**RNA structure landscape of *S. cerevisiae* introns**

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

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

* Corresponding author, Email: rhiju@stanford.edu

# Abstract

Pre-mRNA secondary structure is hypothesized to play widespread roles in regulating splicing of eukaryotic introns, but direct detection of such structure *in vivo* has been challenging. Here, we characterize *S. cerevisiae* pre-mRNA secondary structures through transcriptome-wide dimethyl sulfate (DMS) probing experiments, enriching for low-abundance pre-mRNA through splicing inhibition. These data enable evaluation of structures predicted in phylogenetic and mutational studies. The data further reveal the 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. Structure ensemble prediction for introns across the *Saccharomyces* genus suggests that these structural patterns appear in all these yeast species. Together, these findings represent the first transcriptome-wide mapping of intron RNA structures and suggest new ideas and model systems for understanding how pre-mRNA folding promotes splicing efficiency and regulation of gene expression.


# Introduction

In eukaryotic organisms, splicing is a critical post-transcriptional processing step in which stretches of pre-mRNA called introns are removed from transcripts to generate processed coding mRNA sequences. In the splicing reaction, splice sites are selected specifically over sequence-similar cryptic alternatives 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] The process of splicing is facilitated by a large and dynamic macromolecular machine, the spliceosome, which undergoes numerous rearrangements to catalyze these two splicing steps. Though progress has been made identifying splice sites, predicting steady-state splicing levels, and understanding the spliceosome's structure,[1-3] many of the mechanisms that regulate splicing remain unclear. In addition, functional roles for introns beyond splicing have been proposed[4-8] but are not well understood.

The pre-mRNA substrates of the spliceosome are complex RNA sequences that occupy an ensemble of secondary structures including single-stranded and double-stranded regions. Pre-mRNA secondary structure can play a role in mediating splice site selection and splicing efficiency.[9] Moreover, RNA secondary structures like those found in introns can function in numerous other nuclear processes such as RNA editing, RNA cleavage and end-processing, rRNA modification, and regulating gene expression.[10] *S. cerevisiae* has been a critical model system for the study of eukaryotic splicing and provides a system to study the role of pre-mRNA secondary structures, with the catalytic steps of splicing and the spliceosomal machinery highly conserved across eukaryotes from *S. cerevisiae* to humans.[1] In several cases, secondary structures in *S. cerevisiae* introns have been implicated in modulating splicing levels or playing other roles in regulating gene expression.[9] For instance, stems can link the 5' splice site to the branch point to enable efficient splicing,[11-14] and hairpins can lower the effective distance between the branch point and 3' splice site while occluding cryptic splice sites.[15] 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 alternative splicing.[7,16-18]

These cases highlight the potential variety of functions for pre-mRNA secondary structures, and yet, we are missing an experimental survey of structures in introns across any entire transcriptome. Without structural data, it is challenging to assess the ways introns fold *in vivo* for specific cases, and it is difficult to understand any global structural features that might participate in splicing regulation. Though scans for covariation across sequence alignments have been able to identify potentially functional intron structures,[19] 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.[16-18,20]

*In vivo* chemical probing approaches provide an avenue for obtaining necessary structural data on RNA across the transcriptome.[21] However, due to high sequencing coverage requirements, existing structure probing datasets have not been able to obtain reactivity data for many low-abundance pre-mRNAs. Here, we use splicing inhibition to enrich for unspliced pre-mRNA in transcriptome-wide structure probing experiments, assessing the presence of structural motifs in introns *in vivo* and identifying patterns in accessibility and folding that distinguish introns from coding regions. Through these experiments along with structure predictions across yeast species, we provide an atlas of pre-mRNA structural elements that serve as a foundation for further functional characterization.

# Results

**Transcriptome-wide structure probing with splicing inhibition in *S. cerevisiae*.** Structure probing experiments like DMS-MaPseq[21] can evaluate the formation of RNA secondary structures across the transcriptome. However, since splicing proceeds rapidly in yeast,[22] the coverage of pre-mRNA is limited in existing DMS-MaPseq datasets for *S. cerevisiae*, preventing the analysis of reactivity data in 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).[23] By using a strain carrying mutations in HSH155 HEAT repeat domains 15-16 to match sequences from human SF3B1, we enable pladB to bind the U2 snRNP and stall the assembling pre-spliceosome (A-complex spliceosome) 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 more comparable to 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 (r value 0.91, Fig. S2B).[21] Moreover, DMS reactivity values permitted successful classification of accessible (solvent accessible and not base-paired) from inaccessible (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,[21] 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 at least 0.5 R-value between replicate reactivity correlation.

**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[24], and we additionally assigned confidence estimates for helices in the secondary structure by performing nonparametric bootstrapping on reactivity values.[25] Helix confidence estimates from bootstrapping have empirically been useful for improving the quality of stem predictions.[25] 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. 2A, Fig. 2C). Using reactivity data to make DMS-guided structure predictions improved the quality of predicted stems (PPV 62.6%) compared to *de novo* computationally predicted structures (PPV 54.3% for RNAstructure[24], PPV 53.3% for ViennaRNA 2.0[26]). Using a helix confidence estimate threshold of 70% provided a further improvement in the quality of predicted stems (PPV 82.3%) (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. 2B). DMS-guided structure prediction with helix confidence estimation was able to recover stems with this PPV and sensitivity in native secondary structures even for larger control RNAs like the U1 snRNA (Fig. S3).

We explored additional structure prediction approaches to identify potential pseudoknots and alternative conformations in introns using the DMS data. To find potential pseudoknots,

structural motifs that often participate in stable three-dimensional RNA folds, we predicted structures using ShapeKnots[27] guided by DMS data for each intron. No introns included predicted pseudoknots with helix confidence estimates exceeding the 70% confidence threshold. Additionally, for eight intron regions with high coverage (see Methods), we generated structure predictions using DREEM[28] to test for alternate conformations represented by the DMS data. However, reactivity data for all eight regions were best explained by a single structure.

We turned to multi-dimensional chemical mapping (M2-seq[29]) to verify the formation of base-pairs found from DMS-guided secondary structure prediction, since M2-seq identifies base-pairing partners in addition to providing average per-residue accessibility data. For a set of five introns from DMS-MaPseq, we generated a pool of RNA *in vitro* transcribed with sparse errors that were installed via error-prone PCR. These mutations can lead RNA molecules to adopt altered secondary structure ensembles, with a mutated residue in a stem exposing its base-pairing partner in solution and potentially leading to unfolding of the stem. Chemical probing of this RNA pool by DMS, in addition to modifying positions that are accessible in the wildtype RNA, also modifies base-pairing partners for mutated stem residues and any other newly accessible positions in the mutated RNA. Off-diagonal signals in the resulting background-subtracted Z-score plots indicate the presence of a stem, as these appear when one residue's mutation leads to an increase in another residue's accessibility. In the case of the introns in QCR9 and RPL36B, the Z-score plots from *in vitro* M2-seq included these off-diagonal signals, with mutations in some positions resulting in increased DMS reactivity at other distal positions (Fig. 2D, Fig. S4A). When using RNAstructure to predict secondary structures for these constructs using both 1D and 2D reactivity data, the resulting helix confidence estimates from bootstrapping confirm the formation of stable stems in both cases (Fig. 2F, Fig. S4). The Z-scores include two-dimensional reactivity signals for many of these stems (Fig. 2D, Fig. S4A), and the resulting base-pairing probabilities support the formation of these stems as these introns' primary structure (Fig. 2E, Fig. S4B). These secondary structures align with the stems observed *in vivo*, with most high confidence stems *in vivo* also appearing in the *in vitro* M2-seq structures (Fig. 2F-G, Fig. S4C-D). For the introns in RPS11A, RPL37A, and RPS7B, though M2-seq Z-scores did not include visually apparent off-diagonal signals beyond background, helix confidence estimates computed using the 1D and 2D reactivity data showed support for the stems observed *in vivo* (Fig. S5).

