STRUCTURE-FUNCTION RELATIONSHIPS OF LONG NON-CODING RNAS IN LIVING CELLS

Matthew James Smola

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry

Chapel Hill
2015

Approved by:

Kevin Weeks

Howard Fried

Brian Kuhlman

Alain Laederach

Linda Spremulli

ABSTRACT

Matthew James Smola: Structure-Function Relationships of Long Non-Coding RNAs in Living Cells
(Under the direction of Kevin M. Weeks)

From the beginning of the era of molecular biology in the 1960s until the 1980s, RNA was widely regarded as a passive cellular messenger. However, the importance of RNA has been steadily emerging over the last 30 years and we now know that it is often a critical and central component of genetic regulation. Recently, long non-coding RNAs (lncRNA) have become the focus of intense research because of their roles in development and disease. For most functional RNAs, complex structural characteristics underlie the biological function of the molecule. However, the difficulty of *de novo* RNA structure prediction and the relatively low abundance of lncRNA transcripts have been roadblocks to experimental structure probing. As a result, very little is known about the structural features of lncRNAs. In this work, I present experimental and analytical methods that enable chemical structure probing of rare RNA transcripts and identification of stable RNA-protein interaction sites. First, I show that polymerase chain reactions can be used as an enrichment strategy that faithfully maintains structure-probing data. I then outline an analytical framework that enables statistically rigorous detection of RNA-protein interactions in living cells. Finally, I apply these new methodologies to the *Xist* lncRNA and present a data-driven secondary structure model that highlights the extensive structures present throughout the transcript. I then identify nearly 200 specific sites where *Xist* is strongly impacted by the cellular environment and use them to identify several new protein interaction domains within *Xist*. Together, this work provides new experimental and analytical tools, as well as many new insights on the relationship between lncRNA structure and function, that will enable rapid study of lncRNA structures in the future.

"When we recognize our place in an immensity of light-years and in the passage of ages, when we grasp the intricacy, beauty, and subtlety of life, then that soaring feeling, that sense of elation and humility combined, is surely spiritual."

– Carl Sagan

ACKNOWLEDGEMENTS

This work represents the culmination of over two decades of formal education and would not have been possible without the enduring love, support, and encouragement of my family. From the early days of science fair projects, illegible negative signs, and misplaced calculators to more recent experiences in experimental troubleshooting and code debugging, my parents have always been my strongest and most vocal supporters. They are the giants upon whose shoulders I now stand.

At its best, scientific research is immensely fulfilling, though at its worst it can be incredibly frustrating. Through all the highs and lows of graduate school, I have been lucky to be surrounded by my fellow Weeks Lab members. I am grateful for the constant encouragement, advice, and exquisitely well-timed comic relief that all of you have provided over the years.

It has been a privilege to experience graduate school alongside a group of close-knit friends. Kathleen, Fatima, Gregg, Tim, and Hannah, I am grateful to have had each of you in my life these past few years. I am certain that you will each have a positive impact on our world and I cannot wait to see what you accomplish.

I would like to thank Mauro Calabrese for inadvertently sparking my interest in lncRNA biology by proposing the *Xist* project. I am grateful to have you as a mentor and have enjoyed the casual nature of our interactions, especially the many enlightening, exciting, and frank discussions that we have had. Best of luck as you establish your own research group!

Last but not least, I want to express my gratitude to my advisor, Kevin. Thank you for encouraging me to go after bold projects, teaching me to set and meet high standards, and helping me to properly calibrate my "bad science" detector. I look forward to drawing a few more boxes with you before I depart from Chapel Hill.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

1M6	1-methyl-6-nitroisatoic anhydride
1M7	1-methyl-7-nitroisatoic anhydride
A	adenosine
ARE	A-U rich element
BzCN	benzoyl cyanide
C	cytosine
cDNA	complimentary DNA
CMCT	1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-*p*-toluenesulfonate
cryo-EM	cryo-electron microscopy
DMS	dimethyl sulfate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dsRNA	double-stranded RNA
EDTA	ethylenediaminetetraacetic acid
G	guanosine
g	gram
HEPES	*N*-2-hydroxyethylpiperazine-*N′*-ethanesulfonic acid
kb	kilobase
KCl	potassium chloride
kethoxal	1,1-dihydroxy-3-ethoxy-2-butanone
lncRNA	long non-coding RNA
M	molar
MaP	mutational profiling
$MgCl_2$	magnesium chloride

min	minute
ml	milliliter
mM	millimolar
mol	mole
MPS	massively parallel sequencing
ng	nanogram
nM	nanomolar
NMIA	*N*-methylisatoic anhydride
nt	nucleotide
PCR	polymerase chain reaction
pmol	picomole
RMRP	RNase MRP
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT	reverse transcription
SDS	sodium dodecyl sulfate
SHAPE	selective 2′-hydroxyl acylation analyzed by primer extension
SRA	steroid receptor RNA activator
SRP	signal recognition particle
ssRNA	single-stranded RNA
TF	transcription factor
TPP	thiamine pyrophosphate
TSC	trophoblast stem cells
U	uracil
XCI	X chromosome inactivation

X_i	inactive X chromosome
Xist	X-inactive specific transcript
°C	degree Celsius
μ	mean
μg	microgram
μl	microliter
μM	micromolar
σ	standard deviation

CHAPTER 1: THE RELATIONSHIP BETWEEN RNA STRUCTURE AND FUNCTION

Hierarchy of RNA structure

Ribonucleic acid (RNA) has been the focus of an extended scientific renaissance that has been gradually building momentum over the past 25 years (*1*). While originally thought to act as a simple, passive intermediate between genetic information stored in the nucleus and the protein synthesizing components of the cytoplasm, we now know that RNA is actively and critically involved in modulating gene expression through a variety of mechanisms (*2*). These modes of action can be *cis*- or *trans*-acting and can be carried out by the RNA alone or require additional protein partners. It has been discovered that RNA can regulate gene expression through alternative RNA splicing (*3*), RNA interference (*4*, *5*), and by the action of riboswitches (*6*) and long non-coding RNA (lncRNA) (*7*).

Underlying many of these functions is the ability of RNA molecules to adopt complex structures (*8*). RNA structure is hierarchical: structural elements at one level provide the foundation for higher-order structures (*9*). The simplest, primary structure of an RNA is defined by the specific linear sequence of individual nucleotides that comprise the full-length molecule. This sequence can be any combination of the four RNA nucleotides: adenosine (A), cytosine (C), guanosine (G), and uracil (U) (**Fig. 1.1a**). These are often described as the "letters" of the RNA "alphabet." In some roles, RNA function is driven largely by primary structure (e.g. RNA silencing). However, higher levels of RNA structure can both tune these functions and enable more complex activities.

The secondary structure of RNA refers to the ability of individual RNA molecules to fold and interact with themselves. Secondary structures are defined by hydrogen bonding interactions between pairs of nucleotides. Only certain pairs are able to form stable interactions and canonically, these are formed by G-C, A-U, and G-U pairs (**Fig. 1.1b**). Consecutive base pairs form a double helix structure

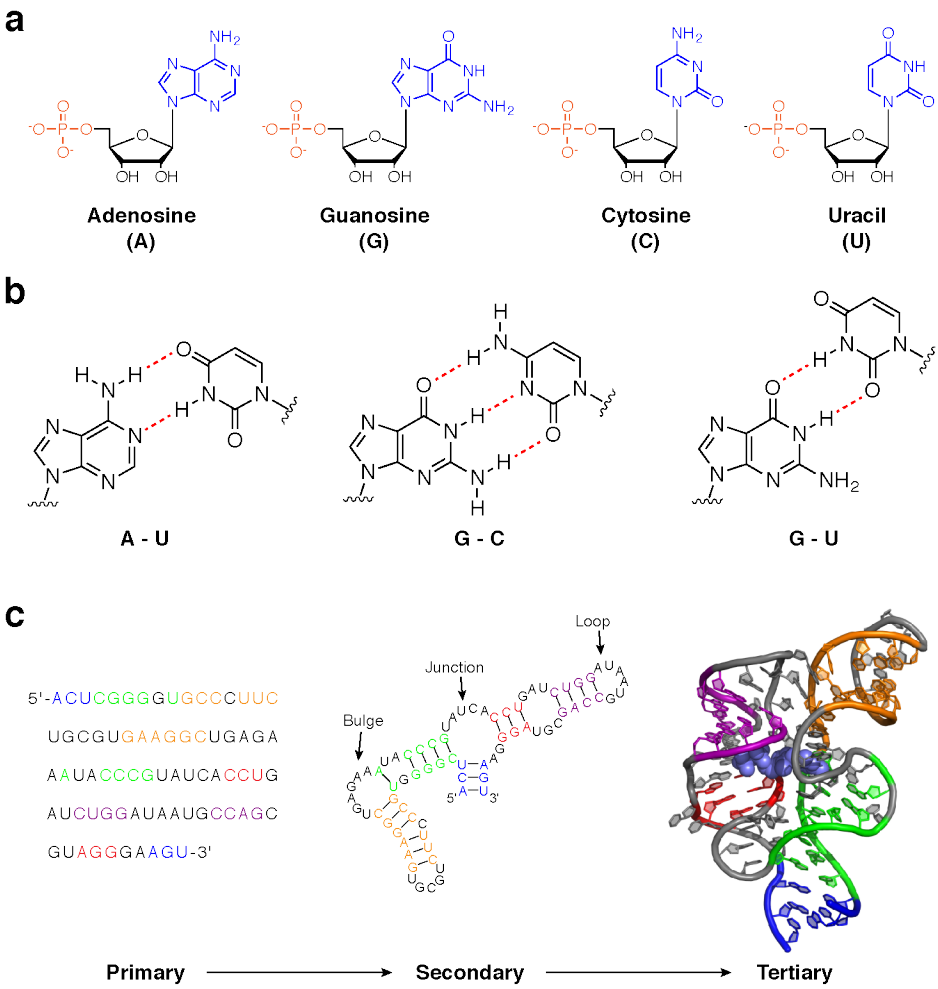


Figure 1.1 Elements and heirarchy of RNA structure. (a) The four RNA nucleotides. All nucleotides have phosphate (orange) and ribose sugar (black) moieties in common, while the nitrogenous base (blue) determines the chemical identity of each nucleotide. (b) The canonical base pairs of RNA. Pairing is faciliated by hydrogen bonding (red dashed lines) between the nitrogenous bases of two nucleotides. A-U (left) and G-C (center) pairs are analogous to those found in DNA, while G-U pairs (right) are unique to RNA. (c) The heirarchy of RNA structure. The primary, secondary, and tertiary structure of the thiamine pyrophosphate riboswitch are shown. Primary structure is defined by the linear sequence of nucleotides (left). Base pairing interactions between distant elements of primary structure (indicated by lines connecting colored nucleotides) lead to helix formation and the secondary structure of an RNA (center). Examples of a bulge, junction, and loop are labeled. In many cases, RNA secondary structure leads to the formation of well-defined three-dimensional structures (right). The precise three-dimensional arrangement of the riboswitch tertiary structure enables recognition and binding of a small ligand, shown as purple spheres (PDB ID: 2GDI). Nucleotides that form helices are color-coded throughout.

