Secondary structural analysis of human lncRNAs

Thesis dissertation

Yizhu Lin

Department of Biological Sciences

Carnegie Mellon University

May 2018

Acknowledgements

Being in graduate school for the past six years in CMU has been a precious learning process for me. I have been extremely lucky to be given opportunities to take part in the cutting-edge technologies and science in RNA biology, at the mean time working with and learning from some of the most talented people I have met in life. There are many people that I would like to thank, who have guided me in science and in life, helped me overcoming difficulties, and made my journey of graduate school a joyful and unforgettable one.

First and foremost, I would like to thank my advisor, Dr. Joel McManus. He has been extremely supportive from the very beginning when I joined lab as a rotation student. He is always passionate about the ongoing science in lab, and always has new ideas keeping me busy to test out. This has been great encouragement and inspirations whenever I encountered difficulties and frustrations in research. He is also meticulous and thorough about every data we gathered and every conclusion we drawn. This habit I learned from him has benefits me a lot and will benefit me long term in the future.

My thesis committee guided me through all these years. I would like to thank Dr. John Woolford, Dr. Carl Kingsford and Dr. Andrea Berman for spending their precious time coming to my committee meetings, providing

insightful comments and pointing me into the right directions for my research. They have exhorted me to keep up the hard work and plan out my own career path.

I have been fortunate to work with wonderful colleges in the past few years. I would like to thank Gemma May for keeping the benchwork in lab working productively, for helping me out trouble shooting my experiments, and generating data in high quality and quantity that I have always enjoyed exploring. Thank you to Dr. Pieter Spealman, Dr. Salini Konikkat and Christina Akirtava for giving me support and being my company. Interesting discussions with you have always been great inspiration for me.

I am very grateful to my family and friends, especially to my parents, who I spent too little time with over the past few years. It is their love, care and education that make me who I am today. They are always supportive and give me the freedom and trust to make my own decisions for my life. Thank you to my friends, both of those who are in US keeping me company, and those who give me support across half the world and hours of time difference. Without them my past six years would not be as enjoyable as it was.

Table of Contents

Abstract

In the past decade, long noncoding RNAs (lncRNAs) have been increasingly recognized as important regulators of gene expression at various levels (1). The human genome encodes thousands of lncRNAs (2), and an increasing number of these lncRNAs have been associated with human diseases (3). lncRNA structures are expected to play essential roles in gene regulatory functions, but our current understanding of them remains limited. Traditional methods for RNA structure determination each has its limitations: biophysical approaches, such as NMR or crystallography, are not feasible for large RNAs which are relatively more flexible; traditional chemical probing methods often focus on small regions of single RNAs (4). To overcome these constraints, we developed a novel method for high-throughput probing of RNA structure using massively parallel sequencing (Mod-seq (5)). Compared to traditional RNA structure probing methods, Mod-seq provides substantial improvements in throughput, allowing rapid and simultaneous probing of the whole transcriptome (5, 6). My thesis work focused on using both experimental methods and computational methods to study the structure of human lncRNAs. I first developed Mod-seeker, an automatic data analysis pipeline for Mod-seq (5, 6). I then focused on studying the structure of lncRNA NEAT1, an essential

component of mammalian nuclear paraspeckles (7, 8). Structure probing and comparative analyses suggest lack of evidence of covariant base-pairs in NEAT1 across mammals. However, a conserved long-range interaction was observed that may contribute to NEAT1's scaffolding function in paraspeckle formation. The experiments described in this thesis suggest that lncRNAs can have conserved cellular functions without maintaining conserved secondary structures, even when they function as structural scaffolds. This work is one of the first attempts to use both chemical probing and computational modelling to study the secondary structure of lncRNAs. The case study of NEAT1 lncRNA structure helps us understand its function in paraspeckle formation and gives insights into the contributions of lncRNA structures towards their functions.

Chapter 1 Introduction

Long non-coding RNAs are a group of RNA molecules that do not encode proteins and are longer than 200 nucleotides. Although lncRNA usually have lower expression levels than mRNA, many are found to be important regulators of gene expression. lncRNA are evolutionarily young and their sequence are often not well-conserved. Like proteins, lncRNA can form secondary and tertiary structures, though lncRNA structures are much more flexible and difficult to study using traditional methods for protein structure determination, such as X-ray crystallography, NMR or Electron Microscopy. Only a few lncRNA have secondary structure models determined by chemical probing approaches. Determining the structure of large, flexible lncRNAs and understanding the function and conservation of lncRNA structures are challenging tasks. In this chapter I first review our current understanding about lncRNA in general, including their expression patterns, possible function mechanisms, evolutionary features and structures. Then I summarize current methods for RNA structure determination. Finally, I introduce lncRNA NEAT1, a scaffolding lncRNA for paraspeckles, which is a good candidate to use as a case study to understand lncRNA structure and its conservation.

1.1 Long non-coding RNAs are important regulators of gene expression

For decades, RNAs were mainly recognized as messengers mediating the transfer of genetic information from DNA to protein. However, as we now have a much more comprehensive annotation of the human genome, we now recognize that a large proportion of transcribed RNAs do not encode proteins. Some of these non-coding RNAs are highly expressed and were relatively well studied, such as ribosomal RNA (rRNA), tRNA, snRNA snoRNA, but many others were newly identified and annotated. Non-coding genes that are longer than 200 nucleotides (nt) are categorized as long non-coding RNA (lncRNA). According to Gencode version 27 release, there are 15,778 lncRNA genes in the human genome, which is more than a quarter of the total number of human genes (2, 9), and comparable to the total number of protein coding genes (19,836, Gencode v27).

Although not coding for proteins, previous research has shown that lncRNA comprise a diverse family of RNAs that can regulate gene expression in multiple ways (10)(Table 1). Some lncRNAs are involved in chromosome regulation. For example, the mammalian Xist (X-inactive specific transcript) gene is located on the X-chromosome and is only expressed on the inactive chromosome. Xist lncRNA spreads across the X chromosome from which it is transcribed, and

mediates the inactivation of that X chromosome (11–13). Another example of a chromosome regulatory lncRNA is HOTAIR (HOX transcript antisense RNA). The HOTAIR gene is located within the HoxC gene cluster in chromosome 12. When transcribed, HOTAIR recruits polycomb repressive complex 2 and silences HoxD genes by regulating their proximal chromatin states (14).

Another class of lncRNAs (antisense lncRNA) regulates transcription by forming duplex or triplexes with other DNA and RNA. For example, the lncRNA ANRIL (antisense non-coding RNA in the INK4 locus) interacts with DNA to regulate transcription (15, 16). In the cytoplasm, lncRNAs can pair with other RNAs and also interact with proteins to stabilize mRNA (TINCR) (17), promote mRNA degradation (1/2-sbsRNAs) (18), up-regulate translation (antisense Uchl1) (19), or inhibit translation (lincRNA-p21) (20, 21).

Still other lncRNAs serve as molecular scaffolds for other RNAs and proteins. Sno-lncRNAs, for instance, have multiple predicted binding sites for Fox family splicing regulatory proteins, suggesting sno-lncRNAs may be involved in alternative splicing regulation by Fox protein sequestration (22). Another example of a scaffolding lncRNA is NEAT1. NEAT1 is not only essential, but is also the seeding component for paraspeckle formation (7, 8, 23, 24). Further studies showed that NEAT1 lncRNA has highly organized spatial composition, likely forming a circular scaffold for other paraspeckle

proteins and RNAs to bind (25, 26), and regulating mRNA editing (27) , retention (28), and protein sequestration (29).

The RNAs discussed above are only a small number of lncRNAs whose functional mechanisms were relatively well-studied. As a newly identified RNA species, our understanding of lncRNA functions is far behind our ability to annotate lncRNAs. A recent study utilized a CRISPRi-based genome-scale screening method to identify functional lncRNA loci in human cell lines (30). They targeted 16,401 lncRNA genes in seven different human cell lines; 499 of these lncRNA loci are identified as functional, as they increase or decrease cell growth. The proportion of lncRNA identified as functional is similar to that of protein-coding genes in some cell types (31). Remarkably, 89% of these lncRNAs only showed phenotypes in one of the seven cell types. It is reasonable to speculate that many other lncRNAs have important functions in other cell types, in different developmental stages, or under alternative growth conditions. Nonetheless, our understanding about lncRNAs' characteristics and functions are still limited and much remains to be learned.

lncRNA	Location	Binding targets	Function
Xist	Nuclear	Chromatin, RBPs	X-chromosome inactivation
HOTAIR	Nuclear	Chromatin, RBPS	Regulate chromatin status
Sno-lncRNA	Nuclear	RBPs	Scaffolding
NEAT1	Nuclear	RBPs	Scaffolding
ANRIL	Nuclear	DNA sequence	Regulate transcription
TINCR	Cytosplasm	mRNA	mRNA degradation
1/2-sbsRNAs	Cytosplasm	mRNA	mRNA degradation
antisense Uchl1	Cytosplasm	mRNA	Regulate translation
lincRNA-p21	Cytosplasm	mRNA	Regulate translation

Table 1. Example of lncRNAs and their functions.

1.2 lncRNA sequences are less conserved than mRNA

The primary sequences of lncRNAs have relatively fast evolutionary rates (32). When using phastCons scores to calculate the nucleotide-level conservation level, lncRNA exons are significantly less conserved than protein-coding exons (9). A recent transcriptome-wide study of 1,898 human lincRNAs (long intergenic non-coding RNAs) in six mammals found that only 80% of them have orthologous transcripts expressed in chimpanzee, 63% in rhesus, 39% in cow, 38% in mouse, and 35% in rat (33). A more recent transcriptome analysis showed that over a thousand human lncRNA have homologs with mammals, and only hundreds beyond mammals (34). For lncRNAs that do have orthologs in other species, the lengths of identified stretches of conserved sequences are also much shorter than those of mRNAs. Compared to mRNAs, lncRNAs undergo frequent rewiring of their exon-intron structure, rapidly losing or gaining sequences (34). Notably, even though the majority of human lncRNAs only have homologs in mammals or in vertebrates, lncRNAs are not unique to vertebrates. Many lncRNA genes are found in other species including *D. melanogaster*, mosquito, bee, some plants and sponges (reviewed in (32)).

Although low sequence conservation is often associated with non-functionality, lncRNAs might be exceptional. As mentioned above, CRISPRi screening in human cell lines identified a proportion of functional lncRNAs, and

many others might be functional but not identified in this study due to high cell-type specificity. The tissue-specific expression pattern of lncRNA is highly conserved among species, with similar levels of regulatory conservation as protein-coding genes (34). Also, different regions of lncRNA genes show different conservation levels. lncRNA promoters are generally more conserved than exons, and almost as conserved as protein-coding gene promoters (35), suggesting conservation of the expression regulation of lncRNA.

The evidence above suggests that lncRNAs are under different selective pressures than mRNAs, and likely to have other forms of conservation other than sequence conservation. Compared to mRNAs, lncRNAs do not need to conserve codon usage or prevent frameshift mutations. It is possible that many lncRNAs are not functional, or their function only rely on short sequences motifs, but the flanking sequences are less important. There are also examples of syntenic positional conservation of lncRNAs, suggesting that transcription through a lncRNA locus is important, but the sequences of the actual lncRNA sequences is less important (36). Secondary or tertiary structural motifs are also considered to be possible constraints of lncRNA conservation (37, 38), which I will discuss in more details in next section.

1.3 RNA structure in lncRNA function and conservation

It is well known that RNA structures are important for the function and conservation of many small noncoding RNAs. One typical example is tRNA, which has three hairpin loops that form the so-called three-leafed clover structure. The three-nucleotide anti-codon, located in the middle loop of tRNA, recognizes the coding sequence in mRNAs. Each tRNA can be charged with its corresponding amino acid, thus allowing the genetic information in mRNAs to be faithfully translated into proteins via ribosomes. tRNA structure is highly conserved across almost all species and is crucial to its function (39). Mostly found in bacteria, riboswitches represent another group of small RNAs with important conserved structures. A riboswitch can switch between two different structural conformations, usually in response to the presence of its ligand, thus regulating the activity of its host mRNA (40). Another group of structured RNAs are ribozymes. Similar to protein enzymes, ribozymes have catalytic activities, which rely on their structure. The most heavily used ribozyme in human cells is ribosomal RNA, whose secondary structures and structural-interactions with ribosomal proteins are now well-characterized and shown to be important for ribosomal function (41–43). Given these examples, it is natural to suspect the same for lncRNA - that RNA structures may play a role in their functions.

Studying the structure of lncRNA is difficult given their large size, typically much lower expression level than mRNA, and relatively flexible structures. Recently, a few studies chemically probed the secondary structure of several lncRNAs (Xist (44), HOTAIR (45), lincRNAp21 (46) and ncSRA (47)), aiming to understand how RNA structure contributes to the function of lncRNAs. These studies generally suggested that the probed lncRNAs are structured and have some level of secondary structure conservation. However, other studies argued there is no statistically significant evidence for structural conservation in these probed lncRNAs (48). A more careful study of lncRNA structure is needed to resolve this controversy regarding function and conservation of lncRNA structures.

1.4 Current methods for studying RNA structures

1.4.1 Traditional RNA structure determination methods

Traditional biophysical methods, such as NMR and X-ray crystallography, have been applied to determine RNA structures. These methods are useful to provide comprehensive, high-resolution structural information on RNA molecules. However, their application is limited to a small number of highly-structured and relatively short RNAs, such as group II introns (49–51), telomerase RNA (52), RNase P (53), and ribosomal RNAs (54). LncRNAs are

generally much less structured and longer, which make crystallization or NMR structure determination almost impossible. Cryo-EM can reveal an ensemble of branching patterns of RNA molecules that is useful to confirm proposed secondary structure models, but often insufficient to reconstruct precise structure models on its own due to limits in spatial resolution (55). Besides, biophysical methods are often time consuming and require extensive effort, making them unsuitable for high-throughput lncRNA structure determination.

Currently, most large RNA structure models are generated by computational predictions. RNA structure prediction methods are usually based on sequence information and minimum free energy models (Mfold (56), RNAstructure (57) ViennaRNA (58, 59) etc.) or partition function models (60, 61) (also included in RNAstructure and ViennaRNA packages) that generate base-pair probability matrices to represent multiple possible structural conformations of an RNA molecule. Other methods are based on finding conserved motifs across species to predict RNA structures (GPRM (62), Pfold (63) etc.). Although easy to perform, computational prediction often results in multiple possible structures that need to be verified and differentiated by experimental methods.

Chemical probing can be used to provide RNA structure information. Small chemical molecules can react with RNA by either cleaving the RNA backbone or covalently modifying RNA bases, so that the reverse transcription of RNA is

blocked (Table 2). By using small chemicals that specifically react with single stranded RNA, followed by reverse transcription and denaturing electrophoresis, we can visualize RNA secondary structure information. Commonly used small chemicals for RNA secondary structure probing include DMS (modifies A and C by methylation), Kethoxal (modifies U) and CMCT (modifies G) (64, 65). Later, a chemical probing method called selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) (66, 67) was developed. SHAPE reagents (e.g. NAI (68) and 1M7 (69)) modify the 2' OH groups of RNA in a base-independent way, such that more flexible RNA backbones are more reactive. This provides more comprehensive information of RNA secondary structure than DMS probing. Furthermore, DMS and the SHAPE reagent NAI can be applied both *in vitro* and *in vivo*, allowing RNA structures to be captured in native conditions.

Unlike DMS and SHAPE probing, which probe RNA secondary structure, hydroxyl radicals (70, 71) induce backbone cleavage according to solvent accessibility in a secondary structure-independent manner. Therefore, hydroxyl radical probing provides a method for RNA tertiary structure determination. Although the mechanism of hydroxyl radical-mediated RNA backbone cleavage is still unclear (72), it is possible that the major product will be similar to that formed on DNA (73): an RNA strand with a 3' or 5' phosphate end at the site of cleavage. The most frequent method for hydroxyl radical generation is by

Fenton-Haber-Weiss chemistry. Hydroxyl radicals species are produced from decomposition of hydrogen peroxide (H_2O_2) catalyzed by Fe(II). In solution probing experiments, Fe(II) is chelated to EDTA to prevent its direct binding to the nucleic acid backbone. A reducing reagent, such as ascorbic acid or dithiothreitol, is also needed to recycle Fe(III), which is generated during reaction to Fe(II). This method can only be applied to *in vitro* RNA probing, since hydroxyl radicals are not cell membrane permeable. Alternatively, hydroxyl radicals can also be generated by synchrotron X-ray radiolysis for *in vivo* RNA tertiary structure probing (70).

Reagent	Modification sites	Structural Probed
DMS	Adenine and Cytosine	Secondary structure
CMCT	Guanine	Secondary structure
Kethoxal	Uracil	Secondary structure
1M7	2'-hydroxyl	Secondary structure and flexibility
NAI	2'-hydroxyl	Secondary structure and flexibility
Hydroxyl radical	Backbone cleavage	Tertiary structure

Table 2. Summary of chemical reagents for RNA structure probing.

1.4.2 High-throughput RNA structure probing by sequencing methods

With the development of chemical and enzymatic structure probing techniques and popularization of high-throughput sequencing, several high-throughput structure probing approaches have been developed. Kertesz et al. (74) developed a method for high-throughput RNA secondary structure measurement, PARS (parallel analysis of RNA structure). In the PARS method, two different enzymes are used to digest RNA *in vitro*. RNase V1 cleaves phosphodiester bonds 3' of double-stranded RNA; S1 nuclease cleaves 3' of single-stranded RNA nucleotides. When digesting RNA, both enzymes leave a 5' phosphate at the cleavage point, which facilitates ligating adapters to the digested RNA. RNA fragments generated from random fragmentation or degradation typically have a 5' hydroxyl instead of a 5' phosphate; thus, RNA fragments cleaved by RNase V1 or S1 nuclease can be enriched. These fragments are subjected to library preparation and deep sequencing. The number of stops caused by enzymatic digestion is counted on each single base, and a PARS score is calculated as $\log_2$(V1/S1). A higher score indicates the nucleotide is more likely to be in a double-stranded conformation. Recently, PARS has been applied to transcriptome-wide human RNA secondary structure analysis (75). However, the PARS method has its limitations: it can only work *in vitro*, and enzyme digestion may alter RNA structure when digesting (76).

There have also been several reports of high-throughput methods for transcriptome-wide RNA secondary structure analysis *in vivo* and *in vitro* using DMS. Ding et al. (77) developed a method called Structure-seq. In their method, after DMS treatment, RNA molecules are reverse transcribed using random hexamers (N6) with adapters. The reverse transcription reaction will stop at one nucleotide before the DMS modification site. The single-stranded cDNA product is then ligated with a 3' single strand DNA linker to generate double stranded DNA library using PCR. By comparing the DMS-treated sample and the negative control, they were able to identify DMS modification sites. Rouskin et al. (78) developed a method called DMS-seq. In this method, random fragmentation is applied to DMS-modified RNA molecules. Fragments with sizes between 60-70 bp are then selected for 3' adapter ligation, following by reverse transcription. Single-stranded cDNA products with sizes between 25-45 bp are selected, circularized, and PCR-amplified for sequencing.

Several other modified SHAPE-based high-throughput probing protocols were later developed. In SHAPE-MaP (79), instead of identifying sites of reverse transcription termination, this method uses a different experimental condition for reverse transcription after 1M7 probing, thus introducing mutations at the modification sites. These mutations can then be identified by mutational profiling (MaP).

Our lab independently and contemporaneously developed a high-throughput method for RNA secondary structure probing: Mod-seq (5, 6). In Mod-seq, RNAs from DMS or SHAPE treated cells are randomly fragmented, ligated to specific 5’ and 3’ adapter oligos, and reverse transcribed. One of the challenges in sequencing-based high-throughput methods are that chemically modified RNAs need to be enriched so only RNA fragments that can provide structural information are sequenced. In Mod-seq, this is achieved by ligating a 5' adapter to RNA fragments before reverse transcription. Because reverse transcription prematurely stops at modification sites, the 5’ adapter sequence is excluded in the cDNA product. For RNA fragments without chemical modification, reverse transcription goes through the whole sequence including the 5’ adapter. The cDNA is then circularized and products containing the 5’ adapter sequences are reduced via subtractive hybridization. The remaining cDNA products, which are enriched for modification caused RT stops, are PCR-amplified for high-throughput sequencing. Since Mod-seq works both *in vivo* and *in vitro*, it can also be used to footprint RNA-binding proteins, allowing researchers to identify binding sites in a massively parallel manner (5, 6).

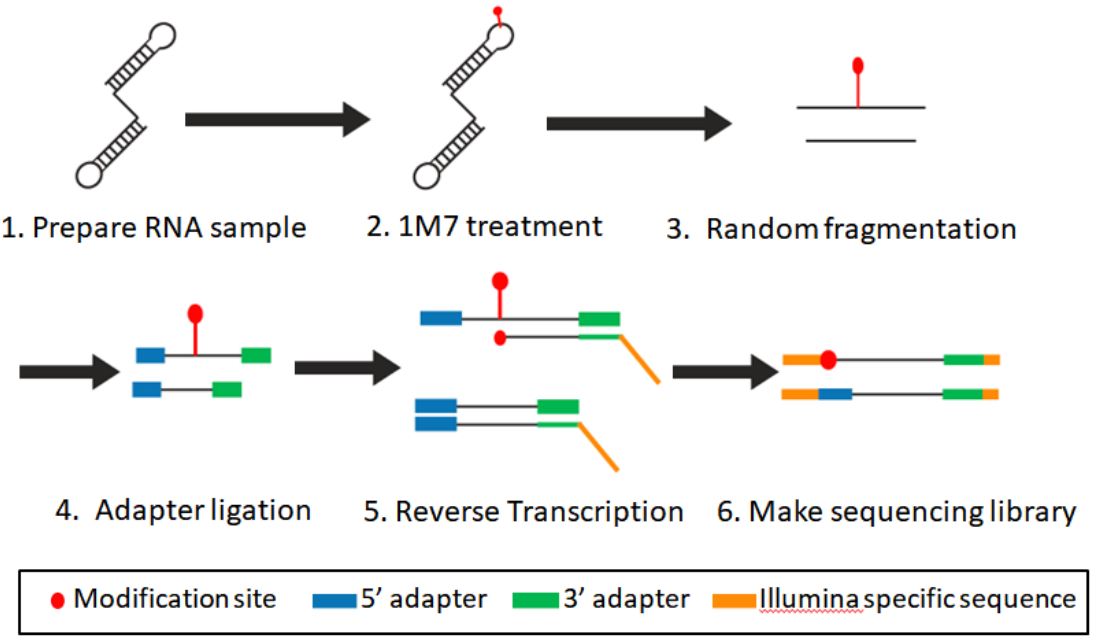


Figure 1. Workflow of the Mod-seq method.

1.4.3 Computational prediction of RNA structure based on experimental data

The experimental data obtained from RNA structure probing must be deciphered for RNA structure determination. One way to do this is to combine probing data and computational RNA structure prediction based on minimum free energy (MFE) models or partition function models with constraints. That is, nucleotides are forced to be double-stranded or single-stranded according to probing data before calculating MFEs. Many webservers are capable of such constrained RNA folding, including Mfold (56), RNAfold in the ViennaRNA (58, 59) package and RNAstructure (57). As chemical probing of RNA secondary structures has become more popular, RNA structure prediction software packages now also incorporate chemical probing data (for example, SHAPE reactivity scores) as continuous thermodynamic parameters, instead of binary (paired vs. unpaired) constraints (80). Incorporating chemical probing data in this way usually increases computational time and memory usage significantly and is often not viable for very large RNAs.

Another approach is to use probing data to choose the most “correct” fold from an ensemble of RNA structures (81, 82). Ding et al. (82) developed the Sfold package, which can sample thousands of possible RNA secondary structures for a single RNA molecule, calculate clustering of these structures, and compute the centroid structure from the clusters. They reported that the

centroid structure models often outperform MFE structure models in terms of positive predictive values (ppv) and sensitivity, and are more similar to structure models generated by comparative analysis. This method can be combined with chemical structure probing, i.e., choosing the structure cluster with the highest consistency with chemical probing data as the accepted structure. This approach was recently implemented by Spasic et al. (83), and can be used to generate alternative RNA structures based on probing data.

The original SHAPE probing quantification method was based on capillary electrophoresis measurements (84, 85). As the development of high-throughput sequencing aided structure probing methods, new challenges arose to process raw sequencing data into quantitative structural information in a fast, automatic manner. Publicly available data analysis software of high-throughput sequencing profiling data is needed to make such methods feasible to labs that lack of bioinformatics expertise. Several bioinformatics pipelines were developed to address this challenge, including SeqFold (86), which is optimized for PARS data, and StructureFold (87), which is available through the Galaxy platform (https://usegalaxy.org). In Chapter 2, I will also describe Mod-seeker, a bioinformatics pipeline I developed for the Mod-seq method.

1.5 NEAT1 as a candidate for understanding the function and conservation of lncRNA structure

NEAT1 is a lncRNA involved in paraspeckle formation (88, 89). Paraspeckles are nuclear bodies located in the nucleous interchromatin space. Though their function and regulatory mechanisms are not completely understood, recent studies showed that paraspeckles are involved in multiple gene regulatory processes, such as mRNA retention, mRNA cleavage, A-to-I editing, and protein sequestration (27–29). Perhaps because of these regulatory functions, NEAT1 is associated with many human diseases, including different types of cancer and neurodegeneration diseases (90–94).

Paraspeckles contain multiple protein and RNA components. NEAT1 lncRNA is the key RNA component of the paraspeckle (88). Human NEAT1 has two isoforms sharing the same transcription start site, but with different length. The long NEAT1 is as long as 23,000 nt, while the short one is 3,700 nt. The short isoform (NEAT1_S) undergoes canonical polyadenylation, while the long isoform (NEAT1_L) is cleaved by RNase_P and forms a triple-helix at 3’ end. (95) Multiple NEAT1 binding proteins are involved in this alternative 3’ end processing, including NUDT21-CDSF6 (CMIm complex) and HNRNPK. The short isoform NEAT1 has 5-8 fold higher expression than the long isoform (96), although more recent studies suggest this may be an artifact due to unbalanced

RNA extraction because paraspeckle NEAT1_L is likely trapped in the protein phase during normal RNA extraction (97). NEAT1 genes are found across mammals. Although the NEAT1 gene sequence is not well conserved, it was shown that both in human and in mouse cells, NEAT1 has conserved function, that it is essential for paraspeckle formation. Knockdown of NEAT1 leads to a significant decrease in paraspeckle formation (7).

NEAT1 is not only essential for paraspeckle formation, but also has a specific spatial organization in paraspeckles. Electron microscopic analysis combined with *in situ* hybridization (EM-ISH) showed that the short NEAT1 (or the 5' end of long NEAT1) and the 3' end of long NEAT1 are localized to the periphery of paraspeckles, while central sequences of long NEAT1 are found within the core of paraspeckles (25). This suggests that NEAT1 RNA may be folded end-to-end and serve as the circular skeleton of the paraspeckle, as shown in Figure 2. Given the scaffold function of NEAT1 and its specific spatial organization, it is reasonable to suspect NEAT1's secondary structure is important for its function. For this thesis, NEAT1 was chosen as a candidate for secondary structure probing and structural conservation analysis. Understanding the structure of NEAT1 is potentially helpful for understanding NEAT1's function in paraspeckle formation and may also provide general insights into lncRNA structures and their conservation.

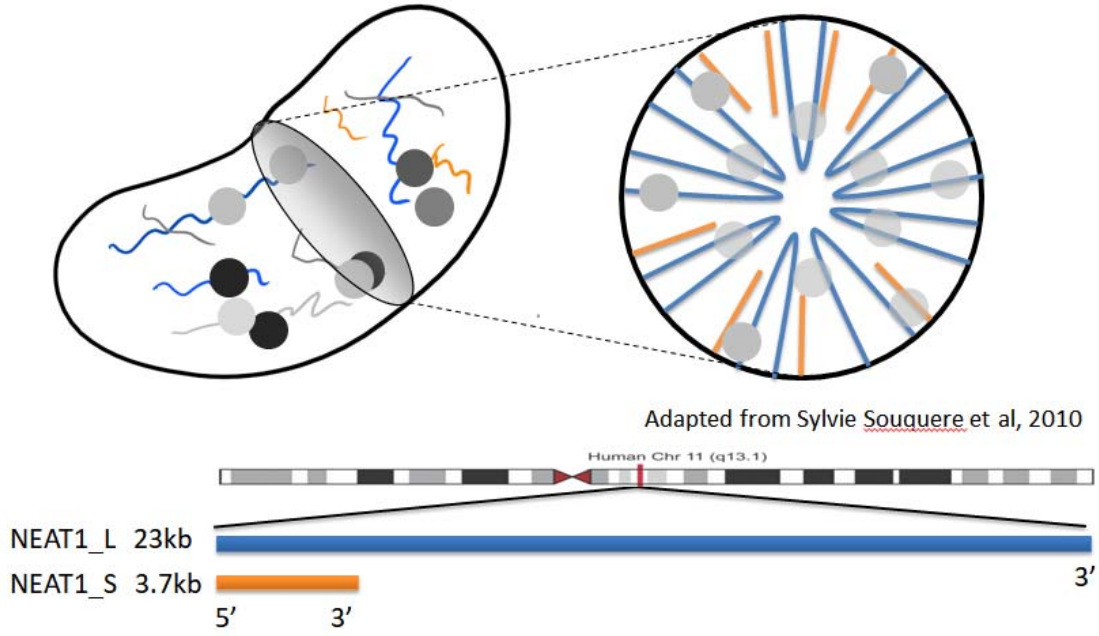


Figure 2. NEAT1 loci and paraspeckle architecture model. NEAT1 has 2 isoforms starting at the same locus; the short isoform is 3.7 k nt long, the long isoform is 22k nt long. The current model of paraspeckle structure suggests long isoform NEAT1 folds end-to-end, forming a circular skeleton as a scaffold for other paraspeckle proteins and RNAs.

