NAR Genomics and Bioinformatics, 2024, **6**, lqae143
https://doi.org/10.1093/nargab/lqae143
Advance access publication date: 24 October 2024
Standard Article

Intronic RNA secondary structural information captured for the human *MYC* pre-mRNA

Taylor O. Eich[1], Collin A. O'Leary[1,2] and Walter N. Moss [1,*]

[1]Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA
[2]Current Address: Department of Biology and Chemistry, Cornell College, Mount Vernon, IA 52314, USA

[*]To whom correspondence should be addressed. Tel: +1 515 294 6214; Email: wmoss@iastate.edu

Abstract

To address the lack of intronic reads in secondary structure probing data for the human *MYC* pre-mRNA, we developed a method that combines spliceosomal inhibition with RNA probing and sequencing. Here, the SIRP-seq method was applied to study the secondary structure of human *MYC* RNAs by chemically probing HeLa cells with dimethyl sulfate in the presence of the small molecule spliceosome inhibitor pladienolide B. Pladienolide B binds to the SF3B complex of the spliceosome to inhibit intron removal during splicing, resulting in retained intronic sequences. This method was used to increase the read coverage over intronic regions of *MYC*. The purpose for increasing coverage across introns was to generate complete reactivity profiles for intronic sequences via the DMS-MaPseq approach. Notably, depth was sufficient for analysis by the program DRACO, which was able to deduce distinct reactivity profiles and predict multiple secondary structural conformations as well as their suggested stoichiometric abundances. The results presented here provide a new method for intronic RNA secondary structural analyses, as well as specific structural insights relevant to *MYC* RNA splicing regulation and therapeutic targeting.

Introduction

The *c-MYC* (*MYC*) gene is master regulator of transcription and is classified as an oncogene due to its dysregulation in over 70% of cancers, providing motivation to focus on the development of therapeutics that target *MYC* (1–3). *MYC* has three constitutive exons separated by two introns that are efficiently spliced out during the maturation of the precursor mRNA (pre-mRNA). RNA secondary structures provide a regulatory role in RNA splicing by promoting contacts for RNA–RNA and RNA–protein interactions to facilitate splicing (4). Secondary structures present within the intronic regions and at the exon-intron junctions regulate splicing through splicing regulatory elements (SREs). SREs may enhance or silence splicing by recruiting or blocking trans-activating factors, such as RNA-binding proteins (RBPs) that primarily recognize single-stranded RNA (5). Due the dynamic and transient nature of RNA structure formation and the variety of RBPs capable of binding to RNA (6), multiple RNA isoforms can be generated as the result of alternative splicing. Alternative splicing occurs in approximately 95% of human genes and is a key step in mRNA maturation and regulation (7). Dysfunctional splicing is prevalent in many human diseases, including cancer, neurodegenerative disorders, immune and infectious diseases, and metabolic conditions (8). Although the *MYC* gene is not known to undergo alternative splicing, the *MYC* protein is highly disordered and has been famously termed 'undruggable' (9). *MYC* and other hard-to-target genes have been under intense study for the development of small molecule drugs and antisense oligonucleotides (ASO) that target RNA structure rather than the protein (6,10,11). To exploit the full potential of RNA structures as drug targets, more accurate models are needed to improve the specificity of RNA targeting therapeutics (10,12). Thus, the expansion of the structural probing data across the intronic regions of genes would facilitate basic research into splicing by making structural data more accessible, which could benefit in the development of targeted splicing inhibitors against intronic RNA structures.

To investigate the RNA secondary structures associated with splicing regulation, methods for interrogating intronic RNA structures are needed (13). RNA structure prediction has benefited enormously through the advent of Next-Generation Sequencing (NGS), as scientific advancements applied to sequencing methods have increased the accuracy of predicted models (14). A challenge remains for high-throughput sequencing (HTS) experiments to obtain coverage for intronic sequences in the context of their splice sites, since the exon-intron junctions are transiently formed and separated during splicing (15,16). The spliceosomal inhibition with RNA probing and sequencing (SIRP-seq) method aims to address this challenge for the *MYC* gene, by stimulating intron accumulation through inhibition of the spliceosome. The biochemical approach at the heart of SIRP-seq is treatment of human cells with the small molecule spliceosome inhibitor pladienolide B. Although the exact molecular mechanism is still under investigation, pladienolide B targets the SF3B1 complex of the spliceosome to disrupt splicing (17,18). Pladienolide B is thought to occupy a hinged pocket that forms in the open state of the SF3B1 subunit of the U2 snRNP, which competes with the branchpoint adenosine and U2 snRNA duplex preventing its transition to a closed state (18). The result of spliceosome

Received: July 1, 2024. **Revised:** September 6, 2024. **Editorial Decision:** October 2, 2024. **Accepted:** October 4, 2024

inhibition with pladienolide B is global intron retention, defective pre-mRNA maturation, and dysregulation of crucial cellular processes (19). However, the predominant effect of SF3B1 inhibition was determined to be exon-skipping, especially in genes containing significantly longer downstream exons than their respective upstream exons (20).

Thermodynamic principles of nucleic acid base pairing underpin the algorithms of many RNA secondary structure prediction programs, allowing for high-throughput analysis (21). The two main bioinformatic programs used in this study for predicting the pre-mRNA secondary structures are RNAfold and ScanFold. The RNAfold program, from the ViennaRNA 2.0 package (22), uses an algorithm that incorporates thermodynamic parameters for free energy and enthalpy change called the 'Turner Rules', derived from optical melting experiments on small model systems (23,24). The RNA folding algorithm implemented in RNAfold is used in the RNA structure discovery program ScanFold (25). ScanFold identifies structures ordered through evolution to contain high thermodynamic stability. It works by calculating a minimum free energy (MFE) structure for every frame in a 120-nucleotide scanning window (25,26). The ΔG° value for each MFE structure is then used to calculate a z-score by comparing the native ΔG° of the MFE structure to the average ΔG° from a distribution of shuffled sequences with an identical nucleotide composition, which is then normalized by the standard deviation of all MFE values. The z-score reflects the number of standard deviations the native MFE is from the shuffled MFE average, which is an estimate of unusual sequence-ordered RNA stability—a potentially evolved property for maintain structure/function.

ScanFold can also incorporate experimentally derived reactivity information as pseudo-energy constraints that are either rewarding or penalizing during RNAfold predictions, depending on the nucleotide reactivity. Reactivity values are provided by RNAframework (27)—a bioinformatics toolkit for analyzing NGS-based mutational profiling data and was applied to DMS-MaPseq experiments. A high reactivity indicates a nucleotide was frequently unpaired and exposed to methylation by dimethyl sulfate (DMS), a chemical probe for single stranded RNA. In contrast, a low reactivity indicates a nucleotide was frequently base paired. RNAframework data are also compatible with the program DRACO (28), a highly complementary analysis tool for identifying heterogeneous conformations from probed RNA transcripts. An additional check for conservation of structure through co-variational analysis was completed using cm-builder (29), an automated pipeline that combines Infernal (30) and R-Scape (31–33).

The major impetus for the development of the SIRP-seq method came from our previous RNA secondary structural analyses of the *MYC* gene. ScanFold previously identified a number of evolutionarily conserved low-z-score structures, which were incorporated into dual luciferase reporter assays to identify functional elements within intron 2 (34) and the 3′UTR of *MYC* (35). To validate identified structures and enrich our knowledge of *MYC* RNA structure, we attempted to use a targeted version of DMS-MaPseq (36) to gain DMS reactivity profiles across *MYC* transcripts. However, the resulting data lacked significant read depth across the intronic regions. Thus, SIRP-seq was utilized to fill this gap in the structural landscape of the MYC gene to gain access to the structural reactivities that are found within several low z-score ScanFold predictions that span the *MYC* introns and exon-intron junctions.

Materials and methods

Cell culture

HeLa cells (ISU Hybridoma Facility) were maintained at 37°C in 5% CO_2 and allowed to reach 70–100% confluency in a 10-cm plate. During passage, old media was removed and 3.0 ml of DPBS was used to gently wash the cells before a ∼3-min incubation with 1.5 ml of 0.025% trypsin to dislodge adherent cells from the plate. Trypsin digestion was quenched by adding at least 4.5 ml of pre-warmed DMEM, supplemented with 10% FBS, 100 U/μg per ml penicillin/streptomycin, and 2 mM L-glutamine (Gibco). Cells were routinely passed at a 1:10 ratio until passage 7, then cells were used for plating. During plating, cells were dislodged from the plate and counted using a hemocytometer. Cells were plated into two standard 24-well treated plates at a seeding density of ∼125000 cells/well, where replicate samples were plated into neighboring wells and allowed to incubate 2 days prior to treatment.

Spliceosome inhibition with pladienolide B

An outline for the SIRP-seq method is given in Figure 1, where intronic RNA structure probing is illustrated for the *MYC* pre-mRNA (Figure 1A). Cell confluency was recorded for each well (90–100%) before treatment. A stock solution of 1 mM pladienolide B was diluted to a working solution (100×) in DMSO, then diluted to 350 nM (1×) into DMEM (Figure 1B). Media was prepared for the untreated samples (0-h) and treated samples (2-h and 4-h). DMSO was used as a vehicle control in the 0-h timepoint instead of pladienolide B. Cells incubated in prepared media for up to 4-h, at 37°C in 5% CO_2.