and, in many RNAs, multiple helical elements are connected by unpaired nucleotides to form loops, bulges, and junctions (**Fig. 1c**). In most cases, the secondary structure of an RNA strongly affects its tertiary structure This highest level of RNA structure is defined in three dimensions and is modulated by the length and placement of helices, loops, bulges, and other structural elements (*9*) (**Fig. 1c**). Tertiary structures can be stabilized by and, in many cases, facilitate interaction with proteins, small molecule ligands, and metal ions (*10*). RNA tertiary structures often play critical roles in biology. For example, riboswitches are a class of RNA elements that fold into well-defined tertiary structures and specifically recognize small metabolites (*6*). The binding and release of these metabolites induces dramatic three-dimensional structure changes that regulate gene expression.

Since RNA structure is closely linked to function, structural characterization of an RNA is often the first step to understanding how it operates. Experimentally determining the tertiary structure of an RNA can be an arduous undertaking, and the computational tools needed to predict such structures *de novo* are both resource-intensive and imperfect. However, even secondary structure can be highly informative of the function of an RNA (*11*), and several efforts to improve RNA secondary structure prediction have been successful (*12*). For this reason, and because secondary structure provides the foundation for tertiary structure, most investigations into RNA structure-function relationships focus on secondary structure. However, until recently, most RNA structure work has focused on abundant RNAs. Although biologically critical in many cases, rarer transcripts have been largely ignored in favor of those that are more easily accessible.

The rise of lncRNAs as effectors of genetic regulation

Recently, advances in sequencing technology have enabled rapid and unprecedented studies on the composition of mammalian DNA genomes and RNA transcriptomes. One of the most striking results of these studies is that while roughly 75% of the human genome is transcribed into RNA (*13*), less than 2% is translated into proteins (*14*). The resulting interpretation is that the remainder of the transcriptome functions in a non-coding capacity. While a handful of non-coding RNAs such as ribosomal RNA,

transfer RNA, and small nuclear RNA have been well-known for years, new families of non-coding RNAs are being continually discovered (*15*). Of particular interest are the long non-coding RNAs (lncRNAs), defined as transcripts longer than 200 nucleotides with little to no coding potential (*7*). This class of RNA is frequently associated with epigenetic regulation of chromatin (*16*), but the list of cellular roles ascribed to lncRNAs now includes transcriptional regulation, modulation of protein activity, organization of protein complexes, and intercellular signaling (*17*). lncRNAs exhibit cell type-specific expression patterns (*18*) and are associated with several diseases, including cancer (*19*). As a result, lncRNAs have been the focus of intense research in the last decade.

In performing various regulatory functions, lncRNAs often interact with protein complexes and genomic DNA (*7*, *17*, *20*, *21*). The nature of these interactions can vary greatly. For example, lncRNAs such as *Xist, HOTAIR,* and *Kcnq1ot1* act as molecular scaffolds to coordinate histone modifying complexes and silence specific regions of the genome (*22-24*) (**Fig. 1.2a**). In other cases, lncRNAs can act as molecular sponges and sequester RNA polymerase II or transcription factors, reducing gene transcription (*25*, *26*) (**Fig. 1.2b-c**). Occasionally the lncRNA itself is not important; rather, the process of lncRNA transcription negatively affects antisense genes at the same locus (*27*) (**Fig. 1.2d**). Alternatively, some lncRNAs have been found to interact directly with genomic DNA, inhibiting the binding of transcription complexes (*28*) (**Fig. 1.2e**). Finally, not all lncRNA actions are repressive. For example, *Evf-2* recruits a transcriptional activator to specific DNA regulatory elements, leading to increased transcription at those loci (*29*) (**Fig. 1.2e**).

Although it is clear that lncRNAs regulate gene expression at transcriptional, post-transcriptional, and epigenetic levels, there are many unanswered questions regarding lncRNA structure and function: Why are lncRNAs so long? Do lncRNAs have well-defined structures? If so, to what extent does the cellular environment modulate structure? What features govern lncRNA-protein interactions? While lncRNAs are often described as having modular structure (*20*), only two lncRNAs have been structurally characterized to date: the steroid receptor RNA activator (SRA) and *HOTAIR* (*30*, *31*). These lncRNAs

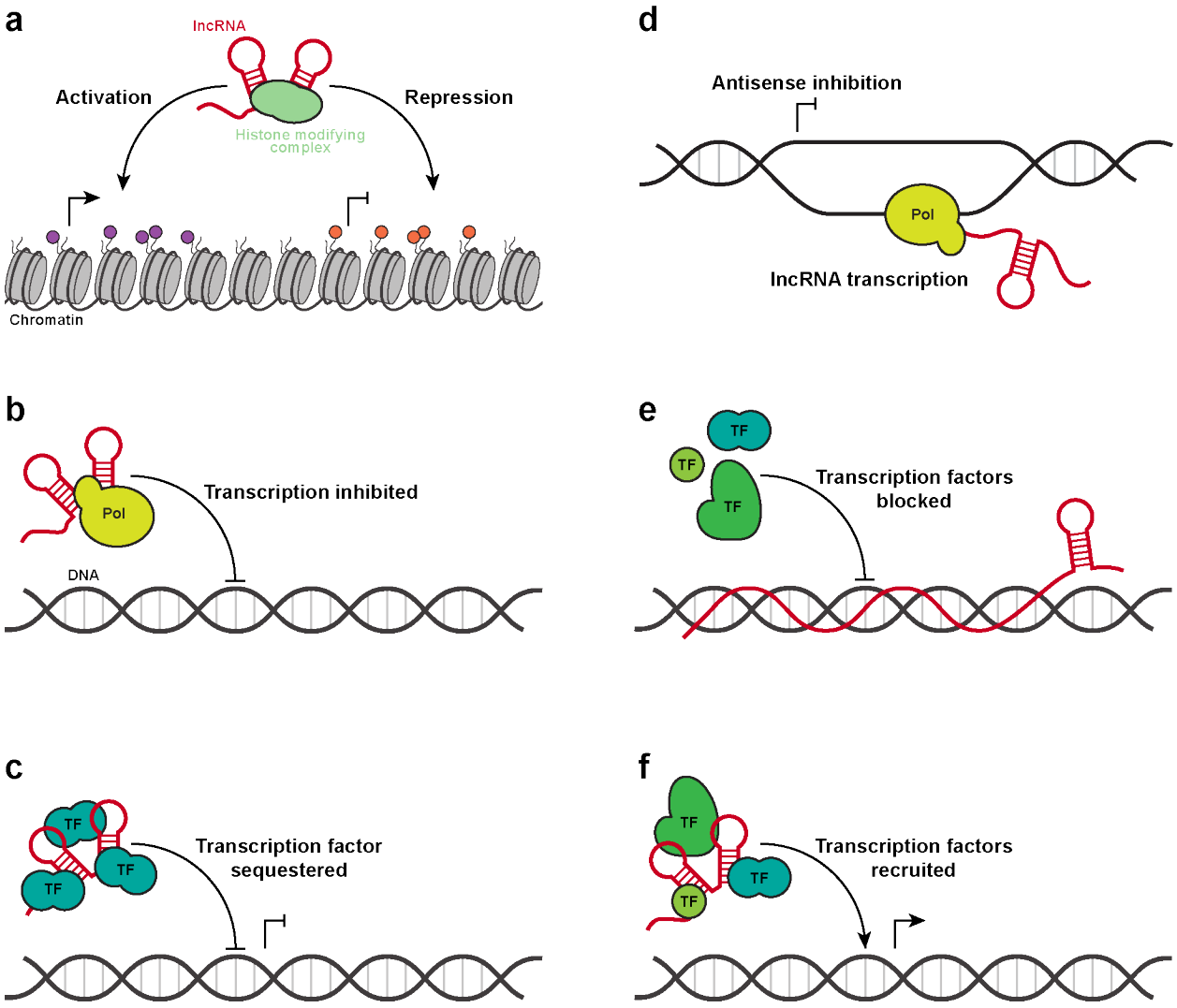


Figure 1.2 Example mechanisms of lncRNA gene regulation. (a) Recruitment of histone modifying complexes by lncRNAs leads to deposition of activating (left, purple) or repressing (right, orange) histone modifications. (b) Direct interaction between lncRNAs and RNA polymerase (Pol) inhibits transcription. (c) lncRNAs sequester transcription factors (TF) and prevent gene expression. (d) Transcription of lncRNA inhibits expression of antisense genes at the lncRNA locus. (e) lncRNAs interact with genomic DNA to block transcription factors and repress gene expression. (f) lncRNAs enhances gene expression by coordinating the recruitment of TFs.

are 0.87 and 2.1 kilobases (kb) long, respectively, and indeed adopt modular structures. However, they represent simple systems relative to other lncRNAs. For example, the *Xist* lncRNA is roughly an order of magnitude greater in length and associates with more than 50 proteins in order to silence an entire chromosome in a process called X-chromosome inactivation (XCI) (*24*, *32-35*). Although XCI has been researched for more than 50 years (*36*) and the central role of *Xist* in this process was discovered 20 years ago (*24*, *35*), the structural architecture of this lncRNA has yet to be determined. As an archetype of lncRNA-mediated genome silencing, structural characterization of *Xist* would provide an excellent foundation for understanding structure-function relationships in lncRNAs.

