by Wilcoxon ranked-sum test. D) Distribution of zipper stems across introns in the *Saccharomyces* genus. Heatmap green values indicate a predicted zipper

stem, white indicates no predicted zipper stem, and gray values indicate deleted introns. Ohnologous introns are combined into a single row, and a zipper stem is annotated if present in either ohnolog. The species represented in this figure are abbreviated: *E. gossypii* (agos), *C. glabrata* (cgla), *E. cymbalariae* (ecym), *K. africana* (kafr), *K. lactis* (klac), *K. naganishii* (knag), *V. polyspora* (kpol), *L. thermotolerans* (kthe), *L. waltii* (kwal), *N. castellii* (ncas), *N. dairenensis* (ndai), *S. kudiavzevii* (skud), *S. mikatae* (smik), *S. uvarum* (suva), *T. blattae* (tbla), *T. delbrueckii* (tdel), *T. phaffii* (tpha), *Z. rouxii* (zrou).