My thesis work focuses on using both experimental methods and computational methods to study the structure of human lncRNAs. The rest of this dissertation is arranged as follows. In Chapter 2, I will describe Mod-seeker, an automatic data analysis pipeline for high-throughput RNA secondary structure probing method Mod-seq. This is one of the first open-sourced data analysis packages for high-throughput sequencing based RNA secondary structure chemical probing. In Chapter 3, I will focus on determining the structure of lncRNA NEAT1, an essential component of mammalian nuclear paraspeckles. In Chapter 4, I explore various computational methods to study the conservation of lncRNA secondary structure by identifying covariant base-pairs in NEAT1. Finally, in Chapter 5, I summarize my research on lncRNA structure, and particularly, the structure, long-range RNA-RNA interactions and structural conservation of NEAT1. I will also discuss both remaining and newly identified challenges in the field of lncRNA structure study. This thesis is the first to probe the secondary structure of lncRNA NEAT1. The comparative analysis of NEAT1 secondary structure provides new insight regarding the structural conservation of lncRNA. Even though flexible lncRNA do not have strong evidence for covariant base-pairs, they do have conserved structural features such as conserved single stranded regions or conserved long-range RNA-RNA interactions which may be important for their functions.

Chapter 2 Automated data analysis of Mod-seq data using Mod-seeker

The work presented in this chapter was published in the original Mod-seq paper (Talkish et al., 2013) for which I am a co-author, and a Methods in Enzymology paper (Lin et al., 201X) for which I am the first-author.

High-throughput sequencing based massive parallel RNA secondary structure probing methods are preferable than traditional PAGE gel-based methods or capillary electrophoresis-based methods, for they can be applied to very long RNAs or transcriptome-wide RNA structure profiling. They also give quantitative measurement as sequencing reads coverage, allowing for further data processing to achieve higher signal-to-background ratio. However, analyzing high-throughput sequencing data requires certain computational effort and expertise. New algorithms that are specifically optimized for these methods are in need to provide accurate RNA structural information. In this chapter I will describe a software package, Mod-seeker, that I implemented for Mod-seq structure probing method. Mod-seeker is an open source package that uses Mod-seq sequencing reads as input, generates SHAPE reactivity scores and Integrative Genomics Viewer (IGV) track files as output for easy data visualization and other downstream analyses such as SHAPE data aided RNA

secondary structure modeling. This work was published as part of the original Mod-seq paper (5) and the Mod-seq protocol method paper (6).

2.1 Principal of the Mod-seq method

Mod-seq (5, 6) combines RNA secondary structure chemical probing with high-throughput sequencing to determine RNA secondary structures in large scale or for long RNAs. In Mod-seq, DMS or SHAPE reagents-treated RNA molecules are purified and then randomly fragmented. Both 5’ adapters and 3’ adapters were ligated to the RNA fragments. The 5’ adapter is used as a marker to distinguish modification stops from 5’ ends generated from random fragmentation. The 3’ adapter is used so universal primer that hybridize to 3’ adapter can be used in primer extension. For RNA fragments containing chemical modification sites, reverse transcription prematurely stops at modification sites during primer extension, thus, the 5’ adapter is excluded from the cDNA product. In RNA fragments without chemical modification, reverse transcription goes through the whole sequence, thus the 5’ adapter sequences are present in the cDNA products. cDNAs are then circularized and products containing 5’ adapter sequences are reduced via subtractive hybridization. The remaining cDNA products, which are enriched for modification caused RT stops, are PCR-amplified for high-throughput sequencing. The presence of the 5’

adapter sequence at the beginning of Illumina sequencing reads indicates full-length reverse transcription products that do not contain chemical modification sites. Consequently, such products must be subtracted from the analysis. Thus, Mod-seq allows for reduced background for higher signal-to-noise ratios with proper data processing and analysis.

2.2 Mod-seeker data analysis pipeline

Mod-seeker contains two separate scripts. “Mod-seeker-map.py” is used to count the number of modifications at each position in each gene from each sample. In this script, sequencing reads are first trimmed to remove 3’ and 5’ adapters using Cutadapt (98). During adapter trimming, reads beginning with a 5’ adapter are removed from further analysis, as the presence of the 5’ adapter sequence indicated there is no chemical modification on this RNA molecule. The remaining trimmed reads are aligned to the reference sequence using Bowtie (99), Bowtie2 (100) or Tophat (101) per the user’s choice, and short 5’ mismatches indicating untemplated nucleotides introduced during reverse transcription are removed. Reads are then mapped to annotated genes using samtools (102) and bedtools (103). Finally, Mod-seeker-map.py counts the number of modifications at each position by tallying reads whose sequence initiates 3’ to each nucleotide. In the final output files (“CountMod” files), each

gene with modifications is represented by two lines, where the first line is a summary of the gene and the second line records space-separated counts of modifications at each position.

The second script, “Mod-seeker-stats.py” finds statistically significant sites of modifications by comparing chemically-treated samples with no-treatment controls. This script uses the Cochran-Mantel-Haenszel test (104, 105) with two or more replicates, or a chi-squared test for cases with no replicates. The p-values from these statistical tests are then corrected for multi-testing using Benjamini-Hochberg control (106) to calculate adjusted p-values. In addition to p-values from the statistical tests, the output file will also report odds ratios as a measurement of the modification level. The odds ratio is calculated as shown in equation (1), for a gene with length n, the odds ratio of position is:

$$OR_i = \frac{T_i / \sum_{j=1}^{n} T_j}{C_i / \sum_{j=1}^{n} C_j}, \quad (1)$$

where T_i is the count of modifications at position i in chemical-treated sample, and C_i is the count of modifications at position i in control sample. Additional data processing can be applied for further analysis. For example, the odds ratios can be log-transformed and rescaled to mimic SHAPE scores as described in (66, 107), and serve as input for RNAstructure to predict RNA secondary structures.

2.3 Evaluating normalization methods in Mod-seeker

The odds ratio in equation (1) was chosen as the preferred metric in the Mod-seeker pipeline for a variety of reasons. First, log(odds ratio) can be statistically tested by the Fisher's exact test, or proximately by the chi-square test. This will generate reliable p-values for Mod-seeker to quantitatively identify nucleotides with statistically significant signal of chemical modification. Second, this normalization method was validated on ribosomal RNAs and has better performance than other tested methods.

Receiver operator curve (ROC) analysis of different data normalization methods on *S.cer* ribosomal RNAs are shown in Figure 3 and summarized in Table 3. Ribosomal RNA was chosen in this analysis since it is a large RNA whose size is similar to other lncRNAs and whose structure is determined by crystallography. Nucleotides are classified into modified and unmodified groups based on their SHAPE reactivity scores under a certain threshold, the sensitivity (true positive rate) and specificity (1 – false positive rate) varies as the threshold changes. The area under curve (AUC) of ROC is commonly used as a measurement of the performance of such binary classifiers, where a perfect classifier will have AUC close to 1.0, and higher AUC is an indicator of better performance. The odds ratio, and the log-transformed odds ratio have better performance than the other metrics, and log transformation increases

performance slightly. Also, Mod-seq has better performance than a similar rRNA structure probing experiment using traditional SHAPE probing detected by capillary electrophoresis (hSHAPE score) (42).

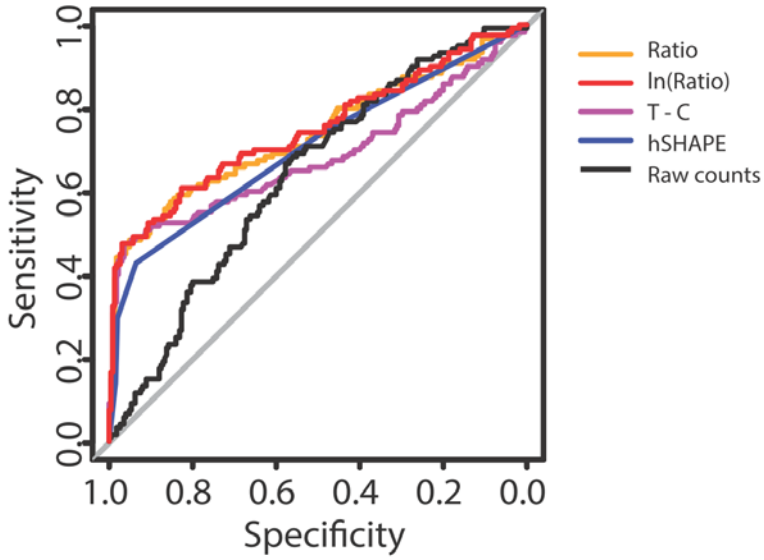


Figure 3. Comparison of ROC performance for different data normalization methods on *S. cerevisiae* 18S rRNA.

Normalization Method	Area Under Curve (AUC) in ROC analysis	
	18s rRNA	25s rRNA
$Ratio_i = \frac{T_i / \sum_{i=1}^{n} T_i}{C_i / \sum_{i=1}^{n} C_i}$	0.740	0.742
$\ln(Ratio_{i)} = \frac{\ln(T_i)}{\sum_{i=1}^{n} \ln(T_i)} - \frac{\ln(C_i)}{\sum_{i=1}^{n} \ln(C_i)}$	0.748	0.749
$T_i - C_i$	0.684	0.713
hSHAPE score	0.705	0.704

Table 3 Summary of the ROC performance of different data normalization methods.

2.4 Discussion

Mod-seeker is a data analysis pipeline designed and optimized for the Mod-seq parallel structure probing method. Mod-seeker takes raw sequencing reads from Mod-seq as input, and selects for reads lacking the 5' adapter, which indicates that reverse transcription stopped due to a chemical modification on a template nucleotide. Mod-seeker requires sequencing reads from both the treated and control samples to calculate the normalized odds ratios as SHAPE reactivity scores. At least two replicates are required in order to perform statistical tests and call significantly modified nucleotides. RNA structure information generated from Mod-seq and analyzed by Mod-seeker was shown

to be highly consistent with known RNA structures (5), and the SHAPE reactivity scores calculated using Mod-seeker can improve RNA secondary structure modeling accuracy (108). Compared to other publicly available packages that can process high-throughput sequencing based RNA structure SHAPE probing data, such as SeqFold (86) and StructureFold (109), Mod-seeker requires both treated sample and untreated samples as background control. This allows for higher signal-to-noise ratio and can distinguish true chemical modification sites from other reverse transcription stops introduced by factors such as random fragmentation of RNA, alternative transcription start sites, and premature falloff of reverse transcriptase.

Future improvements on Mod-seeker can be made for versatility and easier integration with other bioinformatics tools. Mod-seq was shown to be able to identify potential binding sites for RNA-binding proteins, by comparing SHAPE profiles with and without proteins present. This function, however, is not included in the current version of Mod-seeker. Also, Mod-seeker requires several pre-installed packages, such as cutadapt, bowtie2, bedtools, and samtools. Although these are commonly used packages in general bioinformatics analyses, installing each of these individually might be overwhelming for researchers lacking related experiences. Integration with a python package manager such as pip will be extremely useful in this case.

Integration with other bioinformatics toolsets, such as R-bioconductor will also be useful.

Chapter 3 Structural analyses of NEAT1 lncRNAs suggest long-range RNA interactions that may contribute to paraspeckle architecture

The work presented in this chapter was published in NAR (Lin et al., 2018), with some modifications.

Yizhu Lin[1], Brigitte F. Schmidt[2], Marcel P. Bruchez[1,2,3], and C. Joel McManus[1,*]

[1] Department of Biological Sciences, [2]Molecular Biosensor and Imaging Center, [3]Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, 15213,

* To whom correspondence should be addressed. Tel: 412-268 9407; Email: mcmanus@andrew.cmu.edu.

ABSTRACT

Paraspeckles are nuclear bodies that regulate multiple aspects of gene expression. The long non-coding RNA (lncRNA) NEAT1 is essential for

paraspeckle formation. NEAT1 has a highly ordered spatial organization within the paraspeckle, such that its 5’ and 3’ ends localize on the periphery of paraspeckle, while central sequences of NEAT1 are found within the paraspeckle core. As such, the structure of NEAT1 RNA may be important as a scaffold for the paraspeckle. In this study, we used SHAPE probing and computational analyses to investigate the secondary structure of human and mouse NEAT1. We propose a secondary structural model of the shorter (3,735 nt) isoform hNEAT1_S, in which the RNA folds into four separate domains. The secondary structures of mouse and human NEAT1 are largely different, with the exception of several short regions that have high structural similarity. Long-range base-pairing interactions between the 5’ and 3’ ends of the long isoform NEAT1 (NEAT1_L) were predicted computationally and verified using an *in vitro* RNA-RNA interaction assay. These results suggest that the conserved role of NEAT1 as a paraspeckle scaffold does not require extensively conserved RNA secondary structure and that long-range interactions among NEAT1 transcripts may have an important architectural function in paraspeckle formation.

INTRODUCTION

Long non-coding RNAs (lncRNAs) are defined as non-protein coding RNAs that are longer than 200 nucleotides. In the human genome, more than

thirteen thousand lncRNAs have been annotated (9), making up a large proportion of human genes. lncRNAs are involved in gene regulatory functions through diverse mechanisms including chromatin binding (Xist) (12), regulating gene transcription in *cis* (ANRIL) (16), and scaffolding of nuclear bodies (NEAT1). Intriguingly, although many lncRNA have important conserved functions, they usually have relatively low sequence conservation (9). This is counterintuitive, as sequence conservation is often assumed to be required for genes with important functions (110). One possible explanation is that lncRNA preserve higher order conservation, such as conservation of secondary structure (base pairing interactions) or tertiary structure (three-dimensional shape of folded RNA).

Large RNAs fold into secondary structures, which then influence their three-dimensional tertiary structures. Resolving the secondary structures of lncRNAs *in vivo* is a difficult task due to their large size and low abundance in cells. High-throughput *in vivo* structure probing using reverse transcription truncation (-seq) methods requires extreme sequence depth for low abundance lncRNAs. Till now, there is only one human lncRNA, Xist, whose structure has been probed *in vivo* (111). Furthermore, lncRNAs are expressed in alternative isoforms and bound by a variety of RNA-binding proteins *in vivo*, both of which can obscure interpretation of chemical modification patterns. *In vitro* structure probing interrogates an RNA's inherent folding potential without interference by

bound proteins or alternative transcript isoforms. Although this simplifies the task, the large size of lncRNA still poses a significant challenge, and only a few lncRNA structures have been experimentally characterized *in vitro* (48) (HOTAIR (45), Xist (44, 112) and ncSRA (47) RepA (113) and lincRNAp21 (46)).

NEAT1 is an especially interesting lncRNA for structural study. It is a key structural component of paraspeckles and is essential for paraspeckle formation. Paraspeckles are nuclear bodies located in the nucleus interchromatin space. Though paraspeckle functions and regulatory mechanisms are not completely understood, recent studies showed they are involved in multiple gene regulatory processes, such as mRNA retention, mRNA cleavage, A-to-I editing (88) and protein sequestration (29). These regulatory functions are responsible for several cellular responses and shown to be associated with the pathology of multiple cancers and neurodegenerative diseases (94, 114, 115). Deletion of NEAT1 in mice disrupts development of female reproductive tissues, underscoring the biological importance of this lncRNA (116, 117).

NEAT1 has two isoforms that share the same transcription start site but have different termination sites. In humans, the short isoform NEAT1_S is 3,735 nt long with a polyA tail. The long isoform, which is essential for paraspeckle formation, is 22,741 nt in length and has a non-polyadenylated 3' end produced by RNase P cleavage (8, 95). The expression level of NEAT1_S is estimated

to be at least five-fold higher than NEAT1_L, and even higher in many tissues and cell types (89, 96). Though less abundant, NEAT1_L is considered to be the key isoform for paraspeckle formation. Targeted knock down of NEAT1_L leads to loss of paraspeckles, while *de novo* paraspeckle formation can be rescued by transient expression of NEAT1_L (23, 95). Intriguingly, NEAT1_S can be found outside of the paraspeckle in tissue culture cells, suggesting it may have independent biological functions (118). The two-isoform gene structure and the function of NEAT1 in paraspeckle formation were observed in both humans and mice. However, the sequence of NEAT1 is not well conserved between human and mouse. This suggests higher-order conservation of NEAT1 RNAs, such as secondary structural conservation or conserved RNA-protein interactions.

Interestingly, evidence has emerged indicating that the specific structural conformation of NEAT1 might be important for paraspeckle architecture. EM-ISH (electron microscopy-*in situ* hybridization) studies using DNA probes to the 5’ and 3’ ends of NEAT1_L RNA showed that NEAT1_L has a highly ordered spatial organization within the paraspeckle (114). The 5’ and 3’ ends of NEAT1_L were localized to the paraspeckle periphery, while the central region of NEAT1_L was found within the paraspeckle core. Since the 5’ end of NEAT1_L is identical to NEAT1_S, the short isoform NEAT1_S should also localize to the periphery of paraspeckle. Based on these observations, an

ultrastructural paraspeckle model was proposed with two salient features. First, NEAT1_L folds end-to-end. Secondly, multiple folded NEAT1_L and NEAT1_S molecules are regularly organized in the cross sections of paraspeckle, forming a circular skeleton. However, the actual secondary structure of NEAT1 has not yet been characterized. The nature of the spatial organization of NEAT1 and its contribution to paraspeckle architecture is yet to be understood.

Here, we combined high-throughput RNA structure probing (Mod-seq) (5) with computational analyses to investigate the structural features of NEAT1. Mapping and comparing the structures of human and mouse NEAT1_S revealed two short regions of similar SHAPE reactivity, and phylogenetic comparisons found relatively little evidence for conservation of RNA secondary structure. Computational analysis identified putative long-range RNA-RNA base paring interactions between NEAT1_L's 5' and 3' ends, which commonly exist in all analyzed mammals NEAT1 sequence. We propose that the NEAT1 lncRNA has maintained its function as a paraspeckle scaffold with little structural conservation, and identify a strong propensity for long-range intramolecular base-pairing that may contribute to scaffolding the paraspeckle.

MATERIAL AND METHODS

In vitro transcription

hNEAT1_S and mNEAT1_S plasmids were generously provided by Dr. Gérard Pierron (25) and Dr. Lingling Chen (119), respectively. PCR primers were designed for both full length NEAT1 RNA and short segments, and the SP6 promoter sequence was included in the forward primers. The DNA template for *in vitro* transcription was amplified from the plasmids using Phusion high-fidelity polymerase and purified by agarose gel extraction. The RNA was *in vitro* transcribed using Promega RiboMAX large scale RNA production systems (SP6), as described in the manufacturer's instructions. Briefly, 200-500 ng cDNA template, 4 µL 5X SP6 buffer, 4 µL 25 mM rNTPs and 2 µL SP6 enzyme mix were mixed in a 20 µL reaction and incubated at 37 °C for 3.5 hours. 0.5 µL RQ1 RNase-Free DNase (1u/µL) were added to each reaction and incubated at 37 °C for 15 min to destroy DNA template. 0.5 µL proteinase K (20 mg / ml) was then added to reaction and incubated at 37 °C for 1 hour to destroy SP6 transcriptase and RQ1 DNase.

Non-denaturing purification of RNA

A non-denaturing purification was adapted from Somarowthu et al. (45) to maintain the co-transcriptionally folded structure for SHAPE probing

experiments. Briefly, after proteinase K treatment, the RNA was diluted with 200 µL 1X SHAPE buffer (111mM NaCl, 111 mM HEPES, 6.67 mM $MgCl_2$), transferred to Amicon Ultra 100K column and centrifuged at 14,000 g for 10 min to concentrate the RNA sample to approximately 30 µL. This dilution / concentration step was repeated for a total of two rounds. The purified RNA was then collected by centrifuging the column upside down 2 min at 1,000g. The RNAs were verified on a TapeStation. The RNAs were kept on ice and were immediately used for SHAPE probing

1M7 Synthesis Procedure

We synthesized 1M7 using a novel procedure. In brief, 2-Amino-4-nitrobenzoic acid was converted to 2-((Ethoxycarbonyl)amino)-4-nitrobenzoic acid through the addition of ethyl chloroformate by reflux for 1 hr. This product was converted to 7-Nitro-1H-benzo[d][1,3]oxazine-2,4-dione by heating at 65°C in the presence of thionylchloride for 30 minutes, cooled to room temperature and washed with chloroform. The 7-Nitro-1H-benzo[d][1,3]oxazine-2,4-dione dissolved in DMF was then treated with potassium carbonate and iodomethane, similar to published methods (29), yielding an orange precipitate containing both 1M7 and a hydrolyzed contaminant (as determined by NMR). Pure 1M7 (light yellow in color) hydrolizes to 2-(methylamino)-4-nitrobenzoic acid (orange in color). Published synthesis methods describe an orange product that is likely

contaminated with the hydrolysis product. We purified 1M7 by fractional crystallization from ethyl acetate/hexane where the contaminant crystallized first to yield (40%) of orange crystals, mp 256-258°C. 1M7 crystallized second to yield (50%) of light yellow crystals, mp 206- 208°C. 1M7 was resuspended in DMSO at 65 mM and stored at -80 °C. The solution retained a light yellow color that turned bright orange when mixed with the RNA sample in SHAPE buffer.

In vitro SHAPE probing with 1M7

RNA secondary structure probing was performed using 1M7 as the SHAPE reagent, as described in Mortimer et al. (69). 2 pmoles RNA product were diluted in 13.3 µL 1 x SHAPE buffer, incubated at 37 °C for 5 min. 1.7 µL 1M7 (65 mM, in DMSO) were then added into each reaction, continue incubation at 37 °C for 70 s. The control samples were incubated with same volume of DMSO instead of 1M7. 1M7 probed RNA was then purified using ethanol precipitation method.

Mod-seq library preparation and data processing by mod-seeker pipeline

Probed RNA samples were pooled together for Mod-seq library preparation. At least 2 replicates were sequenced for 1M7 treated samples and negative control samples (Supplementary Table S1). Mod-seq libraries were

generated as previously described (6) and sequenced with an Illumina Miseq sequencer. Sequencing reads were aligned to hNEAT1 or mNEAT1 sequences and replicates were combined for further analysis after checking for correlations. The SHAPE reactivity score is calculated using the equation: SHAPE Reactivity = Normalized Count(Treated) – α * Normalized Count(Ctrl), as described in Spitale et al.(120). Parameter α was set to 0.35 by using *in vitro* transcribed and probed *Tetrahymena* P4P6 domain (121) (Supplementary Figure S1) as a positive control.

RNA secondary structure modeling

RNA secondary structure models with or without SHAPE probing constraints were generated using RNAstructure software (Linux text interface 64bit, version 5.8.1; default parameters) (57). SHAPE reactivity scores were used as constraints for RNA secondary structure predictions. To generate RNA secondary structure models of NEAT1 segments, partition functions (60) were first calculated with the “partition” command in RNAstructure; the “max expect” structures (122) were used as RNA structure models, which was calculated using the “MaxExpect” command. For full length hNEAT1_S and mNEAT1_S structure modeling, partition function predictions are computationally intense, so minimum free energy structures were instead calculated with the “Fold”

command in RNAstructure. Structure models were stored in ct files and visualized with VARNA (v3.92) (123).

Comparing structures of full length NEAT1 and 3S shotgun segments

To compare structures of full length NEAT1 and segments, we calculated Pearson's correlations of their SHAPE reactivity scores between segments and the corresponding regions in full length NEAT1_S. A similar correlation analysis was done in sliding windows with a window size of 60 nt and a step size of 1 nt.

Infernal alignment and covariation analysis

To identify conserved secondary structure in NEAT1_S, we first used Infernal (default parameters) (124) to generate improved multiple alignments of regions in NEAT1_S as described in Chillon and Pyle (45). Multiple alignments of 99 vertebrates were downloaded from UCSC genome browser database (125), where 64 sequences have alignments to human NEAT1_S region. Covariation models were built using Infernal cmbuild on 8 sequences including hNEAT1_S and mNEAT1_S, and then calibrated with cmcalibrate. Improved multiple alignments across 64 species were then generated using cmsearch and cmalign. Finally, covariant base pairs were identified with both R2R (126) using a 15% threshold (45, 113) and R-scape using default parameters (48). To compare R-scape results from NEAT1 to those of well-characterized

structured RNAs, we subsampled sequence alignments to have similar numbers of sequences in each alignment (~50) and pairwise sequence identity (average ~68%). For covariation score analysis, R-scape's default scoring metric (APC G-test statistics) was used. With Infernal improved alignments of hNEAT1_S and mNEAT1_S, we calculated Pearson's correlation coefficients of SHAPE reactivity scores in each region after aligning SHAPE scores to their sequence alignment.

Generating synthetic NEAT1 alignments with random mutations

For each Infernal aligned region, the hNEAT1_S sequence was used as an ancestor sequence to build random synthetic alignments. In each round of sequence generation, 2 child sequences were generated from their parent sequence, where point mutations were introduced at random for each nucleotide position with a fixed mutation rate (probability). After 7 rounds, 128 sequences were generated. 50 out of 128 sequences were randomly selected to build each synthetic alignment. This simulation was repeated 100 times each with mutation rates ranging from 0.5% to 5% to generate random null alignment models with average pairwise identity ranging from 60% to 95%. These null alignments were used directly for R2R analyses, or realigned with Infernal before R2R analyses.

RNA-RNA interaction prediction

Prediction of long-range interactions in NEAT1 was done with RNAduplex (59, 127). The sequence of NEAT1_S and the rest of NEAT1_L sequence (after trimming off NEAT1_S sequence) were used as input. In sliding window analyses, NEAT1_L sequence was separated into 120 nt long windows with a step size of 40 nt. The pairwise minimum free energy of each duplex was then predicted using RNAduplex using default parameters.

In vitro gel shift assay

NEAT1 segment templates were generated by PCR from genomic DNA (HEK genomic DNA for hNEAT1 and mouse kidney genomic DNA for mNEAT1). After *in vitro* transcription with SP6, the predicted interacting NEAT1 segments were treated with RQ DNase and purified with phenol chloroform extraction and ethanol precipitation as described in RiboMax SP6 kit (Promega). An RNA gel shift experiment was adapted from Gavazzi et al. (128). Briefly, 2 pmol of each RNA segment were mixed in 8 µL H_2O, incubated at 90 °C for 2 min and then chilled on ice. 4 µL 3x pairing buffer (50 mM Sodium Cacodylate, 40 mM KCl, 0.5/2/6 mM $MgCl_2$) and 0.25U SUPERase-in was added into each reaction and incubated at 37 °C for 30 min. RNA duplexes were then assayed by agarose electrophoresis. The duplexes were electrophoresed through a 3% agarose

gel in TBM buffer (45 mM Tris, 43 mM borate, 2 mM $MgCl_2$, pH 8.3) for 1 hour at 4 °C.

eCLIP data analysis

eCLIP RNA-binding protein binding site data was downloaded from ENCODE (129) in narrowPeak format. Protein binding sites on NEAT1 were filtered using bedtools intersect. To map the binding sites of TARDBP on NEAT1_S structure, each nucleotide in NEAT1_S was assigned an eCLIP score that equals to the highest signal value among all peaks covering that nucleotide. A nucleotide that has no crosslinking has a score of zero. hNEAT1_S structure model was then visualized by VARNA and colored by eCLIP scores. For hierarchy clustering analysis, eCLIP score on each nucleotide was filtered such that it has enough signal enrichment (signal value greater than 3), and is statistically significant (p-value smaller that 1e-5), and has significant binding sites in both replicates. The mean scores of the two replicates were then used in clustering analysis, where correlation was used as distance matrix with average-link clustering algorithm.