Methods for interrogating RNA secondary structure

For many years prior to the realization that RNA can form functional structures, most biochemical structure studies were performed on proteins using high-resolution (and labor-intensive) methods such as x-ray crystallography. For many proteins, the diverse array of hydrophobic and hydrophilic surfaces enables formation of suitable crystals. Unfortunately, the relatively uniform surface features and flexibility of RNA create significant challenges for crystallographers, so much so that many alternative structure-probing methods have been developed.

The goal of most RNA structure probing strategies is to discriminate, experimentally, base-paired nucleotides from single-stranded nucleotides. Generally, this is accomplished by using structure-selective RNases or small chemical probes. RNases are protein enzymes that recognize and cleave RNA; several RNases with structure-specific behaviors have been identified (*37*). Although once widely used for RNA studies, the activity of these bulky enzymes can be biased by factors unrelated to RNA structure, decreasing the accuracy of experimental results (*37*).

Small molecule reagents have become a popular alternative to enzymatic structure probing. The small size, relatively simple chemistry, and structural selectivity of these reagents enable accurate interrogation of RNA structure. Generally, RNA structure probes react selectively with flexible nucleotides, allowing for the discrimination of constrained and flexible nucleotides. Reagents used in

early experiments react with the nucleobase moieties of RNA and include dimethyl sulfate (DMS), which reports on A and C nucleotides (*38*, *39*) (**Fig. 1.3a**); 1,1-dihydroxy-3-ethoxy-2-butanone (kethoxal), which reports on G nucleotides (*40*) (**Fig. 1.3b**); and 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-*p*-toluenesulfonate (CMCT), which reports on G and U nucleotides (*41*) (**Fig. 1.3c**). While together these reagents yield structure information on all four RNA nucleotides, a major disadvantage is that obtaining a complete dataset for all four nucleobases requires optimization of reaction conditions for each reagent, including incubation time, buffer conditions, reagent concentration, and reaction quenching. In many cases, researchers settle for either DMS or CMCT, as these each yield data for two of the four RNA bases.

In the last decade, a suite of chemical reagents that react specifically with the 2′-hydroxyl moiety of flexible RNA nucleotides has been developed (*42-44*) (**Fig. 1.3d**). Unlike base-specific probes, these acylating reagents report on all four RNA bases since every nucleotide carries a 2′ hydroxyl. This approach to RNA chemical probing, known as SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension), typically makes use of isatoic anhydride derivatives, including 1-methyl-7-nitroisatoic anhydride (1M7), 1-methyl-6-nitroisatoic anhydride (1M6), and *N*-methylisatoic anhydride (NMIA). 1M7 is generally considered the "workhorse" of SHAPE reagents and accurately reports on local nucleotide flexibility (*43*), while 1M6 and NMIA are often used in combination with 1M7 to reveal nuanced details of RNA structure (*44*, *45*).

SHAPE reagents have many advantages over other chemical probes. They are self-inactivating via competing hydrolysis with water and thus require no specific quench step. They are unaffected by the presence of small molecules or proteins, and are compatible with most biologically relevant *in vitro* solution conditions. SHAPE reagents also perform well in complex environments, such as those within virus particles (*46-51*) and living cells (*52-55*).

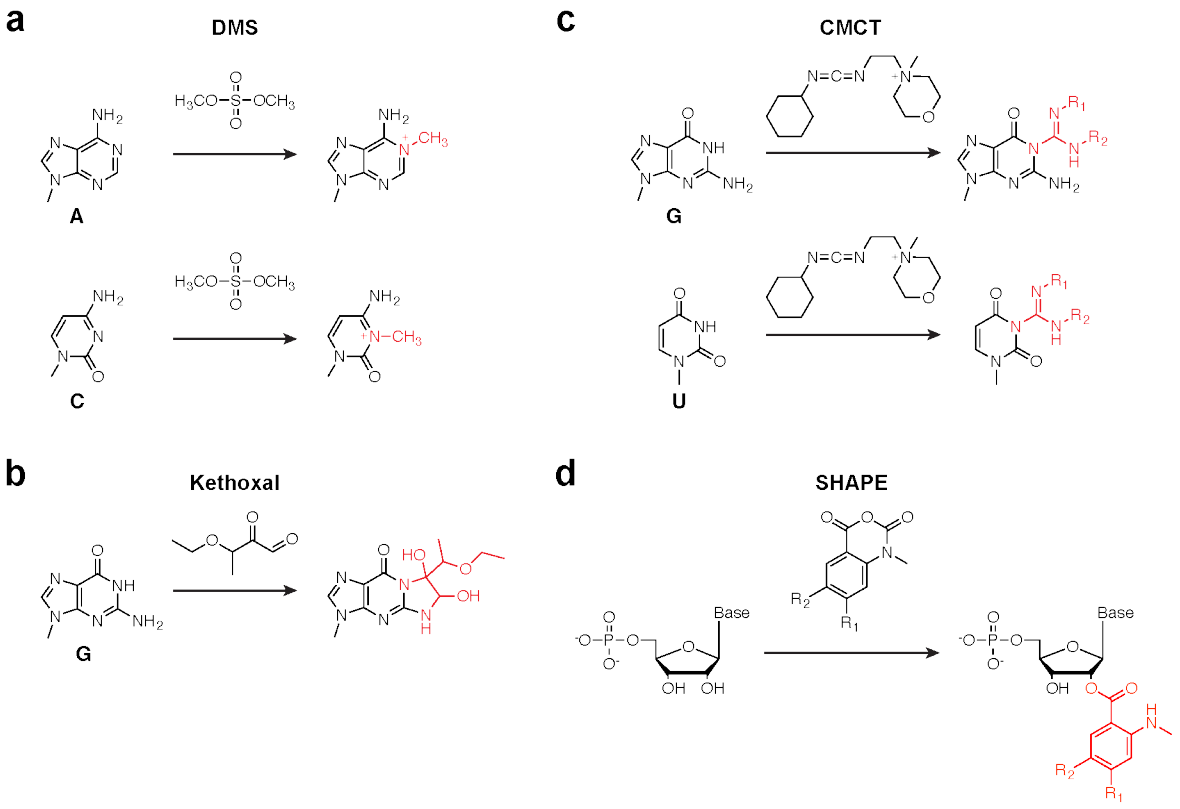


Figure 1.3 Examples of RNA structure-probing reagents. (a) DMS modification of A and C nucleotides yields addition of a methyl group on the base-pairing edge of the nucleobase. (b) Reaction of G nucleotides with kethoxal produces a bulky cyclic adduct that blocks base pairing. (c) CMCT modification of G and U nucleotides produces an extensively bulky adduct. Each R group contains a 6-membered aliphatic ring. (d) SHAPE reagents react with the 2′-hydroxyl to form a bulky 2′-*O*-adduct and are thus capable of probing all four nucleotides simultaneously. The R_1 and R_2 groups can vary based on the application.

Perhaps one of the greatest advantages of SHAPE reagents over other chemical probes is their ability to accurately constrain computational predictions of RNA secondary structure. For very small RNAs, computer algorithms can use thermodynamic parameters to model generally accurate structures. However, as the size of the RNA increases, the number of possible base-pair combinations also increases and computational prediction accuracy decreases dramatically. Since SHAPE chemical probing data report on RNA structure, they can be used to guide structure prediction software and result in highly accurate structure models (*45*, *56*, *57*). In addition, structure modeling of large RNA transcripts has been robustly automated such that data-driven structure models and analysis can be easily generated within a few hours (*50*, *58*).

Coupling RNA structure probing to massively parallel sequencing

Until very recently, the results of RNA chemical probing experiments were read out by annealing a labeled primer to the RNA of interest and extending the primer with reverse transcriptase to create a complimentary DNA (cDNA) strand. Under normal conditions, reverse transcriptase enzymes are blocked by chemical adducts or RNase cleavage sites. Thus, when a sample of RNA molecules is probed and analyzed by primer extension, the length and abundance of each cDNA species relates to the position and intensity of chemical modification or cleavage, respectively. Using labeled primers, these cDNAs are resolved and quantified using gel or capillary electrophoresis.

While a small amount of sample multiplexing is possible with this approach, experiments read out by electrophoresis are usually limited to studying a single RNA at a time. Due to signal decay, multiple primers must be used to study large RNAs, which introduces the additional challenge of accurately merging data from individual reactions into a contiguous data set. As the number of primer extension reactions increases, the material requirements also increase; a 2009 study of the ~9,000 nucleotide HIV RNA genome required nearly 300 μg of RNA purified from 19 L of virus culture (*47*).

The recent and rapid development of massively parallel sequencing (MPS) technology has dramatically changed how nucleic acids are sequenced and analyzed. Instead of sequencing a single DNA

of interest, modern instruments sequence sample libraries containing millions of individual DNA fragments simultaneously. Most commercial platforms require specific adaptor sequences on the 5′ and 3′ ends for initial recognition by the instrument. Thus, it would appear as though all that is needed to couple RNA structure probing with MPS technology is to construct sequencing libraries that quantitatively preserve the 3′ termini of cDNAs generated during primer extension. Several research laboratories have developed methods that aim to achieve this type of massively parallel structure probing (**Table 1.1**). However, most of these methods rely on ligation steps that are biased in unpredictable ways, making them difficult to correct (*71*, *72*). As a result, sequencing libraries generated in this fashion may not preserve the original chemical probing data with high fidelity.

To date, only one approach, called mutational profiling (MaP), offers an alternative strategy for quantifying the position and intensity of RNA structure probe reactivity. Instead of performing conventional primer extension in which reverse transcriptase is blocked by adducts, MaP exploits the ability of reverse transcriptase enzymes to proceed through chemical adducts in the presence of divalent manganese. Under MaP conditions, when reverse transcriptase encounters a site of modification it is more likely to incorporate a nucleotide non-complimentary to the original RNA sequence in the nascent cDNA strand. These sequence mutations are thus permanently and securely embedded within the cDNA, rendering MaP experiments impervious to common downstream library preparation biases. MaP has been coupled with SHAPE and DMS chemical probing (*50*, *70*). Data generated by the SHAPE-MaP approach are of as high (or higher) quality than data obtained by prior gold-standard capillary electrophoresis methods (*50*) and, unlike other massively parallel structure probing approaches, can be used to accurately predict RNA secondary structures (*50*, *58*).