Dimethyl sulfate (DMS) probing of HeLa cells

After the pladienolide B incubation, treatment media was discarded, and replicates were simultaneously probed with DMS following a slightly modified version of the DMS-MaPseq protocol (36,37). Briefly, DMS preferentially methylates at specific sites to the Watson-Crick face of unpaired adenines and cytosines. Methylation at the Watson–Crick face interferes with base pair interactions to induce a mutation during reverse transcription, where a thermostable group II intron reverse transcriptase (TGIRT) can be utilized to promote mutational readthrough to avoid truncations. Mutational readthrough allows for the readout of multiple methylation sites on a single complementary DNA (cDNA). Mutations in the cDNA correspond to reactive bases during sequencing alignment, allowing for the high-throughput collection of experimentally obtained reactivity information from sequencing datasets. The 2% (v/v) DMS solution was prepared by carefully transferring a concentrated stock solution of DMS (>99%) into a prepared solution of 25% ethanol in 75% DPBS and carefully vortexed. The prepared 2% DMS solution was added directly to cells for a ∼1 min incubation, followed by immediate quenching with two aliquots of 0.053M DTT/DPBS solution (5-times molar excess of DTT is required to quench residual DMS). Before adding the second aliquot of quenching solution, the quenched media was removed. Immediately

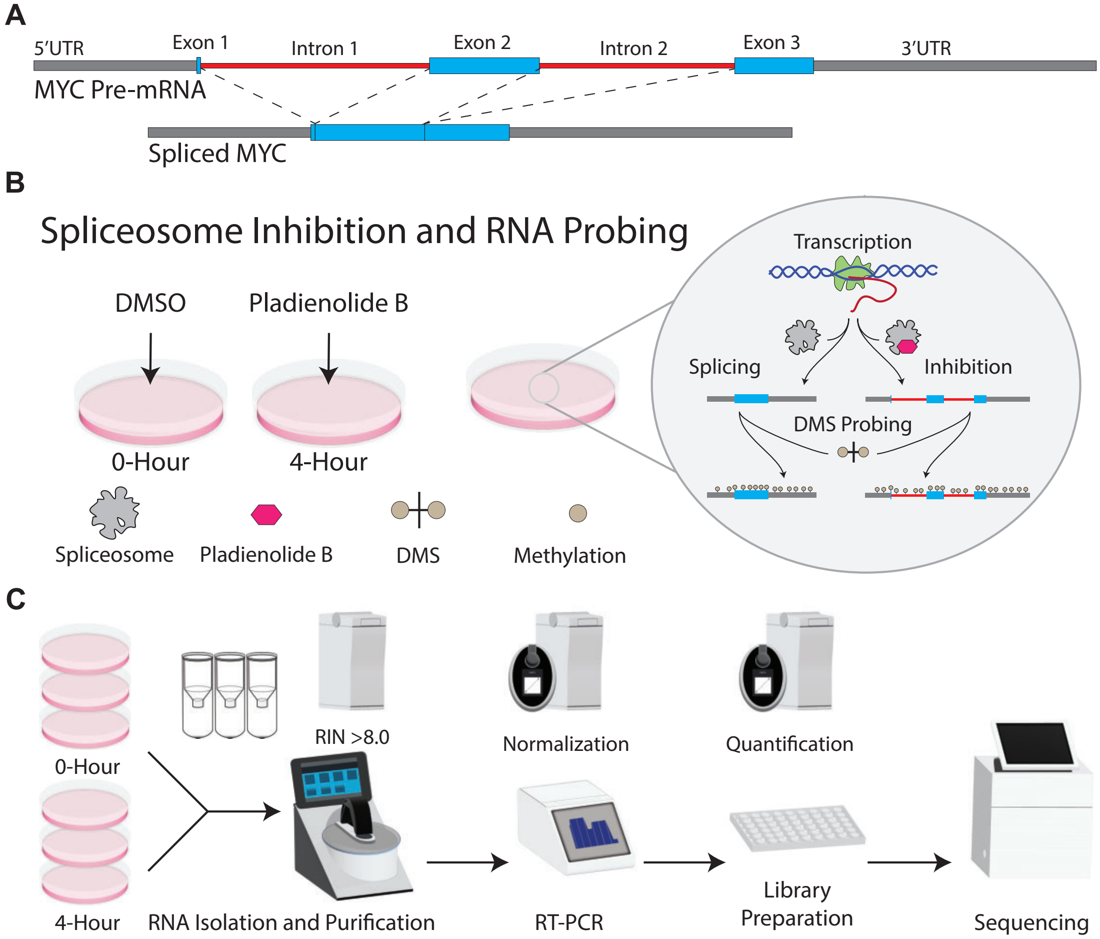


Figure 1. Overview of the SIRP-seq method for targeted DMS-MaPseq of intronic regions of genes, applied here to study the introns of *MYC.* (**A**) Cartoon of the pre-mRNA of MYC, containing a 5′UTR, three exons, two introns, and a 3′UTR, which is constitutively spliced into mRNA without undergoing alternative splicing. (**B**) Overview of pladienolide B treatment and the inhibition of the spliceosome, where untreated (DMSO) samples undergo splicing but treated (pladienolide B) samples experience spliceosome inhibition and accumulation of MYC with retained introns. (**C**) The samples are processed as described in methods, in sections: ‘Isolation and purification of total RNA probed with DMS’, ‘Reverse transcription’, ‘PCR validation of intron retention’,’ Library processing and sequencing’. Quality control steps are shown for ‘Normalization’ and ‘Quantification’, indicated by icons for a 2100 Agilent Bioanalyzer and a Qubit® 2.0 Fluorometer.

following removal of the second quenching solution, TRIzol reagent was used to extract the probed RNA, and was stored overnight at −20°C.

Isolation and purification of total RNA probed with DMS

Stored RNA samples were thawed on ice before processing. A Direct-zol Plus RNA mini-prep kit protocol (Zymo) was used to isolate the DMS probed RNA from the TRIzol reagent. Following the manufacturer’s protocol, samples were prepared by adding equal volumes of 100% ethanol, mixed, and transferred into the provided mini-prep columns. Samples were spun, DNase treated, and washed as described in the protocol. A spectrum was collected using a NanoDrop One (ThermoFisher) to measure the concentration of the total eluted RNA. Total RNA was stored in the −20°C overnight before further RNA integrity analysis. An Agilent 2100 Bioanalyzer was used to determine an RNA integrity number (RIN) using the RNA nano protocol for total eukaryotic RNA. All samples that were used in downstream processing achieved RIN scores greater than 8.0 (Supplementary File S1). All quality control steps with the Agilent 2100 bioanalyzer and qubit® 2.0 fluorometer were completed at the DNA Facility at Iowa State University.

Reverse transcription

Following the DMS-MaPseq protocol (37), A 25 μl reaction for each replicate was run using 550 ng of input RNA, the TGIRT enzyme, RNase-out, dNTPs, DTT and a 1:1 ratio of random hexamer primers to poly-dT primers. The resulting reactions containing the cDNA were diluted to 55 ul with sterile deionized H_2O. The final concentration of cDNA in each sample was obtained using the ssDNA kit for a Qubit® 2.0 Fluorometer before normalizing samples to ∼9 ng/μl prior to PCR amplification.

PCR validation of intron retention

The cDNA was used for the targeted amplification of *MYC* pre-mRNA, where tiled primer sets were designed to amplify products with at least a 50 nt overlap, including products that spanned across the exon-intron junctions and intronic fragments. The primers used for the amplification of introns were designed with primerBLAST (38). A 20 ul Q5 (NEB) reaction was prepared for each primer set. During optimization, several GC-rich target amplicons required GC enhancer and variable cycling conditions. See Supplementary File S2 for a detailed breakdown of PCR completed, including primer sequences used (Supplementary Table S1), suggested cycling conditions used for optimizations (Supplementary Table S2), and gel imaging results (Supplementary Tables S3–S4 and Supplementary Figures S1–S4). Validated PCR products were processed with the Zymo Research Clean and Concentrate kit before measuring the total amount of dsDNA using the DNA high sensitivity kit on the Qubit® 2.0 Fluorometer. Quantified products were pooled by combining an equal amount (w/v) of each product into a single pool for each replicate, where the total input DNA for each pool of amplified cDNA from replicates ranged from 20 to 94 ng.

Library processing and sequencing

The Illumina DNA library preparation protocol was used to prepare nine libraries. During preliminary work, three replicate libraries for each of the timepoints (0-h, 2-h and 4-h) were prepared. Adapter sequences were ligated to fragmented sequences during the library tagmentation step. Using the number of cycles determined through quantification of the normalized pool of PCR products, the tagmented DNA was PCR amplified to generate the complete libraries for each sample pool. After completing the clean-up step, libraries were quantified using high sensitivity protocols via Qubit® 2.0 Fluorometer and an average fragment size was determined using the DNA high sensitivity chip for the Agilent 2100 bioanalyzer (Supplementary File S1). Sequencing was completed in the Moss laboratory, where paired end reads (150 × 150 nt) were obtained as FASTQ files on the iSeq100 benchtop sequencer (Illumina). A quality control step was completed for the FASTQ files using FASTQC, which provides a quality score distribution. The average quality per read exceeded a score of 30 in all FASTQ files, indicating a very low error rate in read determination (<0.01%).

Sequencing data analysis with RNAframework

The processed paired-end reads were then trimmed using Cutadapt to remove the adapter sequences and low-quality reads. The nextera transposase adapter sequence used for trimming was 'CTGTCTCTTATA'. RNAframework extracts reactivity information by aligning sequences containing the sites of mutation to a reference sequence. Using the reference sequence for *MYC*, the rf-index module was used to generate a reference index via Bowtie2-build utility. Trimmed fastq files were mapped to the index using the rf-map module. Output alignments were sorted and converted automatically into BAM format via Bowtie2. The BAM files were processed as individual and merged datasets. Merged datasets were used to represent the data for the 0-h and 4-h timepoints to generate global reactivity profiles for the *MYC* pre-mRNA.The data processed with RNAframework, DRACO, ScanFold, and RNAfold can be found in Supplementary File S3. Once reactivities were pre-processed, the rf-count module creates a mutational map (MM) by calculating the per-base mutational rate, per-base read coverage, and total number of reads covering the transcript and stored as an RNA Count (RC) file. The RC files were normalized with rf-norm using the Zubradt scoring method for calculating reactivities (37), called with the '-sm 4' flag, following a 2–8% normalization scheme, called with the '-nm 1' flag. The 2–8% method for signal normalization takes the top 10% of the highest reactivity values, then the top 2% of raw reactivities are removed, and then all reactivities are divided by the average value of the remaining 8% (39). The only nucleotides considered during reactivity normalization and reactivity profile generation were adenosines and cytosines, which was called using the '-rb AC' flag. Normalized reactivity data were output into XML format before implementation into the rf-fold module, which uses the ViennaRNA 2.0 package (22) for predicting secondary RNA structures that were output as connectivity table (CT) format. For DRACO analyses, MM files were generated during the rf-count step that were input for alternative structure deconvolution with DRACO, which were output as JSON files. JSON files were processed into RC files using the rf-json2rc module before following the same normalization step as previously described. Normalized DRACO conformational data were output with a text file that indicated the relative stoichiometry of each conformation found within a given window along with an XML file for each conformation. XML files were used to generate reactivity files using a custom python script, which extracts the reactivity information and creates the file that is used for pseudo-energy predictions. XML files were converted into the CT format using the rf-fold module. A more in-depth explanation all the parameters that can be used for processing mutational profiling datasets are available at https://rnaframework-docs.readthedocs.io/en/latest/.