RESULTS

In vitro secondary structure probing of human NEAT1_S

We first used Mod-seq (5) (Figure 4) to probe the *in vitro* structure of the 3,735 nt human NEAT1 short isoform (hNEAT1_S). Large RNAs often adopt multiple structural folds after heat denaturation and refolding *in vitro*. To avoid this, we purified *in vitro* transcribed NEAT1_S under non-denaturing conditions designed to preserve its co-transcriptionally folded structure (45). hNEAT1_S RNA was probed with 1M7 (69), and modification sites were identified using Mod-seq. SHAPE reactivity scores for each nucleotide were then calculated as previously described (120), where higher scores suggest structural flexibility (Supplementary Figure S2). Although modeling long RNA structures with Mod-seq has not been validated, Mod-seq measures SHAPE reactivity accurately (Supplementary Figure S1) and SHAPE reactivity data have been used to model many long RNA secondary structures (44–48, 112, 113, 130, 131).

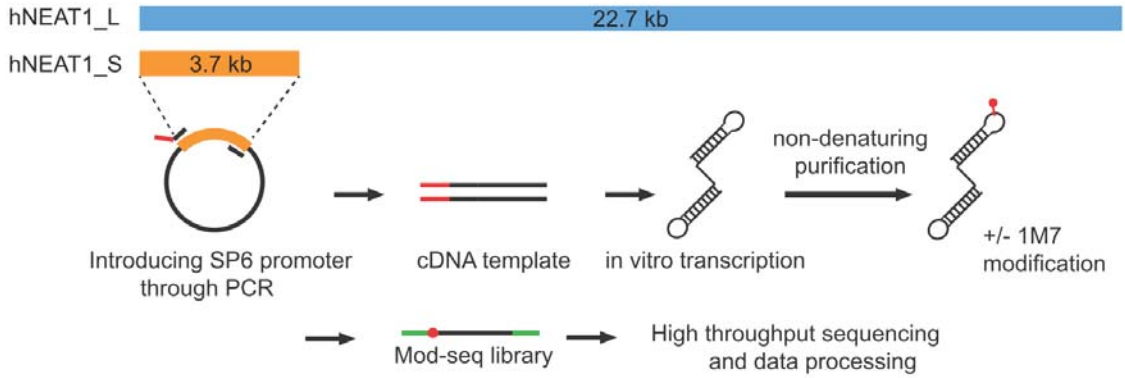


Figure 4. Overview of NEAT1 secondary structure probing. cDNA templates of NEAT1 regions were generated by PCR using primers that incorporated the SP6 promoter sequence. NEAT1 RNA was then generated by SP6 *in vitro* transcription. After non-denaturing RNA purification, RNAs were probed with the SHAPE reagent, 1M7. The negative controls were treated with DMSO only. Mod-seq libraries were then made and sequenced to an average combined depth of ~ 100 reverse transcriptase stops per nucleotide. SHAPE reactivity was calculated by comparing reverse transcriptase stops from 1M7 treated and untreated control samples

We investigated the domain structure of NEAT1_S using an approach similar to the 3S shotgun method (132). In this approach, full length NEAT1_S was divided into 13 overlapping ~500 nt segments (Figure 5A and Supplementary Table S2). Each segment was *in vitro* transcribed and SHAPE probed individually using the same non-denaturing method that we used in full length NEAT1_S probing. If nucleotides within a segment exhibit similar SHAPE reactivity to that seen in the context of full length RNA, they likely form base-pairs within a sub-domain with relatively independent and stable local structure. The similarity of SHAPE scores between each segment and full length NEAT1_S was measured by Pearson's correlation (Figure 5B), finding that most regions appear to have stable local structures. To identify boundaries between local structures, we also evaluated Pearson's correlations in 60-nucleotide sliding windows across NEAT1_S (Figure 5C). These results indicate that hNEAT1_S has primarily local base-pairing interactions when prepared under non-denaturing conditions.

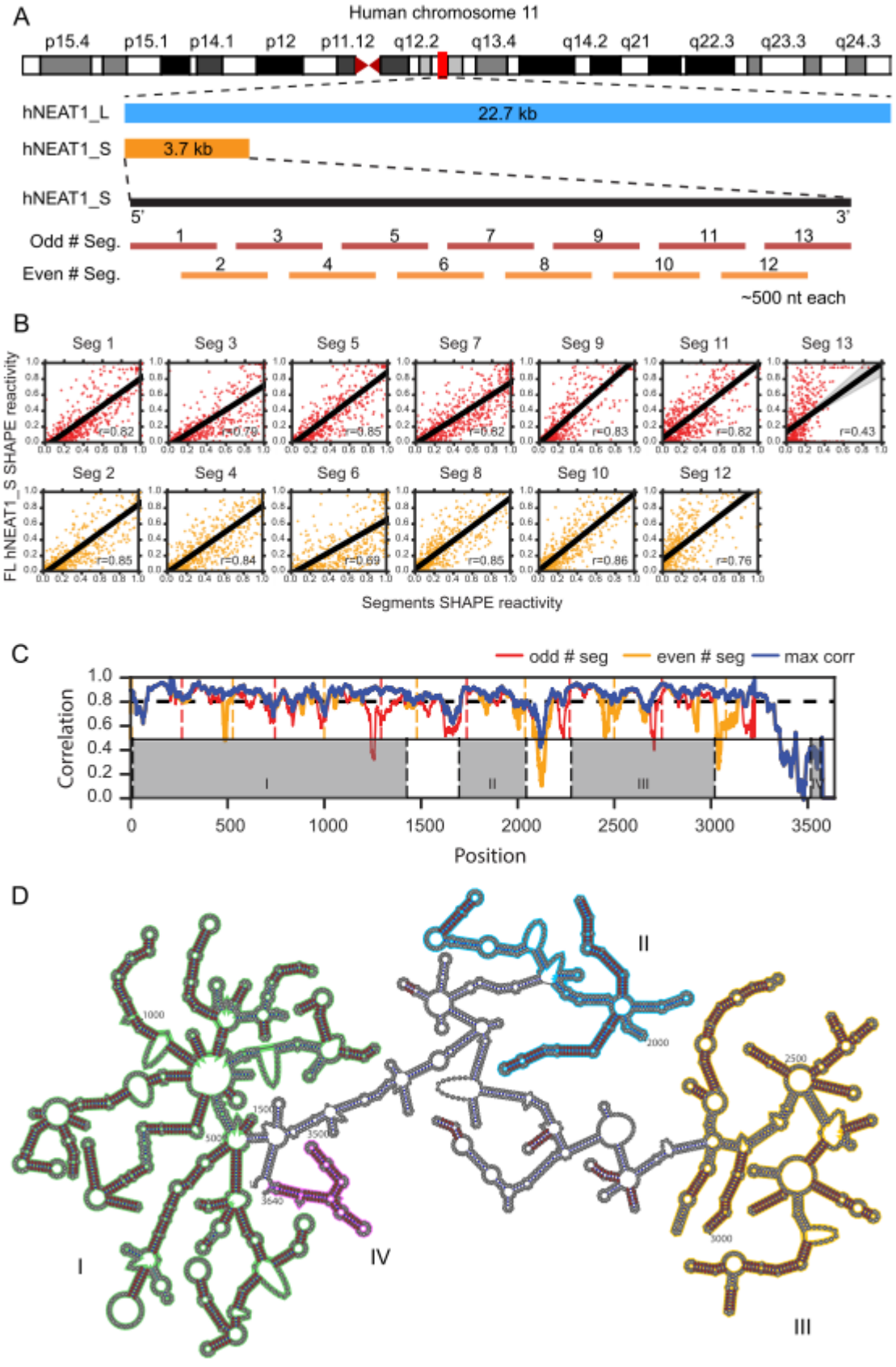

A
Human chromosome 11
p15.4 p15.1 p14.1 p12 p11.12 q12.2 q13.4 q14.2 q21 q22.3 q23.3 q24.3
hNEAT1_L
22.7 kb
hNEAT1_S
3.7 kb
hNEAT1_S
5'
3'
Odd # Seg.
1 3 5 7 9 11 13
Even # Seg.
2 4 6 8 10 12
~500 nt each
B
Seg 1 Seg 3 Seg 5 Seg 7 Seg 9 Seg 11 Seg 13
Seg 2 Seg 4 Seg 6 Seg 8 Seg 10 Seg 12
FL hNEAT1_S SHAPE reactivity
Segments SHAPE reactivity
C
odd # seg
even # seg
max corr
Correlation
Position
I
II
III
IV
D
I
II
III
IV

Figure 5. Identification of local stable structures in hNEAT1_S. A). Illustration of gene locus of hNEAT1_S and hNEAT1_L. The secondary structure of full length hNEAT1_S and 13 ~500 nucleotide sub-segments were probed *in vitro*. B). Scatter plots showing the correlation of SHAPE reactivity scores in each segment with the corresponding region in full length hNEAT1_S. C). Pearson's correlations of SHAPE reactivity scores between full length hNEAT1_S and each segment were calculated using a 60-nucleotide sliding window with 1 nucleotide step size. The correlations of hNEAT1_S and even number segments are shown in red, while the correlations of hNEAT1_S and odd number segments are shown in orange. The blue line indicates the larger correlation of the two (odd vs even segments). Odd and even segment boundaries are marked as upper dashed lines. The lower dashed lines indicate boundaries of identified structural domains. D). Secondary structure models in hNEAT1_S. Shared base-pairs between full length hNEAT1 and the 500 nucleotide sub-segments are marked in red. The four structural domains are highlighted with colors.

To identify stable local sub-domains of hNEAT1_S, we compared the secondary structure models of each segment with the 100 lowest free energy structures of full length hNEAT1_S and searched for shared base-pairs (Figure 5D). 696 shared base-pairs were identified in total, accounting for 57.7% of all base pairs in the full length hNEAT1_S structure. By manually clustering adjacent shared base-pairs, we demarcated 4 domains in hNEAT1_S that have relatively stable local structures, as highlighted by colors (Figure 5D). Domain I encompasses most of the 5’ end of NEAT1_S, while domains II, III and IV are more separated. Domain IV marks a folded 3' end. The separation of domains is also observed in the sliding window correlation analysis (Figure 5C), where the correlation of SHAPE reactivity scores is higher within each domain, but drops in junction regions between domains. These results support a model in which NEAT1 folds into a modular multi-domain RNA.

Phylogenetic analyses of NEAT1 secondary structure conservation

We used phylogenetic analyses to investigate the conservation of the NEAT1_S structure. We first used Infernal (124) to generate improved mammalian multiple alignments of NEAT1_S using our SHAPE-constrained structure model. As it is possible that only small subdomains of NEAT1_S have conserved structure, we applied Infernal to compact helical regions from the domains defined using the 3S shotgun procedure (see methods; Table 4). For

12 of 14 subdomains, Infernal identified at least 40 out of 64 mammalian species with significant alignment to human NEAT1_S. Two regions in domain III (nt 2470-2609 and nt 3199-3316) had only 12 and 25 alignments, respectively, and the former one only had alignments within primates.

We applied R2R (126) and R-scape (48) to evaluate the conservation of NEAT1_S secondary structure. R2R classifies base-pairs as covarying if at least one compensatory mutation is present in an alignment, given there are less non-canonical base pairs than a user-defined threshold. R-scape uses a background null distribution to identify statistically significant covariant base-pairs, but performance depends on the number of alignments used and their average pairwise identity. Some lncRNAs have covariant base-pairs identified by R2R (45, 113) but many failed the statistical tests in R-scape (48). Similarly, R2R identified many more covariant base pairs than R-scape on NEAT1_S (Figure 11 I and J, see details in Chapter 4). However, R2R may be too liberal and / or R-scape too conservative for analysis of NEAT1_S structural conservation. Further analyses suggest R2R is prone to false-positive covariation calls on NEAT1_S (Figure 11 D and E, see details in Chapter 4), and that R-scape has reasonably strong performance on well-structured RNAs (tRNA, riboswitches, TERC, etc.) after matching alignment number and pairwise identity to that of NEAT1_S (Figure 12, see details in Chapter 4). NEAT1_S alignments had higher R-scape co-variation scores than random null

alignments (Figure 13 , see details in Chapter 4), however NEAT1_S had relatively fewer significant covariant base pairs (E value < 0.05; Figure 12, see details in Chapter 4).These results suggest that NEAT1_S is under less selective pressure for specific RNA structures than well-known highly-structured RNAs.

SHAPE probing of mouse NEAT1_S identifies few structurally similar regions

Since most human lncRNAs only exist in mammals and are much younger than structured small non-coding RNAs, the R-scape E-value significance threshold of 0.05 may be too stringent for lncRNAs. In addition, it is possible that lncRNAs like NEAT1 have conserved single-stranded regions that would be undetectable using R-scape. To experimentally evaluate the conservation of NEAT1 structure, we compared the *in vitro* structures of human NEAT1_S and mouse NEAT1_S. A secondary structural model of mNEAT1_S was determined using the same pipeline for hNEAT1_S (Supplementary Figure S3). Both full-length mNEAT1_S and 12 overlapping segments (Supplementary Table S2) were *in vitro* transcribed and probed with 1M7, and their SHAPE reactivity profiles were assayed by Mod-seq. We compared the SHAPE reactivity profiles of hNEAT1_S and mNEAT1_S using the Infernal derived mammalian NEAT1_S sequence alignment to align their SHAPE scores. Out

of 10 regions with well-defined sequence alignments, 5 had significantly positive correlations (nt 514 – 680, nt 901- 1036, nt 1037-1268, nt 1269-1467, nt 1710-1833) (Table 1). The nt 514-680 region had the highest correlation (R = 0.43; Figure 6), suggesting higher structural similarity, even though R-scape identified no covariant base pairs in this region. These results show NEAT1 has small regions with evidence for structural similarity, while other regions have much lower structural conservation.

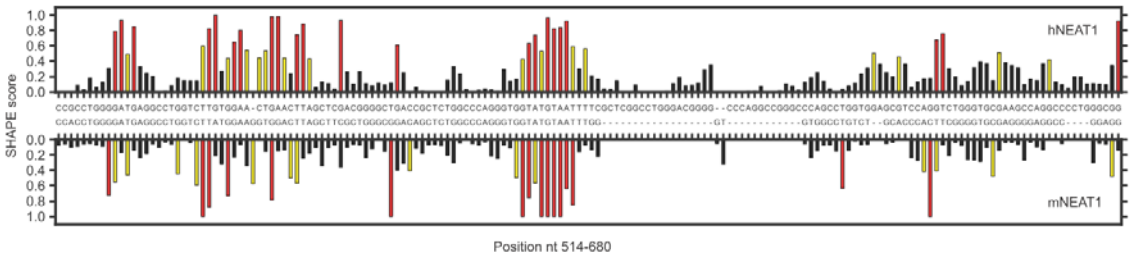


Figure 6. Comparison of SHAPE reactivity profiles. SHAPE profiles for the region of hNEAT1_S found to have the highest correlation with the corresponding region in mNEAT1_S (nts 514-680). SHAPE scores (see methods) are plotted for hNEAT1 (upper) and mNEAT1 (lower).

Long-range RNA-RNA interactions in NEAT1

Previous studies have reported that the 5’ and 3’ ends of NEAT1 are co-localized in the paraspeckle periphery, and speculated that this is a consequence of interactions among RNA-binding proteins (25), We investigated the possibility that long-range RNA-RNA interactions might contribute to colocalization. We used RNAduplex, a software package for predicting structure upon hybridization of two RNA, with hNEAT1_S sequence and the remaining 19,006 nt sequence of hNEAT1_L to identify potential long-range interactions. Surprisingly, RNAduplex predicted a large interaction of almost the entire short hNEAT1 with the 3' end of long hNEAT1. The prediction is similar in mouse NEAT1, with mNEAT1_S predicted to form a duplex with the 3' end sequence of mNEAT1_L (Figure 7A and Figure 7B). To further investigate the potential for long-range interactions, we separated human and mouse NEAT1_L sequences into 120 nt windows and calculated the minimum free energy of each pair of windows (Figure 7C and Figure 7D). Both in human and mouse, duplex minimum free energy heat maps show darker colors at the edges and corners. These long-range interaction regions in hNEAT1_L and mNEAT1_L have significantly lower minimum free energy (z-scores < -3) than random pairs of NEAT1_L sequences (Supplementary Figure S4A-B). This pattern is consistent across mammals (Supplementary FigureS4B). These

results show that NEAT1 has a conserved inherent capacity to form long-range interactions between its 5’ and 3’ ends.

Based on our windowed analysis of base-pairing potential, we predicted RNA segments most likely to form long-range interactions by searching for the best candidate segment pairs (Supplementary Table S3). Selected RNA-RNA interactions of predicted regions were tested using an *in vitro* RNA-RNA gel shift assay (Figure 7E and Supplementary Figure S5). As predicted, hNEAT1 segment 1 (nt 282 - 546) and hNEAT1 segment 2 (nt 600 - 840) formed a stable duplex structure with segment 3 (nt 20761 - 21120). In mNEAT1, the predicted regions also show RNA-RNA interaction ability, though the interaction seems to be weaker than the tested hNEAT1 segments (Supplementary Figure S5). These results show that sequences in the 5' and 3' ends of NEAT1 can form base-pairing interactions under physiological Mg^{2+} concentration.

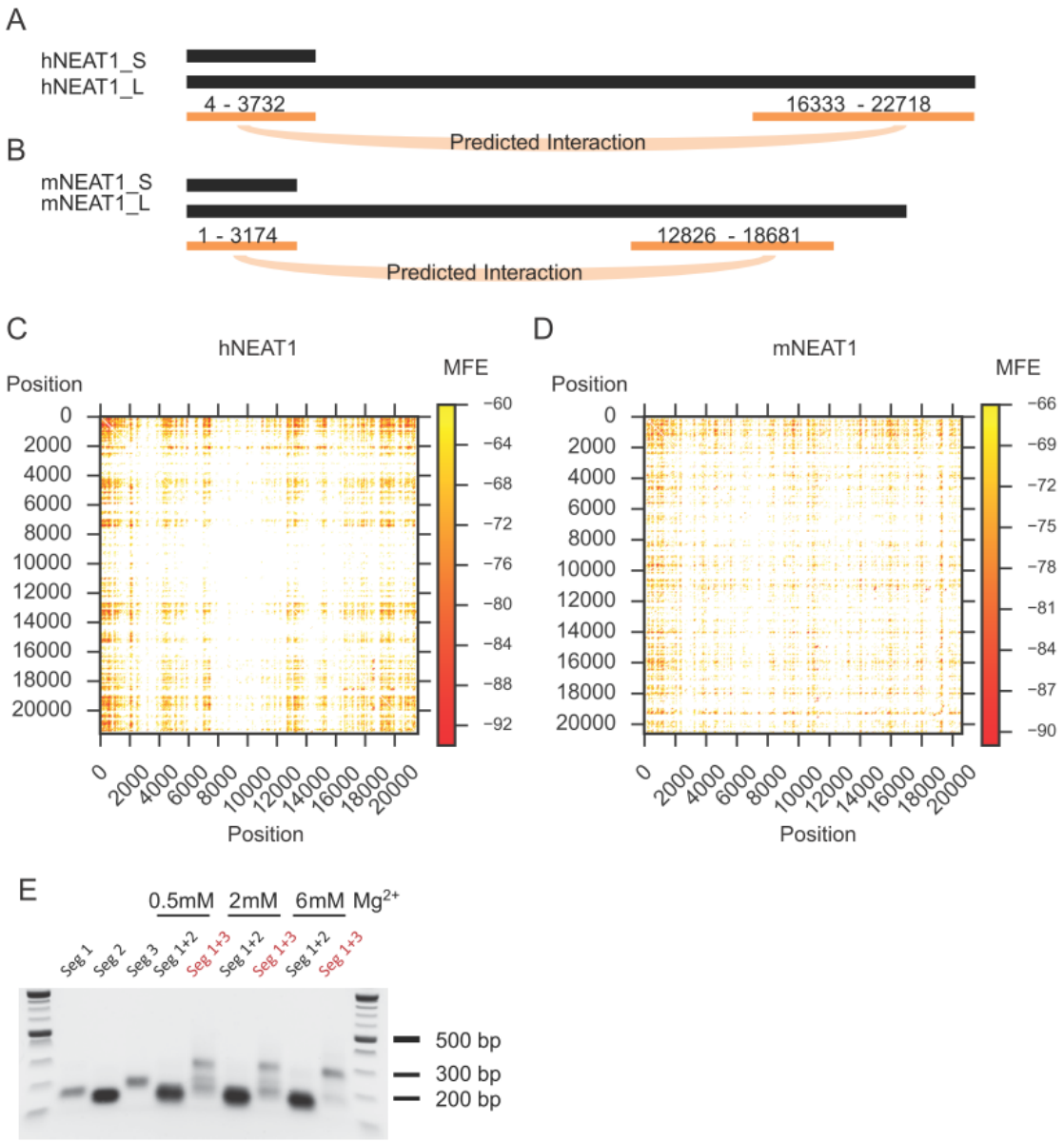


Figure 7. Putative long-range base-pairing interaction in mammalian NEAT1 RNAs. A and B). RNAduplex analyses of NEAT1_S and NEAT1_L predict NEAT1_S is likely to interact with the 3' end of NEAT1_L, in both human and mouse. C and D). RNAduplex analysis of pair-wise 120 nt window regions of NEAT1_L. The heatmaps are colored by the predicted minimum free energy of each RNA duplex. These predicted interactions are significantly stronger than expected by chance along NEAT1 RNAs in

mammals (Supplementary Figure S4). E). *In vitro* gel shift assay shows the predicted interacting RNA segments (seg 1 and seg 3) form a duplex *in vitro*. The duplex product is visible as a band that migrates similar to the 300 nt DNA ladder on the native agarose gel.

Mapping RBP binding sites on the NEAT1_S secondary structure model

A recent study by West et al. (26) investigated the localization of proteins within the paraspeckle. TARDBP was identified as a shell component that co-localizes with the NEAT1_L 3' and 5' ends, while other paraspeckle proteins such as SFPQ, NONO, FUS and PSPC1 were identified as core components expected to associate the with middle region of NEAT1_L. Public eCLIP data generated by the ENCORE project shows four significant clusters of TARDBP binding sites on NEAT1. Two sites are located within NEAT1_S, while one is in 3' end of NEAT1_L (Supplementary Figure S6 and S7). Strikingly, our predicted long-range interacting region in each of the 5' end and 3' end is adjacent to a TARDBP associated region (~40 nt apart). Thus RNA-RNA interactions and NEAT1-TARDBP interactions could act cooperatively to stabilize a NEAT1 circular scaffold within the paraspeckle (Figure 8).

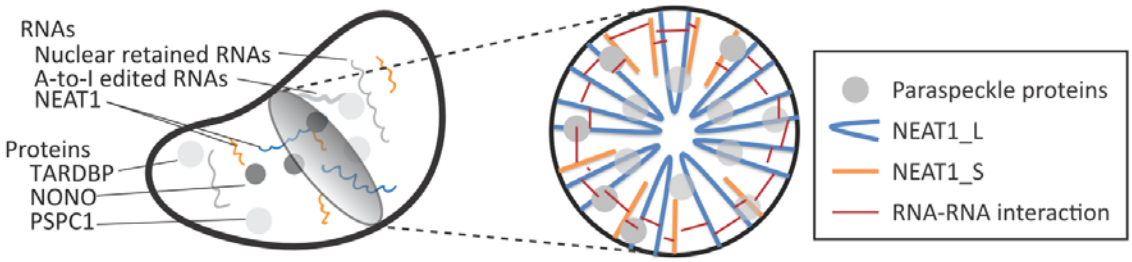


Figure 8. Model of NEAT1's architectural function in scaffolding the paraspeckle. NEAT1_L RNA folds end to end. RNA-RNA interactions between the 5' end and 3' end of NEAT1_L, or between NEAT1_S and NEAT1_L 3'end help form a circular skeleton for the paraspeckle.

We also examined the binding sites of all 160 proteins with available ENCODE eCLIP data. After stringent filtering, 50 out of 160 proteins have significant binding sites on NEAT1_L. Hierarchical clustering analyses of these binding sites are shown in (Supplementary Figure S8). Two other paraspeckle proteins, SFPQ and NONO, are clustered together. These two proteins are known to form dimers and localize to the core region of the paraspeckle, consistent with their eCLIP binding sites.

DISCUSSION

It has been an intriguing mystery that lncRNA often have very little sequence conservation even when they appear to have conserved biological functions. One hypothesis is that secondary structures, rather than primary sequences, are more likely to be conserved in lncRNA. In this study, we compared the structure of human and mouse NEAT1, the lncRNA component of paraspeckles. Our phylogenetic analyses and Mod-seq structure probing results suggest that most of the NEAT1 secondary structure is undergoing evolutionary drift, leaving only a few short regions of structural similarity and very few specific base pairs with significant covariation. Thus, secondary structure conservation alone is not sufficient to explain NEAT1's functional

conservation; other molecular interactions are likely important for scaffolding the paraspeckle.

Previous studies on the organization of NEAT1 within paraspeckles reported that the 5' and 3' ends are co-localized to the paraspeckle periphery. However, the nature of co-localization is not well understood. Our computational analyses and *in vitro* gel shift experiments suggest that the 5' and 3' ends of NEAT1 could form long-range base-pairing interactions. In the 5' end of NEAT1, the regions most likely to form such interactions (nt 282 – 546 and nt 600 – 840) flank a region of highly conserved SHAPE probing (nt 514-680). It's possible that local structures in the interacting segments may be required for long-range interactions with the 3' end of NEAT1_L. Future studies, including targeted mutation around this region, would help evaluate its role in paraspeckle formation. Since NEAT1_S and NEAT1_L share the same transcription start site, the NEAT1_S sequence is identical to the NEAT1_L 5' end sequence. Thus, our predicted intramolecular interaction between the 5' and 3' ends of NEAT1_L could also occur between separate molecules of NEAT1_S and NEAT1_L. Such interactions could form a network of RNA-RNA basepairs that help shape the architecture of the paraspeckle (Figure 8).

Recently, several groups reported high-throughput analysis of RNA-RNA interactions mapped by *in vivo* psoralen crosslinking of RNA helices (PARIS (133), LIGR-Seq (134) and SPLASH (135) methods). Notably, 435 out of 1206

base-pairs (36.1%) in our *in vitro* hNEAT1_S structure model are supported by PARIS data (133), (Supplementary Figure S9). However, only 59 out of 298 PARIS RNA-RNA interactions were also observed in our structure model. This discord likely stems from the fact that PARIS samples a population of alternative or intermediate structures, while SHAPE probing of *in vitro* transcribed NEAT1 assays a homogenous, single RNA transcript. Interestingly, the PARIS data include seven crosslink reads consistent with a long-range base-pairing interaction between the 5' and 3' ends of NEAT1_L (nt 3172-3190 and nt 21219-21264, Supplementary Figure S7). The fact that this is a very small fraction of the total mapped interactions suggests that each NEAT1 molecule may have only few intramolecular interactions in the paraspeckle. Alternatively, as NEAT1_S is expressed 5 to 8-fold more than NEAT1_L and can be localized as single-transcript "microspeckles" outside of the paraspeckle (118), the PARIS data may reflect mostly intermolecular interactions among separate NEAT1_S transcripts. Finally, the AMT psoralen used in PARIS is biased towards crosslinking U residues in adjacent AU pairs (136), such that long-range interactions involving GC pairs would be difficult to identify with PARIS. In addition, some RNA-RNA interactions supported by PARIS may require protein binding in the *in vivo* environment.

Previous work suggested that two other lncRNAs, repA and HOTAIR, have conserved secondary structure supported by co-varying nucleotides in genomic

sequence alignments (45, 113). A more recent computational analysis using R-scape (48) reported that the apparently conserved base-pairing seen in these lncRNAs was no more common than expected by chance. However, R-scape may have suffered from a lack of power due to having too few alignments of lncRNA genes. Our analyses suggest that R-Scape has the power to identify conserved base-pairs in highly structured RNAs, even when applied to a smaller number of alignments with mutation rates similar to those of lncRNAs. Furthermore, our simulations illustrate that using R2R can result in random mutations being interpreted as evidence of co-varying base pairs. Our results suggest R-scape, when properly evaluated for detection power, is an appropriate tool for analysis of lncRNA structural conservation.

As more and more genomes are sequenced, the power to identify significant covariation with tools like R-scape will increase. However, it may be wrong to assume that lncRNA structural conservation is comparable to that of deeply conserved, ancient structured RNAs like tRNA, rRNA, and RNase P RNA. Because lncRNA are relatively young (in evolutionary terms), they may not have yet evolved as many constraints on their secondary and tertiary structure. For example, tRNA must be recognized by multiple processing enzymes and synthetases, in addition to their interactions with the translation machinery, all in the space of ~ 70 nucleotides. In comparison, lncRNAs are much longer and may have fewer sequence and structural-specific interactions.