Research overview

The overarching vision of this work is to address the role of RNA structure within the context of lncRNA function. Using the *Xist* RNA as an example, I highlight how lncRNAs can adopt stable secondary structures and, by comparison of in-cell and *ex vivo* data, show that these structures likely play a critical

Name	Probe	System	Citation	Detection of modifications
PARS	RNase S1, RNase V1	Yeast, Human	(*59*)	Reverse transcriptase stops (cDNA 3′ ends)
Frag-seq	RNase P1	Mouse	(*60*)	
SHAPE-seq	1M7	Synthetic, Bacteria	(*61*)	
MAP-seq	1M7, CMCT, DMS	Synthetic	(*62*)	
HRF-seq	Hydroxyl radical	Bacteria	(*63*)	
ChemMod-seq	1M7, DMS	Yeast	(*64*)	
CIRS-seq	CMCT, DMS	Mouse	(*65*)	
Structure-seq	DMS	Plant	(*66*)	
DMS-seq	DMS	Yeast, Human	(*67*)	
Mod-seq	DMS	Yeast	(*68*)	
icSHAPE	NAI	Mouse	(*69*)	
SHAPE-MaP	1M7	HIV, HCV	(*50*)	Mutational profiling (cDNA internal mutations)
RING-MaP	DMS	Synthetic	(*70*)	

Table 1.1 Approaches for massively parallel RNA structure probing. Numerous approaches have been developed in order to merge MPS and RNA structure probing. These various methods employ both nuclease and chemical probes, and most focus on detecting the 3′ ends of cDNA produced when reverse transcriptase encounters a chemical adduct or cleavage site. The MaP approach is unique in that it records chemical modifications as mutations internal to the cDNA sequence.

role in coordinating and modulating protein interactions in cells. In the course of this work, it was necessary to invent new solutions in order to obtain the necessary data. As such, my overall work toward understanding lncRNA structure has led to new experimental and computational methods for obtaining and analyzing SHAPE data.

In Chapter 2, I outline a novel approach for studying the structures of low-abundance RNAs with SHAPE-MaP. To properly characterize RNA structure by chemical probing, it is necessary to sufficiently sample the original pool of RNA molecules. While many RNAs are abundant enough that probing total cellular RNA yields high-quality structure data, the vast majority of transcripts are present at low levels and must be enriched or isolated prior to study. I show that PCR amplification prior to library construction can be used to extract chemical probing data for RNAs of interest. This allows for rapid and efficient analysis of rare, native RNA transcripts without the need for specialized pull-downs and with relatively low input material requirements.

In Chapter 3, I address the challenge of rigorously evaluating changes in SHAPE reactivity between two experimental conditions, focusing on detecting effects of the cellular environment on RNA structure. Using the built-in error estimates of SHAPE-MaP, I outline an analytical framework that identifies strong, significant changes in SHAPE reactivity. I validate this approach, called ΔSHAPE, with the U1 snRNA, 5S rRNA, and signal recognition particle, showing that ΔSHAPE analysis results in robust and correct identification of RNA-protein interaction sites. Finally, I apply ΔSHAPE to the RNA component of RNase MRP and show that this analysis detects previously-known RNA-protein interactions and identifies new contacts unaccounted for by current models.

Chapter 4 focuses on my overall vision of understanding lncRNA structure-function relationships via the *Xist* RNA. I first present a data-driven model of *Xist* and highlight how much of the RNA is involved in stable secondary structures, which may rationalize the conserved length of the RNA among mammals. I show that previously identified functional regions appear to sample a variety of structures, and I identify new portions of the transcript that undergo extensive structural changes in the cellular environment. Using ΔSHAPE analysis, I identify nearly 200 specific sites where *Xist* is strongly impacted

by the cellular environment. Cross-referencing these sites with protein binding databases, I then identify several new protein interaction domains within *Xist*. Overall, this work provides many new insights on the relationship between RNA structure and function in lncRNAs. It shows that lncRNAs can adopt stable structures, that these structures modulate protein interactions in several specific ways, and that RNA chemical probing can be used to identify novel regions of interest even in *Xist*, an RNA that has been extensively studied for twenty years.

Perspective

The notion that form follows function is pervasive in all of biology. At every scale, from opposable thumbs, to the valves and chambers of the heart, down to the intricate atomic arrangement of an enzyme active site, the physical structures of biology are tightly linked to their functions. In this work, I apply principles from molecular biology, biochemistry, and biophysics in order to broaden our understanding of this structure-function relationship in the *Xist* lncRNA. In doing so, I have developed broadly useful approaches for accessing the chemical probing profiles of rare RNAs and for rigorously detecting chemical probing differences *ex vivo* and in living cells. In applying these advances to *Xist*, I have created a useful model that highlights how structure and function converge in the context of this RNA.

We have only recently begun to understand and appreciate the extent to which lncRNAs work to modulate gene expression. My hope is that the work presented here will provide useful methodologies and serve as a starting point and touchstone for future studies. As we work to further understand the intricacies of *Xist*-mediated genome silencing, I anticipate that the model of *Xist* structure-function relationships presented here may act as a "molecular roadmap," highlighting functional regions of the RNA worthy of additional study. I also hope that this work will be a helpful guide as researchers make efforts to understand the relationship between structure and function in other lncRNAs.

REFERENCES

1. T. R. Cech, RNA World research--still evolving. *RNA*. **21**, 474–475 (2015).

2. P. A. Sharp, The Centrality of RNA. *Cell*. **136**, 577–580 (2009).

3. Z. Wang, C. B. Burge, Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*. **14**, 802–813 (2008).

4. A. Fire *et al*., Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*. **391**, 806–811 (1998).

5. V. Ambros, The functions of animal microRNAs. *Nature*. **431**, 350–355 (2004).

6. B. J. Tucker, R. R. Breaker, Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol*. **15**, 342–348 (2005).

7. J. L. Rinn, H. Y. Chang, Genome Regulation by Long Noncoding RNAs. *Annu. Rev. Biochem*. **81**, 145–166 (2012).

8. J. A. Cruz, E. Westhof, The Dynamic Landscapes of RNA Architecture. *Cell*. **136**, 604–609 (2009).

9. N. B. Leontis, A. Lescoute, E. Westhof, The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol*. **16**, 279–287 (2006).

10. S. E. Butcher, A. M. Pyle, The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc. Chem. Res*. **44**, 1302–1311 (2011).

11. Y. Wan, M. Kertesz, R. C. Spitale, E. Segal, H. Y. Chang, Understanding the transcriptome through RNA structure. *Nat. Rev. Genet*. **12**, 641–655 (2011).

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

13. S. Djebali *et al*., Landscape of transcription in human cells. *Nature*. **489**, 101–108 (2012).

14. E. S. Lander *et al*., Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921 (2001).

15. J. S. Mattick, Non-coding RNA. *Hum. Mol. Genet*. **15**, R17–R29 (2006).

16. B. K. Dey, A. C. Mueller, A. Dutta, Long non-coding RNAs as emerging regulators of differentiation, development, and disease. *Transcription*. **5**, e944014 (2014).

17. S. Geisler, J. Coller, RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol*. **14**, 699–712 (2013).

18. T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, J. S. Mattick, Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A*. **105**, 716–721 (2008).

19. J. R. Prensner, A. M. Chinnaiyan, The emergence of lncRNAs in cancer biology. *Cancer Discov*.

1, 391–407 (2011).

20. T. R. Mercer, J. S. Mattick, Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* **20**, 300–307 (2013).

21. A. Fatica, I. Bozzoni, Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* **15**, 7–21 (2014).

22. J. L. Rinn *et al.*, Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell.* **129**, 1311–1323 (2007).

23. C. Kanduri, Kcnq1ot1: A chromatin regulatory RNA. *Semin. Cell Dev. Biol.* **22**, 343–350 (2011).

24. Y. Marahrens, B. Panning, J. Dausman, W. Strauss, R. Jaenisch, Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev.* (1997).

25. P. D. Mariner *et al.*, Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell.* **29**, 499–509 (2008).

26. T. Kino, D. E. Hurt, T. Ichijo, N. Nader, G. P. Chrousos, Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Sci. Signal.* **3**, ra8 (2010).

27. P. A. Latos *et al.*, Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science.* **338**, 1469–1472 (2012).

28. I. Martianov, A. Ramadass, A. Serra Barros, N. Chow, A. Akoulitchev, Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature.* **445**, 666–670 (2007).

29. J. Feng, The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* **20**, 1470–1484 (2006).

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

31. S. Somarowthu *et al.*, HOTAIR Forms an Intricate and Modular Secondary Structure. *Mol. Cell.* **58**, 353–361 (2015).

32. C. Chu *et al.*, Systematic Discovery of Xist RNA Binding Proteins. *Cell.* **161**, 404–416 (2015).

33. A. Minajigi *et al.*, A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* (2015), doi:10.1126/science.aab2276.

34. C. A. McHugh *et al.*, The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature.* **521**, 232–236 (2015).

35. G. D. Penny, G. F. Kay, S. A. Sheardown, S. Rastan, N. Brockdorff, Requirement for Xist in X chromosome inactivation. *Nature.* **379**, 131–137 (1996).

36. M. F. Lyon, Gene Action in the X-chromosome of the Mouse (Mus musculus L.). *Nature.* **190**, 372–373 (1961).

37. C. Ehresmann *et al*., Probing the structure of RNAs in solution. *Nucleic Acids Res*. **15**, 9109–9128 (1987).

38. P. Brookes, P. D. Lawley, The reaction of mono-and di-functional alkylating agents with nucleic acids. *Biochem. J*. **80**, 496–503 (1961).