Merging individual datasets for RNAframework and DRACO analysis

Sequenced cDNA libraries were processed with RNAframework to obtain an overall DMS reactivity for each nucleotide in the transcript by merging individual samples into a compiled dataset for each timepoint. The compiled BAM files were used to generate the MM files that contained all the data for heterogeneous RNA conformations for DRACO to deconvolute and parse out distinct reactivity profiles during conformational assignment. This is done by using co-mutational patterns that are collected in a process that uses spectral clustering, which is ideal for identifying the optimal number of heterogeneous RNA conformations from mutational profiling experiments. The clustering method DRACO uses relies on several unpaired bases to be on a single RNA conformation and for the same read to co-mutate multiple times to be assigned to a cluster (28). Overlapping windows with an equal number of clusters are combined before reconstructing reactivity profiles, which will only be generated from sequenced reads containing a depth of coverage that exceeds 5000 reads per nucleotide (28).

DRACO analysis predicts conformationally dynamic regions within introns 1 and 2 of MYC

The DRACO algorithm returned multiple regions where distinct reactivity profiles were deconvoluted from identical regions within the merged dataset. Deconvoluted conformations were compared by calculating a positive predicted value (PPV) and sensitivity value using equations (1) and (2) to identify conformational reactivity profiles that resulted in the prediction of distinct structural conformations (40,41). For example, in some instances DRACO identified multiple reactivity profiles for a given region, however PPV and sensitivity analyses revealed the extracted reactivity profiles resulted in identical structural predictions (Supplementary File S4). The PPV and sensitivity values for the compared structures of the DRACO conformations were collected with a python script 'ppv_sens_batch_draco.py'.

$$PPV = \frac{Consistent\ Base\ Pairs}{Total\ \#\ of\ Referenced\ Base\ Pairs} \quad (1)$$

$$Sensitivity = \frac{Consistent\ Base\ Pairs}{Total\ \#\ of\ Experimental\ Base\ Pairs} \quad (2)$$

Conservation and covariation analysis

While checking the sequence conservation across the intronic regions using the Integrative Genome Viewer (IGV), conserved elements that appear in MYC were compared to the 4-h DMS informed ScanFold base pair (BP) tract on a multiple alignment of 20 species from annotation data provided by IGV for the UCSC genome browser (42). ScanFold −2 *z*-score motifs from introns were less conserved than −2 structures in the untranslated regions and exons of MYC, where the opposite was true for −1 motifs. Exonic sequences were almost entirely conserved, while intronic sequences were only partially conserved (Supplementary Figure S5). Covariation analysis required the use of Infernal and R-Scape. Infernal is a program that performs an RNA homology search that builds covariance models using an input multiple sequence alignment, offering increased speed through the implementation of hidden Markov model methods for its homology search and alignment methods (30). R-Scape is a program that runs a test to determine the statistical significance of covariation in structure as it relates to its conservation in an alignment. Together they provide the current 'gold standard' in accessing propensity for an RNA secondary structure to be conserved through evolution of sequence to maintain function. These two tools have been combined into a program called cm-builder to automate their functions (29). To assess the conservation of structural elements predicted by ScanFold, we completed a check for covariation in sequence using a curated database (Supplementary File S5) and utilization of stage *1B* from the Cobretti method (43), which is another program used for automating the process of covariational analyses using motifs output from the ScanFold pipeline. The database was constructed by utilizing the NCBI E-utilities to collect all known genomic sequences for the *MYC* gene, which were then manually curated for the alignment by removing transposable elements and selecting sequence ranges that included approximately 100 nucleotides past exon-intron junctions. *MYC* sequences were filtered and selected from an exhaustive list of reported sequences for all the homologs for *MYC* containing genes, resulting in a finalized list of 576 vertebrate species. A multiple sequence alignment was generated for each intron with the Multiple Alignment using Fast Fourier Transform (MAFFT), which narrowed down the number of aligned sequences to 122 total sequences for intron 1 and 130 total sequences for intron 2 (Supplementary File S5).

Manual curation of alignments used for covariation analyses

Some of the reference sequences accessed from the NCBI nucleotide database for *MYC*, such as the East Asian finless porpoise (Accession: NW_020172779.1), contained transposable elements that required removal prior to their alignment. The manual curation of NW_020172779.1 involved the truncation of sequence surrounding an assembly gap located in intron 1 and the removal of a transposed element containing repetitive sequence within intron 2. The process of removing transposable elements from alignments can be accomplished with the help of guides (44), but often require some preexisting knowledge in bioinformatics to complete. Curation of the sequence for NW_020172779.1 improved the power of statistically significant base pairing detected in an exceptionally low *z*-score ScanFold motif (Motif 42) that was accessed with Infernal and R-Scape.

File reformatting for figure generation

CT files output from rf-fold were converted to dot-bracket-notation (DBN) files. XML files containing reactivity information were converted into REACT files, in which all non-applicable (NA) base reactivities were formatted to −999 prior to ScanFold and RNAfold analyses. All NA readings changed to −999 were replaced with 0.0 and all values >1.0 were set to a maximum value of 1.0 to increase the contrast of moderate reactivities compared to highly reactive nucleotides with values exceeding 1.0. This step was done prior to figure generation VARNA 2D RNA modeling software (45). All nucleotide coordinates were renumbered for the start of each model, which each have a starting coordinate of '1'. The renumbered coordinates used for annotations, including binding sites predicted with RBPmap and known eCLIP binding sites are available in Supplementary File S6 and Supplementary File S7, respectively. Similarly, the transcript coordinates of ScanFold motifs, including locations of covarying pairs detected with R-SCAPE, are listed along with the converted nucleotide positions as they appear in the RNAfold models (Supplementary File S8).

Global RNA secondary structure prediction with RNAfold

RNAfold was selected for modeling of global RNA structure models because it offers compatibility with incorporation of structural reactivities and provides a global refold option for limiting base pairing considerations outside of a given window. When considering the base pairing window used, a maximum base pairing distance of 600 nucleotides was selected, since approximately 99% of base pairs that occur in rRNAs involve pairings less than this distance (39). To generate the RNAfold models, the required module (viennarna/2.5.1) was loaded and sequence folding was completed using the command 'RNAfold –maxBPspan = 600 –shape = Reactivity_filename.react Fasta_filename.fasta'.

Scanning and folding RNA with the ScanFold pipeline

ScanFold can be separated into two steps, a scanning step and a folding step. During the scanning step, a sliding window analysis is performed by dividing the input sequence into multiple overlapping windows with the selected size of 120 nucleotides. For each window, a four metrics are calculated (i) a MFE ΔG°, which reflects the most stable base pairing arrangement for the given sequence using RNAfold; (ii) ΔG° *z*-score, which offers a measure of if the MFE ΔG° is more stable than expected by comparison to shuffled versions for the sequence; (iii) a *P*-value for the ΔG° *z*-score, which is calculated as a fraction of the random ΔG° values that are more stable than the native sequence; (iv) an ensemble diversity, which suggests the propensity for a sequence to adopt multiple distinct or similar secondary structures within the ensemble of potential structures.

During the folding step, the metrics from the scanning step are compiled and all the stable base pairs are listed along with the averaged metrics from each nucleotide of the input sequence. Then a consensus model is built for the entire input sequence by selecting the most ordered/stable arrangement for each nucleotide. This is done algorithmically by selection of the nucleotide that appears most often and with the lowest z-scores than the nucleotide that is being compared. ScanFold was completed for each of the individual replicate and merged datasets using the required modules (python/3.6.5-fwk5uaj, py-biopython/1.70-py3-wos466g) and the command '$python …/path-to ScanFold.py Fasta_name.fasta –out_name Designated_Out_filename –react Reactivity_filename.react –name Fasta_name.fasta –global_refold'. PPV and sensitivity comparisons were completed for the global refold structures output by ScanFold with the python script 'ct_sensitivity_PPV.py'.

Mapping RBP binding sites to RNA with RBPmap

RBPmap is a computational tool that offers an accurate prediction for mapping putative RBP binding sites to input RNA sequence (46). The database used by RBPmap contains experimentally defined protein binding motifs that have been retrieved from the literature scored from a Position Specific Scoring Matrix (PSSM). The algorithm used for motif mapping is based on a Weighted-Rank approach that was originally developed for the SFmap webserver (47). The algorithm used for mapping takes into consideration the likeliness for a motif to cluster and the overall frequency of conservation in a regulatory region. The output is a list of RBPs, including positional information corresponding to the binding motifs selected and target sequences used for predicting their likeliness to bind.