This would explain the observation that these RNAs have generally less conserved structure (48).

Our comparative structural analysis on NEAT1 serves as a case study of lncRNA structural evolution. With the exception of a few short regions, the secondary structure of NEAT1 has changed extensively over evolutionary time. Thus, the conserved function of NEAT1 cannot be explained solely by conserved secondary structure. It is possible that maintaining certain small regions of NEAT1 in single-stranded conformation, is a conserved structural feature. This is consistent with the regions of correlated SHAPE signal we observed in human and mouse NEAT1_S. In addition, there may be non-canonical RNA-RNA interactions in NEAT1 (e.g. pseudoknots) that are not accommodated by most structure modeling software. We propose a model in which a small number of short regions in the NEAT1 RNA have important specific base-pairs, while the rest remains structurally heterogeneous, allowing multiple intermolecular interactions among RNA-binding proteins and separate molecules of NEAT1 RNA.

ACKNOWLEDGEMENTS

We thank Dr. Ling-Ling Chen and Dr. Gérard Pierron for sharing plasmids encoding mouse and human NEAT1 lncRNA, Dr. Andrea Berman for sharing

plasmids encoding *Tetrahymena* ribozyme. We thank Howard Chang and Zhipeng Lu for correspondence regarding PARIS data interpretation. We also thank members of the McManus lab for helpful comments on the manuscript.

DATA ACCESSIBILITY

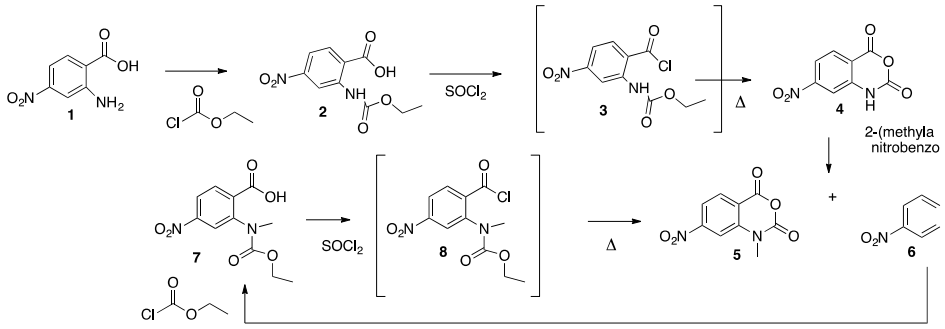
Mod-seq data have been deposited to the NCBI Sequence Read Archive, under accession number SRP128926

FUNDING

This work was supported by a grant from the Kaufman Foundation to CJM, and a financial support from the David Scaife Family Charitable Foundation to MPB.

SUPPLEMENTARY MATERIALS

1M7 Synthesis Procedure



2-((Ethoxycarbonyl)amino)-4-nitrobenzoic acid 2

2-Amino-4-nitrobenzoic acid **1** (5.46 g, 30 mmol) was dissolved in 40 mL dry dioxane. Ethyl chloroformate (12.62 mL, 132 mmol) was added under argon. The reaction mixture was reflux for 1 hr. After cooling to rt the solvent was removed under reduced pressure and the residue was suspended in 50 ml of water, suction filtered and washed with another portion of water to yield 6.86g (90%) of a light brown solid, mp 215-218°C. ^{1}H NMR (500 MHz, DMSO-d_6) δ 14.35 (s, 1H), 10.77 (s, 1H), 9.07 (d, J = 2.4 Hz, 1H), 8.18 (d, J = 8.7 Hz, 1H), 7.87 (dd, J = 8.7, 2.4 Hz, 1H), 4.21 (q, J = 7.1 Hz, 2H), 1.29 (t, J = 7.1 Hz, 3H).^{13}C NMR (126 MHz, DMSO) δ 168.73, 153.18, 150.76, 142.22, 133.24, 121.07, 116.35, 113.00, 61.86, 39.72, 14.70.

7-Nitro-1H-benzo[d][1,3]oxazine-2,4-dione 4

2-((Ethoxycarbonyl)amino)-4-nitrobenzoic acid **2** (5.08 g/20 mmol) was suspended in 10 mL thionylchloride and heated to 65 °C. Gas evolution was observed, while the reaction mixture turned from a pasty slurry to a more liquid consistency. About 15 min into heating the reaction mixture started to solidify. Heating was continued until the gas evolution stopped - about 30 min. The reaction mixture was cooled to rt. Chloroform (30 mL) was added and the solid were filtered off. The pale yellow solid was washed with 20 ml chloroform and dried to yield 2.5 g (60%) of 7-Nitro-1H-benzo[d][1,3]oxazine-2,4-dione **4**, mp 256-258 °C. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.14 (d, J = 8.1 Hz, 1H), 8.16 (dd, J = 8.6, 2.3 Hz, 1H), 7.97 (dd, J = 8.6, 2.3 Hz, 1H), 7.90 (d, J = 2.1 Hz, 1H). ^{13}C NMR (126 MHz, DMSO) δ 159.20, 152.38, 147.03, 142.55, 131.34, 117.73, 115.89, 110.68.

1M7 5 and 2-(methylamino)-4-nitrobenzoic acid 6

7-Nitro-1H-benzo[d][1,3]oxazine-2,4-dione **4** (1.04 g, 5 mmol) and anhydrous postassium carbonate (828 mg, 6 mmol) was placed in a 50 ml three neck flask. Under argon 10 mL of dry DMF as added followed by iodomethane (0.4 ml, 6.45 mmol). The reaction mixture was stirred at rt overnight. The reaction mixture was poured into ice-cold 1N HCl (50 ml). The orange precipitate was filtered off and washed sequentially with water and then ether.

NMR- analysis revealed that the product 1M7 **5** was still contaminated with the orange byproduct 2-(methylamino)-4-nitrobenzoic acid **6**. 1M7 **5** was purified by fractional crystallization from ethyl acetate/hexane where the byproduct **6** crystallizes first to yield 400 mg (40%) of orange crystals, mp mp 256-258°C. 1M7 **5** crystallizes second to yield 560 mg (50%) of light yellow crystals, mp 206- 208°C. 1M7 **5** ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.26 (d, *J* = 8.5 Hz, 1H), 8.13 (d, *J* = 2.0 Hz, 1H), 8.07 (dd, *J* = 8.5, 2.0 Hz, 1H), 3.33 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 158.36, 152.88, 147.82, 143.53, 131.60, 118.00, 117.07, 110.33, 32.58.

2-(Methylamino)-4-nitrobenzoic acid **6** ^{1}H NMR (500 MHz, MeOD) δ 8.24 (d, *J* = 8.5 Hz, 1H), 8.11 (d, *J* = 2.0 Hz, 1H), 8.06 (dd, *J* = 8.5, 2.0 Hz, 1H), 3.56 (s,3H). 2.50 (1H, NH). ^{13}C NMR (126 MHz, DMSO) δ 169.08, 152.23, 151.84, 133.72, 115.39, 108.10, 105.27, 29.78. mp 256-258°C.

The byproduct 2-Methylamino-4-nitro-benzoic acid **6** can be converted to 1M7 **5** by a two-step-one pot reaction that involves the reaction with ethyl chloroformate in dioxane to yield derivative **7** and reaction with thionyl chlorid to acid chloride **8** and subsequent cyclization to 1M7 **5**.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. The secondary structure of the *Tetrahymena* ribozyme P4P6 domain is shown, with nucleotides colored by their SHAPE reactivity scores.

Figure S2. The hNEAT1_S secondary structure model is shown with nucleotides colored by their SHAPE reactivity scores. Nucleotides higher SHAPE scores are more likely to be single-stranded.

Figure S3. The secondary structure model of mNEAT1_S is shown. Base-pairs shared between full length mNEAT1 and 3S shotgun segments are marked in red. The four identified domains are highlighted in colors.

Figure S4. Putative long-range interactions in NEAT1 are more stable than expected by chance. (A) The distribution of minimum free energy (MFE) in the sliding window analysis of hNEAT1_L (see Figure 5C and 5D) is shown in blue, while the MFE distribution from randomly shuffled 120 nt long NEAT1_L sequences is shown in green. (B) RNAduplex predictions show possible long-range interaction between 5’ and 3’ ends of NEAT1_L across mammals. Z-scores were calculated for each segment pair by comparing the actual MFE to the background null distribution (shown in A). The heat maps are colored by the predicted minimum free energy z-scores of each RNA duplex. Lower z-scores (in red) support long-range interactions in mammalian species.

Figure S5. *In vitro* gel shift assay of predicted interacting segments in hNEAT1_L and mNEAT1_L. Both hSeg 1 and 3, hSeg 2 and 3 form a duplex as predicted. The predicted mouse interacting segments (mSeg 1 and 3, mSeg 2 and 3) show only faint gel shift bands.

Figure S6. eCLIP suggested binding sites of TARDBP mapped onto the proposed secondary structure model of hNEAT1_S. Nucleotides are colored by ENCODE eCLIP signal values.

Figure S7. IGV genome browser tracks depicting regions with similar SHAPE scores between hNEAT1_S and mNEAT1_S (SHAPE), regions able to form long-range interaction verified in *in vitro* gel shift assay (Gel Shift), eCLIP identified TARDBP binding sites (TARDBP), regions show long-range crosslinking in PARIS data (PARIS), and annotations of NEAT1_S and NEAT_L.

Figure S8. Clustering analysis of RNA binding proteins' binding sites on hNEAT1. The heatmap is colored by ENCODE eCLIP signal values. eCLIP scores on each nucleotide were filtered to show only nucleotides with high signal enrichment (> 3) and statistical significance (P < 1e-5) in both available replicates.

Figure S9. Comparison of the hNEAT1_S *in vitro* SHAPE probing inferred structure with published *in vivo* PARIS data. Basepairs supported by PARIS data are colored in blue.

SUPPLEMENTARY TABLE LEGENDS

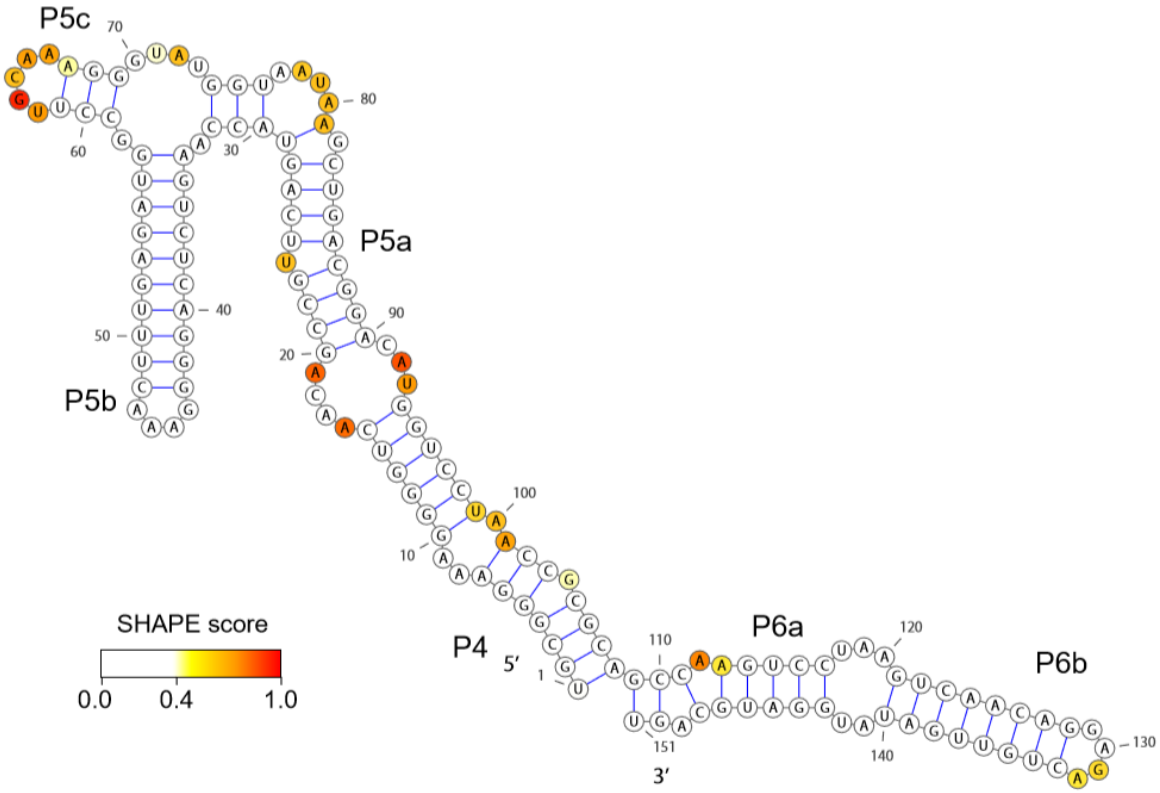
Table S1. Summary of sequencing runs.

Table S2. Positions of hNEAT1_S and mNEAT1_S segments used in the 3S shotgun method.

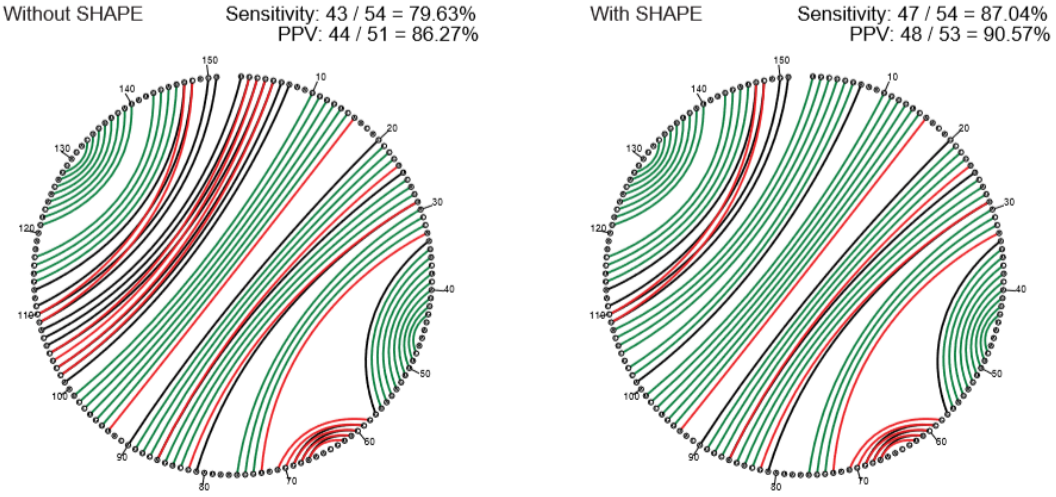
Table S3. Predicted long range interactions for 120 nucleotide long windows in hNEAT1_L and mNEAT1_L. The “Mutual Mini” column shows window pairs whose interaction provides the lowest potential free energy of all pairs involving those windows.

SUPPLEMENTARY FIGURES

Figure S1



B



Pair present in both predicted and accepted structure (Green).
Pair present in predicted structure only (Red).
Pair present in accepted structure only (Black).

Figure S2

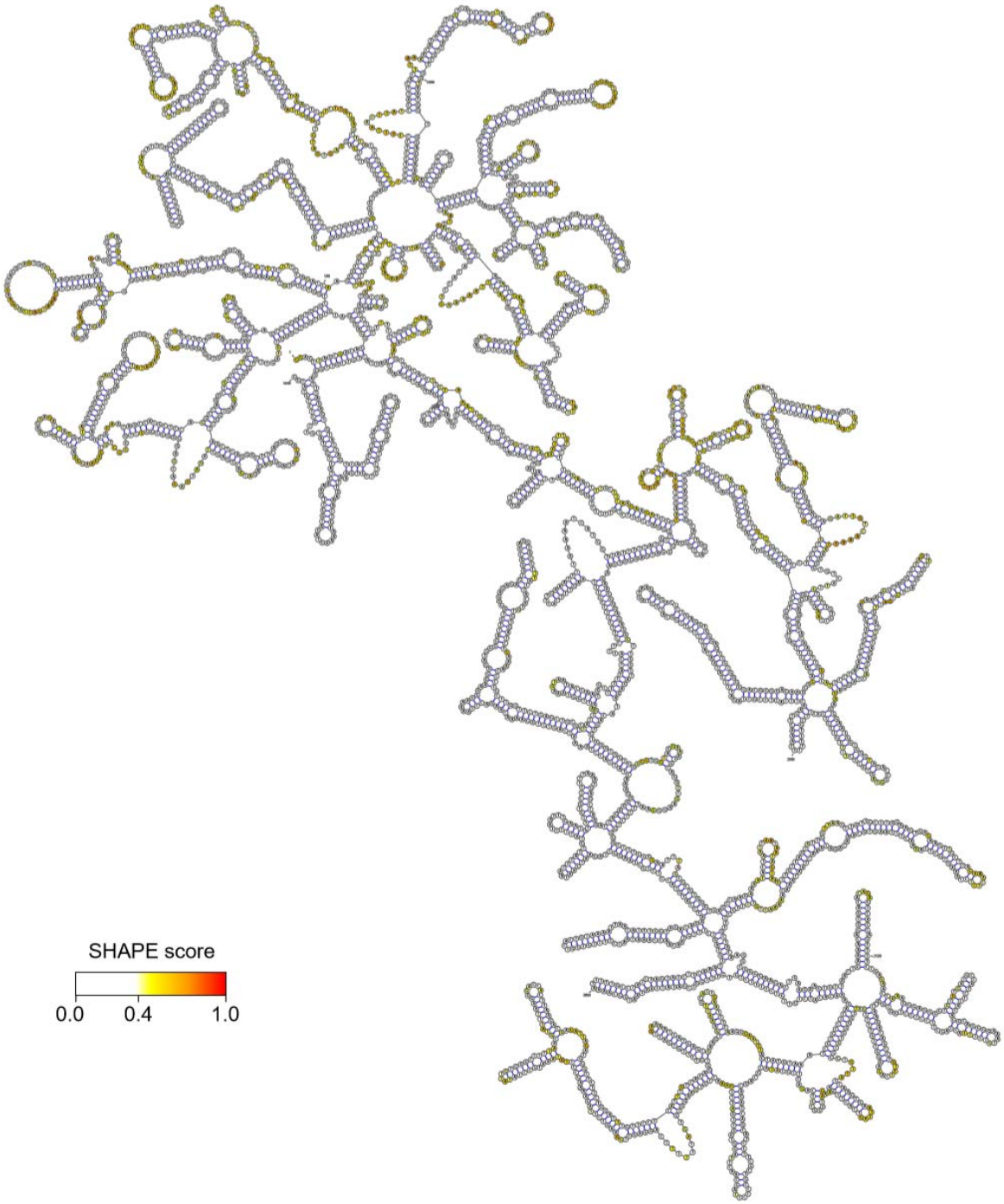


Figure S3

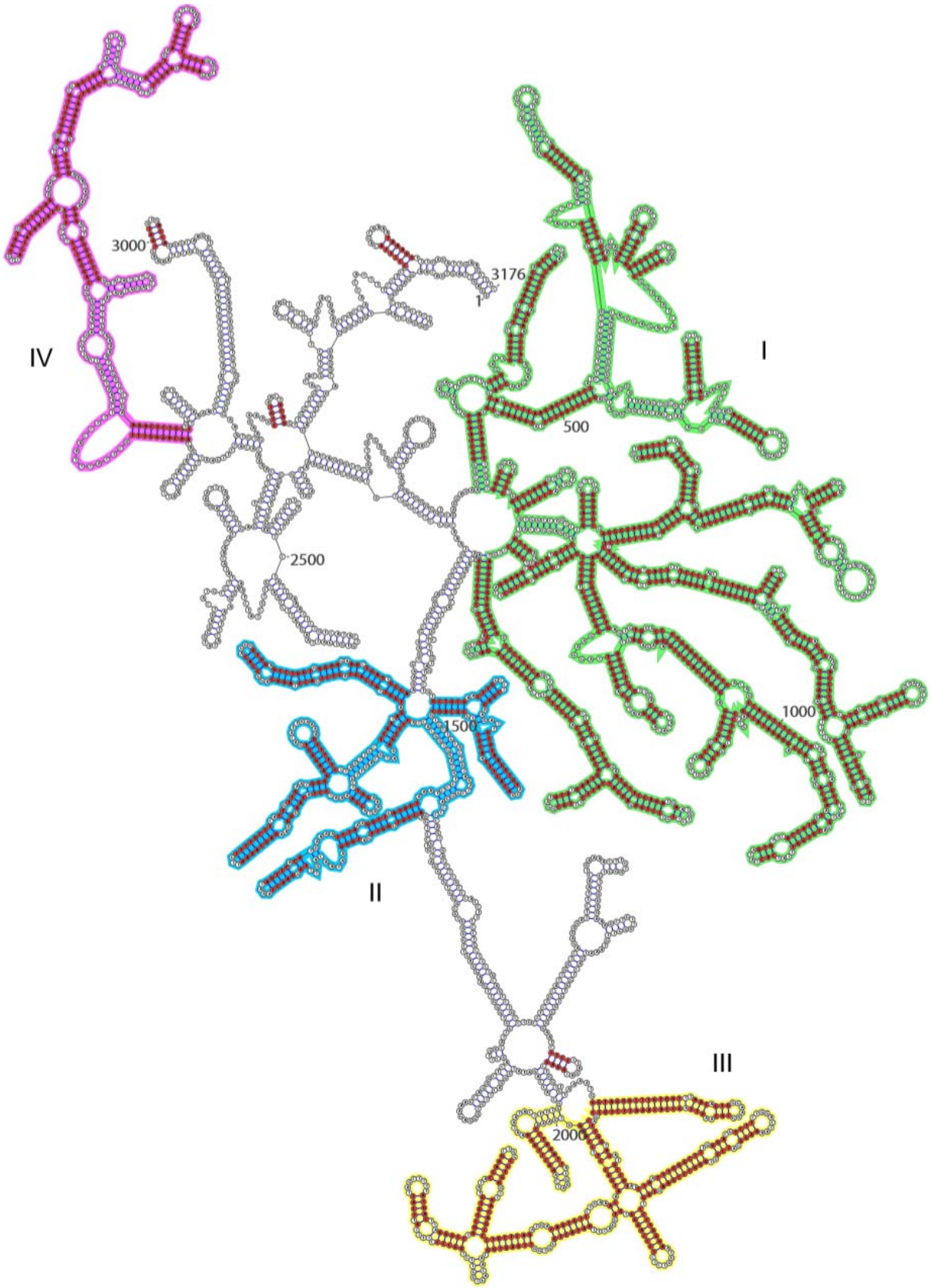


Figure S4

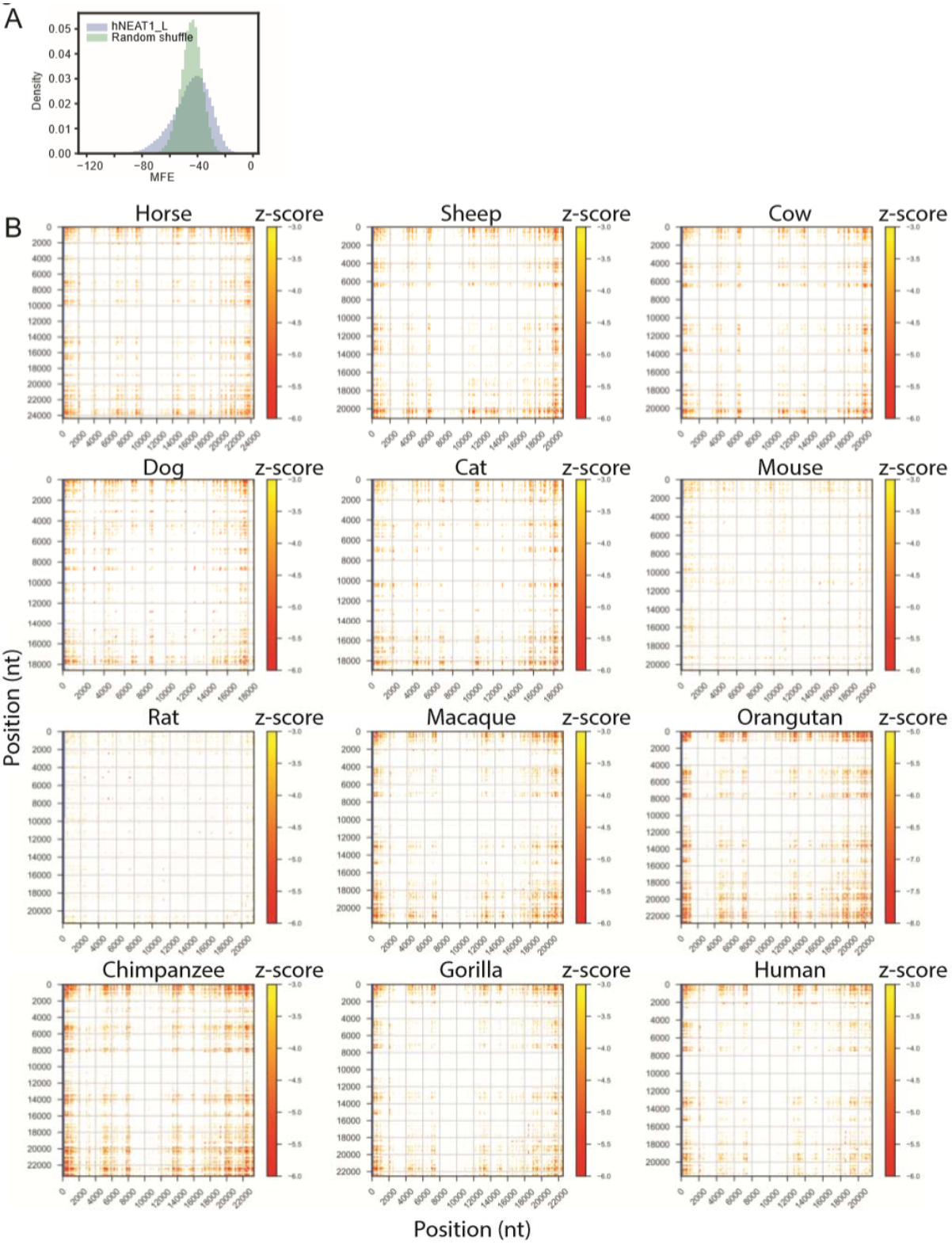

A
hNEAT1_L
Random shuffle
Density
MFE
B
Horse
Sheep
Cow
Dog
Cat
Mouse
Rat
Macaque
Orangutan
Chimpanzee
Gorilla
Human
z-score
Position (nt)
Position (nt)

Figure S5

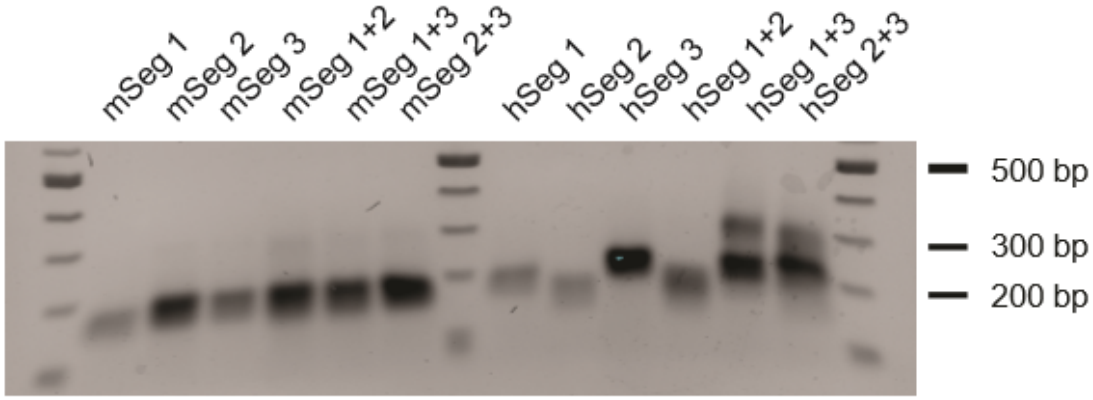

mSeg 1
mSeg 2
mSeg 3
mSeg 1+2
mSeg 1+3
mSeg 2+3
hSeg 1
hSeg 2
hSeg 3
hSeg 1+2
hSeg 1+3
hSeg 2+3
500 bp
300 bp
200 bp