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 done 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. The secondary structures predicted for the RPL26B and RPL28 introns using DMS agreed well between conditions with and without pladB treatment (Fig. S6A-B). Additionally, while some stems in the RPS13 intron disagree between these conditions, all high confidence stems (helix confidence estimate over 70%) were shared between conditions (colored segments in Fig. S6C). These comparisons suggest that the structural profiles adopted by introns in the condition with pladB treatment will 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 various approaches, including functional experiments,[7,11,12,15-18,20] covariation scans that pinpoint functional base-pairs,[19] or *de novo* prediction methods based on conservation in sequence alignments.[30] Reactivity profiles from splicing inhibited DMS-MaPseq enabled us to assess the formation of these proposed secondary structures *in vivo*. We determined whether stems from proposed structures had high helix confidence estimates from our probing data, supporting their formation *in vivo*. We note that 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 or splicing stage. Additionally, we note that our stem predictions include some false positives and false negatives (16.4% false positive rate and 33.8% false negative rate on controls, Fig. 2A-B).

Prior studies point to potential regulatory structures in *S. cerevisiae* introns, with experiments assessing the role of structure using mutants and compensatory mutants. For instance, in the case of RPS17B, a stem linking the 5' splice site to the branch point, termed a "zipper stem", enables efficient splicing despite this intron's weak 5' splice site.[11,12] A stem between the branch point and 3' splice site in RPS23B hides a more proximal cryptic 3' splice site and enables thermosensitive regulation of 3' splice site selection for this intron.[15] In the case of the introns in RPL32, RPS9A, and RPS14B, portions of the intron have been implicated in regulating gene expression by binding of excess protein product to pre-mRNA, leading to gene-specific reduced splicing efficiency.[7,16-18] Finally, structures in RPL18A and RPS22B have been found to mediate the degradation of unspliced pre-mRNA.[20]

In four out of six cases, structures from prior functional experiments received medium or high support from our DMS data, with high loop reactivity, low stem reactivity, and high helix confidence estimates (Fig. 3). For instance, the structure in RPL18A involved in pre-mRNA degradation includes high confidence stems from DMS probing (Fig. 3A). DMS data additionally support the stem in RPS23B co-localizing the branch point and 3' splice site (Fig. 3B), and high confidence stems are identified in the secondary structure proposed in RPS14B to bind the RPS14B protein (Fig. 3C). Finally, high confidence stems are predicted in the RPS9A intron structure proposed to regulate gene expression levels of RPS9A and RPS9B. In some cases, proposed structures are not supported by the probing data, with DMS accessible residues present in these structures' stems. These lower support structures include the zipper stem in RPS17B (Fig. S7A), and the RPL30 structure proposed to regulate gene expression (Fig. S7B). The structure in RPS22B found to trigger pre-mRNA degradation was not assessed, as this intron did not pass our DMS-MaPseq coverage threshold. With most structures in this class showing medium or high DMS support, proposed structures from functional experiments assessing mutants and compensatory mutants are in general validated by our data (Fig. 3G).

Covariation across sequence alignments can identify functional secondary structures by identifying signals for compensatory mutations across related species. We evaluated DMS support for structures with covarying residues identified through R-scape by Gao, et al. (2021).[19] To supplement this set of structures with predicted covariation, we followed this approach to scan for covariation in *S. cerevisiae* introns across the *Ascomycota* phylum, requiring the presence of at least one significant covarying pair in a stem of at least three base-pairs (more stringent cutoffs were used in Gao, et al. 2021[19]). We focused on the 11 introns with length at

least 200 nucleotides where covarying residues were consistent with the predicted minimum free energy structure (Fig. S8). As a positive control for functional secondary structures in introns, we noted the enrichment of snoRNA-containing introns amongst the sequences with predicted covariation. Indeed, out of the eight *S. cerevisiae* introns that encode snoRNAs,[31] seven included significantly covarying residues within the snoRNA, suggesting that the remaining introns with predicted covarying residues may also encode functional structural elements.

We evaluated intron structures that included covarying residues using DMS data, finding that in the majority of cases, covarying base-pairs were supported by the DMS data. Six of the seven snoRNA-containing introns had sufficient coverage from DMS-MaPseq for evaluation. In these six introns, a majority of covarying residues were supported by DMS data, with 20 of the 27 covarying residues present in high-confidence stems from DMS-guided structure prediction (Fig. S9). Beyond snoRNA-containing introns, covarying residues were identified in four introns from our covariation scan and in two introns from Gao, et al. (2021).[19] First, we identified covariation in stems of RPL7A's first intron, and these covarying residues were included in high confidence stems forming a 3-way junction (Fig. 3D). Next, as noted by sequence analysis in Plocik and Guthrie (2012),[7] we found signals for covariation in a hairpin shared by the introns in RPS9A and RPS9B, and covarying base pairs in these two introns align with the secondary structure from DMS-MaPseq (Fig. 3E-F). Only the structure in RPS13 noted to have covarying residues in Gao, et al. (2021)[19] was assigned low helix confidence estimates (22%), suggesting that these residues are not paired in the dominant *in vivo* structure (Fig. S7C). The first intron in RPL7B, found to have covarying residues by Gao, et al. (2021)[19] and in our covariation scan, did not have enough sequencing coverage from DMS-MaPseq for analysis. Overall, the prevalence of DMS-validated covarying residues across these cases suggests that covariation from R-scape[32] can reliably identify structures that form *in vivo* (Fig. 3G).

As a final test of proposed functional structures, we evaluated those identified by Hooks, et al. (2016)[30] through predictions from CMfinder[33], RNAz[34], and Evofold[35], approaches that use sequence alignments to identify structures. These structures showed mixed support from DMS data, with introns in YRA1, PSP2, RPL18B, RPL7A, and RPL28 including some stems with high confidence and other stems with low DMS support (Fig. S10A-E). The intron structures in RPL22B and GLC7 were not predicted from the DMS data (Fig. S10F-G). Proposed intron structures in MPT5 and RPS22B could not be evaluated due to low sequencing coverage. Unlike R-scape, RNAz uses conservation alone rather than covariation, and CMfinder and Evofold do not use phylogeny sequence backgrounds to identify significant covariation. With a higher fraction of predicted structures that were not validated by DMS data from this set, these approaches appear to less reliably identify structures that form *in vivo* compared to R-scape covariation and functional experiments.

**Structural features from structure probing in *S. cerevisiae* pre-mRNA.** DMS-guided structure predictions from the 161 introns passing coverage thresholds support the presence of numerous structural features in *S. cerevisiae* introns *in vivo*. 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,[13] and a zipper stem in RPS17B has been shown to be essential for efficient splicing.[11,12] Zipper stems could lower the entropic cost of key splicing sequences interacting in the spliceosome, and

moreover, could directly interact with interfaces on the U1 snRNP to enhance introns' binding to the spliceosome, as seen in recent cryo-EM structures of the pre-spliceosome.[36,37]

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[38] (Fig. S11A-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 (at least 6 base-pairs in the stem with at least 70% helix confidence estimate) are predicted for 42 of 161 introns with sufficient read coverage for DMS-guided structure inference (Fig. 4A). Reactivity values align with the formation of these zipper stems, with the A and C residues in stems having low reactivity values while neighboring loop residues are highly reactive, as in the case of QCR9, RPL36B, RPS11A, and RPS7B (Fig. 2G, Fig. S4D, Fig. 4A).

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.[15] 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. 4B). 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. 3B, Fig. 4B).

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 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. 4C). 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. 4D). 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. S11C).

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,[39] intron regions have higher Gini coefficients, a measure quantifying the extent to which reactivity values diverge from an even distribution (Fig. 4E). 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 (Fig. 4F). 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. 4G). Finally, intron stems with at least 6 base-pairs in the predicted minimum free energy structure tended to have higher helix confidence estimates on average, indicating that they more often included reactivity values that aligned well with the secondary structure (Fig. 4H). 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 strongly by DMS-guided structural analysis.

The secondary structure patterns enriched in introns may distinguish spliced introns from other 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. S12A). 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 dG) 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. S12B). In contrast to spliced introns, the DMS-guided secondary structures of unspliced decoys were not enriched for these features (Fig. S12B). It is thus possible that authentic *S. cerevisiae* introns are enriched in secondary structure patterns that enable efficient splicing, distinguishing them from decoy sequences that include consensus splice site sequences and yet remain unspliced.