39. P. D. Lawley, P. Brookes, Further studies on the alkylation of nucleic acids and their constituent nucleotides. *Biochem. J*. **89**, 127–138 (1963).

40. M. Litt, V. Hancock, Kethoxal—A Potentially Useful Reagent for the Determination of Nucleotide Sequences in Single-Stranded Regions of Transfer Ribonucleic Acid. *Biochemistry*. **6**, 1848–1854 (1967).

41. N. W. Ho, P. T. Gilham, Reaction of pseudouridine and inosine with N-cyclohexyl-N'-beta-(4-methylmorpholinium)ethylcarbodiimide. *Biochemistry*. **10**, 3651–3657 (1971).

42. E. J. Merino, K. A. Wilkinson, J. L. Coughlan, K. M. Weeks, RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc*. **127**, 4223–4231 (2005).

43. S. A. Mortimer, K. M. Weeks, A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc*. **129**, 4144–4145 (2007).

44. K.-A. Steen, G. M. Rice, K. M. Weeks, Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *J. Am. Chem. Soc*. **134**, 13160–13163 (2012).

45. G. M. Rice, C. W. Leonard, K. M. Weeks, RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA*. **20**, 846–854 (2014).

46. K. A. Wilkinson *et al*., High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol*. **6**, e96 (2008).

47. J. M. Watts *et al*., Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. **460**, 711–716 (2009).

48. C. Gherghe *et al*., Definition of a high-affinity Gag recognition structure mediating packaging of a retroviral RNA genome. *Proc. Natl. Acad. Sci. U.S.A*. **107**, 19248–19253 (2010).

49. E. J. Archer *et al*., Long-Range Architecture in a Viral RNA Genome. *Biochemistry*. **52**, 3182–3190 (2013).

50. N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson, K. M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*. **11**, 959–965 (2014).

51. D. M. Mauger *et al*., Functionally conserved architecture of hepatitis C virus RNA genomes. *Proc. Natl. Acad. Sci. U.S.A*. **112**, 3692–3697 (2015).

52. R. C. Spitale *et al*., RNA SHAPE analysis in living cells. *Nat. Chem. Biol*. **9**, 18–20 (2012).

53. J. Tyrrell, J. L. McGinnis, K. M. Weeks, G. J. Pielak, The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry*. **52**, 8777–8785 (2013).

54. J. L. McGinnis, K. M. Weeks, Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry*. **53**, 3237–3247 (2014).

55. J. L. McGinnis *et al*., In-cell SHAPE reveals that free 30S ribosome subunits are in the inactive state. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2425–2430 (2015).

56. K. E. Deigan, T. W. Li, D. H. Mathews, K. M. Weeks, Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 97–102 (2009).

57. C. E. Hajdin *et al*., Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5498–5503 (2013).

58. M. J. Smola, G. M. Rice, S. Busan, N. A. Siegfried, K. M. Weeks, Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile, and accurate RNA structure analysis. *Nat. Protoc.* **10**, 1643–1669 (2015).

59. M. Kertesz *et al*., Genome-wide measurement of RNA secondary structure in yeast. *Nature*. **467**, 103–107 (2010).

60. J. G. Underwood *et al*., FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*. **7**, 995–1001 (2010).

61. J. B. Lucks *et al*., Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11063–11068 (2011).

62. M. G. Seetin, W. Kladwang, J. P. Bida, R. Das, Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods Mol. Biol.* **1086**, 95–117 (2014).

63. L. J. Kielpinski, J. Vinther, Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic Acids Res.* **42**, e70 (2014).

64. R. D. Hector *et al*., Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. *Nucleic Acids Res.* **42**, 12138–12154 (2014).

65. D. Incarnato, F. Neri, F. Anselmi, S. Oliviero, Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biol.* **15**, 491 (2014).

66. Y. Ding *et al*., In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*. **505**, 696–700 (2014).

67. S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, J. S. Weissman, Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*. **505**, 701–705 (2015).

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

69. R. C. Spitale *et al*., Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*. **519**, 486–490 (2015).

70. P. J. Homan *et al*., Single-molecule correlated chemical probing of RNA. *Proc. Natl. Acad. Sci.*

U.S.A. **111**, 13858–13863 (2014).

71. K. M. Weeks, RNA structure probing dash seq. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10933–10934 (2011).

72. C. A. Raabe, T.-H. Tang, J. Brosius, T. S. Rozhdestvensky, Biases in small RNA deep sequencing data. *Nucleic Acids Res.* **42**, 1414–1426 (2014).

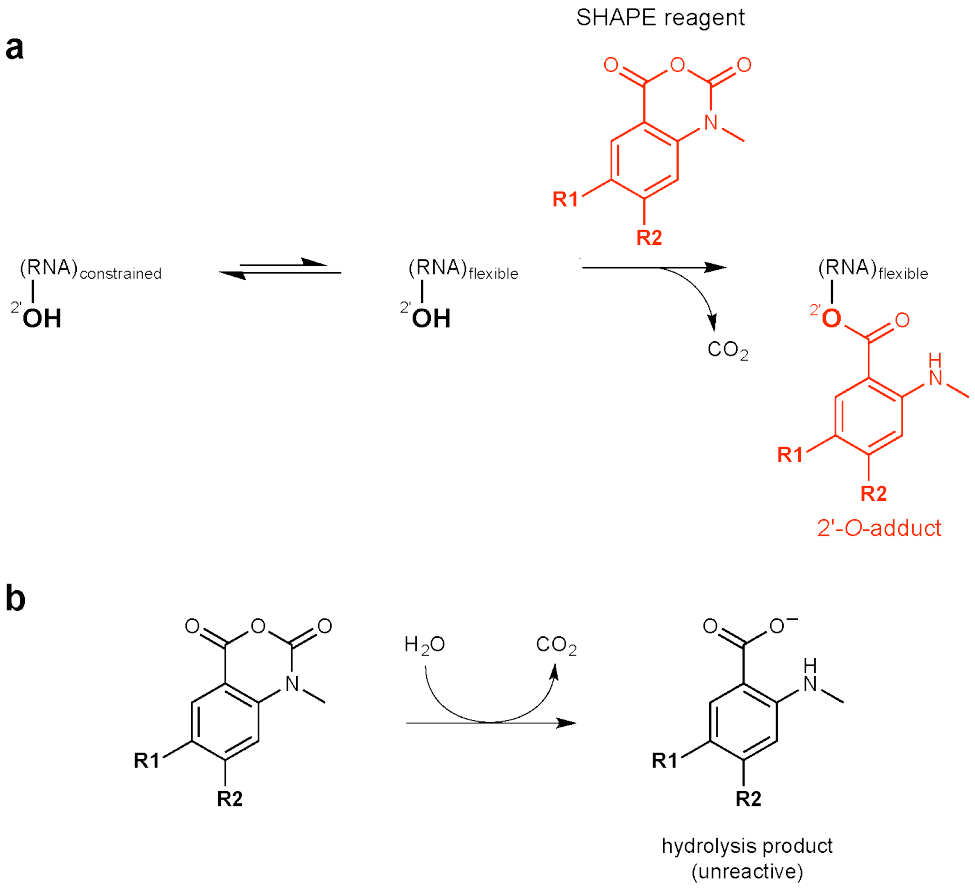
CHAPTER 2: ACCESSING SHAPE-MAP PROFILES OF RARE RNA TRANSCRIPTS[1]

Introduction

RNA plays many fundamental biological roles and interacts with small-molecule ligands, proteins, and other RNAs (*1*). In these roles, RNA molecules must adopt specific secondary and tertiary structures, the details of which are often difficult or impossible to characterize from sequence alone. Chemical probing techniques have proven to be powerful tools for understanding the critical features of RNA structure at both small and large scales. SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) uses small hydroxyl-selective electrophilic reagents to probe the reactivity of the RNA ribose 2′-OH group. SHAPE reactivities are insensitive to base identity and measure local nucleotide flexibility and dynamics (*2-4*) because flexible residues sample a wide range of conformations, a subset of which enhance the reactivity of the 2′-hydroxyl (*5*) (**Fig. 2.1a**).

SHAPE chemistry makes it possible to examine RNA structure in an especially thorough way because, with the exception of some post-transcriptionally modified RNAs, all RNA nucleotides carry a 2′-hydroxyl group. SHAPE reactions are self-inactivating via a competing hydrolysis reaction with water (**Fig. 2.1b**) and thus require no specific quench step. Because few compounds have a net reactivity as high as 55 M water, intrinsic SHAPE reactivities are largely insensitive to the presence of (additional) competing small molecules, ligands, and proteins. SHAPE experiments work robustly when performed in complex environments including those inside virions (*6-8*) and in living cells (*9*, *10*). By careful choice of SHAPE reagent (*9-11*) and experimental design, nucleotide flexibilities can be compared under

[1] The text and figures in this chapter are adapted from a publication written in collaboration with two colleagues. My critical contributions were the development of an amplicon-based strategy to enrich for low-abundance RNAs and the implementation of rapid, transposase-mediated library construction techniques. Elements of this chapter originally appeared in: M.J. Smola, S. Busan, G.M. Rice, N.A. Siegfried, and K.M. Weeks, Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile, and accurate RNA structure analysis. *Nature Protocols*. **10**, 1643-1669 (2015).



c

Reagent	R1	R2	Half-life (s)	Measures	
1M7	H	NO_2	17	***local nucleotide dynamics***	
1M6	NO_2	H	31	***stacking***	together, detect non-canonical and tertiary interactions
NMIA	H	H	260	***slow***	

Figure 2.1 SHAPE chemistry and useful SHAPE reagents. (a) SHAPE reagents react preferentially with the 2′-hydroxyl groups of conformationally flexible RNA nucleotides. (b) Quenching of SHAPE reagents via hydrolysis. (c) Overview of the three most useful SHAPE reagents. 1M7 is the workhorse SHAPE reagent; its reactivity with RNA measures local nucleotide flexibility. 1M6 and NMIA are selective for nucleobases that have one face available for stacking and that achieve a reaction-competent conformation on a slow timescale, respectively. Together, 1M6 and NMIA can be used to detect non-canonical and tertiary interactions in RNA and to increase the accuracy of secondary structure modeling.

different experimental or environmental conditions, including cell-free *versus* in-cell and as a function of ligand and protein binding. SHAPE reactivity information used as constraints in RNA modeling algorithms results in accurate secondary structure models (*12-14*).