Figure S6

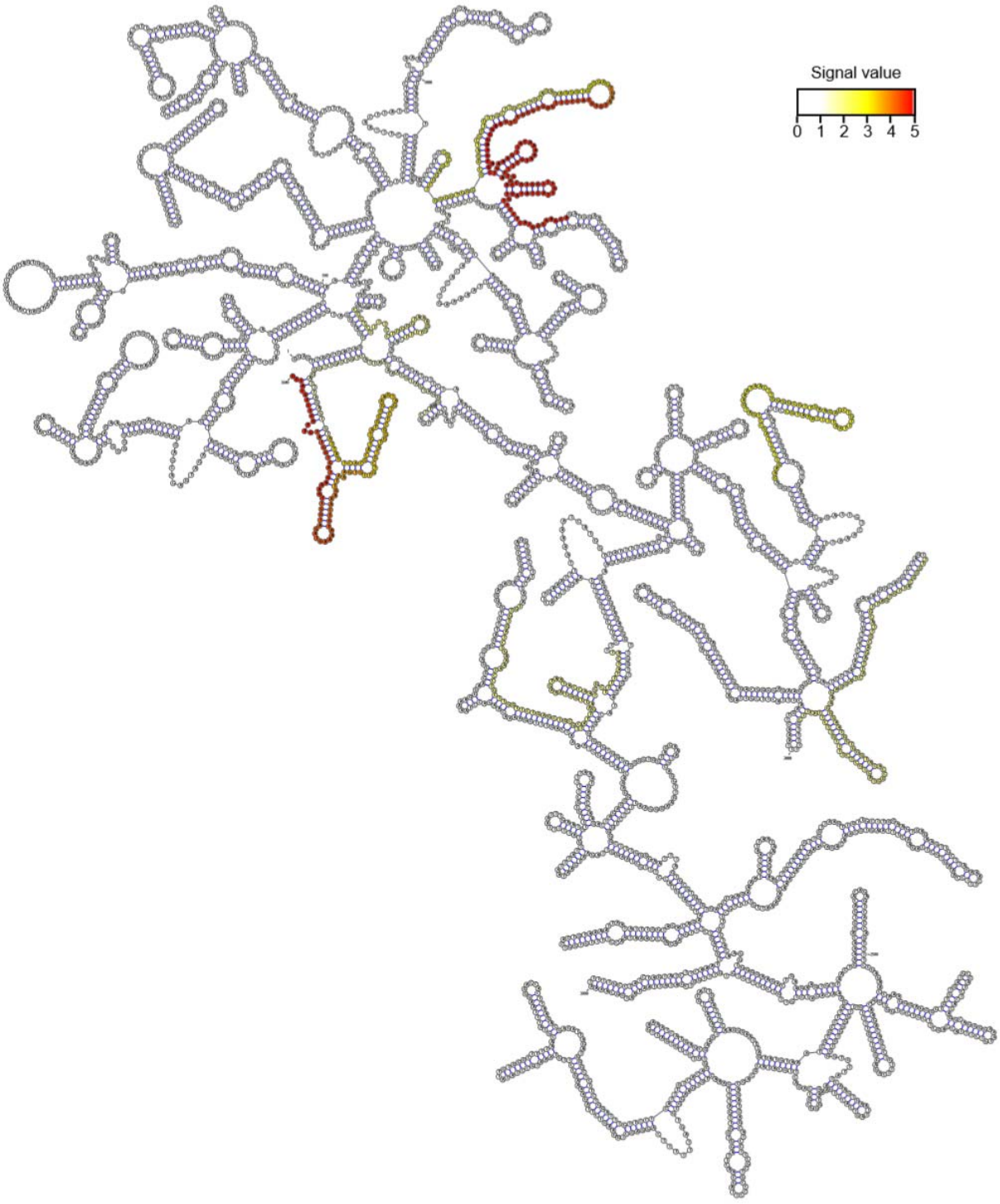


Figure S7

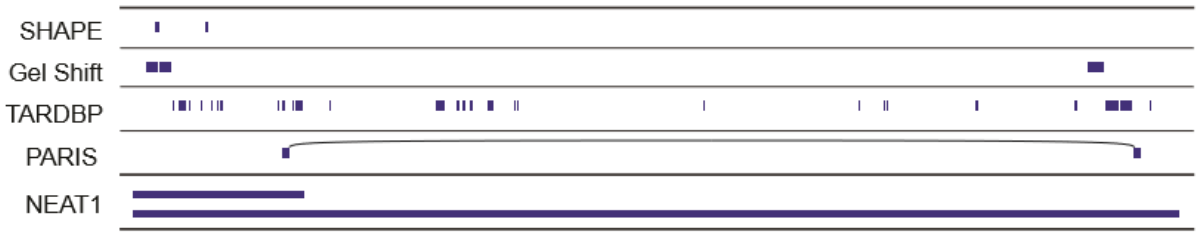


Figure S8

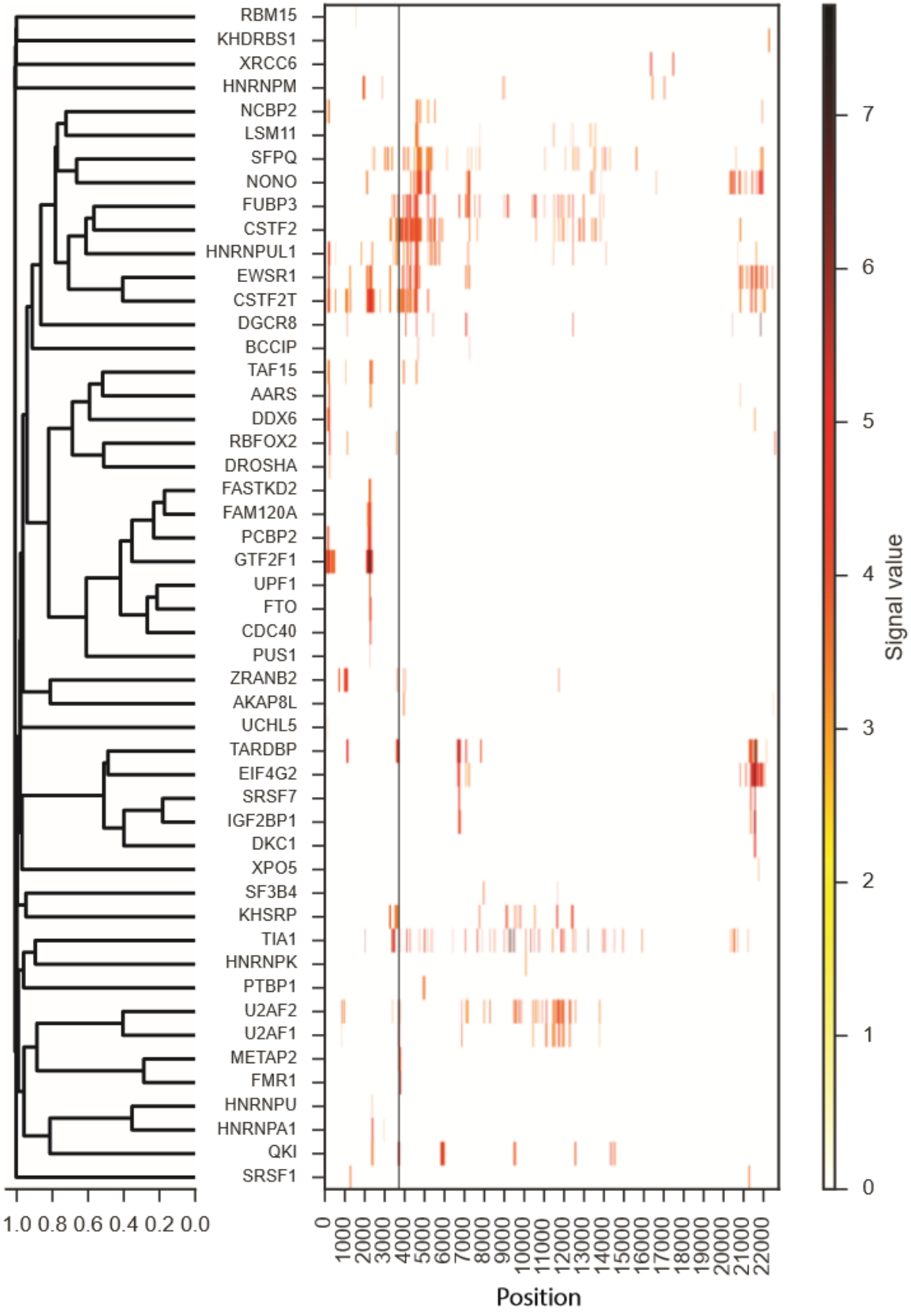
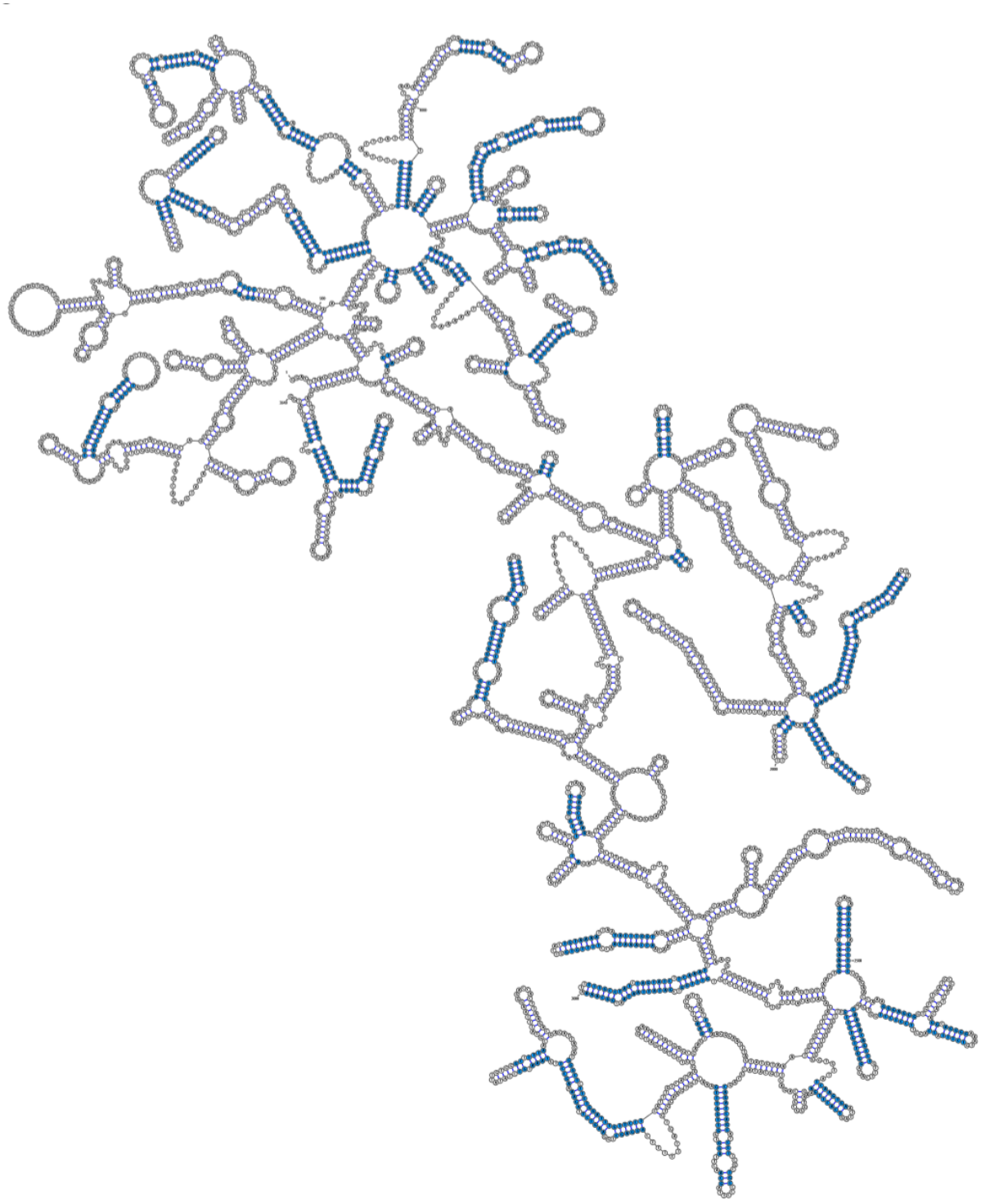

RBM15
KHDRBS1
XRCC6
HNRNPM
NCBP2
LSM11
SFPQ
NONO
FUBP3
CSTF2
HNRNPUL1
EWSR1
CSTF2T
DGCR8
BCCIP
TAF15
AARS
DDX6
RBFOX2
DROSHA
FASTKD2
FAM120A
PCBP2
GTF2F1
UPF1
FTO
CDC40
PUS1
ZRANB2
AKAP8L
UCHL5
TARDBP
EIF4G2
SRSF7
IGF2BP1
DKC1
XPO5
SF3B4
KHSRP
TIA1
HNRNPK
PTBP1
U2AF2
U2AF1
METAP2
FMR1
HNRNPU
HNRNPA1
QKI
SRSF1
1.0 0.8 0.6 0.4 0.2 0.0
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 15000 16000 17000 18000 19000 20000 21000 22000
Position
Signal value
0 1 2 3 4 5 6 7

Figure S9



Chapter 4 NEAT1 as a test case for evaluating secondary structural conservation of lncRNAs.

This chapter expands on the phylogenetic analyses of NEAT1 structure published in Lin et al., 2018, which is included in the supplementary material.

There is an ongoing debate in the field regarding the conservation of lncRNA structure. Only a few lncRNA secondary structures have been chemically probed, including Xist (44, 111), HOTAIR (45), ncSRA (137) lincRNAp21 (46) and RepA (113). In these studies, the authors often claimed that the probed lncRNA has a well-defined secondary structure that shows a significant degree of phylogenetic conservation via the presence of covarying base pairs. However, the covariation analyses performed in these studies generally lacked statistical control. In 2016, another study developed a statistical method (R-scape) testing for significant covarying base-pairs (48). Using R-scape, they suggested that there is no statistically significant conservation for the previously probed lncRNA. Although their results appear convincing, it is possible that R-scape is too conservative. In my study of the NEAT1 lncRNA structure, I used multiple evolutionary simulations and analytical permutations to evaluate the performance of R-scape and other

methods. My results suggest that R-scape, when properly calibrated, is the preferred tool for analysis of lncRNA secondary structural conservation.

4.1 Calibrating NEAT1 structural alignments with Infernal

We used phylogenetic analyses to investigate the conservation of the NEAT1_S structure. We first used Infernal (124, 138) to generate improved mammalian multiple alignments of NEAT1_S using our SHAPE constrained structure model. Infernal uses a small set of reliable sequence alignments, along with a secondary structural model in the alignment region as inputs to build a covariance model (CM) for RNA consensus sequence and structure. This CM model can then be used to calibrate alignments for more distant species. As it is possible that only small subdomains of NEAT1_S have conserved structure, we applied Infernal to compact helical regions from the domains defined using the 3S shotgun procedure (132) (see Chapter 3, methods). Alignments of NEAT1 sequences from 8 species (including hNEAT1 and mNEAT1) were used with these secondary structure models to train Infernal CMs. The calibrated CMs were then used to search against 64 mammalian sequences that have alignment around NEAT1 regions. For 12 of 14 subdomains, Infernal identified at least 40 out of 64 mammalian species with significant alignment to human NEAT1_S. Two regions in domain III (nt

2470-2609 and nt 3199-3316) had only 12 and 25 alignments, respectively, and the former one only had alignments within primates (Table 4).

4.2 R2R is likely to introduce false positives when identifying covariant base pairs

We investigated two methods to evaluate the conservation of NEAT1_S secondary structure - R2R (126) and R-scape (48). R2R was used in previous studies that reported compensatory changes in lncRNA structure models. R2R classifies a base-pair as covarying if at least one compensatory mutation is present in an alignment, given there are fewer mutations that disrupt pairing at that position than a user-defined threshold (15% in (45, 113)). R-scape uses a background null distribution to identify statistically significant covariant base-pairs. It was reported previously that some lncRNAs have covariant base-pairs identified by R2R, but many of them failed the statistical tests in R-scape. Thus, R2R may be too liberal and / or R-scape too conservative for analysis of NEAT1_S structural conservation. To compare these two methods, we first applied both to the Infernal-calibrated NEAT1_S alignment. As expected, R2R identified more covariant base-pairs of consistent half-flips than R-scape (Figure 9 and Figure 10, summarized in Table 4).

hNEAT_S position	mNEAT_S position	length	seq Identity	nbpairs	Nseq (infernal)	pearsonr	p-val
21-322	22-308	319	69.0%	86	53	-0.06	0.36
323-501	309-468	183	68.0%	50	53	0.14	0.08
514-680	481-614	170	79.0%	61	53	0.43	2.7E-07
687-899	601-787	214	68.0%	64	54	0.11	0.14
901-1036	793-925	148	67.0%	45	52	0.32	0.00031
1037-1268	925-1145	232	81.0%	77	57	0.22	0.001
1269-1467	1146-1321	208	75.0%	59	51	0.25	0.001
1710-1833	1570-1692	131	68.0%	37	43	0.35	0.00012
1858-2116	1712-2103	300	47.0%	91	44	-0.06	0.38
2290-2434	No alignment			39	41		
2470-2609	No alignment			47	12		
2610-2949	2315-2622	365	36.0%	111	47	0.083	0.21
3199-3316	No alignment			35	25		
3529-3638	No alignment			38	44		

Table 4. (continued on next page)

hNEAT_S position	Rscape E<0.05	%Rscape E<0.05	R2R conserv	R2R halfflip	R2R covaraint	%R2R conserve	%R2R halfflip	%R2R covaraint	%R2R all
21-322	3	3.49%	0	3	14	0.0%	3.5%	16.3%	19.8%
323-501	4	8.00%	1	2	8	2.0%	4.0%	16.0%	22.0%
514-680	0	0.00%	2	8	12	3.3%	13.1%	19.7%	36.1%
687-899	1	1.56%	1	7	11	1.6%	10.9%	17.2%	29.7%
901-1036	2	4.44%	0	5	12	0.0%	11.1%	26.7%	37.8%
1037-1268	0	0.00%	9	9	20	11.7%	11.7%	26.0%	49.4%
1269-1467	1	1.69%	8	12	12	13.6%	20.3%	20.3%	54.2%
1710-1833	0	0.00%	1	3	4	2.7%	8.1%	10.8%	21.6%
1858-2116	1	1.10%	0	7	11	0.0%	7.7%	12.1%	19.8%
2290-2434	0	0.00%	1	9	3	2.6%	23.1%	7.7%	33.3%
2470-2609	0	0.00%	17	7	6	36.2%	14.9%	12.8%	63.8%
2610-2949	1	0.90%	1	6	21	0.9%	5.4%	18.9%	25.2%
3199-3316	0	0.00%	1	5	7	2.9%	14.3%	20.0%	37.1%
3529-3638	0	0.00%	1	6	8	2.6%	15.8%	21.1%	39.5%

Table 4. Summary of conservation analysis in regions in NEAT1_S.

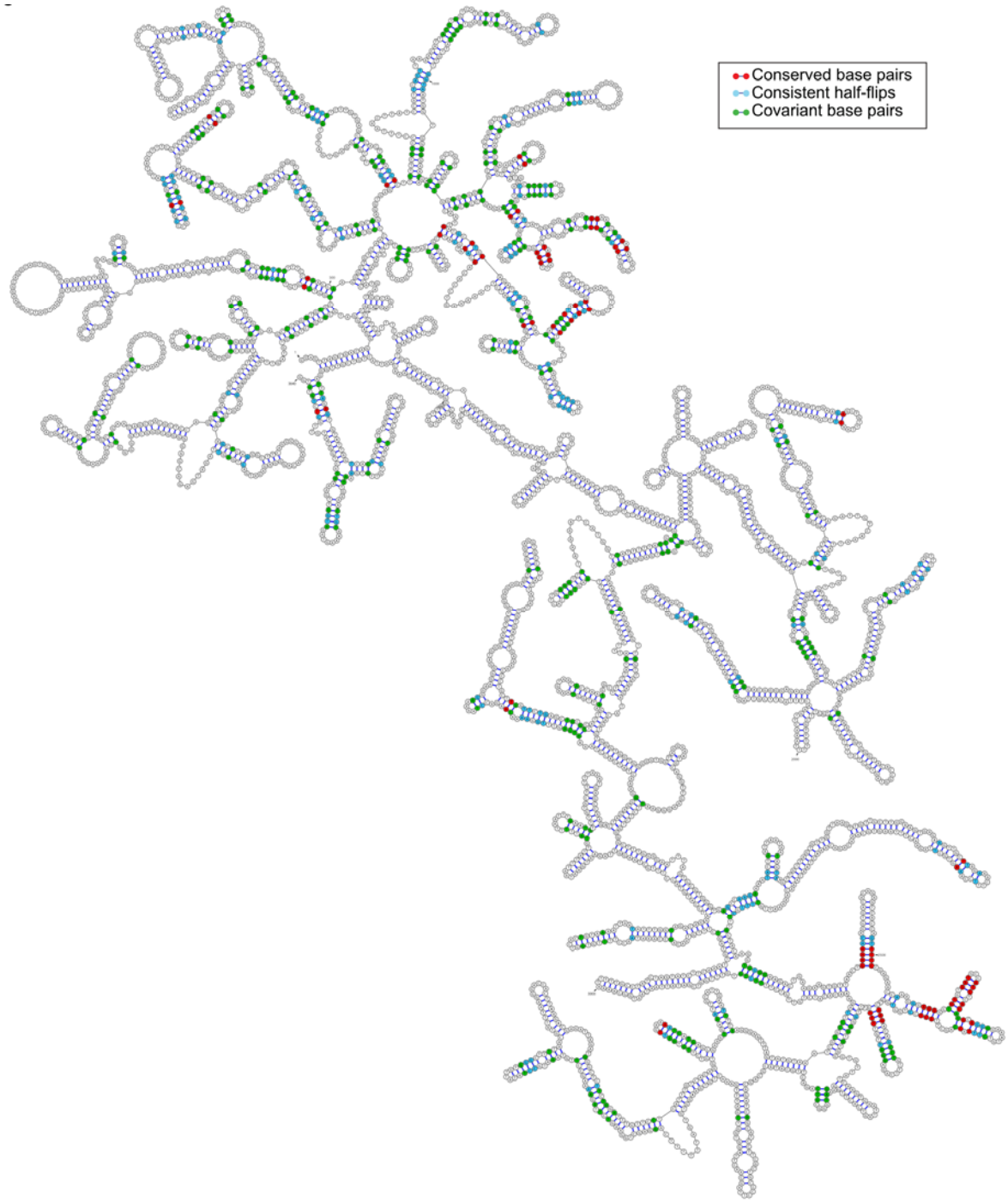


Figure 9. The secondary structure of hNEAT1_S with base-pairs colored by R2R results

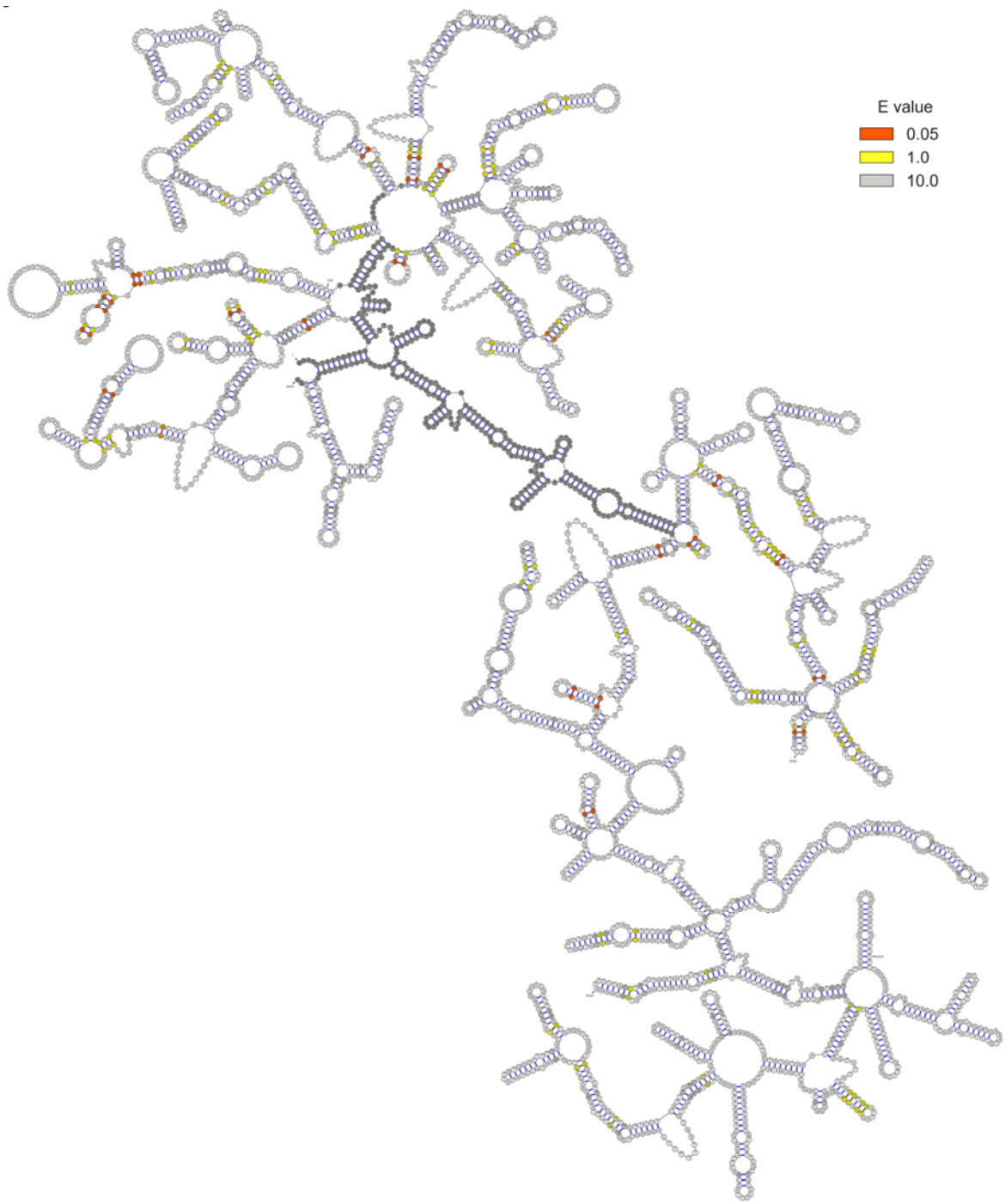


Figure 10. The secondary structure model of hNEAT1_S, with base-pairs colored by R-scape E-values.

To determine which method is preferable, I randomly generated 100 synthetic NEAT1_S sequence alignments for each Infernal-aligned region of NEAT1, using mutation frequencies seen in actual mammalian NEAT1_S alignments (random null models) (Figure 11 A and B). The distribution of pairwise identity in actual NEAT1_S alignment is shown in Figure 11C. When applying R2R (15% threshold) on the random null alignments, we observed a notable amount (~10-20%) of false-positive covariant base pairs (Figure 11 D and E). In contrast, R-scape did not support compensatory mutations from the random null alignment (Figure 11 F and G). These results suggest the R2R approach is prone to false-positive covariation calls.