Annotation data from eCLIP experiments from ENCORE datasets

The ENCODE Consortium and the ENCODE production laboratories associated with the ENCORE dataset were responsible for the generation of data relating to the eCLIP annotation data used in this study (48–50). Furthermore, the eCLIP annotation data obtained from the ENCORE project shows RNA-binding interactions that were experimentally derived from HepG2 and K562 cell lines. A potential limitation to the use of the eCLIP data for this analysis is that HeLa cells may have differential protein expression compared to the reference cell lines. In consideration of the differences in the cell lines used, the eCLIP annotations as they appear in the structure predictions here are only serving as a mode of hypothesis generation for future studies.

G-quadruplex prediction with G4Hunter

The G4Hunter webserver was utilized to determine the propensity for G-quadruplex formation by incorporating G-richness and G-skewness within a sequence to output a propensity score (51). The sequences predicted to form G-quadruplexes within the MYC pre-mRNA sequence were output as a BED file (Supplementary File S9) and compared to potential RBP-binders predicted with RBPmap using IGV (Supplementary Figure S6).

Results and discussion

Pladienolide B increases sequencing coverage across the introns of MYC

The addition of the spliceosome inhibitor pladienolide B resulted in an increase in the amount of detectable *MYC* pre-mRNA (Figure 2). For the merged samples, treated samples returned a higher total number of reads than the untreated samples. In the untreated samples, only partial amplification was achieved for targets that spanned exon-intron junctions (Supplementary Figure S7). The increases in the total number of reads reflect an increased abundance of retained introns, as sequences from retained introns that spanned across all the exon-intron junctions were represented in the treated samples. The 0-h, 2-h and 4-h timepoints were processed together and only the 4-h timepoint was used for structure modeling, as ScanFold predictions were highly similar across timepoints (Supplementary File S9), and the 4-h timepoint contained more coverage across introns compared to the 2-h timepoint (Supplementary Figure S8). The 2-h timepoint was used as an intermediate timepoint to measure the comparable effect on splicing inhibition over time (Supplementary File S2), however due to reduced coverage in sequencing, this timepoint was omitted from the more in-depth analyses. Combined, these results show that treatment with pladienolide B increases the abundance of *MYC* pre-mRNA, which expands the current dataset to include regions that span exon-intron junctions and allow ScanFold and DRACO to be applied to other genes to begin predicting RNA structures for the current 'dark' regions of pre-mRNA.

Secondary structural modeling of MYC introns

DMS reactivities were introduced into the ScanFold folding algorithm as constraints to identify and model motifs sequence-ordered (likely functional) secondary structure with support from experimental probing data. In total, ScanFold predicted ~70 total structures for the *MYC* pre-mRNA with evidence of sequence-ordered thermodynamic stability, where 21 structural motifs had −2 z-scores or less (Figure 2) indicating highly ordered structures. Nucleotide resolution DMS reactivities were processed with RNAframework and conformationally dynamic regions were deconvoluted with DRACO and aligned to the MYC pre-mRNA transcript (Figure 2A), along with base pair arc diagrams for ScanFold motifs and per-nucleotide z-scores for the merged 0-h (Figure 2B) and the merged 4-h timepoints (Figure 2C). Incorporation of DMS reactivities did not significantly impact the overall number of motifs predicted by ScanFold (see Supplementary File S9). Both the 0-h and 4-h timepoint provided DRACO with adequate read depth within the exons to detect multiple conformations within the ensemble of structures. A more detailed overview of the reactivity data from the merged 4-h timepoints can be found in Supplementary Figure S9.

Overall, 35.8% of the windows ScanFold predicts for the *MYC* pre-mRNA are predicted to contain ordered secondary structure with z-scores less than −1, while 11.7% are predicted to be exceptionally stable with z-scores less than −2. This highlights the robust nature of ScanFold predictions, which identify local regions of ordered thermodynamically

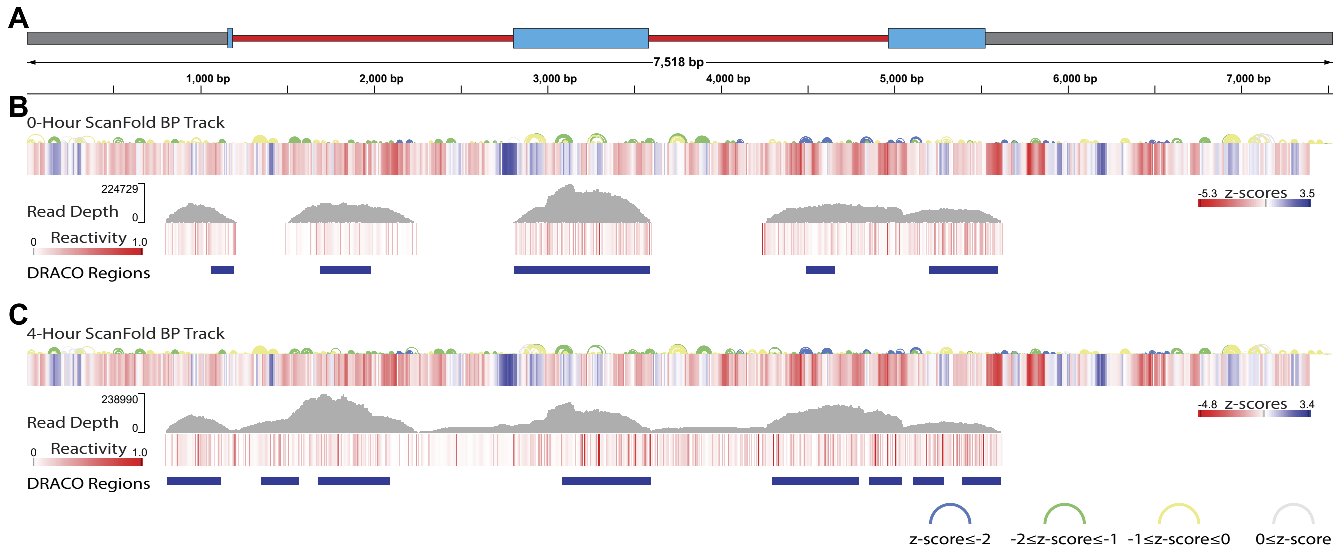


Figure 2. Data summary of the sequencing results for the *MYC* transcript. (**A**) Cartoon representation of MYC pre-mRNA and scale bar. The processed sequencing data was analyzed with RNAframework, DRACO, and ScanFold for reactivity generation, conformational region prediction, and local motif discovery across the targeted portions of MYC pre-mRNA transcript. Tracks are given for (**B**) Merged 0-h timepoints and (**C**) Merged 4-h timepoints. Each timepoint has a ScanFold BP arc diagram for the informed predictions, a heatmap of per nucleotide *z*-scores, and a corresponding *z*-score scale bar, positional max read depth and coverage data for mapped sequences, DMS-MaPseq reactivity data with a heatmap scale bar set from 0 (non-reactive) to 1.0 (reactive) and DRACO regions are displayed as blue bars.

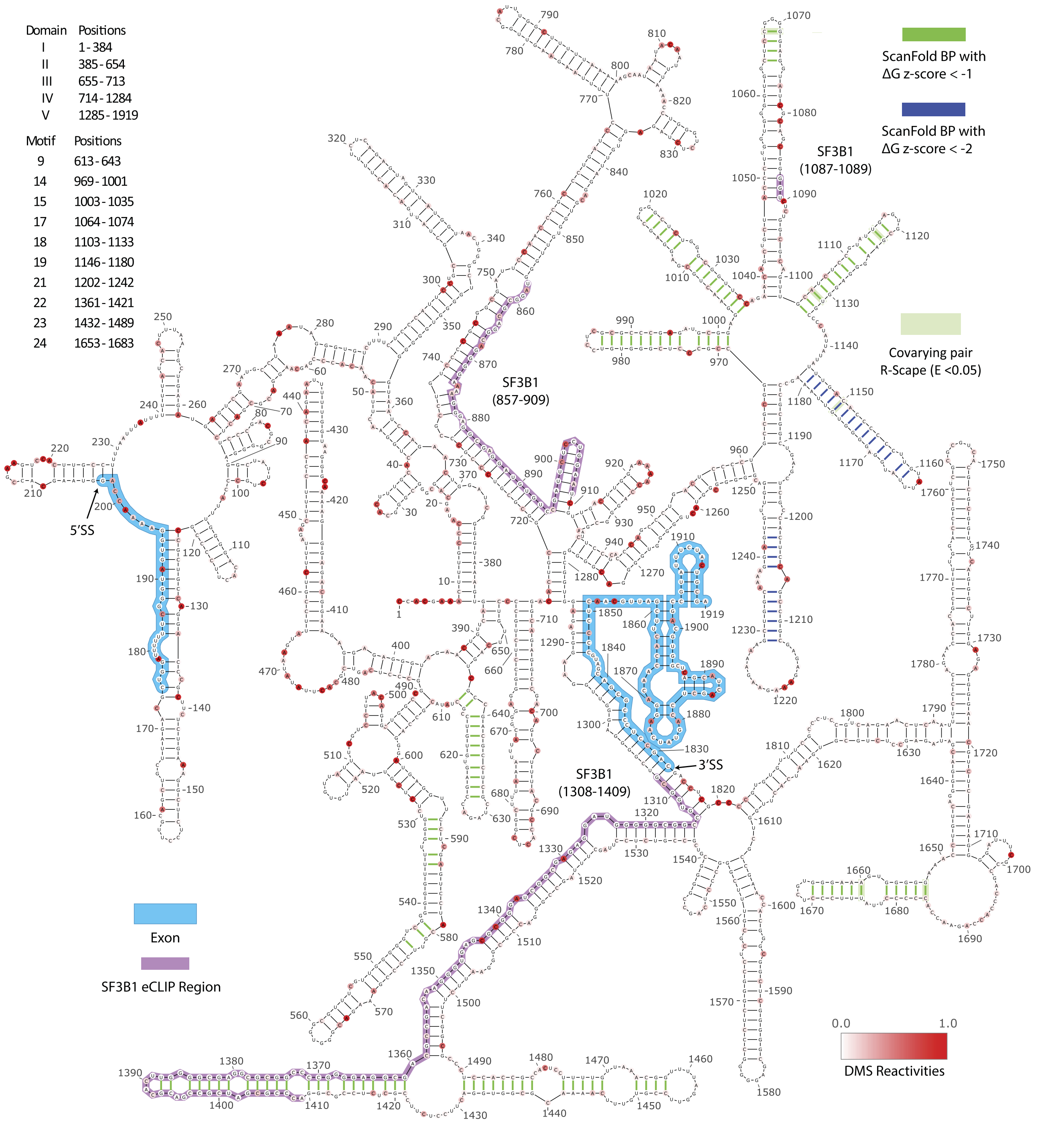
stable RNA secondary structure within a given RNA sequence, and provided a comprehensive overview of the thermodynamic stability of the *MYC* pre-RNA. RNAfold was then used to predict secondary structure models that consider longer range base pairing (beyond the 120 nucleotide ScanFold window). This was done using a maximum base pairing span of 600 nucleotides to enforce a reasonable domain structure, while allowing for longer-range base pairing (39). In addition to the constraints on pairing distance, we included the experimentally derived DMS reactivities as constraints to reflect the in-cell secondary structure of the RNA. The model structures for *MYC* intron 1 and intron 2 can be visualized in Figures 3 and 4, respectively, which include the flanking sequences base paired to the intronic sequences as predicted by RNAfold. Reflecting the robust nature of the consensus low *z*-score motifs deduced by ScanFold, such motifs were preserved within the larger RNAfold models.