To enable identification of sites of SHAPE modification using high-throughput sequencing, we developed a strategy termed mutational profiling or MaP. In MaP, specialized primer extension conditions allow reverse transcriptase to read through SHAPE-modified nucleotides without termination of the nascent cDNA strand (*15*). The bulky 2′-*O*-adduct at modified RNA positions induces incorporation of a nucleotide non-complimentary to the original RNA sequence at the corresponding position in the newly synthesized cDNA. Thus, during reverse transcription, the positions and relative frequencies of SHAPE adducts are directly and permanently encoded as mutations in the cDNA sequence. Two control experiments are performed in parallel (*15*): (*i*) a no-reagent control to characterize the background mutations resulting from the MaP procedure and (*ii*) a denaturing control, in which the RNA is modified roughly evenly along its length, to measure position-specific differences in adduct detection. Ultimately, MaP is largely impervious to the substantial sequence- and structure-based biases that are introduced during construction of the libraries required for massively parallel sequencing. Because the reverse transcriptase enzyme reads through the chemical adducts, the MaP approach is also insensitive to single-strand breaks or background degradation, and does not exhibit signal decay or drop-off, effects that result in significant noise in other high-throughput sequencing-based strategies for detecting chemical modification of RNA.

All current information supports the view that the MaP approach represents a no-compromises strategy for reading out the results of an RNA structure probing experiment by massively parallel sequencing. SHAPE-MaP was validated using a test set of RNAs ranging in size from 75 to 3,000 nucleotides. SHAPE-MaP was then recently used to analyze the entire HIV-1 RNA genome (~9,200 nt) (*15*). The new model recapitulates all previously known and accepted functional motifs and, moreover, identifies multiple new structural motifs including three experimentally validated pseudoknots.

MaP allows RNAs of virtually any size to be analyzed in a single experiment, facilitates high levels of multiplexing, and permits fully automated data analysis (*15*). Because the region under interrogation is completely sequenced in each read, sequence differences are revealed directly; therefore, the effects of sequence polymorphism and co-existing ribosnitches (*15*) can be evaluated in single experiments. The MaP experiment includes a DNA amplification step; therefore, individual RNAs present in scarce amounts, or in complex mixtures, can be examined. In sum, SHAPE-MaP yields robust nucleotide-resolution RNA structural information, enables accurate secondary structure modeling, can deconvolute sequence polymorphisms in a single experiment, readily allows analysis of low-abundance RNAs, and scales gracefully from short RNAs to transcriptome-wide analyses. We anticipate that SHAPE-MaP will contribute to deep understandings of the relationships between RNA structure and function.

Experimental design

SHAPE-MaP yields quantitative SHAPE reactivity data for nearly every position in an RNA by combining the well-validated SHAPE acylation reaction with specialized reverse transcriptase conditions and deep sequencing. Once the MaP step is complete the RNA structure information is converted to DNA sequence information that is largely immune to biases and artifacts introduced during the steps required to convert the original cDNAs into the libraries required by any specific sequencing platform. Here we describe protocols for SHAPE probing, mutational profiling, and three library construction approaches based on Illumina sequencing (**Fig. 2.2**). The relative merits of the library construction approaches are governed by the length and amount of the RNA of interest.

RNA folding and modification

SHAPE-MaP may be applied to RNAs of any length or complexity; however, since SHAPE modification inherently probes an ensemble of RNA molecules, it is critical to give careful thought to ensure that the RNA sample is folded in a biologically relevant and informative state prior to modification. For *in vitro*-transcribed RNA, suitable folding conditions have been described (*2*, *16*, *17*).

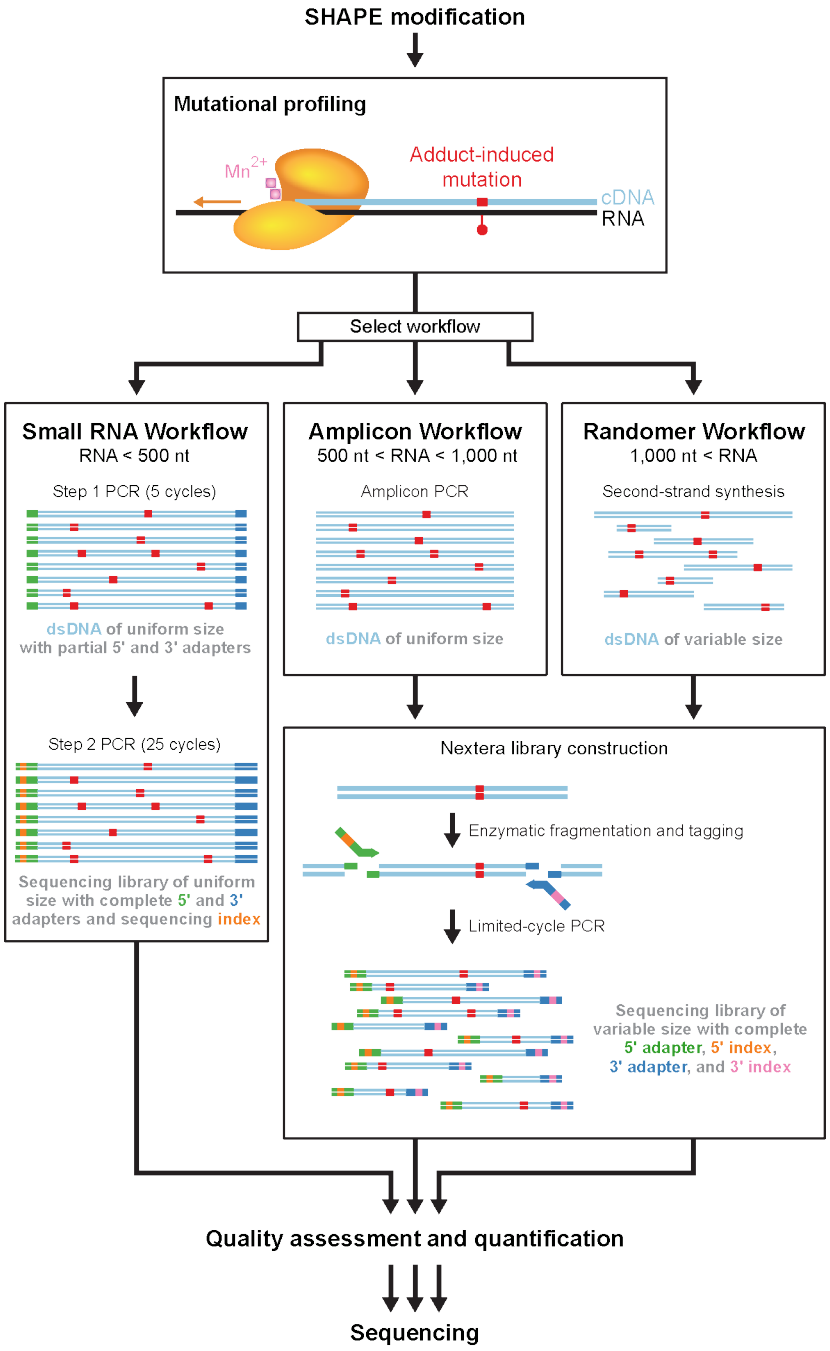


Figure 2.2 Overview of SHAPE-MaP workflows. RNAs are modified with a SHAPE reagent and subjected to reverse transcription under MaP conditions, during which adduct-induced mutations are recorded in the cDNA strand. One of three workflows is then used to construct high-quality libraries for sequencing and recovery of the SHAPE chemical probing information.

Methods for extraction, and purification of large, complex RNAs from virions (*6-8*, *18*) and cells (*10*, *12*) under native-like conditions have been described. RNA should be maintained under conditions likely to retain pre-existing RNA secondary and tertiary structure and use of denaturants, divalent ion chelators, or elevated temperature should be avoided. Direct interrogation of RNA structure directly inside cells by SHAPE is also well-established (*9*, *10*, *19*). This protocol emphasizes simple folding procedures for interrogating native-like and deproteinized RNAs, but any procedure that folds an RNA into an informative state can be used, provided the pH is in the 7.4-8.3 range.

Any SHAPE reagent can be used to modify RNA in a SHAPE-MaP experiment. In this protocol, we emphasize the use 1-methyl-7-nitroisatoic anhydride (1M7). Essentially identical approaches can be used with reagents 1-methyl-6-nitroisatoic anhydride and *N*-methyl-isatoic anhydride (1M6 and NMIA, respectively; **Fig. 2.1c**) (*14*). Reactions with 1M6 and NMIA are selective for nucleotides in which one face of the nucleobase is available for stacking and that undergo relatively slow conformational changes, respectively. These reagent-specific reactivities can be used both to identify residues that participate in non-canonical interactions and to improve RNA secondary structure modeling (*14*, *20*). In addition, the MaP strategy can also be used to follow time-resolved RNA processes, in 1-sec snapshots, using the benzoyl cyanide (BzCN) reagent (*21*). The modest solubility and rapid hydrolysis of these reagents make over-modification of RNA samples virtually impossible.

SHAPE electrophiles are added to the folded RNA (or virus or cell) and then incubated until the reagent has either reacted with RNA or degraded via hydrolysis with water (5 hydrolysis half-lives, **Fig. 2.1b-c**). Two additional reactions are performed in parallel: a no-reagent control and a denaturing control. In the no-reagent control reaction, folded RNA is incubated with solvent only (typically DMSO for SHAPE reagents); this control is used to measures the intrinsic background mutation rate of reverse transcriptase under MaP conditions. In the denaturing control reaction, RNA is suspended in a denaturing buffer containing formamide and incubated at 95 °C prior to modification with SHAPE reagent. Nucleotides are modified relatively evenly in this step, and the resulting site-specific mutation rates directly account for sequence- and structure-specific biases in detection of adduct-induced mutations.

Thus, a complete SHAPE-MaP experiment consists of three reactions: plus-reagent (+), minus-reagent (–), and denaturing control (DC).