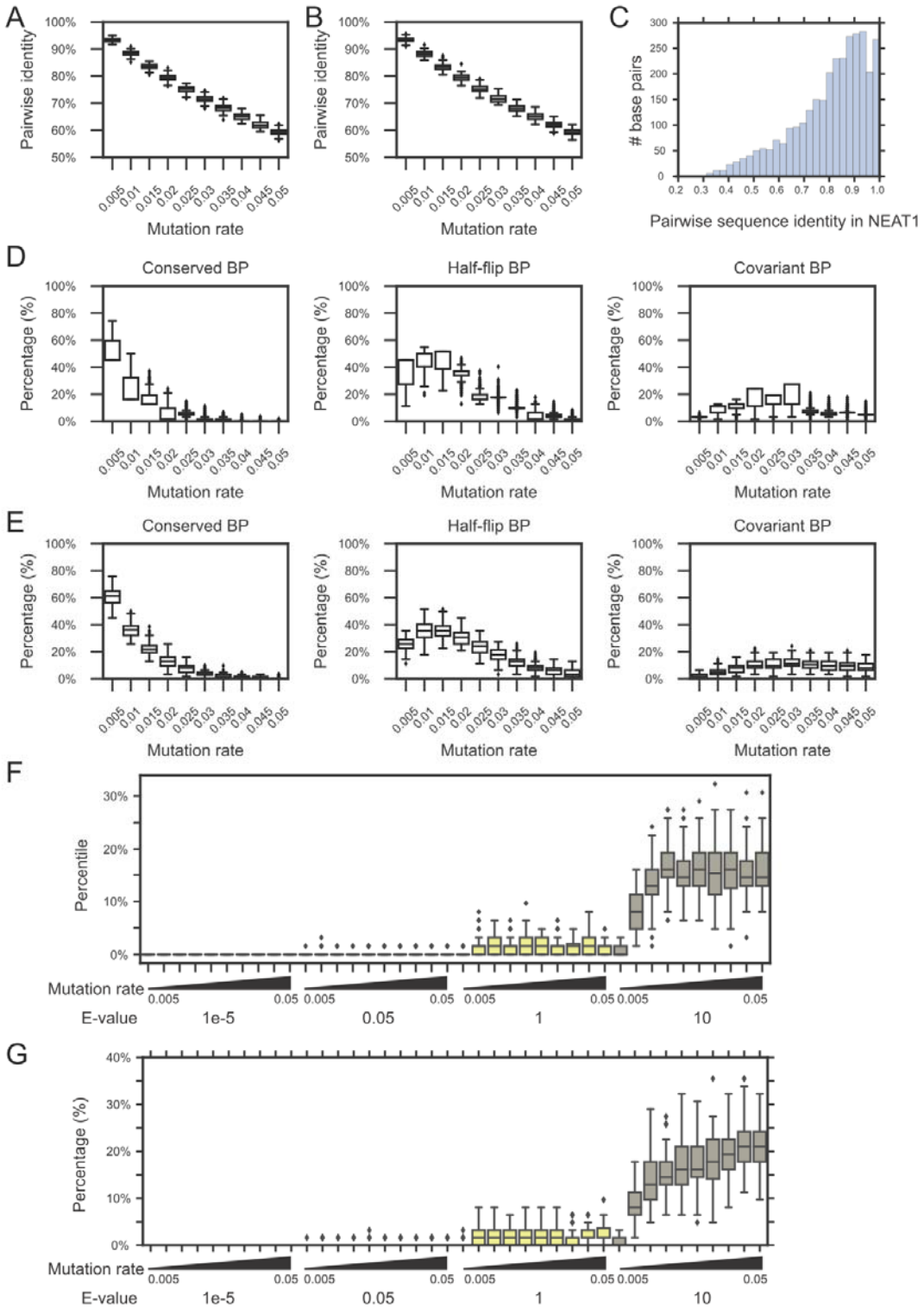


Figure 11. Comparison of the propensity for false-positives using R2R and R-scape using synthetic alignments generated by random mutation. (A) Pairwise identity of synthetic NEAT1_S alignments with different mutation

rates. The box plots show the distribution of average percent identity (y-axis) for 100 synthetic alignments of the hNEAT1_S region 514-680 made at each mutation rate (x-axis, ranging from 0.5% to 5%) (B) Pairwise identity of synthetic NEAT1_S alignments after calibrated with Infernal. (C) The histogram shows the actual distribution of pairwise identity in the Infernal improved alignments of hNEAT1_S region 514-680. (D) R2R analysis (with 15% non-canonical threshold) on synthetic alignments of the hNEAT1_S region 514-680. Boxplots show the distribution of conserved (left), half-flip (middle, e.g. AU->GU pair), and covariant (right) base pair calls by R2R for 100 synthetic alignments at each mutation rate. With the 15% threshold, R2R has a high false positive rate for half-flips and covariant pairs in alignments with pairwise identity ranging from ~65% to ~90% (E) R2R analysis on synthetic alignments of the hNEAT1_S region 514-680, after calibration with Infernal (F) R-scape analysis on synthetic alignments. R-scape identifies very few false positives (G) R-scape analysis on synthetic alignments after calibration with Infernal.

4.3 R-scape suggests NEAT1 is less conserved than other highly structured RNAs

I next examined whether R-scape is too conservative. R-scape is sensitive to the number of sequences in alignments and their pairwise sequence identity, such that the power to detect covariation increases when more mutations are present in the alignment. To evaluate if R-scape could detect conserved structure in our NEAT1_S alignments, I compared R-scape results from NEAT1_S to those of known well-structured RNAs (tRNA, 5S rRNA, riboswitches, telomerase RNA, etc.). I subsampled alignments of these RNAs to similar alignment number (~53) and average pairwise identity (~68%) to NEAT1_S alignments. R-scape identified many (20-65%) conserved base pairs in well-structured RNAs after alignment subsampling (Figure 12). Thus, R-scape has adequate power to detect highly conserved secondary structures when provided alignments similar to our NEAT1_S alignment. NEAT1_S alignments had higher co-variation scores than random null alignments (Figure 13), however NEAT1_S had relatively fewer significant covariant base pairs (E value < 0.05; Figure 10, Figure 12 and Table 4). These results suggest that NEAT1_S is under less selective pressure for specific RNA structures than well-known highly-structured RNAs.

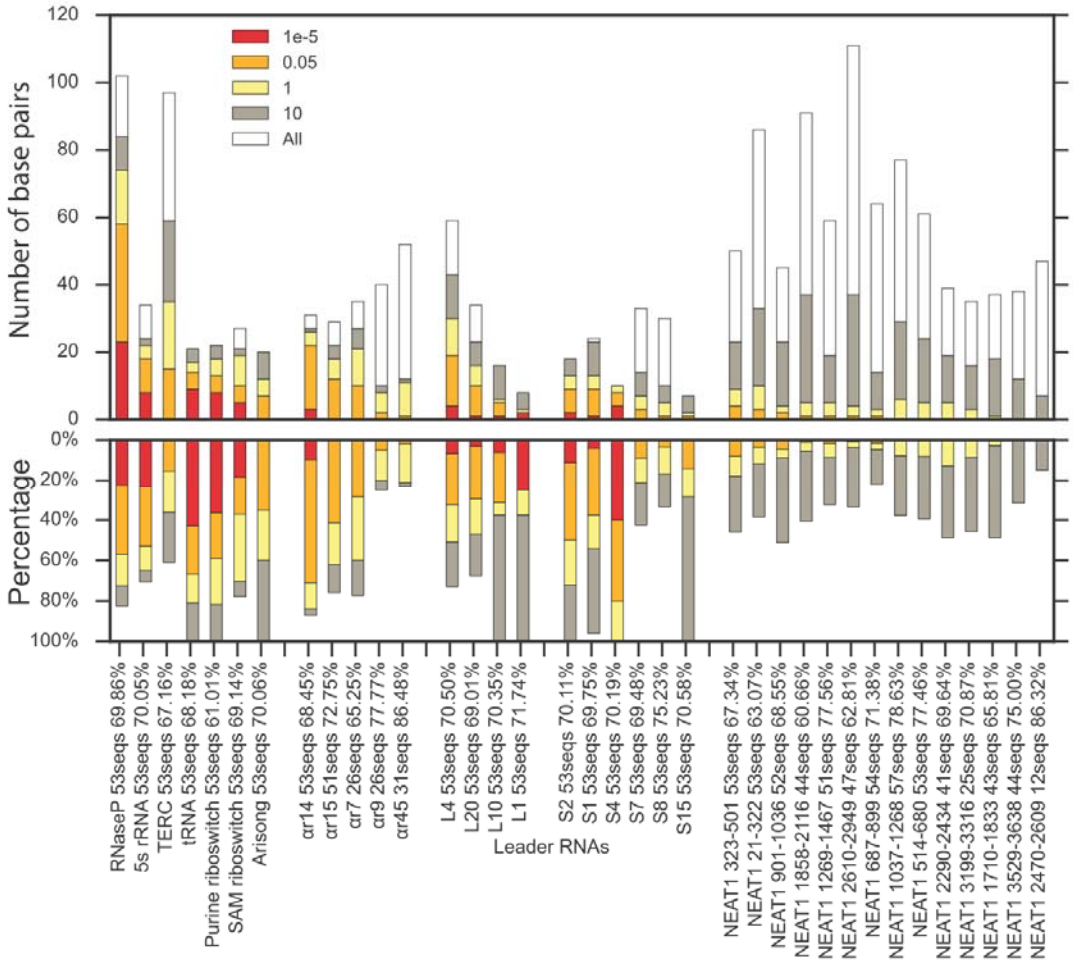


Figure 12. Comparison of R-scape results for Infernal aligned regions of NEAT1_S and known well-structured RNAs. The alignment number of sequences ("seqs") and percent identity are shown for each RNA on the X axis. The sequence alignments of other RNAs were subsampled to have similar numbers of sequences (e.g. RNaseP RNA "53-seqs") and percent sequence identity (e.g. RNaseP RNA, 69.86%) as the NEAT1_S alignment. The bar plots shows counts (upper panel) and percentage (lower panel) of significantly covariant base-pairs with E-values smaller than 1e-5, 0.05, 1, and 10. NEAT1_S has relatively little evidence for co-varying base pairs

compared with most well-structured RNAs (e.g. tRNA, 5S rRNA, TERC RNA, etc).

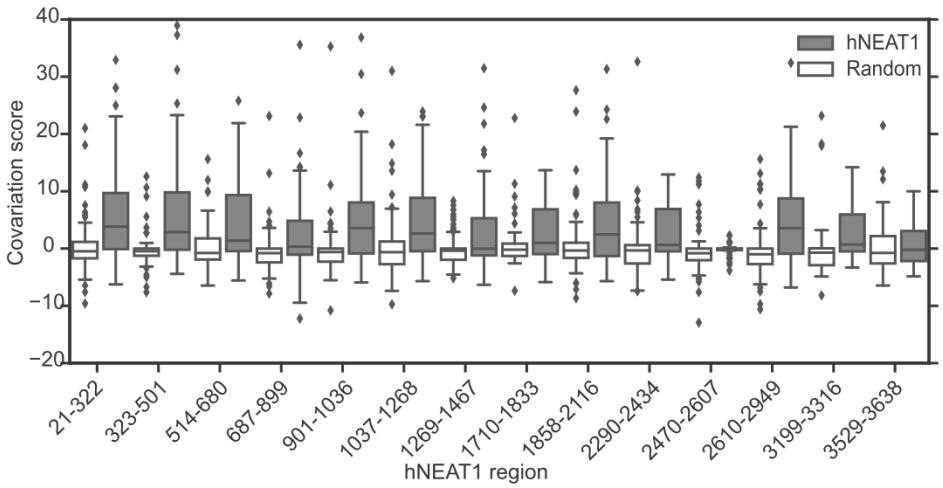


Figure 13. Covariation scores from R-scape analysis on proposed hNEAT1_S secondary structure model using real (non-synthetic) alignments. Most regions of hNEAT1_S have somewhat higher covariation scores than those of random alignments, suggesting a weak tendency towards covariation.

4.4 RNAz suggests low level of conservation in NEAT1_L

As an alternative, we also investigated structural conservation using RNAz (139). RNAz differs from R-scape or R2R that it is not designed for identify covariant base pairs; instead it scans for conserved and thermodynamically stable regions from a multiple sequence alignment. An RNA secondary structure model is not required as input in RNAz analyses. RNAz is also capable of genome-wide screening with an integrated sliding window approach. Thus, we chose to use RNAz to investigate the structural conservation potential in NEAT1_L. RNAz is only compatible with alignments containing 6 or less sequences. Thus, we only used seven closely related primate sequences in RNAz analysis. Only 1 region (nt 513 - 712) out of 80 sliding windows was detected to have conserved and thermostable secondary structure from this analysis, and only 12 out of 520 sliding window regions (probability > 95%) were detected in the long isoform, NEAT1_L (Table 5). Consistent with R-scape analysis, these results indicate that NEAT1_L has low secondary structure conservation even among primates.

RNAz has its limitations in that it screens for conserved, potentially functional non-coding RNA in small sliding windows. Thus, it is impossible for it to detect any long-range RNA-RNA interactions that likely exist in NEAT1_L. Also, it is designed for detecting RNA structures with higher thermodynamical

stability given the same nucleotide composition, which may not be a suitable criterion for lncRNAs that have flexible structures.

4.5 Discussion

In this chapter I explored different computational methods to evaluate the conservation level of NEAT1 RNA secondary structure. Previous studies reported controversial conclusions regarding structural conservation of lncRNAs. Several studies (45, 46, 113) used R2R (126) to identify covariant base pairs in lncRNA secondary structure models and discovered several structurally conserved regions. However, Rivas et al. (48) argued that there is no evidence for structural conservation in these lncRNAs when using R-scape for statistical tests. I first evaluated these two covariant base pairs calling approaches, R2R and R-scape, on simulated multiple sequence alignment. The result suggests R2R is likely to introduce false positives while R-scape showed very low false positive rates with E-value cutoff of 0.05. R-scape, on the other hand, is sensitive to alignment depth and pairwise sequence identity, thus may not give fair comparison when comparing lncRNA to other small structured RNAs. Our analyses showed that even with subsampling to remove these biases, NEAT1_S is still less conserved than small structured RNAs such as tRNA and riboswitches. Consistent with R-scape, RNAz analysis also suggests

a dearth of conserved, thermostable structures in NEAT1_L. However, all these methods are designed to identify specific conserved base-pairs or rigid secondary structures. It is likely that lncRNA such as NEAT1 has alternative structural conservation in the form of conserved single-stranded regions, or conserved long-range RNA-RNA interactions that cannot be identified through these methods. This possibility was further discussed in chapter 3.

This chapter is focused on evaluating the structural conservation of NEAT1. It is very possible that other lncRNAs have different conservation features than NEAT1. Future investigations about other lncRNA structures are necessary in order to for us to understand lncRNA structure conservation comprehensively. lncRNA is a relatively young RNA species in evolution. Many human lncRNAs only have identified orthologs in mammals, and the comparative analysis of lncRNA structure conservation is limited by the number of species with available genomic sequencing data. With the increasing number of sequenced species, we will be able to have a better understanding of lncRNA structure conservation in the future.

Segment	1	2	3	4	5	6	7	8	9
Start(hNEAT1)	513	-	4022	10996	17472	17689	17954	19534	21677
End (hNEAT1)	-	712	4182	11105	17687	17794	18063	19654	21796
Start(aln)	560	600	4200	10720	23200	25560	25840	30000	30975
End(aln)	680	720	4320	10840	23320	25680	25960	30120	31095
Length	120	120	120	120	120	113	120	120	120
Mean pairwise identity	97.83	98	97.67	96.44	78.7	90.98	84.58	95.05	95.68
Mean single sequence MFE	-73.45	-63.98	-33.95	-32.28	-30.16	-32.18	-36.3	-63.58	-46.75
Consensus MFE	-72.89	-59.6	-32.35	-32.28	-22.5	-29.65	-28.82	-56.98	-45.23
Energy contribution	-73	-59.93	-31.8	-31.7	-22.53	-30.65	-28.82	-57.57	-45.53
Covariance contribution	0.11	0.33	-0.55	-0.58	0.03	1	0	0.58	0.31
Combinations/Pair	1.05	1	1.09	1.13	1.05	1	1.12	1.02	1.05
Mean z-score	-3.57	-2.4	-2.49	-1.71	-2.08	-2.55	-2.48	-2.53	-2.18
Structure conservation index	0.99	0.93	0.95	1	0.75	0.92	0.79	0.9	0.97
SVM decision value	3.26	1.75	2.06	1.61	1.8	2.28	1.99	1.89	2
SVM RNA-class probability	99.8%	96.6%	98.1%	95.5%	96.8%	98.7%	97.8%	97.4%	97.9%
GC content	70.4%	69.3%	38.6%	37.1%	46.7%	46.3%	51.0%	65.4%	53.4%

Table 5. Summarize of RNAz identified structural conserved region in NEAT1.

Chapter 5 Discussion and future directions

5.1 Conclusions and discussion

This thesis work is focused on using chemical probing, high-throughput sequencing, and computational methods to study the structure of long noncoding RNAs, a newly identified RNA family that has diverse regulatory functions on gene expression regulation. The development of high-throughput chemical probing methods for RNA secondary structure measurement (e.g. Mod-seq), made it possible for the first time, to probe the structure of large in a fast and accurate manner. Compared to traditional gel-based probing methods, mod-seq not only has much higher throughput, but also generates quantitative data that allows reduced background and higher signal-to-noise ratios with proper data processing and analysis. The first part of my thesis involved developing an automated data analysis pipeline for Mod-seq data analysis, Mod-seeker. This is one of the first publicly available data analysis packages for high-throughput sequencing-based RNA secondary chemical probing data. By making Mod-seeker open source, we hope it will be easier for other researchers to switch from traditional, gel-based probing method to Mod-seq.

We then applied Mod-seq to determine the secondary structure of one specific lncRNA, NEAT1. NEAT1 RNA is a particularly interesting candidate for lncRNA structural study, for it is an essential scaffolding RNA molecule for paraspeckle formation. Previous studies showed that NEAT1 has a highly organized spatial composition in the paraspeckle, implying its structure may be important for its function. We successfully generated a secondary structure model of the 3640 nt full-length “short” isoform of NEAT1. By using the 3S shotgun approach, we found local stable structures in NEAT1_S, and identified four compact structural domains in NEAT1_S. Notably, Li et al. (118) recently discovered that NEAT1_S alone, without the long isoform, is substantially located outside of paraspeckle, forming “microspeckles”. The secondary structure model of NEAT1_S is potentially useful for understanding microspeckle functions. Further computational prediction suggests the long isoform NEAT1 tends to form long-range RNA-RNA interactions between its 5’ and 3’ ends. This long-range interaction was validated *in vitro* using a gel-shift assay and shown to be conserved across mammals. We believe that RNA-RNA long-range interactions within and among NEAT1 molecules, together with RNA-protein interactions, may work cooperatively to form the highly organized paraspeckle architecture.

NEAT1 is also used as a case study for understanding structural conservation in lncRNAs. The extent to which lncRNA have conserved

structures is an on-going debate. Several previous studies probed the secondary structures of some lncRNAs, including HOTAIR, Xist, lincRNAp21 etc., and claimed that these RNAs showed structural conservation. However, in a later publication, a statistical method (R-scape) was developed and applied to these probed lncRNAs, suggesting no statistically significant evidence for covariant base-pairs in these RNAs. In my work, I carefully evaluated two approaches for covariant base-pair identification, R2R and R-scape. By testing on simulated alignments, I found that R2R is likely to introduce false positives in covariant base-pair identification while R-scape does not have this problem due to its stringent statistical controls. On the other hand, R-scape is sensitive to alignment depth and pairwise sequence identity between alignments, which may lead to biases in evaluating conservation level of lncRNAs. Our analysis showed that even after subsampling to correct for this kind of bias, NEAT1 still shows much less covariant base pairs than highly structured small RNAs. RNAz analysis on NEAT1_L also shows few regions with conserved, thermostable structure.

However, lacking evidence of covariant base pairs or thermostable structures does not necessarily means there is no conservation in NEAT1 structure. It is possible than lncRNAs have high-level structural conservation, such as the conserved long-range RNA-RNA interaction tendency in NEAT1 shown in this chapter 3, that does not require specific base-pair conservation.

A recent publication by E. M. Langdon et al. (140) showed that the structure of mRNA can contribute to RNA-RNA interactions, promoting liquid-liquid phase separation to build membraneless compartment in cells. In their proposed model, CLN3 mRNA has regions that can hybridize with BNI1 mRNA, but these regions have low SHAPE reactivity under native condition, and are only exposed when CLN3 is melted. In this example, RNA structural flexibility is crucial to regulate RNA sorting into distinct droplets. As the key structural component in paraspeckle, a membraneless nuclear compartment, it is likely that NEAT1 functions in a similar way through its secondary structure and long-range RNA-RNA interaction capability, but do not have many covariant base pairs that suggesting selection for rigid structures.

5.2 Future directions

In vitro structure probing has its merits in that it interrogates an RNA's inherent folding potential without interference by alternative transcript isoforms or RNA-binding proteins that may obscure interpretation of chemical modification patterns. Nonetheless, *in vivo* structure probing can be complementary to reveal the *in vivo* conformation of lncRNAs. Comparing the *in vitro* chemical modification patterns to the *in vivo* ones can also help to reveal potential protein binding sites in lncRNAs. Several studies have attempted to

perform transcriptome-wide *in vivo* probing of RNAs (68, 120, 141), but obtaining reliable structural information of lncRNAs is challenging due to their relatively low expression levels (9) compared to mRNAs. For a nuclear lncRNA such as NEAT1, this is even more difficult since it takes more time for SHAPE reagents to diffuse into the cell nucleus and compacted, phase-separated nuclear bodies. Recently, Takeshi Chujo et al. (97) reported an improved RNA extraction method for NEAT1, which can increase NEAT1_L extraction by 20-fold. This can potentially be helpful for studying the *in vivo* structure of NEAT1 when combined with *in vivo* structural probing.

Current studies suggest that lncRNA adopts complex alternative structure conformations (133), making it difficult to interpret *in vivo* structure probing patterns. One possible solution for this is combing the SHAPE-MaP (79) method with long-read RNA sequencing techniques, such as PacBio or Nanopore sequencing. In SHAPE-MaP, chemical modification sites are detected by introducing point mutations during the reverse transcription step, instead of introducing RT stops. In principle, this allows the detection of multiple chemical modification signals in a single RNA molecule (79). If SHAPE-MaP can be successfully combined with long read sequencing, we would be able to obtain structural probing information for each lncRNA molecule. This will also be very informative for us to understand the diversity and flexibility in lncRNA structures.

In our NEAT1 structure analysis, we identified several short regions in short NEAT1 with higher conservation levels. We also identified regions that may be involved in long-range RNA-RNA interactions in NEAT1. These regions are potentially important for the functions of NEAT1. Introducing mutations or deletions in these regions using CRISPR/Cas genome editing would be helpful for testing the roles of these structural elements in NEAT1's paraspeckle scaffolding function.

Our investigation of the conservation of the NEAT1 structure suggests that there are few covariant base pairs in NEAT1. However, there are other structural features that are conserved such as conserved single-stranded regions or conserved long-range RNA-RNA interactions. Whether such features are also conserved in other lncRNAs remains to be determined. Also, although using covariant base-pairs as an indicator for conservation is suitable for small RNAs that have compact, stable structures, it may not be ideal regarding lncRNA conservation. Other statistical methods are needed to evaluate the conservation level of lncRNAs that may have more flexible structures.

Appendix

A1. *In vitro* structure probing of sno-lncRNA2

sno-lncRNAs are found to be deleted in an important human disease, Prader-Willi Syndrome (PWS) (22). Deep sequencing analysis shows that these lncRNAs are produced from introns with two imbedded snoRNA genes. snoRNA (small nuceolar RNA) are a class of small noncoding RNA that primarily guide chemical modification, including methylation or pseudouridylation of ribosomal RNAs (142). Sno-lncRNAs are processed by the snoRNA machinery. The sequences between the snoRNAs are not degraded during processing, which leads to the accumulation of lncRNAs flanked by snoRNA sequences at both 3' and 5' ends without 5' cap and 3' poly(A) tails. Sno-lncRNAs are located in nucleus and tend to accumulate near their sites of synthesis. There is no evidence that sno-lncRNAs are precursors of snoRNAs, instead, the relatively highly structured snoRNA found at both ends may contribute to the stabilization of sno-lncRNA. Computational prediction shows that there are multiple potential Fox family splicing regulator binding sites within sno-lncRNAs, suggesting that sno-lncRNAs may be involved in alternative splicing regulation by Fox protein sequestration.

Given that sno-lncRNA2 is a scaffolding RNA molecule for Fox family proteins, we suspect that it may also adopt specific secondary structures. A similar *in vitro* 1M7 SHAPE probing experiment was conducted to study the secondary structure of sno-lncRNA2.

Methods

The Sno-lncRNA2 plasmid is from Ling-Ling Chen lab (22). To obtain linear DNA templates for *in vitro* transcription, sno-lncRNA2 plasmids are cut by corresponding restriction enzymes. *In vitro* transcription is conducted using Promega RiboMAX™ Large Scale RNA Production System-T7 kit. The RNA is then treated with Proteinase K and cleaned up with Amicon® Ultra-0.5 centrifugal filter devices to remove proteins such as polymerases. For 1M7 probing, RNA samples are incubated with 65mM 1M7 for 70 s (or the same volume solvent, DMSO, as negative control). Mod-seq method was conducted as previous described (5, 6).

Results

The predicted sno-lncRNA structure model is shown in Figure 14, colored with normalized SHAPE reactivity scores. SHAPE-aided structure model of sno-lncRNA2 has only 36% similarity when comparing to the sequence–only predicted structure without SHAPE data (Figure 15). We found that all four fox2

binding sites are in a partial single stranded partial double stranded conformation, resembling an internal loop. This suggests there is a potential conserved secondary structure motif for fox2 binding.

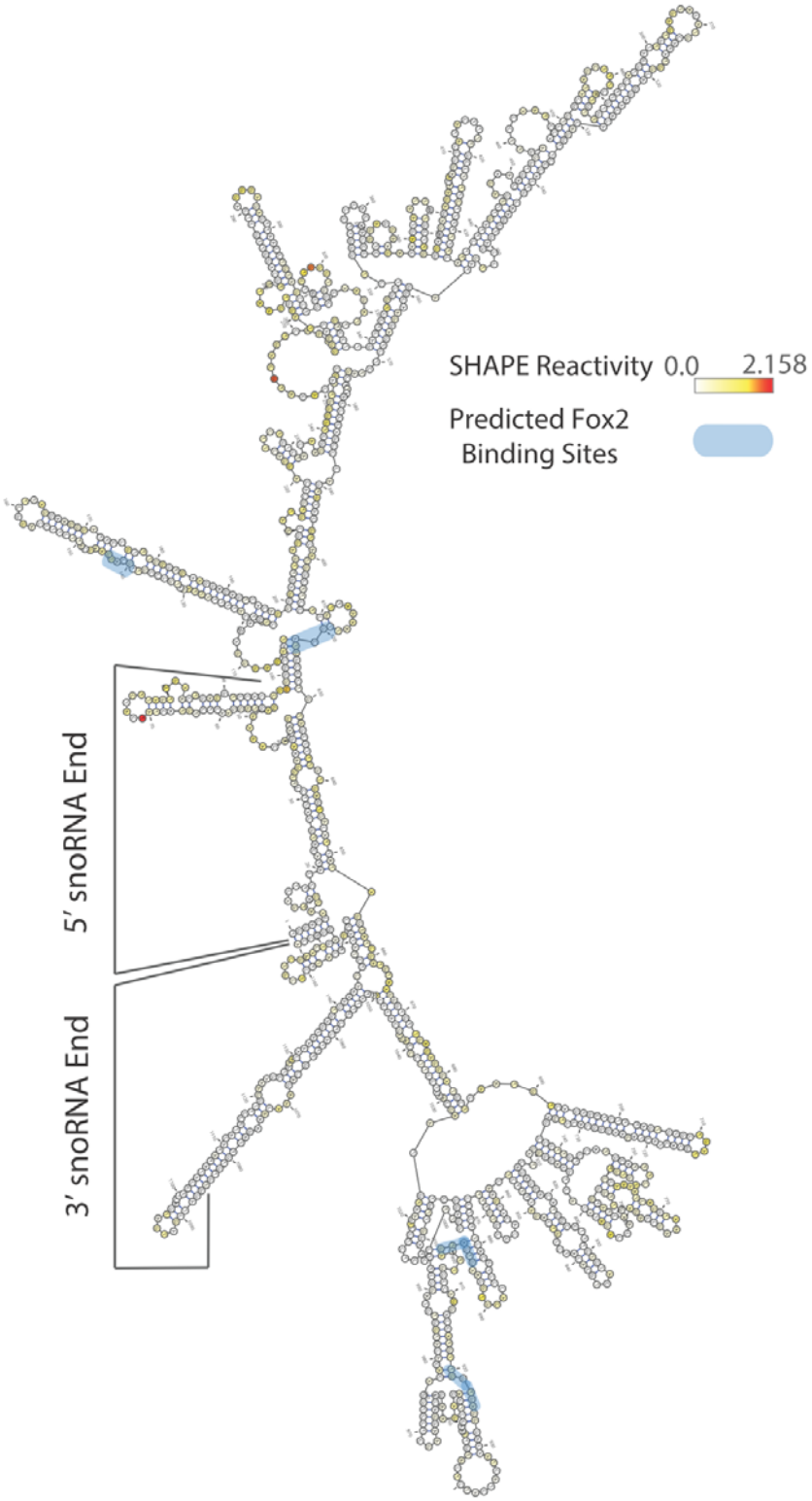


Figure 14 Secondary structure model of sno-lncRNA2 predicted with Mod-seq data. Nucleotides are colored with SHAPE reactivity scores, predicted Fox2 binding sites are marked with blue.