Annotating structural models with RBPmap predictions and eCLIP data

Several mechanisms are proposed for how RNA secondary structure could be influencing splicing (52,53). A recurring theme, however, is the ability of secondary structure to mediate the accessibility of RNA to splicing regulatory RBPs. To assess how our SIRP-seq informed secondary structure models could be affecting RBP interactions, putative binding sites were predicted using the RBPmap tool (46). Here, experimentally defined consensus binding sequences were used to scan for matches in the *MYC* pre-mRNA using the highest stringency level (to eliminate false positives) with the additional application of a conservation filter to weed out interactions without evolutionary support. Without any application of the high stringency and conservation filters, 132 potential binders are predicted with 17 723 sites across the *MYC* pre-mRNA. If we limit our consideration to filtered interactions for RBPs with known functions in regulating splicing, such as hnRNPs, SR proteins, etc. the list reduces to 29 potential binders with 7255 individual sites (Supplementary File S6). Additionally, enhanced cross-linking and immunoprecipitation (eCLIP) experiments accessed via the ENCODE database (54,55) were used to highlight experimentally supported binding sites that overlayed with the RBPmap predictions at key features within the RNAfold models (Supplementary File S7). The exact roles of the suggested RBPs, however, may be context-dependent and even serve opposing roles depending on how splicing is being modulated; however, some interesting RBP sites were predicted, which may be playing roles in splicing regulation.

Structural domains of intron 1

Figure 3 shows a global model for a region of the *MYC* pre-mRNA that spans across both exon-intron junctions of intron 1 from transcript coordinates 988 to 2906. Within the global model for intron 1, structures were characterized into five structural domains. The global model begins at sequence found within the 5′UTR, which was predicted to base pair with intronic sequence from intron 1, and a multibranch loop forms from the sequences of the 5′ UTR, exon 1, and intron 1 (Figure 3; Domain I). The 5′ splice site (SS) was predicted at the base of the fifth branch in between loops from exonic and intronic sequences and nearly 200 nucleotides from other structural domains. Additionally, a binding motif for SRSF9 was predicted (Figure 3; Positions 202 to 206; Supplementary File S6) that span across the 5′SS at the base of the stem, which has been described in the regulation of splice site selection in other pre-mRNAs (56,57). The stem could destabilize to become single stranded if SRSRF9 associates with the splice site, which could allow for increased accessibility for 5′SS selection. Furthermore, the constitutive splicing of *MYC* may be related to the formation of strong splice sites, as alternatively spliced cassette exons were found to have weaker splice sites than those constitutively spliced (58).

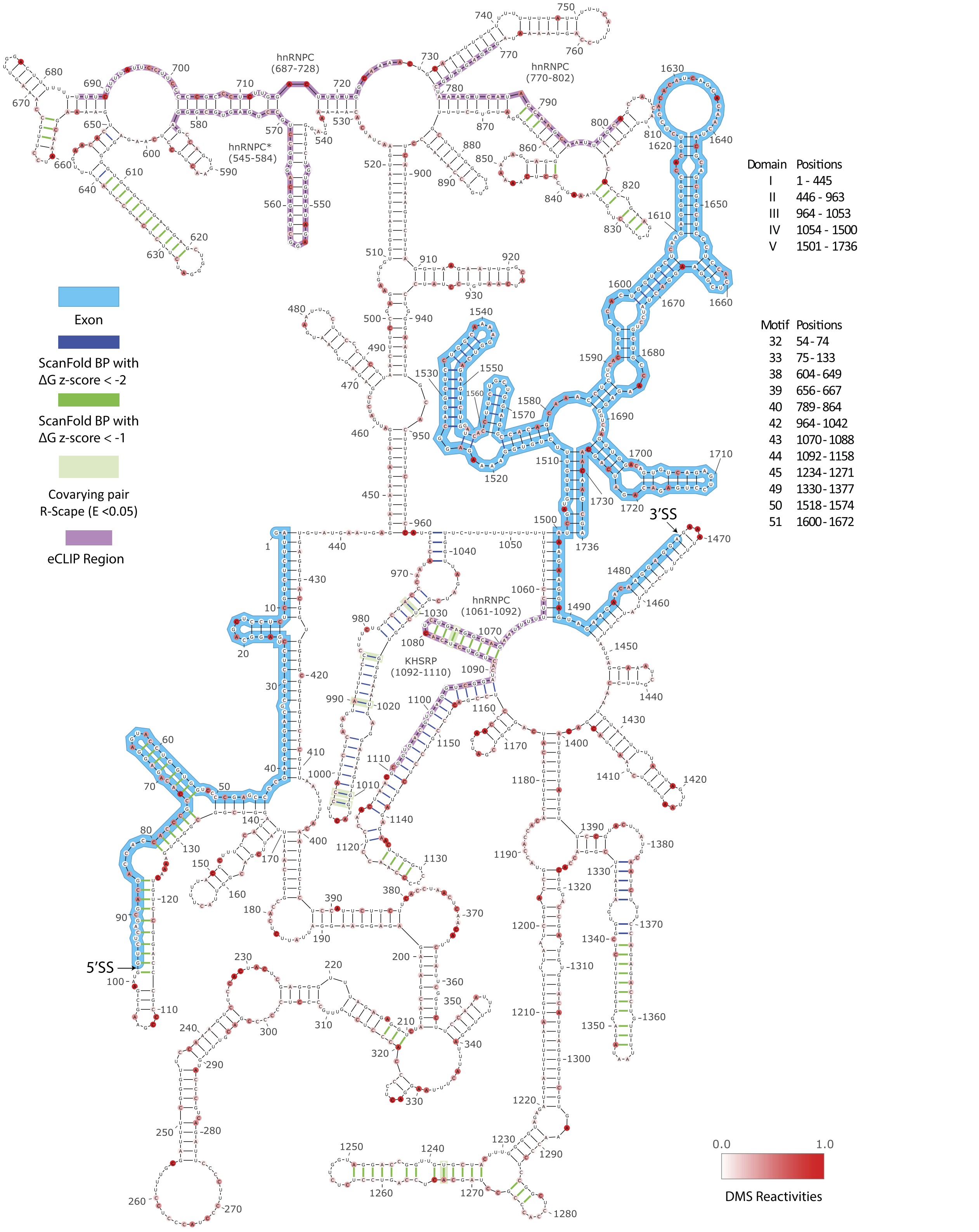


Domain	Positions
I	1 - 384
II	385 - 654
III	655 - 713
IV	714 - 1284
V	1285 - 1919

Motif	Positions
9	613 - 643
14	969 - 1001
15	1003 - 1035
17	1064 - 1074
18	1103 - 1133
19	1146 - 1180
21	1202 - 1242
22	1361 - 1421
23	1432 - 1489
24	1653 - 1683

Figure 3. Secondary structure model for MYC intron 1 predicted with RNAfold using DMS reactivities and 600 bp folding window. The model includes highlighted portions of exon 1 and exon 2 (light blue), the splice sites marked with arrows, annotations for eCLIP binding data are shown (purple), base pairs co-predicted with ScanFold from either −1 (green) or −2 (blue) *z*-score motifs, and statistically significant covarying base pairs (emphasized green bars). Normalized DMS reactivities are presented on a scale from 0.0 to 1.0, the most reactive bases were set to a maximum value of 1.0. The start and end of major structural domains and motifs are given as a region of nucleotide positions in reference to the model beginning at nucleotide 1, which corresponds with transcript coordinate 988.

The GC content of intron 1 is 59% and RNA secondary structures stabilize with increased GC pairing, which may impact splicing by affecting splice site usage (59). RNA secondary structures can extend beyond the Watson-Crick base pairing interactions to include more complex topologies, including RNA G-quadruplexes that may form in the presence of repeated poly-G regions. In Domain II, several stretches of poly-G containing motifs are predicted to bind hnRNPF, hnRNPH1, and hnRNPH2 (Positions: 398–406, 410–415, 397–402, 543–549 and 558–563; Supplementary File S6). The strength of the 5′SS of intron 1 may be related to the number of G-runs and the abundance of hnRNPH binding (60). Previous studies found hnRNPF was linked to exon inclusion that served to regulate an epithelial-mesenchymal transition



Domain	Positions
I	1 - 445
II	446 - 963
III	964 - 1053
IV	1054 - 1500
V	1501 - 1736

Motif	Positions
32	54 - 74
33	75 - 133
38	604 - 649
39	656 - 667
40	789 - 864
42	964 - 1042
43	1070 - 1088
44	1092 - 1158
45	1234 - 1271
49	1330 - 1377
50	1518 - 1574
51	1600 - 1672

Figure 4. Secondary structure model for MYC intron 2 predicted with RNAfold using DMS reactivities and 600 base pair folding window. The model includes highlighted portions of exon 2 and exon 3 (light blue), the splice sites marked with arrows, annotations for eCLIP binding data for hnRNPC and KHSRP (purple), base pairs co-predicted with ScanFold from either −1 (green) or −2 (blue) z-score motifs, and statistically significant covarying base pairs (emphasized green bars). Normalized DMS reactivities are presented on a scale from 0.0 to 1.0, the most reactive bases were set to a maximum value of 1.0. A list of the RBPs predicted to bind with RBPmap are shown. The start and end of major structural domains and motifs are given as a region of nucleotide positions in reference to the model beginning at nucleotide 1, which corresponds with transcript coordinate 3490.