**Structural landscape for *S. cerevisiae* introns.** To understand the distribution of structural patterns in introns across the yeast transcriptome, we visualized the *S. cerevisiae* structural landscape by clustering introns based not on their sequence but on their structure features. For each intron with sufficient DMS-MaPseq coverage, we assembled a full set of their 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. We clustered introns with hierarchical clustering and tSNE using these features, and in Fig. S16, we display secondary structures and reactivity profiles for all introns from DMS-MaPseq, grouped into classes from hierarchical clustering. Additionally, we tabulate the secondary structure features and DMS-guided secondary structures for *S. cerevisiae* introns in Table S2.

The clustered heatmap (Fig. 5A) and tSNE visualizations (Fig. 5B) depict the global distribution of secondary structure features across all introns with sufficient sequencing coverage. From hierarchical clustering, the first class of introns includes the 12 introns that have both a zipper stem and downstream stem, with high average helix confidence estimates (Class 1, blue in Fig. 5A). Class 2 includes 30 introns with zipper stems, and within this class, introns with more stable zipper stems clustered together (orange Fig. 5A). 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. 5A). 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).[40] Class 4 includes structured short introns with at least one high confidence stem (green in Fig. 5A). 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. 5A). 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. 5A). 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. 5A). Finally, Class 8 includes a small set of long introns depleted of structure, including only short stems with low helix confidence (red in Fig. 5A). tSNE clustering separates introns with zipper stems and downstream stems from others and additionally isolates a group of unstructured short introns (Fig. 5B). Of the long introns, a sizable fraction include either a zipper stem or downstream stem, with 51.7% of long introns including at least one of these motifs.

**Computational prediction of enriched structural patterns across the *Saccharomyces* genus.** Our transcriptome-wide structure mapping identified various structural patterns in *S. cerevisiae* introns, including zipper stems, downstream stems, and extended structures, with introns including more structured domains and longer helices than coding regions and unspliced decoy introns. To extend our observations from structure probing data in *S. cerevisiae* to other species in the *Saccharomyces* genus, we turned to computational structure prediction. Secondary structure prediction provides an avenue for rapidly identifying genome-wide structural patterns in introns. We first aimed to evaluate whether *de novo* secondary structure prediction could recapitulate structural features observed through DMS-guided structure analysis. To identify enriched structural patterns, we compared structural features between introns and length-matched controls (either sequences with shifted genomic coordinates or shuffled sequences), using DMS-guided secondary structure predictions and *de novo* structure predictions (Fig. 6A, Fig. S13). For *de novo* structure prediction, as a first approach, we predicted minimum free energy structures using RNAstructure.[24] As a second approach, we generated secondary structure ensembles through stochastic sampling of suboptimal structures,[26] which has previously enabled the study of structural patterns in introns.[12,15]

Structural features from *de novo* secondary structure prediction were largely consistent with patterns from DMS-guided structures for *S. cerevisiae*. Zipper stems and downstream stems in introns are more stable (lower dG values) than those in control sequences, whether using DMS-guided structure prediction, *de novo* minimum-free energy structure prediction, or *de novo* secondary structure ensemble prediction (p-values < 0.01, blue and orange in Fig. 6B). Introns have longer stems than length-matched control sequences and have higher maximum extrusion from ends, again both from *de novo* structure predictions and DMS-guided structure predictions (p-values < 0.01, green and red in Fig. 6B). We additionally measured the secondary structure graph distance between the 5' splice site and branch point sequence positions, comparing between introns and the control sequences (purple in Fig. 6B). From DMS-guided structures, the 5' splice site and branch point were more distant in introns than controls (p-value < 0.01), perhaps due to the requirement for single-stranded nucleotides proximal to these sequences in the A-complex spliceosome. On the other hand, *de novo* structure prediction suggests that introns have lower secondary structure graph distances between the 5' splice site and branch point without the context of the spliceosome (p-values < 0.01), as found by Rogic, et al. (2008),[12] suggesting that introns co-localize key splicing sequences and perhaps lower the entropic cost for

their interaction in the spliceosome. We tested various other structure prediction approaches and found similar structural feature enrichment, comparing against other control sets, using an alternate folding package, and making predictions including exonic context (Fig. S14). In particular, the enrichment for stable zipper stems, higher maximum extrusion from ends, and lower 5' splice site to branch point distances remained significant across all tested *de novo* prediction approaches.

With DMS-guided structure prediction and *de novo* structure prediction pointing to similar intron structural patterns, we proceeded to use *de novo* structure prediction to observe these patterns across species in the *Saccharomyces* genus, focusing on the 20 species in this genus with curated intron alignments from Hooks et al. (2014).[41] 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. 6C), along with higher maximum extrusion from ends and shorter distances between the 5' splice site and branch point (Fig. S15A). 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. S15B).

Secondary structure patterns are maintained across *Saccharomyces* species despite sequence-level divergence between introns, suggesting that these structural patterns may 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. S15C), and most intron regions have low sequence conservation between species, with most intron sequences 50-60% conserved between these species (Fig. S15D). Zipper stem regions in *S. cerevisiae* introns in particular 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 homologous introns (Fig. 6D). 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 have been shown to 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, either by altering splice site choice and splicing kinetics, or by interacting with other nuclear factors and performing orthogonal regulatory functions. Here, we sought to understand the transcriptome-wide 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. Through computational structure prediction, we identified signals for these patterns across the *Saccharomyces* genus.

Structure probing experiments have enabled the measurement of RNA structure transcriptome-wide, but these experiments have high coverage requirements and cannot provide information for low abundance transcripts including many unspliced pre-mRNA molecules. Reactivity data for low abundance introns has previously been obtained by enriching for specific targets of interest. This target enrichment has been achieved by targeted amplification during library preparation[21] or by biotinylated pulldown of an intron of interest.[42] Additionally, prior experiments enriched for nuclear RNA in compartment-specific structure probing experiments.[39,43] We broaden the capture of low abundance targets by using global splicing inhibition to identify transcriptome-wide structural patterns in unspliced pre-mRNA. 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[44] and may instead for instance be free of interaction partners in the nucleus or interacting with the nuclear pore complex,[45] nuclear exosome,[46] or nonsense mediated decay machinery.[47] For the cases where we observed sufficient sequencing coverage from experiments with untreated cells, we found similar structure profiles between the conditions with and without splicing inhibition, suggesting that structures determined with splicing inhibition are relevant for untreated cells.

A pattern that emerged from both computational structure prediction and *in vivo* chemical mapping was the presence of zipper stems between 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*. Past work on RPS17B has demonstrated that these zipper stems can be essential for efficient splicing.[11] Zipper stems may enable efficient splicing by reducing the physical distance between the 5' splice site and branch point; these reduced effective distances have been noted for some pre-mRNA molecules in single-molecule studies of splicing.[48] They may also facilitate efficient splicing by enhancing specific interaction with the spliceosome, with intron helical density seen binding positively charged pockets in the spliceosome in both the E and pre-A complexes.[36,37] Alternatively, formation of intron helices may discourage unfavorable interaction with single stranded RNA elements of the spliceosome.

Observing DMS reactivity across introns in the transcriptome allows us to identify structural patterns that distinguish introns from coding regions. Introns adopted extended structures with longer maximum extrusion from their sequence endpoints, perhaps allowing structural elements to project out of pre-mRNAs captured by the spliceosome. Introns also included more non-uniform reactivity values as measured by Gini coefficients, pointing to the enrichment for stable structures *in vivo* in introns compared to coding regions, as found in earlier compartment-specific SHAPE mapping of RNA structures in human cells.[39] Additionally, introns included longer stems with higher confidence estimates. In most cases, the intron stems formed *in vivo* had similar reactivity profiles and bootstrapping support to those probed *in vitro* by M2-seq, contrasting from previous studies of primarily cytosolic mRNA demonstrating widespread unfolding of RNA structures *in vivo* compared to *in vitro* refolded controls.[49] It is possible that the presence of helicases or the active translation of coding regions may lead to fewer stable extended structures in coding regions relative to introns. However, our computational secondary structure predictions across *Saccharomyces* species suggest that intron sequences have the

capacity to form more stable zipper stems than nearby genomic regions regardless of their translation status or the presence of other factors. While there is evidence across species for depletion of double stranded RNA (dsRNA) in coding mRNA to avoid activating cellular responses against viral dsRNA,[50] this selection against dsRNA is not present in nuclei. Without the evolutionary constraints of adhering to a viable coding sequence, introns may be free to include RNA sequences that form extended structures with stable, long stems.