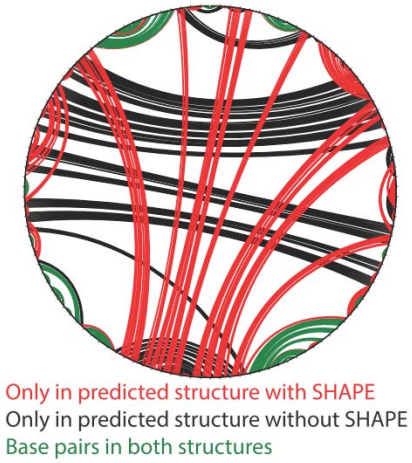


Figure 15 Circle plot comparison of predicted sno-lncRNA2 secondary structure with and without Mod-seq data.

Discussion and future direction

In this study we generated the secondary structure model of sno-lncRNA2 with 1M7 SHAPE probing data. There are several other RNAs in the sno-lncRNA family (sno-lncRNA1,3,4,5), and all have sno-RNA sequence at their 3' and 5' ends, but have different lengths and different sequences in between the sno-RNA ends. Probing the secondary structure of the other sno-lncRNAs will be helpful to discover similar structure features in this RNA family. Also, probing the snoRNAs individually will be helpful to investigate whether the snoRNA ends in sno-lncRNA fold independently or have altered structures. To further study the interaction between sno-lncRNAs and Fox2 proteins, structure probing experiments can be done in the presence of Fox2 protein to see if Fox2 binding will protect the predicted binding sites from SHAPE modification, or if Fox2 binding will alter sno-lncRNA secondary structure.

A2. *In vivo* NAI probing of human lncRNA in K562 cells

In contrast with traditional RNA secondary structure probing methods that use denaturing PAGE to detect chemical modification sites, the Mod-seq method allows high-throughput structure probing of RNAs of any length, making it possible to probe transcriptome-wide RNA structure simultaneously. In early 2014, there were several papers that came out using high-throughput sequencing based chemical probing approaches to study transcriptome-wide RNA structure of yeast, human and Arabidopsis. However, these studies are focused on the structure of mRNA, which is abundant in cells and relatively easier to probe, while the structures of lncRNAs are masked due to their low expression levels compared to mRNAs. In order to reveal lncRNA structure, we applied nuclei isolation and ribosomal RNA removal to reduce rRNA, mRNA and enrich for lncRNA.

Method

For *in vivo* structure probing, 1X107 K562 cells were treated with 100mM NAI (or 10% DMSO for negative control) in 1XPBS for 15 min at 37 ºC. Isolation of cell nuclei was then conducted using the PARIS™ Kit (Ambion®). Nuclear RNA was extracted using TRIzol, and treated with TURBO DNase to destroy genomic DNA. We also did *ex vivo* structure probing, where RNA was isolated

in the same way then treated with the same concentration of NAI. Finally, ribosomal RNA was removed using Ribo-Zero Gold Kit (Epicentre). Sequencing libraries were prepared using Mod-seq method and sequenced with illumina MiSeq.

Result

As shown in Figure 16A, pre-rRNA is enriched in isolated nuclear fraction, absent in cytoplasmic fraction, indicating isolation of nuclei is clean. Also, after using the RiboZero kit, rRNAs are no longer detectable on the TapeStation gel, indicating that most rRNA are successfully removed. Figure 16B is the fluorescent primer extension on nuclear 5S rRNA, where NAI modifications can be observed, suggesting NAI is permeable to the cell nucleus and is able to modify nuclear RNA.

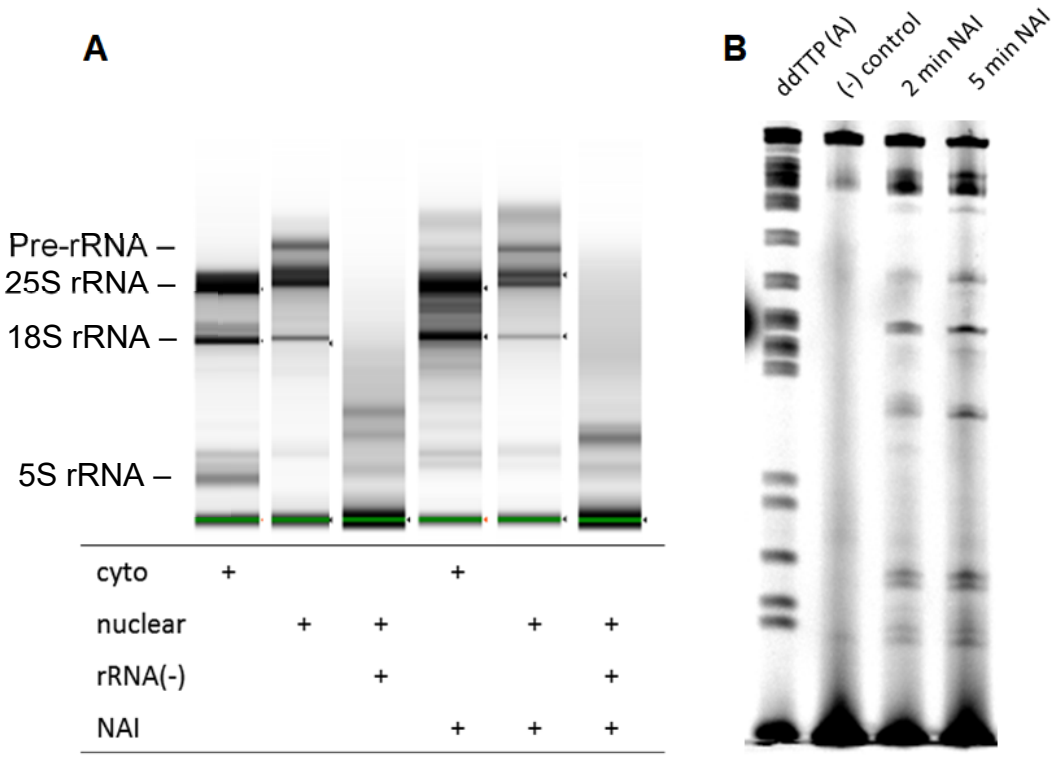


Figure 16. lncRNA enrichment and *in vivo* NAI probing of nuclear RNA. A) Pre-ribosomal RNA is enriched in the nuclear fraction of RNA, indicating successful nuclei isolation. Ribosomal removal process successfully removed most rRNAs. B) Fluorescent primer extension experiment on nuclear 5s rRNA. NAI modification sites can be observed on denaturing PAGE, indicating NAI is cell nuclear permeable and can modify nuclear RNA.

From the pilot MiSeq sequencing, around 7 million reads were aligned to reference human genome. Most of them are aligned to snoRNA and snRNA, only a few aligned to lncRNAs. Read coverage on most lncRNAs was not deep enough to reveal their secondary structure. However, for some abundant ncRNAs, such as RNase P, we were able to obtain enough reads for structure analysis. As shown Figure 17, the structure pattern of *in vivo* and *ex vivo* probed mitochondrial RNase P can be observed and distinguished from the negative controls. Also, the *ex vivo* structure of mitochondrial RNase P is similar to its *in vivo* structure, but more positions are modified moderately with lower SHAPE scores, probably revealing single-stranded regions that are protected by proteins *in vivo*.

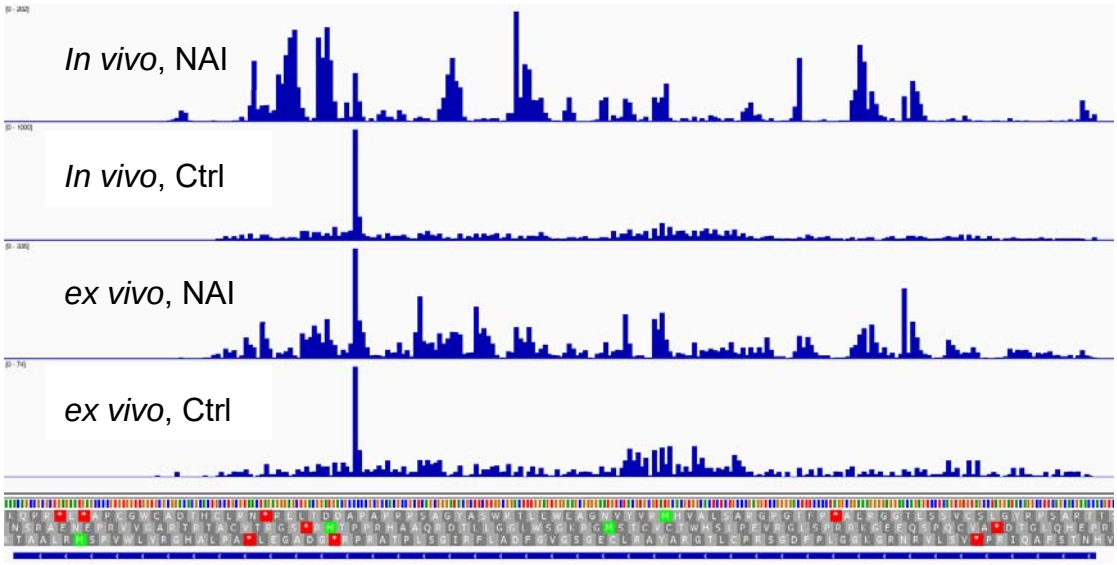


Figure 17. Mod-seq result of *in vivo* and *ex vivo* probed mitochondrial RNase P. The structure pattern of *in vivo* and *ex vivo* probed mitochondrial RNase P can be observed and distinguished from the negative controls. Also, the *ex vivo* structure of mitochondrial RNase P is similar to its *in vivo* structure.

A3. *In vivo* human mRNA structure probing with inhibition of translation elongation

Over the last 2 years, several papers reported transcriptome-wide mRNA secondary structure probing of human mRNAs, using similar high-throughput sequencing protocols combined with either DMS or SHAPE modifications (68, 75, 120, 143). mRNAs are reported as not well-structured, under active unfolding *in vivo*; their structural signatures are also associated with protein binding and structures may change under stress conditions. However, it is unknown whether this lack-of-structure is caused by ribosome binding and scanning or is directly encoded in mRNA sequences.

Method

To investigate this problem, human K562 cells were treated with Harringtonine (HT), a drug that can inhibit translation elongation and leads to ribosomes' accumulation around translation initiation sites (144). After 5 min HT treatment, K562 cells were then incubated with NAI for structure probing.

Result

If the HT treatment successfully inhibited translation, we expected to see protection from NAI modification around translation initiation sites due to

ribosome binding. If the lack-of-structure of mRNA is due to ribosome binding and the unfolding during translation, we expected to see the 5’ ends of mRNA coding regions are better folded in HT treated samples, since during 5 min HT treatment, new-coming ribosomes should be stalled around initiation sites, while those were in elongation state before HT addition should continue elongation, resulting in a ribosome free region at the 5’ end of mRNA coding regions.

The Mod-seq probing data of human transcriptome turned out to be noisy and difficult to interpret. The Pearson correlation of mod-score between 2 replicates are usually lower than 0.5 for HT (-) samples, even lower for HT (+) samples, making it difficult to compare the difference between HT (+) and HT (-). This low correlation is consistent with previous reports stating that mRNAs are not structured, and the even low correlation between HT (+) sample could be evidence for lack-of-structure of ribosome-free mRNAs.

Previous studies also reported three-nucleotide periodicity of secondary structure of mRNAs. We asked if we could confirm this in our data. As shown in Figure 18, we observed three-nucleotide periodicity the in normalized mod-score; but also in NAI (-) sample, which means this periodicity may not directly related to mRNA’s secondary structures. In fact, we observed a strong three-nucleotide periodicity in the frequency of each of the four nucleotides in coding

sequencing, which could contribute to the three-nucleotide periodicity in observed in secondary structure probing data.

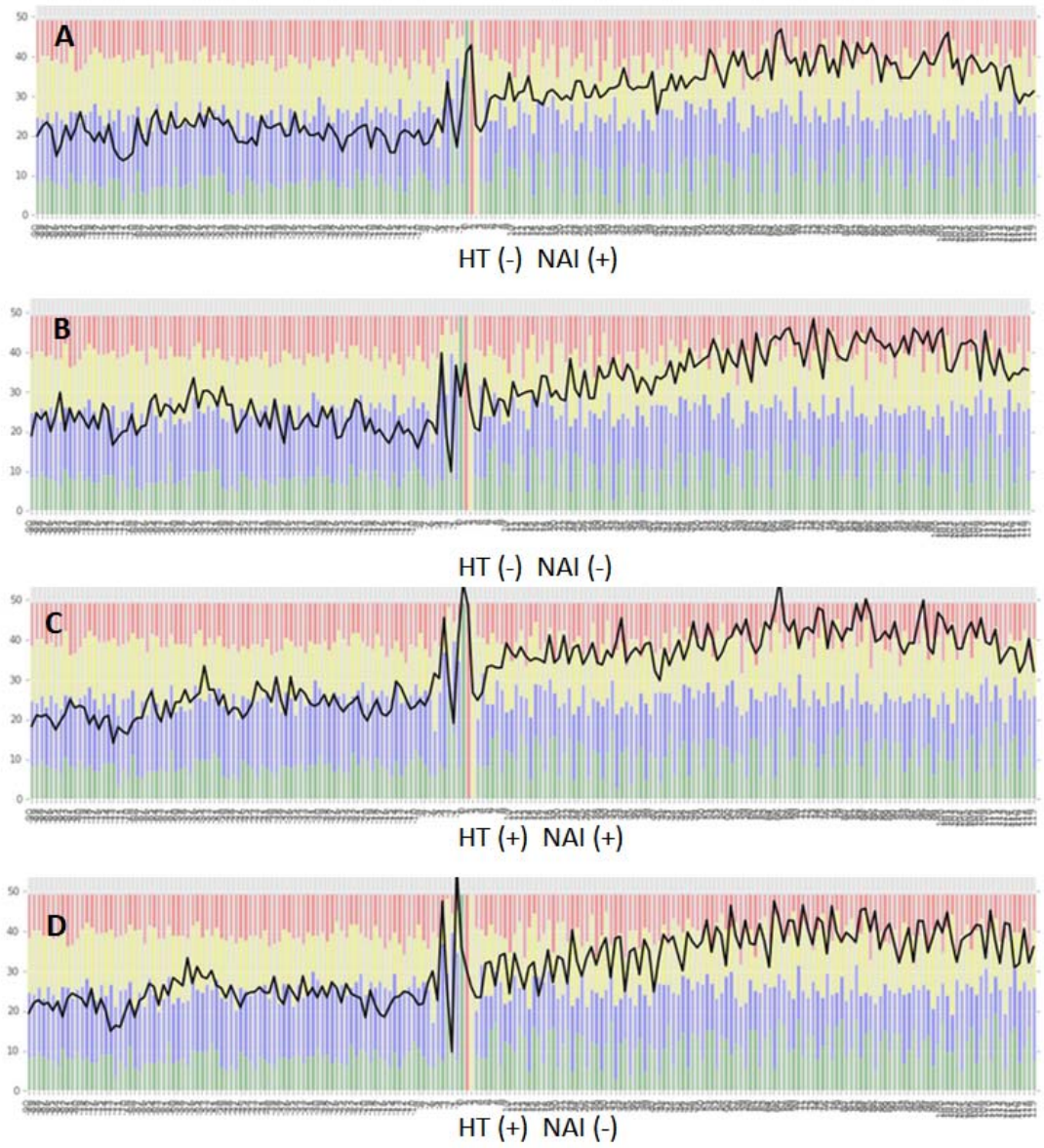


Figure 18. Three-nucleotide periodicity in mRNA CDS. Background was colored by nucleotides frequencies (A-red, C-blue, G-yellow, T-green).

A4. Evolutionary comparison of yeast mRNA secondary structures by Mod-seq

Previous studies on *in vivo* transcriptome wide RNA secondary structure probing showed that mRNAs are under active unfolding *in vivo* (75, 120, 143), and RNA structure rearrangement is affected by multiple factors including translation, interaction with RNA-binding proteins, and RNA modifications. However, it is still an open question how evolution changes mRNA *in vivo* structures. To investigate mRNA's secondary structure evolution, I analyzed Mod-seq DMS probing data of two closely related yeast species, *S. cerevisiae* and *S. paradoxus*.

Method

S. cer and *S.par*. cell cultures were grown in normal condition to log-phase, and were treated with DMS (or no DMS negative control) for 5 min under 37 °C. DMS modification causes RT stops at adenines and cytosines that are in single-stranded structures, and DMS caused RT stops at adenines are usually stronger than those at cytosines. The total RNA was then extracted from yeast cultures and poly-A selected. A cDNA sequencing library was then prepared using Mod-seq protocol.

Result

In this analysis, yeast genes were filtered using two criteria. First, transcripts should have enough coverage in both species. Second, confident annotation of the position of 5’UTR, 3’UTR and coding sequence (CDS) should be available in both species. We ended up with 588 genes that were available for further analysis.

3’UTRs are more structured than coding sequence

The fraction of single-stranded nucleotides in RNA was used as a measurement of RNA’s structure level. If there were more nucleotides modified, more nucleotides were single-stranded, and the RNA is less folded. As shown in Figure 19, consistent with what previous reports, the coding sequences are less structured than 3'UTRs. The 5'UTRs in yeast are usually very short, so the distribution of fraction of single-stranded nucleotides has greater variance. This feature holds in both *S. cer* and *S. par*. Since we conclude that 3' UTRs are more structured, we decided to focus on 3'UTRs for more detailed structural analysis.

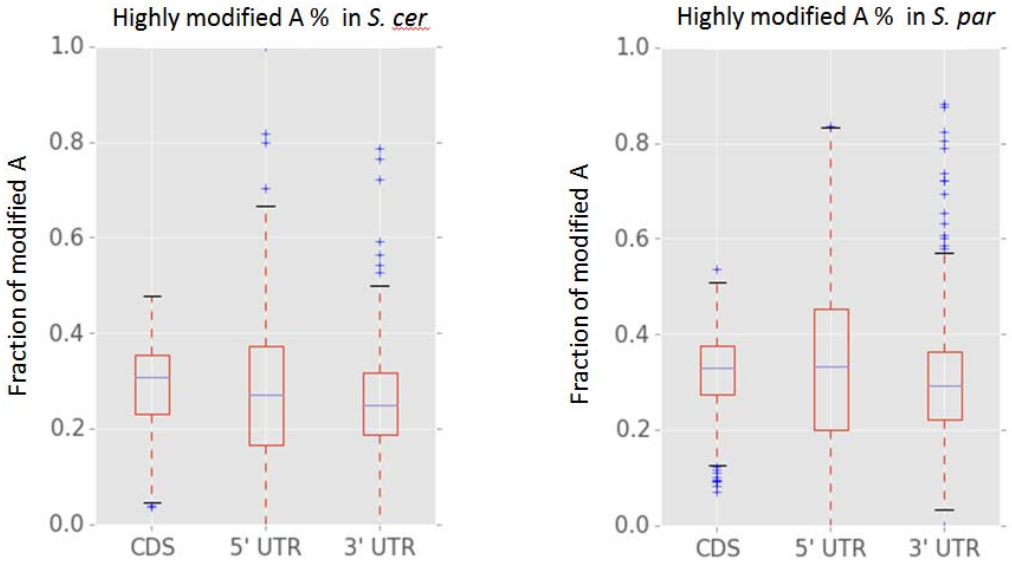


Figure 19. Fraction of single stranded nucleotides in 5'UTR, 3'UTR, and CDS.

Structure level of 3'UTRs is associated with functional/localization features

To find out if specific gene families mRNAs are more likely to have conserved low or high structure level across species, GO enrichment (145) analyses were performed on the most structured and most unstructured 3'UTRs. The most structured RNAs are usually involved in translation activities, likely to be coding for ribosome proteins (this is consistent with the previous report that mRNA with low melting temperatures are enriched for ribosomal protein mRNAs (143)); This enrichment is the same in *S. par.* In *S. cer*, the most unstructured RNAs are associated with membrane proteins (p-val = 0.07). While in *S. par*, no significant enrichment was found of the most unstructured 3'UTRs.

One explanation for this is the most structured RNAs are also the housekeeping RNAs, they tend to be conserved and hold stable structures across species. For the unstructured RNAs, their structures tend to be more dynamic and may change under different conditions.

Comparing orthologous mRNA structures in *S. cerevisiae* and *S. paradoxus* by measurement of distance between BPPMs

The fraction of single-stranded nucleotides is only a rough measurement of RNA structure level. We would like to compare RNA secondary structure models pair-wise in more detail. However, there was no convenient tool available for this analysis. I decided to develop a new algorithm for secondary structure comparison by calculating distance between BPPMs (base-pair probability matrix) (60, 81, 146, 147). BPPM is a representation of multiple possible RNA structures. BPPM is calculated by first generating multiple viable structures, then calculating the probability of each two nucleotides base-pairing with each other, and the probabilities are stored in a 2D upper matrix. Traditionally, BPPM was used to calculate the most likely structure.

Recently, an algorithm called riboSNitches was developed to identify SNPs that are associated with significant RNA structure changes (148). In riboSNitches, BPPM was used to represent RNA structures and the distance between two BPPMs was used to measure the structural difference between two RNA molecules with one SNP difference. A similar approach was adopted here to measure the structural difference between two orthologous RNAs. First, each of the two RNAs was fed into structure prediction software RNAstructure to calculate BPPM; meanwhile, the sequences of the two RNAs were aligned

using the Needleman-Wunsch global alignment algorithm (149). Given the alignment information, columns of zeros were inserted to the BPPM at gap positions, while sites of matched or mismatched nucleotides were kept unchanged. This alignment of matrices resulted in two BPPMs with the same dimension and each cell in matrix were matched up. Finally, the distance between the two matrices was calculated based on each nucleotide's probability of pairing with either an upstream nucleotide, a downstream nucleotide, or un-paired.

To use our Mod-seq probing data to improve RNA secondary structure modeling, the mod-scores are converted into structure constraints and used as an input at the structure prediction step (RNAstructure). This calculation was conducted for all 588 pairs of 3'UTRs from *S. cer* and *S. par.*

Mod-seq probing improve RNA secondary structure modeling

Distances between BPPMs with and without structure probing data as constraints were calculated. As seen in Figure 20, most points were above the y=x line, indicating with DMS probing, the predicted structures were more diverse between *S. cer* and *S. par*. This diversity of secondary structure is consistent with what was found in previous *in vivo* probing data that mRNA tends to be dynamic and adopt different structures under different conditions even in a single species.

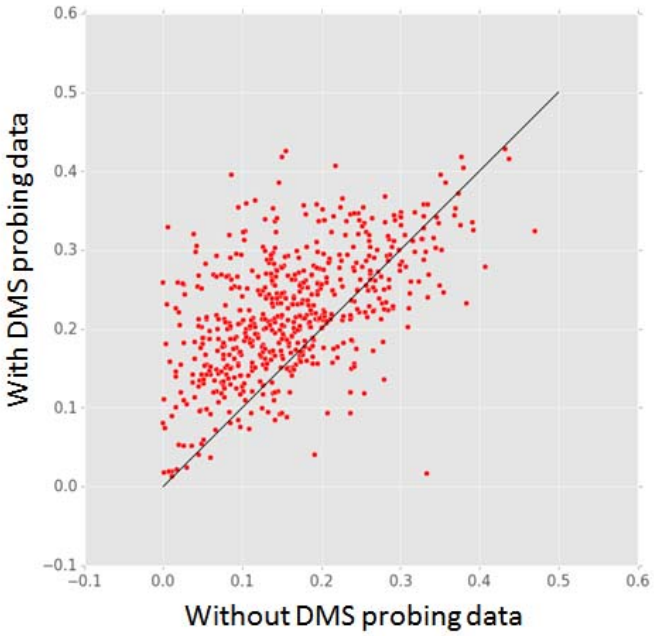


Figure 20. Distances between *S. cer* and *S. par* 3'UTR BPPMs with and without Mod-seq probing data as structure prediction constraints.

Comparing secondary structure conservation with sequence conservation

Next, we would like to see how conserved secondary structure was compared to sequence conservation. As shown in Figure 21, RNA with high sequence conservation may or may not have conserved secondary structure; but if the sequence has greater variation, it was less likely for the RNAs from two species to maintain the same secondary structure.

Secondary structure conservation at RBP binding sites

RNA structures *in vivo* are associated with protein binding. To study the structure at RBP binding sites, we did detailed analysis on protein binding sites identified by gPAR-CLIP (150). 64,594 RBP binding sites were identified, of which 2,236 were located in the 588 3'UTRs included in this study. We first asked if protein binding will affect Mod-seq probing by protecting bound RNA from DMS modification. As shown in Figure 22, when using the fraction of modified as a measurement of percentage of nucleotides in single-stranded structure and accessible for modification, no significant protection effect was seen. In fact, RBP binding sites can either be more modified or less modified; the distribution of modified nucleotides has a greater variance in RBP sites, which may be due to the fact that RBP protection sites are usually short (around 20 nt). To further investigate the RBP binding protection effect, RBP regions'

modification rates need to be compared to randomly sampled RNA segments that have same length with the RBP binding sites.

Figure 23 shows the distribution of distances between RBP binding sites BPPMs from two species, secondary structures in RBP binding sites tend to be more conserved when comparing with the 3'UTR average distance, or comparing to a randomly sampled short RNA segments that has the same length with the RBP sites ($p=0.0002$ by t-test).

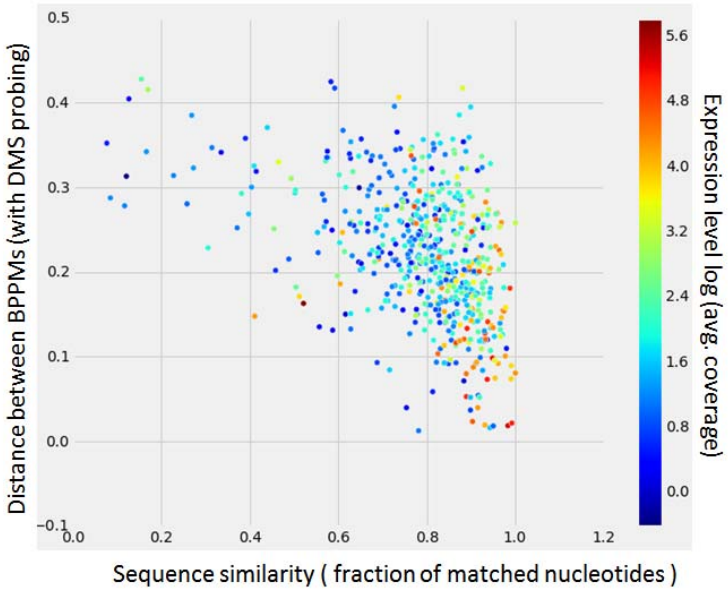


Figure 21. Comparison of secondary conservation and sequence conservation.

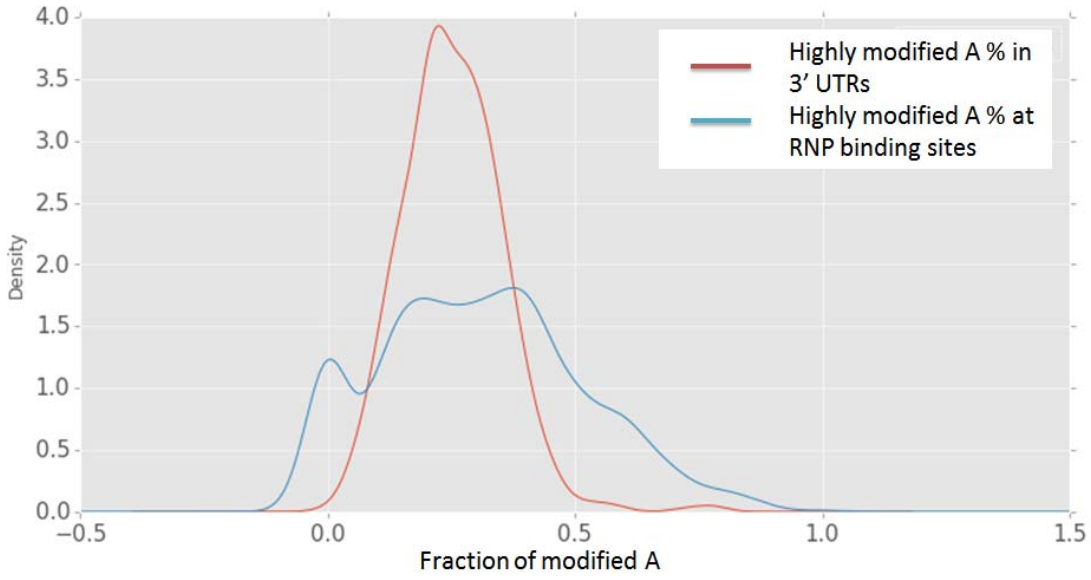


Figure 22. Fraction of single-stranded nucleotides in RBP binding sites.