(EMT)-associated CD44 isoform switch in a G-quadruplex dependent manner (61). G-quadruplex structures have been investigated for their roles in gene regulation and are considered targets with therapeutic potential in multiple diseases (62). The G-quadruplexes predicted within the poly-G regions of intron 1 (Supplementary Figure S6) may offer a regulatory function in the splicing of *MYC*.

The hairpin in the middle of intron 1 forms a smaller domain and lacks RBP binding sites (Figure 3; Domain III). Hairpins lacking RBP interactions may still provide a role in shortening the distances between the splice sites to modulate splicing. The next domain was enriched with RBPmap predicted binding sites (Figure 3; Domain IV; Supplementary File S7). Here, RNA structure is likely providing a scaffolding role for RBPs and hnRNP complexes to associate (63). Associated RBPs may co-localize RNA motifs for splicing related activity (64), while certain hnRNPs can assemble to form the core of 40S hnRNP particles, capable of liquid-liquid phase separation to promote biomolecular condensate formation, a process that may support RNA compaction in a crowded molecular environment (63). Two branches from a multibranch loop split the domain into two sub-domains, a left sub-domain (Figure 3; Positions: 714–911) and a right sub-domain (Figure 3; Positions: 912–1284). Shown in the left sub-domain, there is eCLIP binding data for SF3B1 within the stem and three RBPmap predicted binding sites for SF1 (Figure 3; Positions: 794–800, 804–810, and 814–820; Supplementary File S6). In the right sub-domain, a short region of eCLIP binding data for SF3B1 is shown and six ScanFold motifs containing four statistically significant co-varying base pairs form within a secondary multibranch loop (Figure 3; motif 17, motif 18, motif 19). The abundance of RBPs predicted to bind within the secondary multibranch loop containing ScanFold motifs may indicate a variety of RNA-protein interactions between RBPs, multi-subunit complexes, and RNA structure are likely responsible for regulating the splicing of intron 1 of *MYC*.

The main stem of the next domain (Figure 3; Domain V) contains base pairing between intron 1 and exon 2, leading to a multibranch loop where the domain is split amongst the branches into two sub-domains, a left sub-domain (Figure 3; Positions: 1285–1536) and a right sub-domain (Figure 3; Positions: 1537–1851). In the left sub-domain, two ScanFold motifs (Figure 3; motif 22, motif 23) formed a hammerhead-like structure. Additionally, eCLIP binding data for SF3B1 begins at nucleotides base paired to the exon-intron junction of the 3′SS and extends into motif 22, which may represent part of the RNA structure the SF3B1 subunit interacts with during branchpoint selection. If this is the case, RNA structure could potentially facilitate the localization of the 5′SS to the 3′SS for exon ligation. On the opposite end of the structure, motif 23 contained RBPmap binding sites for HuR (Figure 3; Positions: 504–519; Supplementary File S6). HuR binding is thought to increase the stabilization of pre-mRNA when bound to both intronic and 3′UTR regions of *MYC* (65,66). RNA structures that form to reduce HuR binding may serve a role in controlling the stability of mature *MYC* mRNA, as the half-life of *MYC* mRNA is relatively short (∼10 min), which may allow for fine tuning of its expression for transcriptional control (67). Disruption of structures that contain HuR binding motifs may have an opposite effect and allow for prolonged *MYC* transcriptional regulation activity. In the right sub-domain, one ScanFold motif (Figure 3; Domain V; motif 24) was supported by two statistically significant covarying base pairs. In the main stem, a site is predicted for SRSF1 (Figure 3; Positions: 1298–1304; Supplementary File S6), which is located near the positions that are predicted to base pair at the 3′SS junction.

Structural domains of intron 2

Figure 4 shows a global model for a region of the *MYC* pre-mRNA that spans across both exon-intron junctions of intron 2, containing nucleotides from 3490 to 5225. Within the global model for intron 2, structures were characterized into five structural domains. The global model of intron 2 begins a stem formed through base pairing between sequence from exon 2 and intron 2 until multibranch loop forms, containing a branch where the 5′SS is located (Figure 4; Domain I). The 5′SS was predicted within a ScanFold motif near an asymmetric bulge (Figure 4; motif 33). The downstream RBPmap site for hnRNPK (Figure 4; Positions 102–109; Supplementary File S6) may indicate that an RBP binds within the loop to destabilize the hairpin and increase the accessibility of the 5′SS. Destabilization of RNA structure may allow conformational switching to occur, revealing motifs that were previously sequestered into structure for RBP recruitment.

The binding of hnRNPC has been demonstrated to occur at locations distant from the splice sites, perhaps to facilitate the co-localization of splice sites and the splicing machinery (68). Previous work has demonstrated a functional role of hnRNPC binding to intron 2 of *MYC*, where it was suggested to affect either splicing fidelity or translational output (34). ScanFold predicts three motifs within low *z*-score regions at the boundaries hnRNPC binding (Figure 4; motif 38, motif 39, motif 40). This may suggest that the sequence is highly ordered to preserve the formation of structures that facilitate hnRNPC recruitment, along with other potentially important RBPs for splicing regulation.

The likely area of hnRNPC binding contains four out of the five eCLIP sites found within *MYC*, along with several sites of overlapping RBPmap predictions (Figure 4; Positions: 682–609; Supplementary File S6) in Domain II. Additionally, a 50-nt windowed Pearson Correlation and a 50-nt window ROC analysis was completed, where reactivity values were compared between replicate samples from the 4-h timepoint (Supplementary Figure S10A). There was low coverage obtained across the hnRNPC binding region, and reactivity values did not correlate between replicates. Since multiple RBPs were predicted to bind to motifs sharing the same sequence, conformational switching caused by RBP associations and destabilization of RNA structure was potentially revealed through the detection of DRACO regions downstream of hnRNPC binding, into the next domain. An additional feature may involve hnRNPC binding and co-localization of the splice sites are structural elements.

Near the middle of intron 2, a ScanFold motif was predicted (Figure 4; Domain III; motif 42). A stable motif in the middle of the intron could facilitate the intron to fold on itself to bring the two splice sites together. This motif was conserved in over 84 species, including humans, and contained five statistically significant covarying base pairs. Additionally, there was a high correlation (>0.9) between the nucleotide reactivity values across replicates within these ScanFold motifs determined through windowed correlation analyses (Supplementary Figure S10B). Previous work identified a functional relationship in a structure that represents motif 42 and connected its function to splicing the splicing of intron 2 (34), where the function of motif 42 seemed to be connected to an asymmetric bulge. When the bulge was deleted, there was an increase in the detectable amount of splicing determined in luciferase reporter assays (34). RBPmap predicts KHSRP to bind in the asymmetrical bulge region of motif 42 (Figure 4; Positions: 976–981; Supplementary File S6). Additionally, eCLIP data for hnRNPC and KHSRP are found in downstream ScanFold motifs in the next structural domain.

The stem of the next domain forms through base pairing between intron 2 to exon 3, which are separated by a multibranch loop (Figure 4; Domain IV) containing two ScanFold motifs. The first motif contained a single statistically significant base pair and eCLIP binding data for hnRNPC (Figure 4; motif 43), while the second motif contained eCLIP binding data for KHSRP (Figure 4; motif 44). KHSRP is an RNA binding protein with various post-transcriptional regulatory roles including modulation of mRNA splicing (69).

A ScanFold motif was predicted within the middle branch of the multibranch loop and contained a single statistically significant base pair (Figure 4; motif 45). Within the last branch of the multibranch loop, intronic sequence containing the poly-Y tract is base paired with sequence from exon 3. Multiple overlapping RBP sites are predicted to bind to sequences within the hairpin that forms containing the 3′SS, showing multiple RBPs are required to stabilize spliceosome machinery to the 3′SS. Consensus sequences for the branch point adenosine selection have been identified using statistical analyses that searched the 3′ end of introns and revealed its prevalence is greater between 15–100 nucleotides upstream from the 3′SS, although the exact distance is gene specific and varies depending on intron length (70,71). The structures predicted within the primary multibranch loop may play a role in recruiting RBPs and/or the assembly of complexes near the 3′SS. Downstream of the 3′SS of exon 3 (Figure 4; Domain V), ScanFold demonstrated robust predictions for two motifs (Figure 4; motif 50, motif 51), as both motifs were recapitulated in their global refolds (see Supplementary File S10).

Reactivities in base pairs

The secondary structure models presented here represent the minimum free energy structure predicted by RNAfold when informed with experimentally acquired DMS reactivities. The DMS reactivity values inform RNA folding algorithms by penalizing base pairing at highly reactive nucleotides. Despite the energy penalty the reactivity values offer to structure folding algorithms, the minimum free energy models predicted multiple reactive bases to base pair. Since RNA is a dynamic molecule capable of adopting multiple structural conformations and structural switching between conformations, nucleotide breathing may occur to allow momentary shifts in base pairing and modification by DMS (72,73). This effect, along with transient RNA-protein interactions that would perturb structure, potentially away from a structure closer to the global minimum free energy state, could contribute to variability across replicates.