Scans for covariation can provide candidate functional structures maintained through evolution. Covariation has been pinpointed thus far in a handful of yeast introns (RPS9A,[7] RPS9B, RPS13,[19] RPL7B,[19] and RPL7A). However, some functional structures in *S. cerevisiae* introns may be missed due to the limited power of existing sequence alignments for identifying significantly covarying residues,[51] with high sequence variation and large-scale deletions often present between aligned intron sequences. Since *S. cerevisiae* introns include stable secondary structures beyond those in random control sequences or coding regions, it is possible they harbor additional functional structures. In the future, expanded sequence alignments may highlight additional candidate functional structures by capturing a broader set of species or including more genome sequences as they become available.

The structural elements identified here in *S. cerevisiae* pre-mRNA provide an atlas of candidates for in-depth functional work. These include broadly present zipper stems, downstream stems, extended structures with long stems (for instance in the RPL28 intron), and structures with covarying residues (for instance in introns in RPS9A, RPS9B, and RPL7A). Interrogating the functions for these elements will require customized experiments for individual introns, as they could play varied roles in regulating splicing or in a host of other nuclear processes. Features like zipper stems and downstream stems which are located close to splice sites could have functional roles in recruiting the spliceosome and facilitating efficient splicing, while structures extruded away from splice sites could play a role in recruiting other nuclear factors while the intron core is buried within the spliceosome. Intron secondary structures can regulate splicing by sequestering alternative splice sites,[52,53] occluding exonic splicing enhancers,[54] co-localizing selected splice sites,[55] or mediating protein interactions that influence splicing patterns.[56] Additionally, these intronic structures could have regulatory functions in pathways orthogonal to splicing, much like the snoRNAs encoded in *S. cerevisiae* introns.[31] In fact, RNA secondary structures in yeast associate with RNA-binding proteins involved in varied processes including transcription, tRNA and rRNA processing, ribosome biogenesis and assembly, and metabolic processes.[57] Furthermore, secondary structures in introns are poised to regulate gene expression in auto-regulatory circuits as seen previously in the cases of RPS14B[17] 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. These structures are candidates for further functional characterization, and experiments using compensatory mutants could pinpoint their roles. 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.

## 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.

## Data availability

All sequencing data are available at the Gene Expression Omnibus (GEO) accession number GSE********. 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.

## References


1. Wilkinson, M.E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu Rev Biochem* **89**, 359-388 (2020).
2. Jaganathan, K. et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e24 (2019).
3. Xiong, H.Y. et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
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. Warf, M.B. & Berglund, J.A. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* **35**, 169-78 (2010).
10. Scharfen, L. & Neugebauer, K.M. Transcription regulation through nascent RNA folding. *J Mol Biol*, 166975 (2021).
11. Charpentier, B. & Rosbash, M. Intramolecular structure in yeast introns aids early steps of in vitro spliceosome assembly. *RNA* (1996).

12. 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).
13. Parker, R. & Patterson, B. *Architecture of Fungal Introns: Implications for Spliceosome Assembly*, (1987).
14. 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).
15. 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).
16. Vilardell, J. & Warner, J.R. Regulation of splicing at an intermediate step in the formation of the spliceosome. *Genes Dev* **8**, 211-20 (1994).
17. 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).
18. 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).
19. Gao, W., Jones, T.A. & Rivas, E. Discovery of 17 conserved structural RNAs in fungi. *Nucleic Acids Res* **49**, 6128-6143 (2021).
20. 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).
21. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods* **14**, 75-82 (2017).
22. Wallace, E.W.J. & Beggs, J.D. Extremely fast and incredibly close: cotranscriptional splicing in budding yeast. *RNA* **23**, 601-610 (2017).
23. 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).
24. Reuter, J.S. & Mathews, D.H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
25. Kladwang, W., VanLang, C.C., Cordero, P. & Das, R. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* **50**, 8049-56 (2011).
26. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
27. 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).
28. Tomezsko, P.J. et al. Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature* **582**, 438-442 (2020).
29. 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).
30. 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).
31. 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).

32. 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).
33. Yao, Z., Weinberg, Z. & Ruzzo, W.L. CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445-52 (2006).
34. 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).
35. Pedersen, J.S. et al. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**, e33 (2006).
36. Li, X. et al. A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375-380 (2019).
37. Zhang, Z. et al. Structural insights into how Prp5 proofreads the pre-mRNA branch site. *Nature* **596**, 296-300 (2021).
38. Watkins, A.M., Rangan, R. & Das, R. FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure* **28**, 963-976 e6 (2020).
39. Sun, L. et al. RNA structure maps across mammalian cellular compartments. *Nat Struct Mol Biol* **26**, 322-330 (2019).
40. Neuveglise, C., Marck, C. & Gaillardin, C. The intronome of budding yeasts. *C R Biol* **334**, 662-70 (2011).
41. Hooks, K.B., Delneri, D. & Griffiths-Jones, S. Intron Evolution in Saccharomycetaceae. *Genome Biology and Evolution* **6**, 2543-2556 (2014).
42. Bubenik, J.L. et al. RNA structure probing to characterize RNA-protein interactions on a low abundance pre-mRNA in living cells. *RNA* (2020).
43. Liu, Z. et al. In vivo nuclear RNA structurome reveals RNA-structure regulation of mRNA processing in plants. *Genome Biol* **22**, 11 (2021).
44. 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).
45. Galy, V. et al. Nuclear Retention of Unspliced mRNAs in Yeast Is Mediated by Perinuclear Mlp1. *Cell* **116**, 63-73 (2004).
46. Bousquet-Antonelli, C., Presutti, C. & Tollervey, D. Identification of a Regulated Pathway for Nuclear Pre-mRNA Turnover. *Cell* **102**, 765-775 (2000).
47. 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).
48. Abelson, J. et al. Conformational dynamics of single pre-mRNA molecules during in vitro splicing. *Nat Struct Mol Biol* **17**, 504-12 (2010).
49. 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).
50. 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).
51. Rivas, E., Clements, J. & Eddy, S.R. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics* **36**, 3072-3076 (2020).
52. Tan, J. et al. A Disease-Causing Intronic Point Mutation C19G Alters Tau Exon 10 Splicing via RNA Secondary Structure Rearrangement. *Biochemistry* (2019).

53. 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).
54. 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).
55. Graveley, B.R. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **123**, 65-73 (2005).
56. Fish, L. et al. A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements. *Science* **372**(2021).
57. 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).
58. Plaschka, C., Lin, P.C., Charenton, C. & Nagai, K. Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature* **559**, 419-422 (2018).
59. 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).
60. 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).
61. Talkish, J. et al. Rapidly evolving protointrons in Saccharomyces genomes revealed by a hungry spliceosome. *PLoS Genet* **15**, e1008249 (2019).
62. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
63. Busan, S. & Weeks, K.M. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA* **24**, 143-148 (2018).
64. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
65. 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).
66. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
67. Ben-Shem, A. et al. The structure of the eukaryotic ribosome at 3.0 A resolution. *Science* **334**, 1524-9 (2011).
68. 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).
69. 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).
70. Li, X. et al. CryoEM structure of Saccharomyces cerevisiae U1 snRNP offers insight into alternative splicing. *Nat Commun* **8**, 1035 (2017).
71. Kalvari, I. et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**, D192-D200 (2021).
72. Lan, P. et al. Structural insight into precursor tRNA processing by yeast ribonuclease P. *Science* **362**(2018).