Mutational Profiling (MaP)

After SHAPE modification of RNA, reverse transcriptase is used to create a mutational profile. This step encodes the positions and relative frequencies of SHAPE adducts as mutations in the cDNA sequence. Mutational profiling is efficient, with roughly 50% of SHAPE adducts detected as mutations in the cDNA (*15*). Whereas the reverse transcription reaction conditions are the same for any RNA, the researcher may choose one of two options regarding the type of DNA primer used. RNAs that are small enough to be sequenced end-to-end in a single massively parallel sequencing read (currently read lengths up to 600 nts are possible) can be subjected to reverse transcription with sequence-specific DNA primers. Specific primers can also be used when a specific sub-region of a large RNA is of interest (**Fig. 2.2, small RNA and amplicon workflows**). Use of gene- or region-specific primers also makes it possible to: (*i*) analyze a specific, relatively rare RNA in a complex mixture of RNAs or (*ii*) interrogate very rare, low abundance, RNAs. For analysis of large RNAs or the constituents of entire transcriptomes or multi-component ribonucleoprotein and long noncoding RNA assemblies, random primers facilitate even coverage of complex RNA states in a single experiment (**Fig. 2.2, randomer workflow**). Following mutational profiling with appropriate primers, one of three workflows is used for library construction, as detailed below.

Library construction and sequencing

The Small RNA workflow is ideally suited for RNAs or sub-regions of large RNAs that are short enough to be completely sequenced by a single unpaired sequencing read or by two mated paired-end sequencing reads. After reverse transcription with sequence-specific primers, purified cDNA is "tagged" with incomplete platform-specific adapters in a limited-cycle PCR reaction. The resulting dsDNA product is purified and further amplified in a second PCR reaction that completes the platform-specific adapter

including sequence indices for sample multiplexing. After purification, sequencing libraries are of uniform size and each DNA molecule contains the entire sequence of interest (**Fig. 2.3a**).

The amplicon workflow is well suited for large, low-abundance RNAs or sub-regions of larger RNAs that cannot be sequenced end-to-end by a single deep sequencing read. After reverse transcription, purified cDNA is amplified via PCR with sequence-specific primers. The resulting dsDNA is then enzymatically fragmented and tagged with platform-specific adaptors and multiplexing indices. Sequencing libraries constructed in this way are of variable size, with each molecule containing a fragment of the original amplicon (**Fig. 2.3b**). Typically, when the amplicon workflow is used to construct a sequencing library, reverse transcription is primed with sequence-specific primers. However, if the researcher wishes to generate a sequencing library for a specific region of an RNA that was previously reverse transcribed with random primers, the amplicon workflow allows for targeted "re-construction" of libraries.

The randomer workflow can be used to construct sequencing libraries when the RNA of interest is large (greater than ~500 nt) and reasonably pure; for example, in the case of a viral RNA genome. In addition, the randomer workflow also is appropriate for analysis of very complex systems including complete RNA transcriptomes. After reverse transcription with appropriate random primers, purified cDNA is converted to dsDNA and then enzymatically fragmented and tagged with platform-specific adapters and multiplexing indices. The resulting sequencing library is of variable size, and each molecule corresponds to a fragment of the original RNA (**Fig. 2.3c**). After construction of high-quality SHAPE-MaP libraries by either of the approaches described here, the library is subjected to sequencing with a massively parallel sequencing instrument. The MaP approach is fully compatible with any platform with a high per-nucleotide calling accuracy.

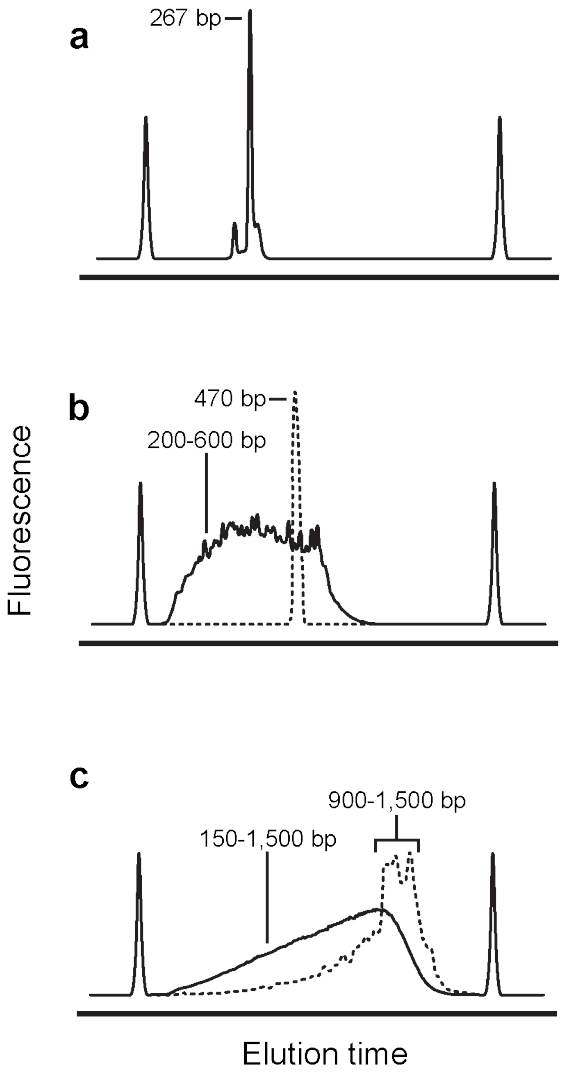


Figure 2.3 Representative library size distributions as a function of workflow. (a) Bioanalyzer electropherogram of a TPP riboswitch library produced with the small RNA workflow. The small peak to the left of the major product is unconverted step 1 PCR product. (b) A library (solid line) constructed from a single amplicon (dashed line) via the amplicon workflow. The library contains some DNAs slightly larger than the original amplicon because platform-specific adaptors are added to near-full length fragments. (c) A library (solid line) constructed via the randomer workflow. The sizes of dsDNA produced by second-strand synthesis (dashed line) set the upper limit on the library size.

Results

Small RNA workflow – TPP riboswitch

A SHAPE profile for the aptamer domain of the TPP riboswitch was readily obtained using the small RNA workflow. Using these data, secondary structure modeling for this riboswitch RNA improved from a base pair prediction accuracy of 73%, obtained using a nearest-neighbor thermodynamic algorithm alone, to 96%, using SHAPE-directed modeling (*15*). Observed reactivities correspond closely to those expected based on the local nucleotide flexibilities for the ligand-bound RNA (**Fig. 2.4a-c**). Reactive nucleotides fall in conformationally flexible single-stranded regions, especially the L3 loop and the J2-4 and J3-2 strands. Overall, relatively few nucleotides are reactive by SHAPE, consistent with the highly constrained conformation of this RNA. SHAPE-MaP also reveals fine differences corresponding to changes induced upon binding by the TPP ligand (**Fig. 2.4b-d**). Ligand interactions induce a large structural organization in the L5 loop and in the J3-2 elements in the ligand-binding pocket.

Amplicon workflow – Mouse ribosomal RNA

The utility of the amplicon workflow is two-fold: (*i*) it can be used to enrich for low-abundance RNAs and (*ii*) it allows for focused sequencing of a specific region of interest within a large RNA. As an example of this second use, primers targeting the *Mus musculus* 18S rRNA 3′ domain were used to generate SHAPE-MaP data for this region using the amplicon workflow. Even though the amplicon workflow introduces 20-30 additional cycles of PCR relative to traditional library preparations (e.g. the randomer workflow), SHAPE reactivity data are very similar between the two approaches. The amplicon approach retains information about the original RNA structure (**Fig 2.5a**) and, over 740 nucleotides of the mouse 18S rRNA, data from the amplicon and randomer workflows correlate strongly, with R = 0.85 (**Fig. 2.5b**). Thus, the amplicon workflow is a viable and useful alternative to other enrichment or targeting techniques and is only possible because of the internal nature of mutations generated with the MaP approach.