Reactivity data across replicates show variation

Conformational shifting creates heterogeneous RNA structures, even for homogeneous sequences. This structural variability, compounded by splicing-induced changes in base pair partners near splice sites, may explain the observed differences in nucleotide reactivity between replicates. When looking at the individual reactivity values across replicates from the same timepoint, reactivities from intronic regions contained high correlation when comparing replicates 2 and 3 from the 0-h timepoint (Supplementary Figure S11A). In contrast, the reactivity values from replicates in the 4-h timepoint show differences in reactivity profiles across the entire transcript (Supplementary Figure S11B). Moderately higher reactivities were obtained in treated samples than in untreated samples, and nucleotides appeared more reactive in single stranded regions near the splice sites. Although it is uncertain the exact cause of the increase in reactivities near the splice sites, the effect may be caused by increased solvent accessibility surrounding the splice sites when the SF3B1 subunit is inhibited, and branch point recruitment is blocked. RNA-protein binding can also contribute to variability between replicates, as multiple RBPs may bind to the same sequence, promoting alternative structural conformations to form. When an RBP becomes associated with an RNA motif, the surrounding RNA structures may destabilize. The collapse of an RNA structure may reveal a previously sequestered RNA motif for the accommodation of a new RBP. Structural rearrangements may promote the release of bound RBPs and allow new structures to form that may reveal SREs for establishing localizations of spliceosomal machinery during splicing.

Stability of Intronic RNA may influence detection

Amplification of products located in the center of intron 1 and a region that crosses the 3′SS junction of intron 2 were also in regions of high thermodynamic stability. However, this observation lacked consistency across all replicates (Supplementary File S2), and may vary based on primers used for targeting. The amplification of intronic regions from the 0-h timepoint was initially confounding, thus, RNA stability was suggested to have an impact on the availability of some intronic sequences. The presence of incompletely processed intronic sequences may also be related to the increased burden to splice all transcripts that are upregulated during *MYC* hyperactivation (19). Additionally, a comprehensive analysis on widespread intron retention was performed in vertebrates (74), revealing that the 3′ end of transcripts experienced intron retention at a higher rate across genes with multiple exons. Lastly, to check that the products containing intronic regions from the 0-h timepoints were obtained from the intronic RNA rather than overlapping non-coding RNAs or active transposons, a search was conducted using Dfam (75) and RNA-Central for genome annotations (76). Results of these analyses supported that the mapped reads were obtained exclusively from sequences from the *MYC* pre-mRNA.

Conclusion

The SIRP-seq method was utilized for biochemical structure probing for intronic regions in the human *MYC* pre-mRNA and may be applied to other genes of interest. The DMS reactivities collected here revealed structural insights into previously unexplored intronic regions. This data not only helped validate local motifs predicted by ScanFold, but also informed our global folding model of the *MYC* pre-mRNA. The global model presents an averaged picture for the *MYC* pre-mRNA, provides a structural context within the exon-intron boundaries, reveals potential SREs, and multiple potential regulatory sites previously predicted and shown to bind RBPs relevant to splicing. We identified multiple regions with evidence of ordered thermodynamic stability, including our most significant finding, motif 42, which contained additional support by evolutionary comparisons that revealed covariation in sequence for five statistically significant base pairs. These results provide leads for additional work to elucidate structure/function relationships of identified secondary structures and interactions that can provide insights into the splicing of *MYC*. Additionally, while this current study focused on *MYC*, the SIRP-seq protocol can be adapted to any spliced RNA target sensitive to pladienolide B treatment. Thus, this work can have broad utility in the study of RNA secondary structure's roles in splicing regulation. It is important to point out, however, that while SIRP-seq offers a powerful tool for probing intronic secondary structures, the use of spliceosomal inhibition, necessary for the technique, perturbs the cellular state. This disruption alters the stoichiometry of important interactions, potentially affecting results. This highlights the importance of using integrative approaches for RNA structure determination, which consider multiple lines of evidence for/against various models.

Data availability

The raw sequencing data is available in the Sequencing Read Archive (SRA) under the BioProject accession PRJNA1122119.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

Van Tompkins: Ordered reagents, provided insight during data interpretation, and maintained facilities.

Warren Rouse: Provided insight with experimental work and during data analysis, maintained facilities.

Jake Peterson: Provided insight during data analysis.

Author contributions: Taylor O. Eich: Carried out all biochemistry experiments and computational analyses, wrote the initial drafts and final draft of the manuscript, and created all figures for the manuscript.

Collin A. O'Leary: Contributed significantly to the design of the biochemical and computational methods used, provided guidance for computational analyses, and revised the manuscript.

Walter N. Moss: Conceptualization of the project, revised the manuscript, provided guidance for computational analyses, and manually curated alignments for covariation analyses.

Funding

National Institute of General Medical Sciences [R01GM133810 to W.N.M.]. Funding for open access charge: Parks Library, Iowa State University.

Conflict of interest statement

None declared.

References

1. Madden,S.K., de Araujo,A.D., Gerhardt,M., Fairlie,D.P. and Mason,J.M. (2021) Taking the Myc out of cancer: toward therapeutic strategies to directly inhibit c-Myc. *Mol. Cancer*, **20**, 3.
2. Chen,H., Liu,H. and Qing,G. (2018) Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduct. Target Ther.*, **3**, 5.
3. Koh,C.M., Sabo,A. and Guccione,E. (2016) Targeting MYC in cancer therapy: RNA processing offers new opportunities. *Bioessays*, **38**, 266–275.
4. Bartys,N., Kierzek,R. and Lisowiec-Wachnicka,J. (2019) The regulation properties of RNA secondary structure in alternative splicing. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1862**, 194401.
5. Hiller,M., Zhang,Z., Backofen,R. and Stamm,S. (2007) Pre-mRNA secondary structures influence exon recognition. *PLoS Genet.*, **3**, e204.
6. Spitale,R.C. and Incarnato,D. (2023) Probing the dynamic RNA structurome and its functions. *Nat. Rev. Genet.*, **24**, 178–196.
7. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
8. Kim,H.K., Pham,M.H.C., Ko,K.S., Rhee,B.D. and Han,J. (2018) Alternative splicing isoforms in health and disease. *Pflugers Arch.*, **470**, 995–1016.
9. Llombart,V. and Mansour,M.R. (2022) Therapeutic targeting of "undruggable" MYC. *EBioMedicine*, **75**, 103756.
10. Falese,J.P., Donlic,A. and Hargrove,A.E. (2021) Targeting RNA with small molecules: from fundamental principles towards the clinic. *Chem. Soc. Rev.*, **50**, 2224–2243.
11. Singh,N.N., Luo,D. and Singh,R.N. (2018) Pre-mRNA splicing modulation by antisense oligonucleotides. *Methods Mol. Biol.*, **1828**, 415–437.
12. Costales,M.G., Childs-Disney,J.L., Haniff,H.S. and Disney,M.D. (2020) How we think about targeting RNA with small molecules. *J. Med. Chem.*, **63**, 8880–8900.
13. Broseus,L. and Ritchie,W. (2020) Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput. Struct. Biotechnol. J.*, **18**, 501–508.
14. Strobel,E.J., Yu,A.M. and Lucks,J.B. (2018) High-throughput determination of RNA structures. *Nat. Rev. Genet.*, **19**, 615–634.
15. Lee,S., Zhang,A.Y., Su,S., Ng,A.P., Holik,A.Z., Asselin-Labat,M.L., Ritchie,M.E. and Law,C.W. (2020) Covering all your bases: incorporating intron signal from RNA-seq data. *NAR Genom. Bioinform.*, **2**, lqaa073.
16. Mishra,A., Siwach,P., Misra,P., Dhiman,S., Pandey,A.K., Srivastava,P. and Jayaram,B. (2021) Intron exon boundary junctions in human genome have in-built unique structural and energetic signals. *Nucleic Acids Res.*, **49**, 2674–2683.
17. Kotake,Y., Sagane,K., Owa,T., Mimori-Kiyosue,Y., Shimizu,H., Uesugi,M., Ishihama,Y., Iwata,M. and Mizui,Y. (2007) Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nat. Chem. Biol.*, **3**, 570–575.
18. Cretu,C., Gee,P., Liu,X., Agrawal,A., Nguyen,T.V., Ghosh,A.K., Cook,A., Jurica,M., Larsen,N.A. and Pena,V. (2021) Structural basis of intron selection by U2 snRNP in the presence of covalent inhibitors. *Nat. Commun.*, **12**, 4491.
19. Hsu,T.Y., Simon,L.M., Neill,N.J., Marcotte,R., Sayad,A., Bland,C.S., Echeverria,G.V., Sun,T., Kurley,S.J., Tyagi,S., *et al.* (2015) The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature*, **525**, 384–388.
20. Wu,G., Fan,L., Edmonson,M.N., Shaw,T., Boggs,K., Easton,J., Rusch,M.C., Webb,T.R., Zhang,J. and Potter,P.M. (2018)

Inhibition of SF3B1 by molecules targeting the spliceosome results in massive aberrant exon skipping. *RNA*, **24**, 1056–1066.