73. Rodgers, M.L. et al. Conformational dynamics of stem II of the U2 snRNA. *RNA* **22**, 225-36 (2016).
74. 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).
75. Rivas, E. RNA structure prediction using positive and negative evolutionary information. *PLoS Comput Biol* **16**, e1008387 (2020).
76. Benson, D.A. et al. GenBank. *Nucleic Acids Res* **41**, D36-42 (2013).
77. Cherry, J.M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**, D700-5 (2012).
78. Bernhart, S.H. et al. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol* **1**, 3 (2006).
79. 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).
80. Cheng, C.Y., Chou, F.C. & Das, R. Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol* **553**, 35-64 (2015).

## 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.); and the National Institutes of Health grant R35GM122579 (R.D.).

## 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 used in the project. P.P. tested alternate structure prediction approaches. R.R., M.A., and R.D. interpreted the results and wrote the manuscript with input from all authors.

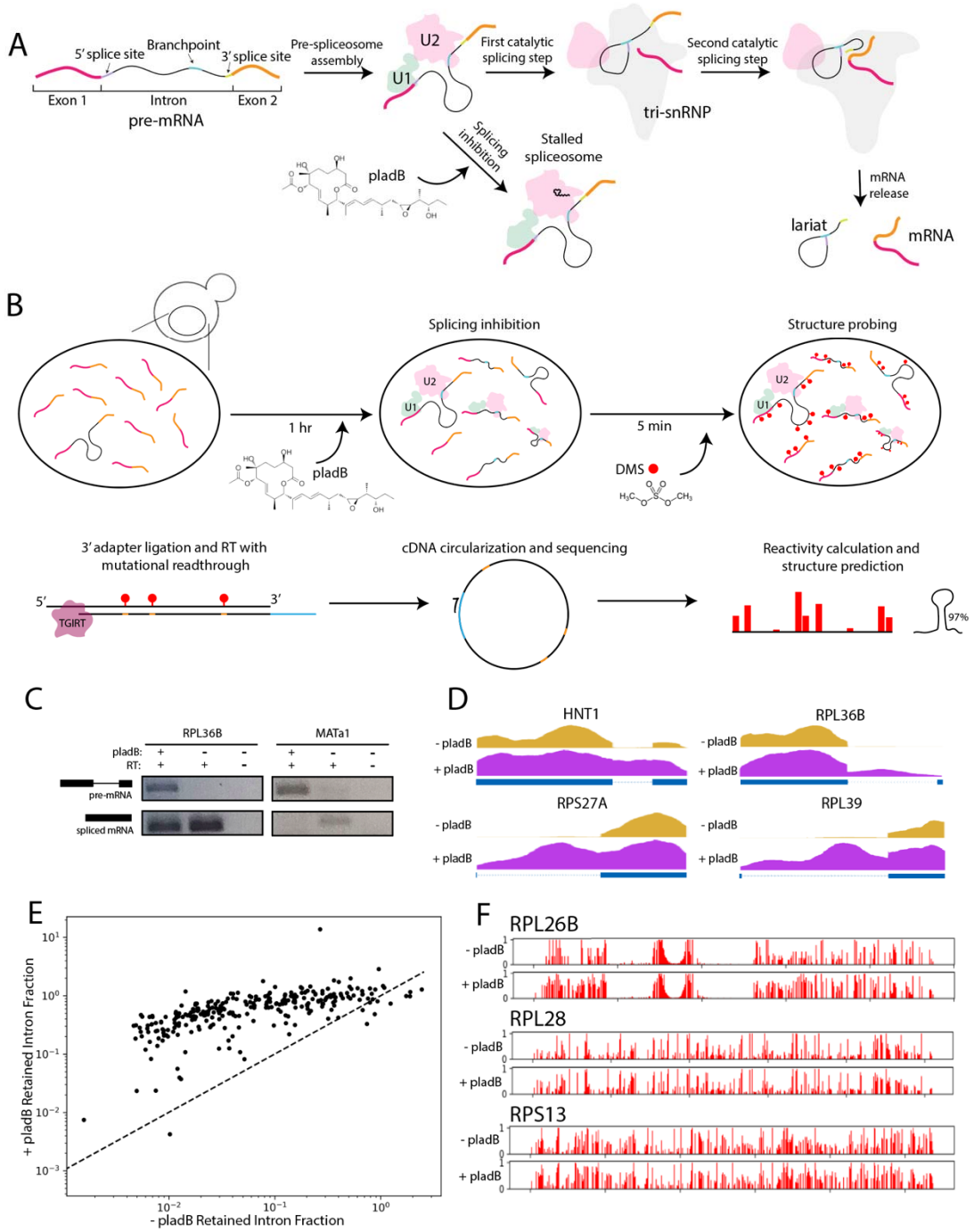


**Figure 1**: **Splicing inhibition by pladB allows for accumulation of pre-mRNA.** A) Schematic of RNA splicing. B) Schematic for splicing inhibition followed by DMS-MaPseq experiment. C) Accumulation of pre-mRNA shown for RPL36B and MATa1 by RT-PCR. D) Read coverage across intron-containing pre-mRNA 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.

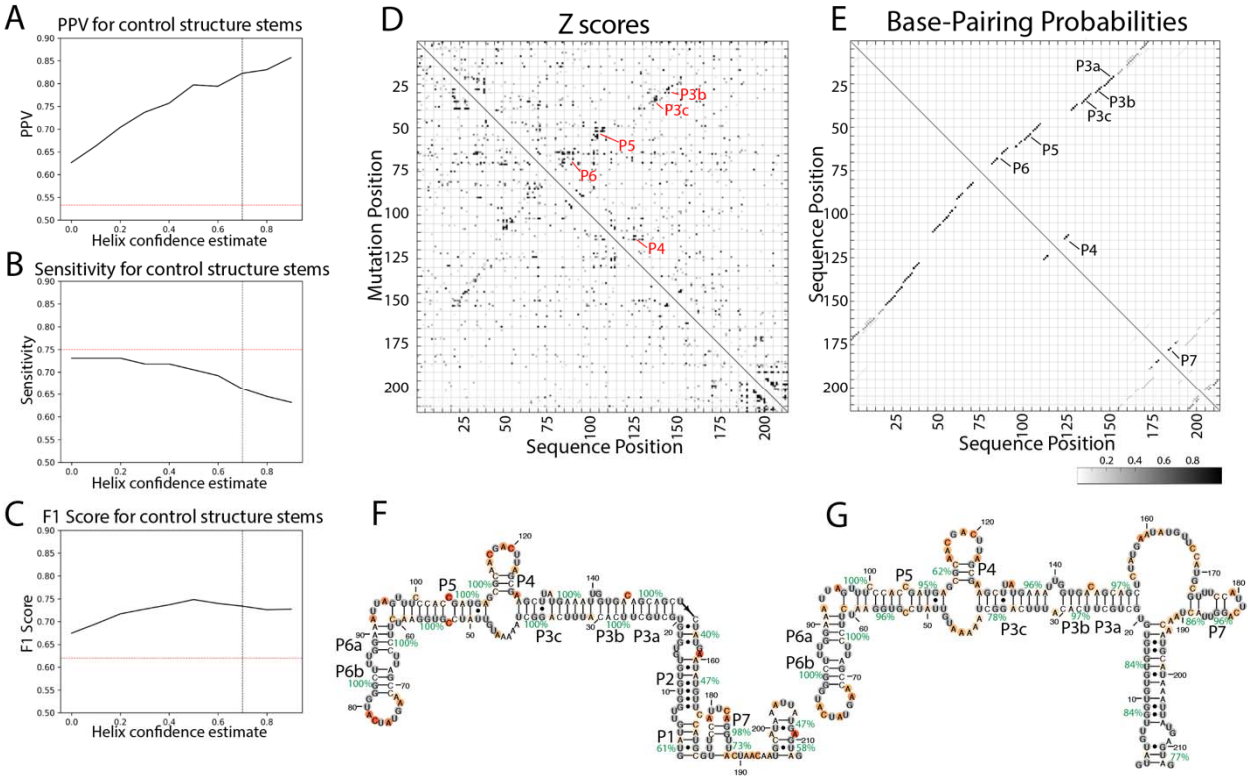


**Figure 2: Assessing intron stem formation with evaluation of control structures and multidimensional chemical mapping.** A) PPV, B) sensitivity, and C) F1 score for structure prediction of a set of control RNA structures (rRNA's, tRNA's, snRNA's, and mRNA's; Table S1), using RNAstructure guided by DMS with various helix confidence estimate cutoffs for identifying stems. The black dotted line represents the helix confidence estimate 0.7 chosen in this paper. The red dotted line represents the PPV, sensitivity, and F1 score for Vienna RNA structure prediction without using DMS data. D) *In vitro* M2seq Z scores for the intron in QCR9, with peaks representing helices annotated in red. E) *In vitro* chemical reactivity base-pairing probabilities for QCR9 using 1D and 2D chemical reactivity from M2seq, with peaks representing helices annotated in black. Secondary structure predictions guided by 1D and 2D DMS probing data for the intron in QCR9 F) from *in vitro* M2seq, and G) from *in vivo* DMS-MaPseq.