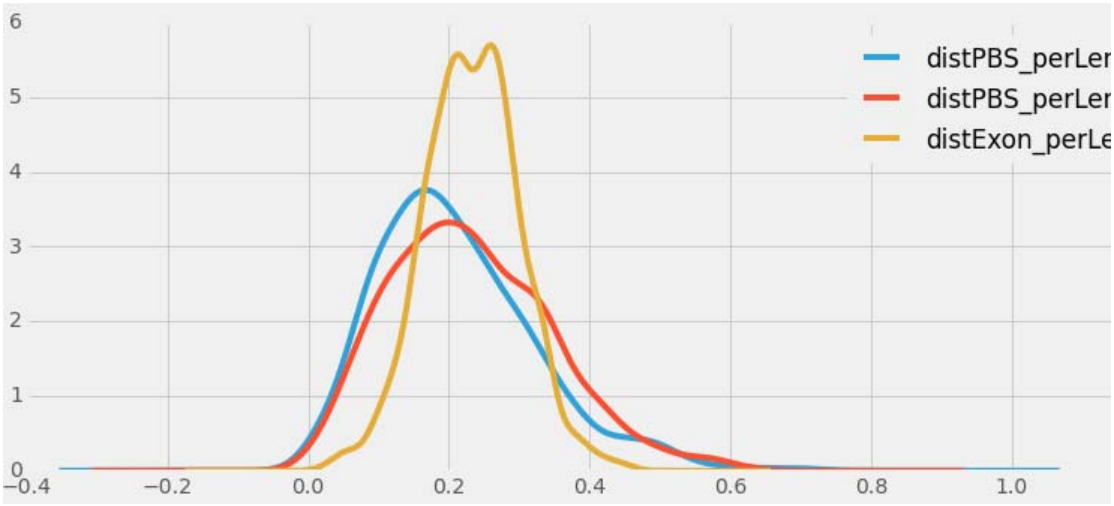


Figure 23. Secondary structure conservation of RBP binding sites.

Future plan

Secondary structure features are likely to have regulatory functions in gene expression. To investigate how the conservation of secondary structure is related to the conservation of mRNA's translation efficiency, I propose to compare Mod-seq probing data with published translational efficiency data (151) to see if conserved secondary structures are associated with conserved translational efficiency, and if changes in secondary structures are associated with changes in translational efficiency.

More detailed structural analyses of RBP binding sites, such as refinement of secondary structure modeling on local context around the binding sites, and searching for motifs that have conserved secondary structures in RBP sites are also called for. To test if the identified structural motifs are functional, deletion or mutations that interrupt secondary structures can be introduced and expression level of corresponding RNA can be assayed *in vivo*. If the RBP is known, *in vitro* gel shift assay can be used to verify if the protein binding is secondary structure dependent.

References

1. Djebali,S., Davis,C. a, Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F., *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–8.

2. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–74.

3. Hung,T. and Chang,H.Y. Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol.*, **7**, 582–5.

4. Brunel,C. and Romby,P. (2000) Probing RNA structure and RNA-ligand complexes with chemical probes. *Methods Enzymol.*, **318**, 3–21.

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

6. Lin,Y., May,G.E. and McManus,C.J. (2015) Mod-seq: A high-throughput method for probing RNA secondary structure 1st ed. Elsevier Inc.

7. Clemson,C.M., Hutchinson,J.N., Sara,S.A., Ensminger,A.W., Fox,A.H., Chess,A. and Lawrence,J.B. (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of

paraspeckles. *Mol Cell*, **33**, 717–726.

8. Sunwoo,H., Dinger,M.E., Wilusz,J.E., Amaral,P.P., Mattick,J.S. and Spector,D.L. (2009) MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.*, **19**, 347–59.

9. Derrien,T., Johnson,R., Bussotti,G., Tanzer, a., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel, a., Knowles,D.G., *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, **22**, 1775–1789.

10. Fang,Y. and Fullwood,M.J. (2016) Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics. Proteomics Bioinformatics*, **14**, 42–54.

11. Herzing,L.B., Romer,J.T., Horn,J.M. and Ashworth,A. (1997) Xist has properties of the X-chromosome inactivation centre. *Nature*, **386**, 272–275.

12. Simon,M.D., Pinter,S.F., Fang,R., Sarma,K., Rutenberg-Schoenberg,M., Bowman,S.K., Kesner,B.A., Maier,V.K., Kingston,R.E. and Lee,J.T. (2013) High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, **504**, 465–469.

13. Penny,G.D., Kay,G.F., Sheardown,S.A., Rastan,S. and Brockdorff,N. (1996) Requirement for Xist in X chromosome inactivation. *Nature*, **379**, 131–7.

14. Rinn,J.L., Kertesz,M., Wang,J.K., Squazzo,S.L., Xu,X.,

Brugmann,S.A., Goodnough,L.H., Helms,J.A., Farnham,P.J., Segal,E., *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, **129**, 1311–1323.

15. Pasmant,E., Sabbagh,A., Masliah-Planchon,J., Ortonne,N., Laurendeau,I., Melin,L., Ferkal,S., Hernandez,L., Leroy,K., Valeyrie-Allanore,L., *et al.* (2011) Role of noncoding RNA ANRIL in genesis of plexiform neurofibromas in neurofibromatosis type 1. *J Natl Cancer Inst*, **103**, 1713–1722.

16. Congrains,A., Kamide,K., Ohishi,M. and Rakugi,H. (2013) ANRIL: Molecular Mechanisms and Implications in Human Health. *Int J Mol Sci*, **14**, 1278–1292.

17. Kretz,M., Siprashvili,Z., Chu,C., Webster,D.E., Zehnder,A., Qu,K., Lee,C.S., Flockhart,R.J., Groff,A.F., Chow,J., *et al.* (2013) Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, **493**, 231–235.

18. Gong,C. and Maquat,L.E. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3’ UTRs via Alu elements. *Nature*, **470**, 284–288.

19. Carrieri,C., Cimatti,L., Biagioli,M., Beugnet,A., Zucchelli,S., Fedele,S., Pesce,E., Ferrer,I., Collavin,L., Santoro,C., *et al.* (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, **491**, 454–457.

20. Yoon,J.H., Abdelmohsen,K., Srikantan,S., Yang,X., Martindale,J.L., De,S., Huarte,M., Zhan,M., Becker,K.G. and Gorospe,M.

(2012) LincRNA-p21 suppresses target mRNA translation. *Mol Cell*, **47**, 648–655.

21. Huarte,M., Guttman,M., Feldser,D., Garber,M., Koziol,M.J., Kenzelmann-Broz,D., Khalil,A.M., Zuk,O., Amit,I., Rabani,M., *et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, **142**, 409–419.

22. Yin,Q.F., Yang,L., Zhang,Y., Xiang,J.F., Wu,Y.W., Carmichael,G.G. and Chen,L.L. (2012) Long noncoding RNAs with snoRNA ends. *Mol Cell*, **48**, 219–230.

23. Mao,Y.S., Sunwoo,H., Zhang,B. and Spector,D.L. (2011) Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat. Cell Biol.*, **13**, 95–101.

24. Naganuma,T. and Hirose,T. (2013) Paraspeckle formation during the biogenesis of long non-coding RNAs. *RNA Biol.*, **10**, 456–61.

25. Souquere,S., Beauclair,G., Harper,F., Fox,A. and Pierron,G. (2010) Highly ordered spatial organization of the structural long noncoding NEAT1 RNAs within paraspeckle nuclear bodies. *Mol Biol Cell*, **21**, 4020–4027.

26. West,J.A., Mito,M., Kurosaka,S., Takumi,T., Tanegashima,C., Chujo,T., Yanaka,K., Kingston,R.E., Hirose,T., Bond,C., *et al.* (2016) Structural, super-resolution microscopy analysis of paraspeckle nuclear body organization. *J. Cell Biol.*, **214**, jcb.201601071.

27. Jiang,L., Shao,C., Wu,Q.-J., Chen,G., Zhou,J., Yang,B., Li,H., Gou,L.-T., Zhang,Y., Wang,Y., *et al.* (2017) NEAT1 scaffolds RNA-binding

proteins and the Microprocessor to globally enhance pri-miRNA processing. *Nat. Struct. Mol. Biol.*, 10.1038/nsmb.3455.

28. Chen,L.L. and Carmichael,G.G. (2009) Altered Nuclear Retention of mRNAs Containing Inverted Repeats in Human Embryonic Stem Cells: Functional Role of a Nuclear Noncoding RNA. *Mol. Cell*, **35**, 467–478.

29. Hirose,T., Virnicchi,G., Tanigawa,A., Naganuma,T., Li,R., Kimura,H., Yokoi,T., Nakagawa,S., Bénard,M., Fox,A.H., *et al.* (2014) NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol. Biol. Cell*, **25**, 169–83.

30. Liu,S.J., Horlbeck,M.A., Cho,S.W., Birk,H.S., Malatesta,M., He,D., Attenello,F.J., Villalta,J.E., Cho,M.Y., Chen,Y., *et al.* (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science (80-.).*, **355**, eaah7111.

31. Horlbeck,M.A., Gilbert,L.A., Villalta,J.E., Adamson,B., Pak,R.A., Chen,Y., Fields,A.P., Park,C.Y., Corn,J.E., Kampmann,M., *et al.* (2016) Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife*, **5**, 1–20.

32. Ulitsky,I. (2016) Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.*, **17**, 601–614.

33. Washietl,S., Kellis,M. and Garber,M. (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.*, **24**, 616–28.

34. Hezroni,H., Koppstein,D., Schwartz,M.G., Avrutin,A., Bartel,D.P.

and Ulitsky,I. (2015) Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep.*, **11**, 1110–1122.

35. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P., *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

36. Kornienko,A.E., Guenzl,P.M., Barlow,D.P. and Pauler,F.M. (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.*, **11**, 59.

37. Novikova,I. V, Hennelly,S.P., Sanbonmatsu,K.Y. and Rna,K. (2012) Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? *Bioarchitecture*, **2**, 189–199.

38. Torarinsson,E., Yao,Z., Wiklund,E.D., Bramsen,J.B., Hansen,C., Kjems,J., Tommerup,N., Ruzzo,W.L. and Gorodkin,J. (2008) Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res*, **18**, 242–251.

39. Fujishima,K. and Kanai,A. (2014) tRNA gene diversity in the three domains of life. *Front. Genet.*, **5**, 142.

40. Montange,R.K. and Batey,R.T. (2008) Riboswitches: Emerging Themes in RNA Structure and Function. *Annu. Rev. Biophys.*, **37**, 117–133.

41. Yusupov,M.M., Yusupova,G.Z., Baucom, a, Lieberman,K., Earnest,T.N., Cate,J.H. and Noller,H.F. (2001) Crystal structure of the

ribosome at 5.5 A resolution. *Science*, **292**, 883–96.

42. Leshin,J.A., Heselpoth,R., Belew,A.T. and Dinman,J. (2011) High throughput structural analysis of yeast ribosomes using hSHAPE. *RNA Biol*, **8**, 478–487.

43. Konikkat,S. and Woolford,J.L. (2017) Principles of 60S ribosomal subunit assembly emerging from recent studies in yeast. *Biochem. J.*, **474**, 195–214.

44. Fang,R., Moss,W.N., Rutenberg-Schoenberg,M. and Simon,M.D. (2015) Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. *PLoS Genet.*, **11**, 1–29.

45. Somarowthu,S., Legiewicz,M., Chillón,I., Marcia,M., Liu,F. and Pyle,A.M. (2015) HOTAIR Forms an Intricate and Modular Secondary Structure. *Mol. Cell*, **58**, 353–361.

46. Chillón,I. and Pyle,A.M. (2016) Inverted repeat Alu elements in the human lincRNA-p21 adopt a conserved secondary structure that regulates RNA function. *Nucleic Acids Res.*, **44**, 9462–9471.

47. Novikova,I. V., Hennelly,S.P. and Sanbonmatsu,K.Y. (2012) Structural architecture of the human long non-coding RNA , steroid receptor RNA activator. *Nucleic Acids Res.*, **40**, 5034–5051.

48. Rivas,E., Clements,J. and Eddy,S.R. (2016) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, **14**, 45–48.

49. Chan,R.T., Robart,A.R., Rajashankar,K.R., Pyle,A.M. and Toor,N.

(2012) Crystal structure of a group II intron in the pre-catalytic state. *Nat Struct Mol Biol*, **19**, 555–557.

50. Toor,N., Keating,K.S., Taylor,S.D. and Pyle,A.M. (2008) Crystal Structure of a Self-Spliced Group II Intron. *Science (80-.).*, **320**, 77–82.

51. Marcia,M. and Pyle,A.M. (2012) Visualizing Group II Intron Catalysis through the Stages of Splicing. *Cell*, **151**, 497–507.

52. Mitchell,M., Gillis,A., Futahashi,M., Fujiwara,H. and Skordalakes,E. (2010) Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol*, **17**, 513–518.

53. Reiter,N.J., Osterman,A., Torres-Larios,A., Swinger,K.K., Pan,T. and Mondragon,A. (2010) Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature*, **468**, 784–789.

54. Schuwirth,B.S., Borovinskaya,M.A., Hau,C.W., Zhang,W., Vila-Sanjurjo,A., Holton,J.M. and Cate,J.H. (2005) Structures of the bacterial ribosome at 3.5 A resolution. *Science (80-.).*, **310**, 827–834.

55. Garmann,R.F., Gopal,A., Athavale,S.S., Knobler,C.M., Gelbart,W.M. and Harvey,S.C. (2015) Visualizing the global secondary structure of a viral RNA genome with cryo-electron microscopy. 10.1261/rna.047506.114.

56. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406–3415.

57. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**,

129.

58. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res*, **36**, W70-4.

59. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F., Hofacker,I.L., Thirumalai,D., Lee,N., Woodson,S., *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

60. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

61. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna*, **10**, 1178–1190.

62. Hu,Y.J. (2003) GPRM: A genetic programming approach to finding common RNA secondary structure elements. *Nucleic Acids Res*, **31**, 3446–3449.

63. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, **31**, 3423–3428.

64. Caprara,M. (2013) RNA Structure Determination Using Chemical Methods. *Cold Spring Harb. Protoc.*, **2013**, pdb.prot078485-pdb.prot078485.

65. Xu,Z. and Culver,G. (2013) RNA Structure Experimental Analysis – Chemical Modification. *Methods Enzymol.*, **530**, 363–380.

66. Wilkinson,K. a, Merino,E.J. and Weeks,K.M. (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–1616.

67. Low,J.T. and Weeks,K.M. (2010) SHAPE-directed RNA secondary structure prediction. *Methods*, **52**, 150–158.

68. Spitale,R.C., Crisalli,P., Flynn,R. a, Torre,E. a, Kool,E.T. and Chang,H.Y. (2013) RNA SHAPE analysis in living cells. *Nat. Chem. Biol.*, **9**, 18–20.

69. Mortimer,S. a. and Weeks,K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.*, **129**, 4144–5.

70. Adilakshmi,T., Lease,R.A. and Woodson,S.A. (2006) Hydroxyl radical footprinting in vivo: mapping macromolecular structures with synchrotron radiation. *Nucleic Acids Res*, **34**, e64.

71. Ding,F., Lavender,C.A., Weeks,K.M. and Dokholyan,N. V (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods*, **9**, 603–608.

72. Costa,M. and Monachello,D. (2014) Probing RNA folding by hydroxyl radical footprinting. *Methods Mol Biol*, **1086**, 119–142.

73. Balasubramanian,B., Pogozelski,W.K. and Tullius,T.D. (1998)

DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 9738–9743.

74. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–7.

75. Wan,Y., Qu,K., Zhang,Q.C., Flynn,R. a, Manor,O., Ouyang,Z., Zhang,J., Spitale,R.C., Snyder,M.P., Segal,E., *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.

76. Westhof,E. and Romby,P. (2010) The RNA structurome: high-throughput probing. *Nat. Methods*, **7**, 965–7.

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

78. Rouskin,S., Zubradt,M., Washietl,S., Kellis,M. and Weissman,J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–5.

79. Siegfried,N. a, Busan,S., Rice,G.M., Nelson,J. a E. and Weeks,K.M. (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*, **11**, 959–65.

80. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary

structure. *Proc. Natl. Acad. Sci.*, **101**, 7287–7292.

81. Ding,Y.Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.

82. Ding,Y., Chan,C.Y. and Lawrence,C.E. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.

83. Spasic,A., Assmann,S.M., Bevilacqua,P.C. and Mathews,D.H. (2018) Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res.*, **46**, 314–323.

84. Karabiber,F., McGinnis,J.L., Favorov,O. V and Weeks,K.M. (2013) QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA*, **19**, 63–73.

85. Vasa,S.M., Guex,N., Wilkinson,K.A., Weeks,K.M. and Giddings,M.C. (2008) ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA*, **14**, 1979–90.

86. Ouyang,Z., Snyder,M.P. and Chang,H.Y. (2013) SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.*, **23**, 377–87.

87. Tang,Y., Bouvier,E., Kwok,C.K., Ding,Y., Nekrutenko,A., Bevilacqua,P.C. and Assmann,S.M. (2015) StructureFold: genome-wide RNA secondary structure mapping and reconstruction *in vivo*. *Bioinformatics*, **31**, 2668–2675.

88. Bond,C.S. and Fox,A.H. (2009) Paraspeckles: nuclear bodies built on long noncoding RNA. *J. Cell Biol.*, **186**, 637–644.

89. Nakagawa,S., Naganuma,T., Shioi,G. and Hirose,T. (2011) Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J. Cell Biol.*, **193**, 31–39.

90. Batista,P.J. and Chang,H.Y. (2013) Long noncoding RNAs: Cellular address codes in development and disease. *Cell*, **152**, 1298–1307.

91. Imamura,K., Imamachi,N., Akizuki,G., Kumakura,M., Kawaguchi,A., Nagata,K., Kato,A., Kawaguchi,Y., Sato,H., Yoneda,M., *et al.* (2014) Long Noncoding RNA NEAT1-Dependent SFPQ Relocation from Promoter Region to Paraspeckle Mediates IL8 Expression upon Immune Stimuli. *Mol. Cell*, **53**, 393–406.

92. Zhang,Q., Chen,C.Y., Yedavalli,V.S. and Jeang,K.T. (2013) NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. *MBio*, **4**, e00596-12.

93. Torres,M., Becquet,D., Blanchard,M.P., Guillen,S., Boyer,B., Moreno,M., Franc,J.L. and Fran??ois-Bellan,A.M. (2016) Circadian RNA expression elicited by 3'-UTR IRAlu-paraspeckle associated elements. *Elife*, **5**, 1–23.

94. Nishimoto,Y., Nakagawa,S., Hirose,T., Okano,H.J., Takao,M., Shibata,S., Suyama,S., Kuwako,K.-I., Imai,T., Murayama,S., *et al.* (2013) The long non-coding RNA nuclear-enriched abundant transcript 1_2 induces paraspeckle formation in the motor neuron during the early phase of amyotrophic lateral sclerosis. *Mol. Brain*, **6**, 31.

95. Naganuma,T., Nakagawa,S., Tanigawa,A., Sasaki,Y.F., Goshima,N. and Hirose,T. (2012) Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J*, **31**, 4020–4034.

96. Sasaki,Y.T.F., Ideue,T., Sano,M., Mituyama,T. and Hirose,T. (2009) MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 2525–30.

97. Chujo,T., Yamazaki,T., Kawaguchi,T., Kurosaka,S., Takumi,T., Nakagawa,S. and Hirose,T. (2017) Unusual semi‐extractability as a hallmark of nuclear body‐associated architectural noncoding RNAs. *EMBO J.*, 10.15252/embj.201695848.

98. Martin,M. (2013) Cutadapt removes adapter sequences from high-throughput sequencing reads [Miyashita mitsunori].

99. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

100. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

101. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562–578.

102. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

103. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

104. Cochran,W.G. (1954) Some Methods for Strengthening the Common χ 2 Tests. *Biometrics*, **10**, 417.

105. Mantel,N. and Haenszel,W. (1959) Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI J. Natl. Cancer Inst.*, **22**, 719–748.

106. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

107. Wu,Y., Shi,B., Ding,X., Liu,T., Hu,X., Yip,K.Y., Yang,Z.R., Mathews,D.H. and Lu,Z.J. (2015) Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Res.*, **43**, 7247–7259.

108. Lin,Y., Schmidt,B.F., Bruchez,M.P. and McManus,C.J. (2018) Structural analyses of NEAT1 lncRNAs suggest long-range RNA interactions that may contribute to paraspeckle architecture. *Nucleic Acids Res.*, 10.1093/nar/gky046.

109. Stephenson,W., Keller,S., Santiago,R., Albrecht,J.E., Asare-Okai,P.N., Tenenbaum,S. a, Zuker,M. and Li,P.T.X. (2014) Combining

temperature and force to study folding of an RNA hairpin. *Phys. Chem. Chem. Phys.*, **16**, 906–17.

110. Graur,D., Zheng,Y., Price,N., Azevedo,R.B.R., Zufall,R.A. and Elhaik,E. (2013) On the immortality of television sets: 'Function' in the human genome according to the evolution-free gospel of encode. *Genome Biol. Evol.*, **5**, 578–590.

111. Smola,M.J., Christy,T.W., Inoue,K., Nicholson,C.O., Friedersdorf,M., Keene,J.D., Lee,D.M., Calabrese,J.M. and Weeks,K.M. (2016) SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc. Natl. Acad. Sci. U. S. A.*, 10.1073/pnas.1600008113.

112. Maenner,S., Blaud,M., Fouillen,L., Savoye,A., Marchand,V., Dubois,A., Sanglier-Cianf??rani,S., Van Dorsselaer,A., Clerc,P., Avner,P., *et al.* (2010) 2-D structure of the a region of Xist RNA and its implication for PRC2 association. *PLoS Biol.*, **8**, e1000276.

113. Liu,F., Somarowthu,S. and Marie Pyle,A. (2017) Visualizing the secondary and tertiary architectural domains of lncRNA RepA. *Nat. Chem. Biol.*, 10.1038/nCHeMBIO.2272.

114. Yu,X., Li,Z., Zheng,H., Chan,M.T. V. and Wu,W.K.K. (2017) NEAT1: A novel cancer-related long non-coding RNA. *Cell Prolif.*, **50**, e12329.

115. Sunwoo,J.-S., Lee,S.-T., Im,W., Lee,M., Byun,J.-I., Jung,K.-H., Park,K.-I., Jung,K.-Y., Lee,S.K., Chu,K., *et al.* (2017) Altered Expression of the Long Noncoding RNA NEAT1 in Huntington's Disease. *Mol.*

Neurobiol., **54**, 1577–1586.

116. Nakagawa,S., Shimada,M., Yanaka,K., Mito,M., Arai,T., Takahashi,E., Fujita,Y., Fujimori,T., Standaert,L., Marine,J.-C., *et al.* (2014) The lncRNA Neat1 is required for corpus luteum formation and the establishment of pregnancy in a subpopulation of mice. *Development*, **141**, 4618–27.

117. Standaert,L., Adriaens,C., Radaelli,E., Van Keymeulen,A., Blanpain,C., Hirose,T., Nakagawa,S. and Marine,J. (2014) The long noncoding RNA Neat1 is required for mammary gland development and lactation. *RNA*, **20**, 1844–1849.

118. Li,R., Harvey,A.R., Hodgetts,S.I. and Fox,A.H. (2017) Functional dissection of NEAT1 using genome editing reveals substantial localisation of the NEAT1_1 isoform outside paraspeckles. *RNA*, 10.1261/rna.059477.116.

119. Hu,S., Xiang,J., Li,X., Xu,Y., Xue,W., Huang,M., Wong,C.C., Sagum,A., Bedford,M.T., Yang,L., *et al.* (2015) Protein arginine methyltransferase CARM1 attenuates the paraspeckle- mediated nuclear retention of mRNAs containing IR Alus. *Genes Dev.*, 10.1101/gad.257048.114.

120. Spitale,R.C., Flynn,R. a., Zhang,Q.C., Crisalli,P., Lee,B., Jung,J.-W., Kuchelmeister,H.Y., Batista,P.J., Torre,E. a., Kool,E.T., *et al.* (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, 10.1038/nature14263.

121. Guo,F., Gooding,A.R. and Cech,T.R. (2004) Structure of the

Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. *Mol Cell*, **16**, 351–362.

122. Lu,Z.J., Gloor,J.W. and Mathews,D.H. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–13.

123. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–5.

124. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

125. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

126. Weinberg,Z. and Breaker,R.R. (2011) R2R--software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, **12**, 3.

127. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

128. Gavazzi,C., Isel,C., Fournier,E., Moules,V., Cavalier,A., Thomas,D., Lina,B. and Marquet,R. (2013) An in vitro network of intermolecular interactions between viral RNA segments of an avian H5N2 influenza A virus: Comparison with a human H3N2 virus. *Nucleic Acids Res.*, **41**, 1241–1254.

129. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K., *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 1–9.

130. Watts,J.M., Dang,K.K., Gorelick,R.J., Leonard,C.W., Bess Jr,J.W., Swanstrom,R., Burch,C.L. and Weeks,K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.

131. Pollom,E., Dang,K.K., Potter,E.L., Gorelick,R.J., Burch,C.L., Weeks,K.M. and Swanstrom,R. (2013) Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs. *PLoS Pathog*, **9**, e1003294.

132. Novikova,I. V, Dharap,A., Hennelly,S.P. and Sanbonmatsu,K.Y. (2013) 3S: shotgun secondary structure determination of long non-coding RNAs. *Methods*, **63**, 170–7.

133. Lu,Z., Zhang,Q.C., Lee,B., Flynn,R.A., Smith,M.A., Robinson,J.T., Davidovich,C., Gooding,A.R., Goodrich,K.J., Mattick,J.S., *et al.* (2016) RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell*, **165**, 1–13.

134. Sharma,E., Sterne-Weiler,T., O'Hanlon,D. and Blencowe,B.J. (2016) Global Mapping of Human RNA-RNA Interactions. *Mol. Cell*, **62**, 1–9.

135. Aw,J.G.A., Shen,Y., Wilm,A., Sun,M., Lim,X.N., Boon,K.-L., Tapsin,S., Chan,Y.-S., Tan,C.-P., Sim,A.Y.L., *et al.* (2016) In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol. Cell*, **62**, 1–15.

136. Cimino,G.D., Gamper,H.B., Isaacs,S.T. and Hearst,J.E. (1985) Psoralens as photoactive probes of nucleic acid structure and function: organic chemistry, photochemistry, and biochemistry. *Annu. Rev. Biochem.*, **54**, 1151–93.

137. Novikova,I. V., Hennelly,S.P. and Sanbonmatsu,K.Y. (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.*, **40**, 5034–5051.

138. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

139. GRUBER,A.R., FINDEIß,S., WASHIETL,S., HOFACKER,I.L. and STADLER,P.F. (2009) RNAZ 2.0:IMPROVED NONCODING RNA DETECTION. In *Biocomputing 2010*. WORLD SCIENTIFIC, pp. 69–79.

140. Maharana,S., Wang,J., Papadopoulos,D.K., Richter,D., Pozniakovsky,A., Poser,I., Bickle,M., Rizk,S., Guillén-Boixet,J., Franzmann,T., *et al.* (2018) RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science (80-.).*, **7432**, eaar7366.

141. Wildauer,M., Zemora,G., Liebeg,A., Heisig,V. and Waldsich,C. (2014) Chemical probing of RNA in living cells. *Methods Mol Biol*, **1086**, 159–176.

142. Cavaillé,J. and Bachellerie,J.P. (1998) SnoRNA-guided ribose

methylation of rRNA: structural features of the guide RNA duplex influencing the extent of the reaction. *Nucleic Acids Res.*, **26**, 1576–87.

143. Wan,Y., Qu,K., Ouyang,Z., Kertesz,M., Li,J., Tibshirani,R., Makino,D.L., Nutter,R.C., Segal,E. and Chang,H.Y. (2012) Genome-wide Measurement of RNA Folding Energies. *Mol. Cell*, **48**, 169–181.

144. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.

145. Mi,H., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.*, **8**, 1551–1566.

146. Ponty,Y. (2007) Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy. *J. Math. Biol.*, **56**, 107–127.

147. Waldispühl,J. and Clote,P. (2007) Computing the Partition Function and Sampling for Saturated Secondary Structures of RNA, with Respect to the Turner Energy Model. *J. Comput. Biol.*, **14**, 190–215.

148. Corley,M., Solem,A., Qu,K., Chang,H.Y. and Laederach,A. (2015) Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. *Nucleic Acids Res.*, **43**, 1859–68.

149. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

150. Freeberg,M.A., Han,T., Moresco,J.J., Kong,A., Yang,Y.-C.,

Lu,Z.J., Yates,J.R. and Kim,J.K. (2013) Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae. *Genome Biol.*, **14**, R13.

151. McManus,C.J., May,G.E., Spealman,P. and Shteyman,A. (2014) Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.*, **24**, 422–30.