21. Zhang,J., Fei,Y., Sun,L. and Zhang,Q.C. (2022) Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat. Methods*, **19**, 1193–1207.
22. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
23. Turner,D.H. and Mathews,D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, D280–D282.
24. Mittal,A., Turner,D.H. and Mathews,D.H. (2024) NNDB: an expanded database of nearest neighbor parameters for predicting stability of nucleic acid secondary structures. *J. Mol. Biol.*, **436**, 168549.
25. Andrews,R.J., Roche,J. and Moss,W.N. (2018) ScanFold: an approach for genome-wide discovery of local RNA structural elements-applications to Zika virus and HIV. *PeerJ*, **6**, e6136.
26. Zuber,J., Schroeder,S.J., Sun,H., Turner,D.H. and Mathews,D.H. (2022) Nearest neighbor rules for RNA helix folding thermodynamics: improved end effects. *Nucleic Acids Res.*, **50**, 5251–5262.
27. Incarnato,D., Morandi,E., Simon,L.M. and Oliviero,S. (2018) RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic Acids Res.*, **46**, e97.
28. Morandi,E., Manfredonia,I., Simon,L.M., Anselmi,F., van Hemert,M.J., Oliviero,S. and Incarnato,D. (2021) Genome-scale deconvolution of RNA structure ensembles. *Nat. Methods*, **18**, 249–252.
29. Manfredonia,I., Nithin,C., Ponce-Salvatierra,A., Ghosh,P., Wirecki,T.K., Marinus,T., Ogando,N.S., Snijder,E.J., van Hemert,M.J., Bujnicki,J.M., *et al.* (2020) Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.*, **48**, 12436–12452.
30. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
31. Rivas,E. (2020) RNA structure prediction using positive and negative evolutionary information. *PLoS Comput. Biol.*, **16**, e1008387.
32. Rivas,E., Clements,J. and Eddy,S.R. (2020) Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*, **36**, 3072–3076.
33. Rivas,E., Clements,J. and Eddy,S.R. (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, **14**, 45–48.
34. Tompkins,V.S., Xue,Z., Peterson,J.M., Rouse,W.B., O'Leary,C.A. and Moss,W.N. (2024) Identification of MYC intron 2 regions that modulate expression. *PLoS One*, **19**, e0296889.
35. O'Leary,C.A., Andrews,R.J., Tompkins,V.S., Chen,J.L., Childs-Disney,J.L., Disney,M.D. and Moss,W.N. (2019) RNA structural analysis of the MYC mRNA reveals conserved motifs that affect gene expression. *PLoS One*, **14**, e0213758.
36. Tomezsko,P., Swaminathan,H. and Rouskin,S. (2021) DMS-MaPseq for genome-wide or targeted RNA structure probing in vitro and in vivo. *Methods Mol. Biol.*, **2254**, 219–238.
37. Zubradt,M., Gupta,P., Persad,S., Lambowitz,A.M., Weissman,J.S. and Rouskin,S. (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods*, **14**, 75–82.
38. Ye,J., Coulouris,G., Zaretskaya,I., Cutcutache,I., Rozen,S. and Madden,T.L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinf.*, **13**, 134.
39. Low,J.T. and Weeks,K.M. (2010) SHAPE-directed RNA secondary structure prediction. *Methods*, **52**, 150–158.
40. Mathews,D.H. (2019) How to benchmark RNA secondary structure prediction accuracy. *Methods*, **162-163**, 60–67.
41. O'Leary,C.A., Van Tompkins,S., Rouse,W.B., Nam,G. and Moss,W.N. (2022) Thermodynamic and structural characterization of an EBV infected B-cell lymphoma transcriptome. *NAR Genom. Bioinform.*, **4**, lqac082.
42. Raney,B.J., Barber,G.P., Benet-Pages,A., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Fischer,C., Navarro Gonzalez,J., Hickey,G., *et al.* (2024) The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res.*, **52**, D1082–D1088.
43. Peterson,J.M., O'Leary,C.A., Coppenbarger,E.C., Tompkins,V.S. and Moss,W.N. (2023) Discovery of RNA secondary structural motifs using sequence-ordered thermodynamic stability and comparative sequence analysis. *MethodsX*, **11**, 102275.
44. Goubert,C., Craig,R.J., Bilat,A.F., Peona,V., Vogan,A.A. and Protasio,A.V. (2022) Correction: A beginner's guide to manual curation of transposable elements. *Mob. DNA*, **13**, 15.
45. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
46. Paz,I., Kosti,I., Ares,M. Jr., Cline,M. and Mandel-Gutfreund,Y. (2014) RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **42**, W361–W367.
47. Akerman,M., David-Eden,H., Pinter,R.Y. and Mandel-Gutfreund,Y. (2009) A computational approach for genome-wide mapping of splicing factor binding sites. *Genome Biol.*, **10**, R30.
48. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
49. Luo,Y., Hitz,B.C., Gabdank,I., Hilton,J.A., Kagda,M.S., Lam,B., Myers,Z., Sud,P., Jou,J., Lin,K., *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.*, **48**, D882–D889.
50. Hitz,B.C., Jin-Wook,L., Jolanki,O., Kagda,M.S., Graham,K., Sud,P., Gabdank,I., Strattan,J.S., Sloan,C.A., Dreszer,T., *et al.* (2023) The ENCODE Uniform Analysis Pipelines. bioRxiv doi: https://doi.org/10.1101/2023.04.04.535623, 06 April 2023, preprint: not peer reviewed.
51. Brazda,V., Kolomaznik,J., Lysek,J., Bartas,M., Fojta,M., Stastny,J. and Mergny,J.L. (2019) G4Hunter web application: A web server for G-quadruplex prediction. *Bioinformatics*, **35**, 3493–3495.
52. Taylor,K. and Sobczak,K. (2020) Intrinsic regulatory role of RNA structural arrangement in alternative splicing control. *Int. J. Mol. Sci.*, **21**, 5161.
53. Rubtsov,P.M. (2016) [Role of pre-mRNA secondary structures in the regulation of alternative splicing]. *Mol. Biol. (Mosk)*, **50**, 935–943.
54. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K., *et al.* (2018) The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
55. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.Y., Cody,N.A.L., Dominguez,D., *et al.* (2021) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **589**, E5.
56. Cloutier,P., Toutant,J., Shkreta,L., Goekjian,S., Revil,T. and Chabot,B. (2008) Antagonistic effects of the SRp30c protein and cryptic 5′ splice sites on the alternative splicing of the apoptotic regulator Bcl-x. *J. Biol. Chem.*, **283**, 21315–21324.
57. Ha,J., Jang,H., Choi,N., Oh,J., Min,C., Pradella,D., Jung,D.W., Williams,D.R., Park,D., Ghigna,C., *et al.* (2021) SRSF9 regulates cassette exon splicing of caspase-2 by interacting with its downstream exon. *Cells*, **10**, 679.
58. Koren,E., Lev-Maor,G. and Ast,G. (2007) The emergence of alternative 3′ and 5′ splice site exons from constitutive exons. *PLoS Comput. Biol.*, **3**, e95.
59. Zhang,J., Kuo,C.C. and Chen,L. (2011) GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics*, **12**, 90.
60. Xiao,X., Wang,Z., Jang,M., Nutiu,R., Wang,E.T. and Burge,C.B. (2009) Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat. Struct. Mol. Biol.*, **16**, 1094–1100.

61. Huang,H., Zhang,J., Harvey,S.E., Hu,X. and Cheng,C. (2017) RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. *Genes Dev.*, **31**, 2296–2309.
62. Cadoni,E., De Paepe,L., Manicardi,A. and Madder,A. (2021) Beyond small molecules: targeting G-quadruplex structures with oligonucleotides and their analogues. *Nucleic Acids Res.*, **49**, 6638–6659.
63. Domanski,M., Dedic,E., Perez,M.E., Clery,A., Campagne,S., Uldry,A.C., Braga,S., Heller,M., Rabl,J., Afanasyev,P., *et al.* (2022) 40S hnRNP particles are a novel class of nuclear biomolecular condensates. *Nucleic Acids Res.*, **50**, 6300–6312.
64. Das,U., Nguyen,H. and Xie,J. (2019) Transcriptome protection by the expanded family of hnRNPs. *RNA Biol.*, **16**, 155–159.
65. Mukherjee,N., Corcoran,D.L., Nusbaum,J.D., Reid,D.W., Georgiev,S., Hafner,M., Ascano,M. Jr., Tuschl,T., Ohler,U. and Keene,J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, **43**, 327–339.
66. Talwar,S., Jin,J., Carroll,B., Liu,A., Gillespie,M.B. and Palanisamy,V. (2011) Caspase-mediated cleavage of RNA-binding protein HuR regulates c-Myc protein expression after hypoxic stress. *J. Biol. Chem.*, **286**, 32333–32343.
67. Dani,C., Blanchard,J.M., Piechaczyk,M., El Sabouty,S., Marty,L. and Jeanteur,P. (1984) Extreme instability of myc mRNA in normal and transformed human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 7046–7050.
68. Konig,J., Zarnack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D.J., Luscombe,N.M. and Ule,J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.
69. Xu,J., Wang,D., Ma,H., Zhai,X., Huo,Y., Ren,Y., Li,W., Chang,L., Lu,D., Guo,Y., *et al.* (2022) KHSRP combines transcriptional and posttranscriptional mechanisms to regulate monocytic differentiation. *Blood Sci.*, **4**, 103–115.
70. Gao,K., Masuda,A., Matsuura,T. and Ohno,K. (2008) Human branch point consensus sequence is yUnAy. *Nucleic Acids Res.*, **36**, 2257–2267.
71. Harris,N.L. and Senapathy,P. (1990) Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis. *Nucleic Acids Res.*, **18**, 3015–3019.
72. Kenyon,J., Prestwood,L. and Lever,A. (2014) Current perspectives on RNA secondary structure probing. *Biochem. Soc. Trans.*, **42**, 1251–1255.
73. Jones,A.N., Pisignano,G., Pavelitz,T., White,J., Kinisu,M., Forino,N., Albin,D. and Varani,G. (2020) An evolutionarily conserved RNA structure in the functional core of the lincRNA Cyrano. *RNA*, **26**, 1234–1246.
74. Braunschweig,U., Barbosa-Morais,N.L., Pan,Q., Nachman,E.N., Alipanahi,B., Gonatopoulos-Pournatzis,T., Frey,B., Irimia,M. and Blencowe,B.J. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.*, **24**, 1774–1786.
75. Storer,J., Hubley,R., Rosen,J., Wheeler,T.J. and Smit,A.F. (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA*, **12**, 2.
76. RNAcentral Consortium (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.*, **49**, D212–D220.

Received: July 1, 2024. **Revised:** September 6, 2024. **Editorial Decision:** October 2, 2024. **Accepted:** October 4, 2024