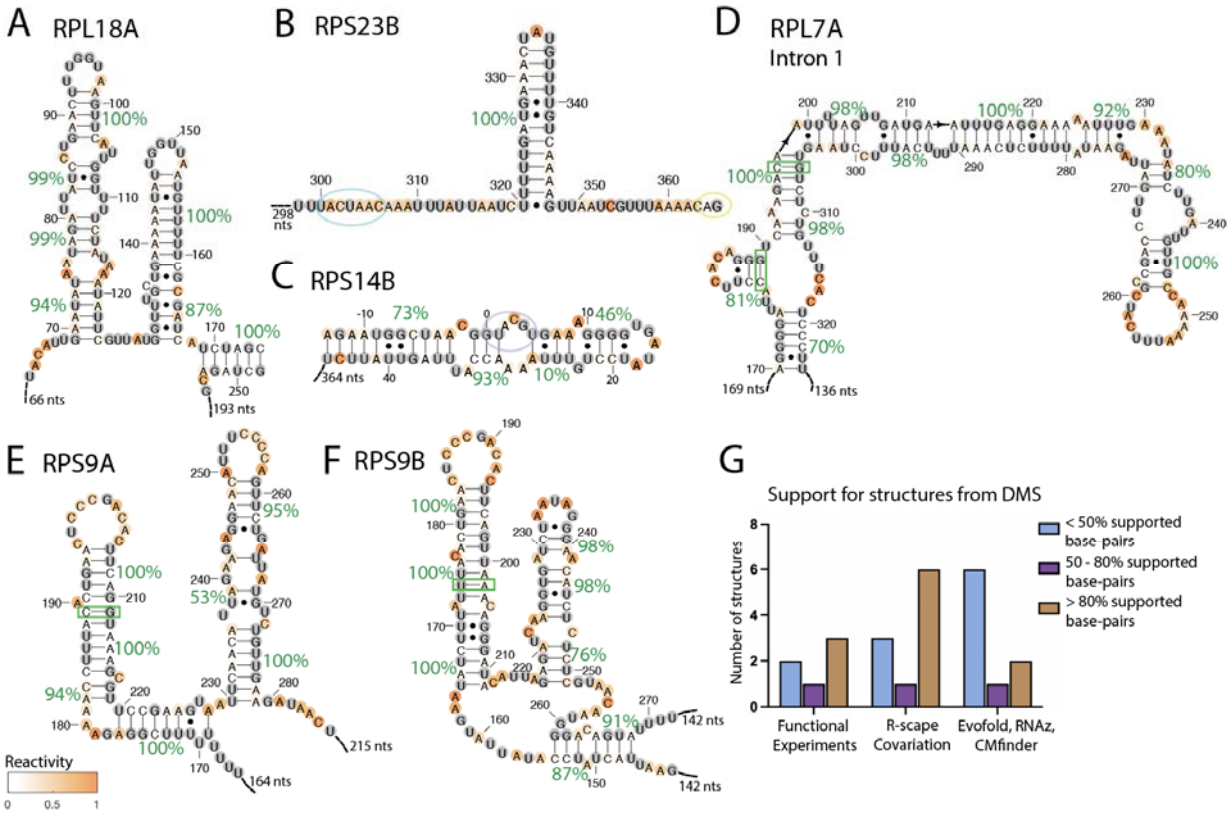


**Figure 3: Proposed functional structures with support from DMS reactivity for *in vivo* formation.** Helix confidence estimates and covariation for intron structures reported in A) RPL18A[20] B) RPS23B[15], C) RPS14B[17], D) the first intron in RPL7A, E) RPS9A, and F) RPS9B[7]. Secondary structures 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 (often including the assessment of compensatory mutations), R-scape scans for covariation in multiple sequence alignments (MSAs), and in other approaches using sequence alignments to pinpoint conserved structures (Evofold, RNAz, and CMfinder). A base-pair is supported if included in a high-confidence 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).

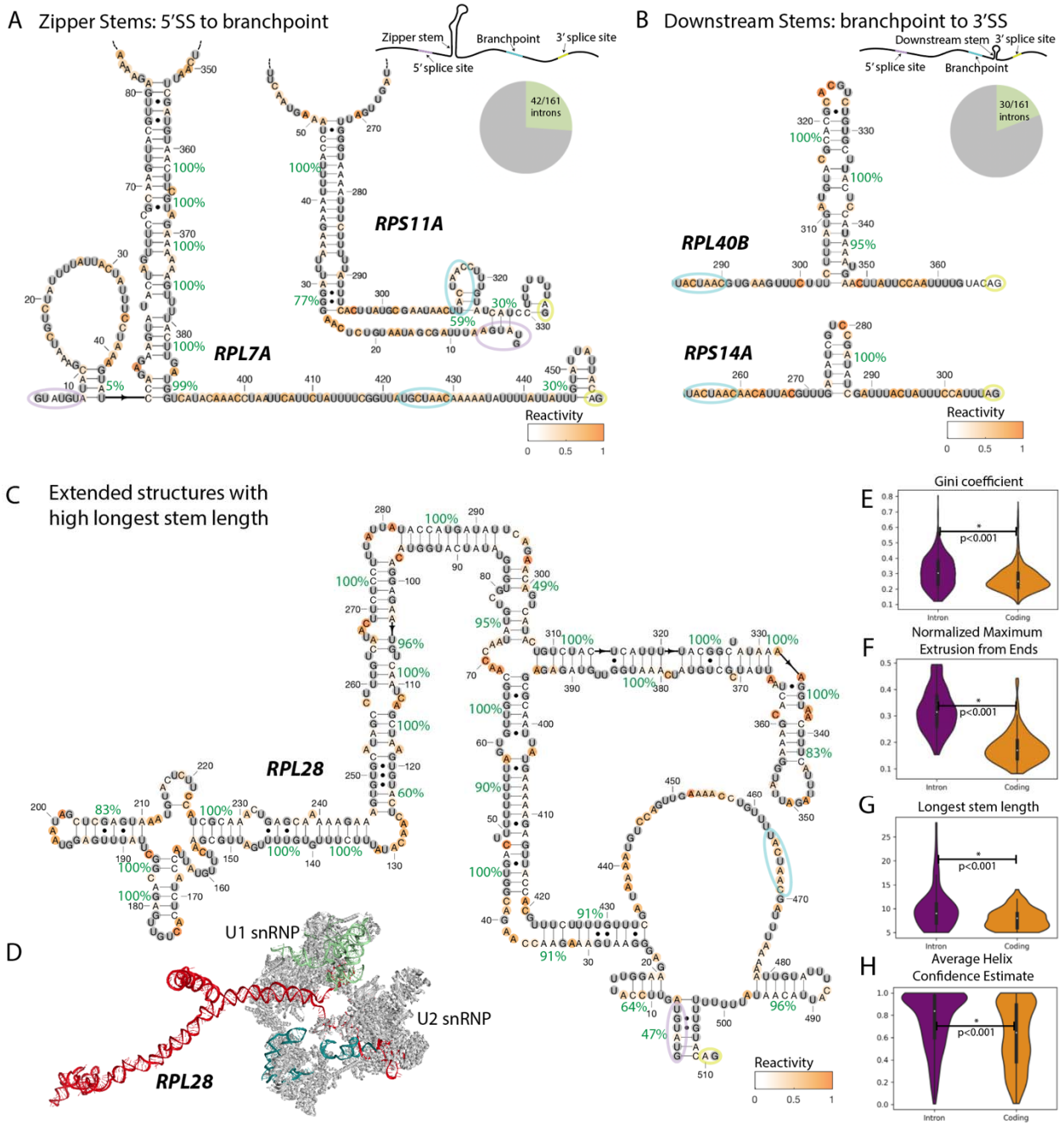


**Figure 4: Structural insights from DMS probing of *S. cerevisiae* introns.** A) Reactivity support for zipper stems in RPL7A and RPS11A. B) Reactivity support for stems connecting the branch point and 3' splice site in RPL40B and RPS14A. C) Secondary structure for RPL28 predicted by RNAstructure guided 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)[58], modeled using the secondary structure derived from DMS-MaPseq. E)-H) Comparisons of secondary structure features between introns and coding regions. P-values were computed using Wilcoxon ranked sum tests to compare classes.

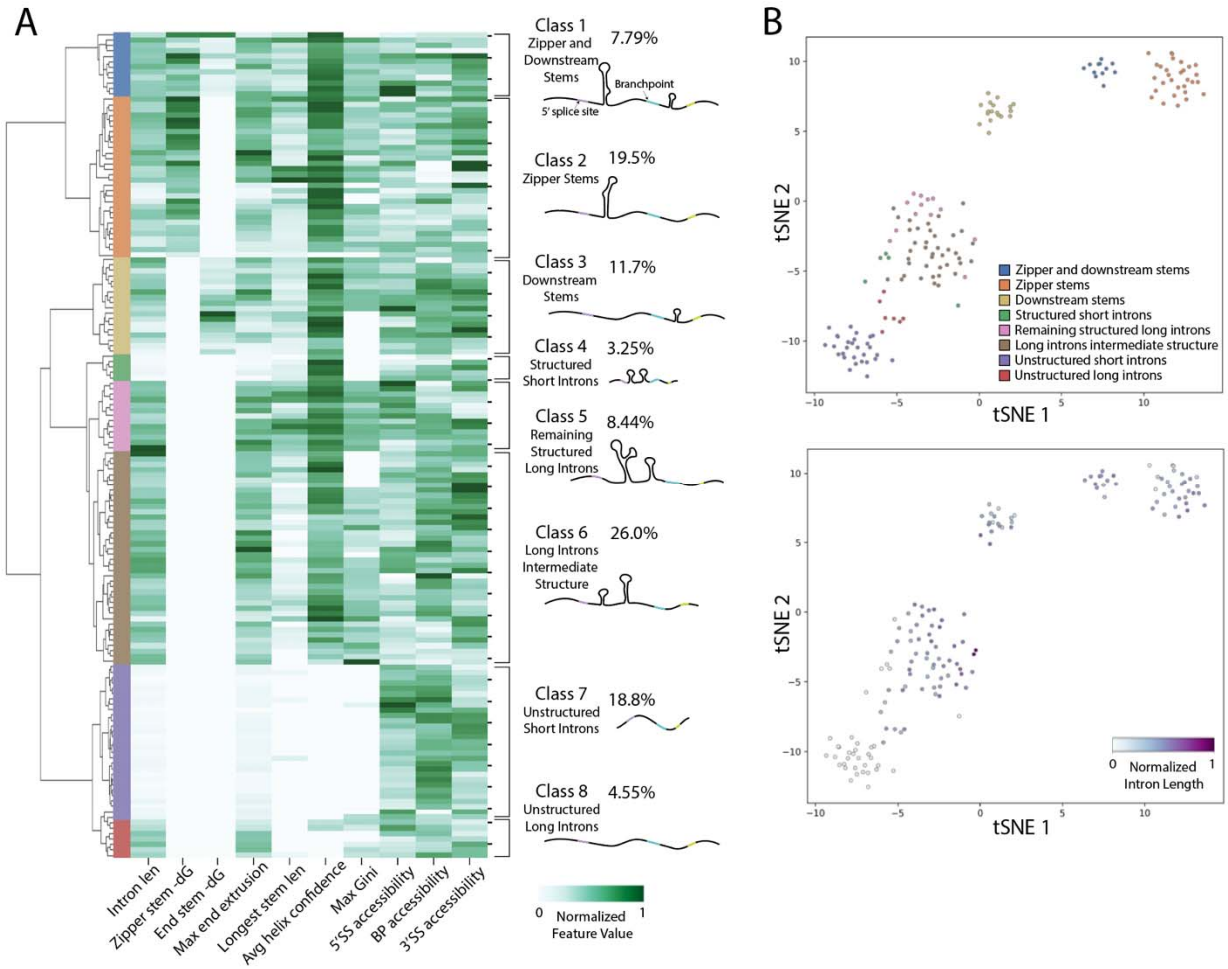


**Figure 5**: **Structural landscape for *S. cerevisiae* introns.** A) Heatmap and dendrogram summarizing intron structural classes, with hierarchical clustering based on secondary structure features. B) tSNE plots for introns, colored by the clusters from hierarchical clustering (top), or the length of the intron normalized to the maximum intron length in *S. cerevisiae* (bottom). 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 and tSNE decomposition.

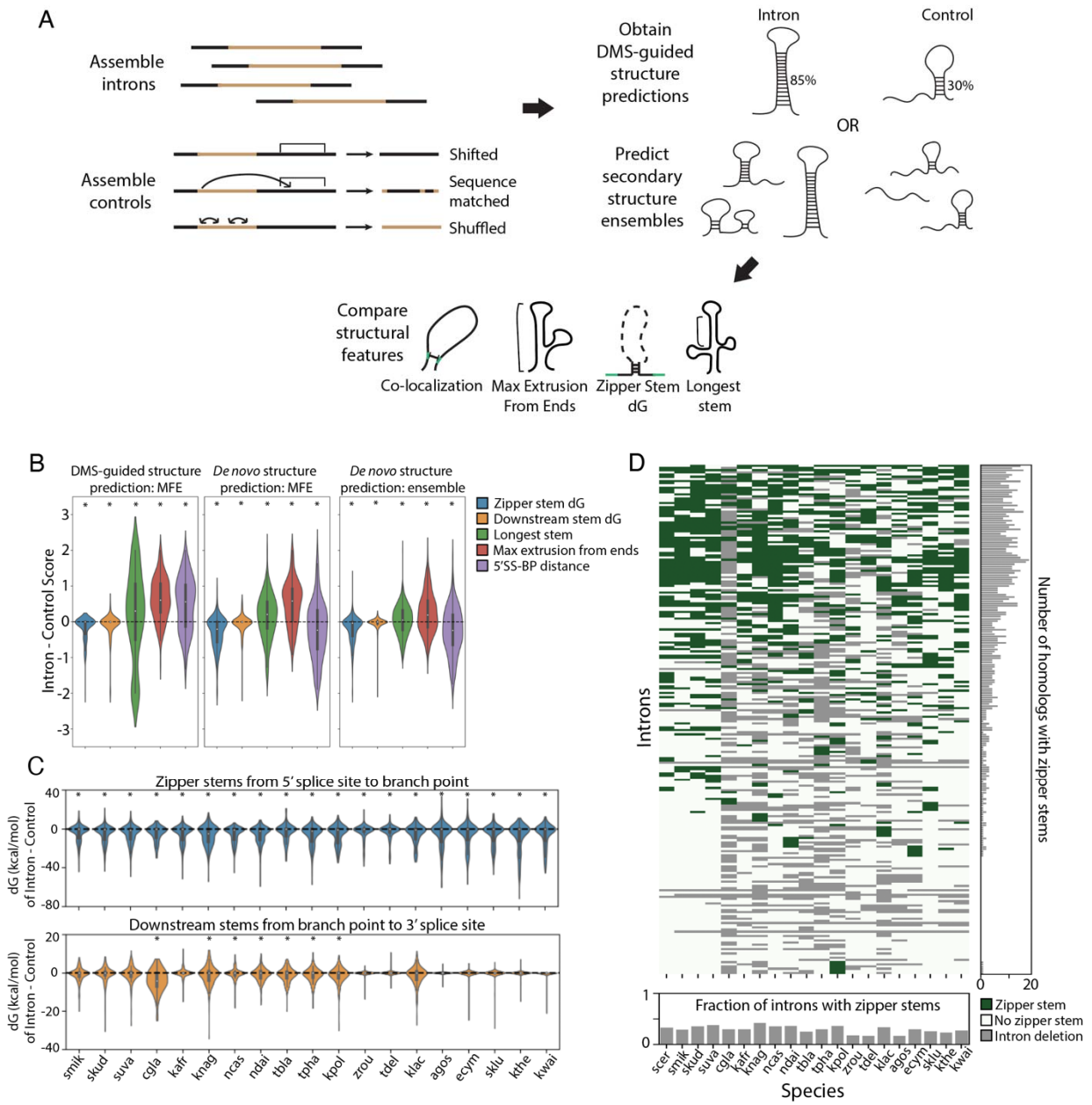


**Figure 6: *De novo* secondary structure feature prediction for *S. cerevisiae* and the *Saccharomyces* genus.** A) Workflow for comparisons between introns' secondary structure ensembles and those of control sequences. B) Comparison of secondary structure feature enrichment between introns and control 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).

# Online Methods

**Strains, media, and growth conditions.** The strain YOH001 was constructed from the background strain JRY8012,[59] which includes three deletions of ABC transporter genes (*prd5::kan*$^r$, *snq2::kan*$^r$, *yor1::kan*$^r$) to reduce drug efflux. YOH001 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 S3). Strains were grown at 30 $^oC$ on YPD plates and in YPD liquid medium. Single colonies of YOH001 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, and 20 mL for the condition without pladB treatment. We prepared a 1 mM stock solution of Pladeinolide B (pladB, Cayman Chemicals) by diluting 100 μg of pladB in 186.3 μL DMSO. We treated cultures with 5 μM pladB using 50 μL of 1 mM pladB for every 10 mL of culture, and we incubated cultures at 30 $^oC$ for 1 hour with shaking. The condition without splicing inhibition was treated with an equal volume of DMSO. 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 $^oC$, 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. For this second wash, cells were separated into 4 1.5 mL Eppendorf tubes for every 10 mL culture and spun down for 2 minutes at 1500g at 4 $^oC$. 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 $^oC$ to reduce RNA fragmentation.

**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 $^oC$ for 1 hour, added 2 μL of 50 mM EDTA, incubated the reactions at 65 $^oC$ 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 $^oC$, 40 minutes at 46 $^oC$, and 1 minute at 95 $^oC$. 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 RPS36B, and RR067 and RR070 for MATa1; Table S3). The reactions were denatured 98 $^oC$ 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 replicate 1, 200 μg RNA for replicate 2, 47.2 μg RNA for the no-modification condition, and 53.2 μg RNA for the condition without splicing inhibition. We pooled 108 50-mer oligos (RR-rRNAdep-1-108; Table S3) 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 and 10X TURBO DNase buffer in a 167 μL reaction, and incubating at 37 $^{o}$C for 30 minutes. We then purified the reaction using an RNA Clean and Concentrator-5 column without size selection, 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 replicate 1, 4.58 μg RNA for replicate 2, 1.64 μg RNA for the no-modification control, and 1.57 μg RNA for the condition without splicing inhibition. 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 S3) 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 S3) 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 $^oC$ for 10 minutes. We then added 2 μL each of ExoCIP A and ExoCIP B (NEB) and incubated for another 10 minutes at 37 $^oC$, followed by 1 minute of heat inactivation at 80 $^oC$. 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 S3). The reactions were denatured at 98 $^oC$ for 30 seconds, followed by 9 cycles of denaturing at 98 $^oC$ for 10 seconds, annealing at 70 $^oC$ for 30 seconds, and extending 72 $^oC$ for 30 seconds, and ending with a final extension period at 72 $^oC$ 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 S3), using the iTaq Universal SYBR Green Supermix (Bio-Rad). Based on this qPCR quantification, we carried out 2 more cycles of PCR for replicates 1 and 2 and the no-modification control, and 5 more cycles for the -pladB control. 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 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 replicate 2 was sequenced with a NovaSeq S4 partial lane, also with paired-end reads of length 150.

**DMS-MaPseq sequencing data analysis.** We obtained 557 million reads for replicate 1, 1.30 billion reads for replicate 2, 375 million reads for the no-modification control, and 169 million reads for the control without splicing inhibition. We used UMI-tools[60] to extract UMI tags from reads by matching to the expected pattern from our RT primers, and we used cutadapt to remove adapter sequences including indexing primer sequences and the cloning linker. We then aligned sequencing reads to various 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.[61] 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[62] using the alignment parameters used in ShapeMapper 2[63]: --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[64]. Alignments were sorted and indexed using samtools. Reads with matching mapped positions and UMI tags were then deduplicated using the UMI-tools[60] dedup function with default parameters for paired-end reads.

Mutational frequencies, coverage values, and normalized reactivities were obtained by processing alignment files using RNAframework[65] 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. 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 used annotations for the ribosomal protein-coding genes from Hooks, et al. 2014.[41] We excluded eight snoRNA-containing introns from further analysis,[31] 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)[21] 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[66] using alignment parameters stated in Zubradt, et al. (2017)[21]: -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)[21].

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)[49]. We used the high-resolution *S. cerevisiae* ribosome structure with PDB ID 4V88,[67] 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[68] 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)[21].

**DMS-guided structure prediction and validation.** For introns meeting the coverage requirement of 3942 as discussed above, we performed secondary structure prediction by RNAstructure[24] guided by DMS reactivity with 1000 bootstrapping iterations without pseudoknot predictions, using the package Biers[29] 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)[67], and DSSR[68] 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)[69], and the U1 snRNA secondary structure was obtained from Li, et. al. (2017)[70]. Four tRNA ground truth structures were obtained from Rfam-derived secondary structures.[71] Finally, secondary structures for mRNA segments in HAC1, ASH1, RPS28B, and SFT2 were obtained from Zubradt, et al. (2017)[21]. We did not include RNA's with known pseudoknots in the control set (for instance, the RNase P RNA[72]), 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.[73] 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[26] 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[27] which allows for pseudoknot prediction, guiding predictions with DMS data in 100 bootstrapping iterations with default parameters. We additionally made predictions using DREEM[28] for the following eight intron regions that included predicted structures and had 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 S3). 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 S3), 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 S3), 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 S3) 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.[29] 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[63] 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[29] 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[24] 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.[74] Structure from Hooks, et al. 2016[30] 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[75] 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[19] with thresholds as follows. We obtained all complete and chromosome-level assemblies of genomes in the Ascomycota phylum from NCBI GenBank[76] (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[19] in flanked mode with a noncoding threshold of 50. We then used R-scape[32] to predict covarying residues with a 0.05 E-value threshold, and we generated structures with CaCoFold.[75] 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[26]. The set of snoRNA-containing introns was obtained from the *S. cerevisiae* snoRNA database[31], and snoRNA gene coordinates were obtained from the Saccharomyces Genome Database.[77]

**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.[38] As test constructs, we used variants of the intron in RPS17B since a zipper stem in this sequence has been characterized.[12] 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 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).[58] 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.[36] 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 various 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-complexspliceosome 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[78] as implemented in Vienna 2.0[26]. 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. 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 networkx package[79] 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 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. 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. 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)[58], 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.[80] 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.[38] 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 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.

***De novo* computational structure prediction and structure metrics.** We predicted secondary structure ensembles for *S. cerevisiae* introns and various 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/arnie) to call structure prediction executables from Vienna 2.0[26] and RNAstructure[24] 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.[41] 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.