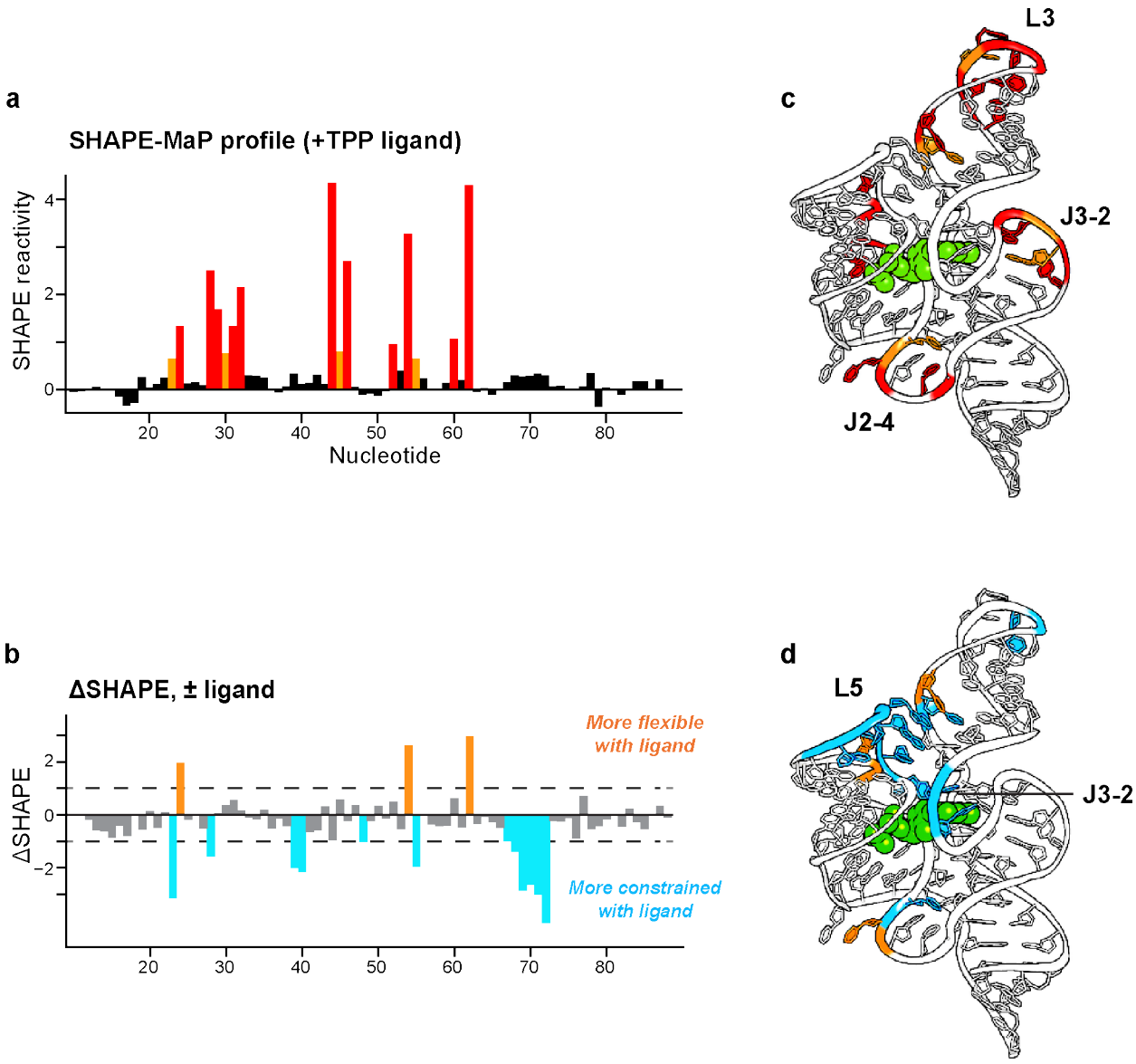


Figure 2.4 Example of results obtained with the small RNA workflow. (a) SHAPE profile of the TPP riboswitch produced using the small RNA workflow. (b) Difference SHAPE profile illustrating conformational changes induced in the TPP riboswitch upon ligand binding. (c) SHAPE-MaP reactivities superimposed on the structure (PDB: 2GDI) of the ligand-bound TPP riboswitch. Red, orange, and black correspond to high, moderate, and low reactivities, respectively, and correspond to reactivities shown in (a). (d) Visualization of ligand-induced conformational changes on the TPP riboswitch structure. Reactivity changes (orange and blue) are the same as shown in (b).

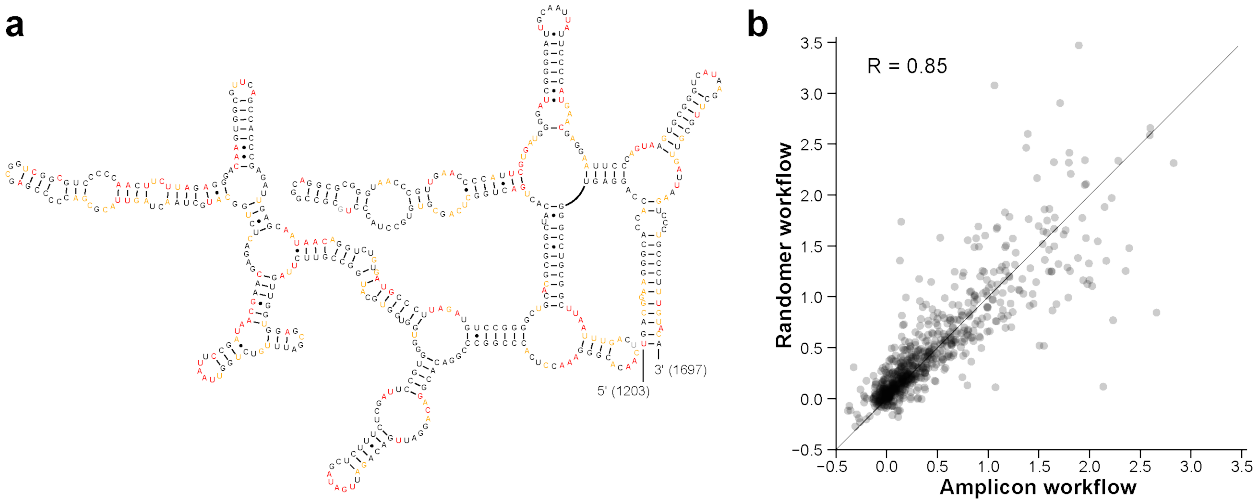


Figure 2.5 Example of results obtained with the amplicon workflow. (a) SHAPE reactivity values plotted on the 3′ domain of the mouse 18S rRNA. Black, orange, and red colors correspond to low, medium, and high reactivities. High reactivities occur predominantly in single-stranded regions, indicating that SHAPE reactivities obtained via the amplicon workflow accurately retain structural information. (b) Correlation between reactivity values obtained by the amplicon and randomer workflows. These data represent 740 nucleotides of the 18S rRNA and correlate with R = 0.85, showing that there is little difference between the two workflows.

Randomer workflow – Bacterial ribosomal RNA

Large RNAs like the bacterial small and large ribosomal subunit RNAs (16S and 23S, respectively) are readily examined by applying the randomer workflow to total *E. coli* RNA (**Fig. 2.6**). Using random primers, both RNAs can be studied simultaneously with fully automated analysis involving approximately 3 days of hands-on experimental effort. The major post-processing requirement is that the per-nucleotide hit level be sufficiently high to permit full recovery of the underlying SHAPE data. In general, the hit level should be 5 or greater, corresponding to a read depth of 1-2,000 (*15*).

The 23S rRNA subunit alone represents ~2,900 nucleotides of SHAPE reactivity information after computational data processing (**Fig. 2.6a**). Comparing the SHAPE reactivities for domain IV of the 23S rRNA with the accepted sequence covariation-derived structural model (**Fig. 2.6b-c**) shows good agreement. Regions involved in canonical base pairs have low SHAPE reactivity, indicating that they are structurally constrained. Conversely, single-stranded loop and bulge regions have high SHAPE reactivity, indicating structural flexibility. Because of the inherent scalability of the MaP approach, these data – spanning several thousand nucleotides – are as accurate at single-nucleotide resolution as are data from a short RNA, like the TPP riboswitch.

Conclusion

In sum, SHAPE-MaP yields quantitative nucleotide-resolution RNA structural information, accessible via one (or more) of several convenient experimental workflows. SHAPE data enable accurate secondary structure modeling, allow for the identification of well-determined regions within large RNAs, facilitate discovery of novel functional RNA motifs, make possible deconvolution of sequence polymorphisms in a single experiment, detect diverse effects of ligand and protein binding, readily allow analysis of low-abundance RNAs, and scale gracefully from short RNAs to transcriptome-wide analyses, including in cells. We anticipate that SHAPE-MaP will contribute to deep understandings of the relationships between RNA structure and function.

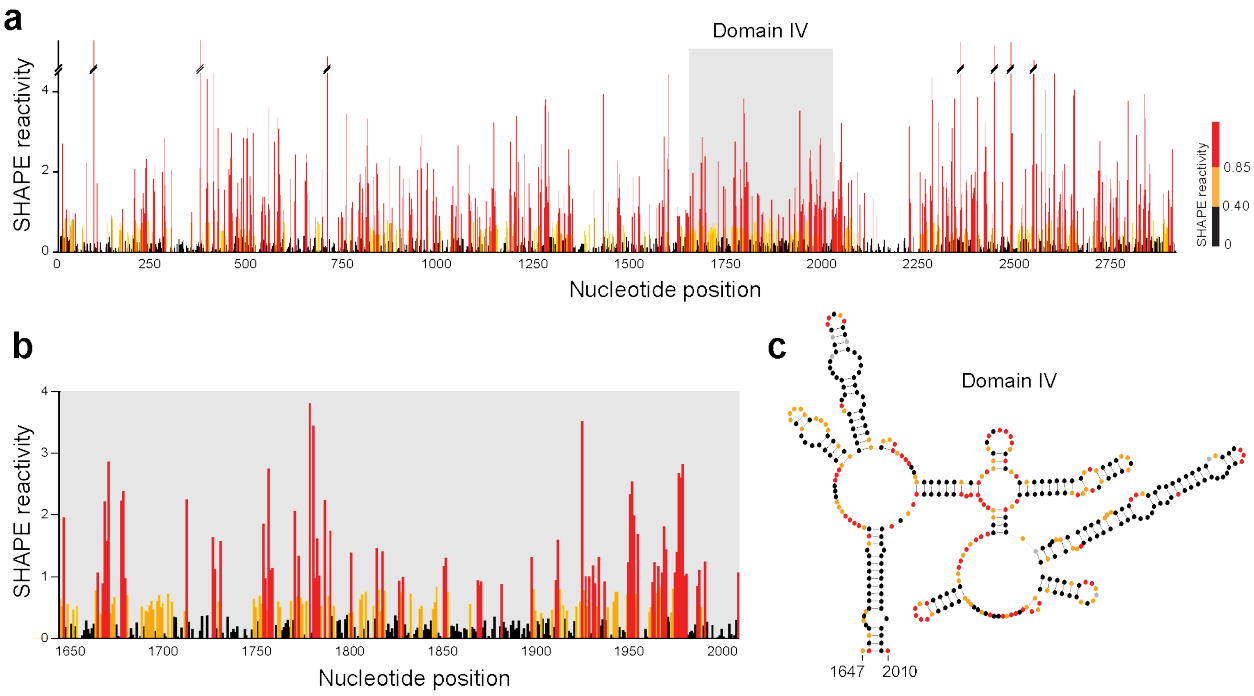


Figure 2.6 Example of results obtained with the randomer workflow. (a) SHAPE reactivities across the entire *E. coli* 23S rRNA obtained in a single experiment. (b) Expanded view of SHAPE reactivities for Domain IV of the 23S rRNA. (c) Accepted secondary structure of Domain IV colored by SHAPE reactivity. Reactive nucleotides (orange and red) occur predominantly in single-stranded regions.

Methods

SHAPE-MaP data for the TPP riboswitch and *E. coli* 16S rRNA were obtained previously (*22*) using the small RNA and randomer workflows, respectively. Data for the mouse 18S rRNA was obtained as described below.

RNA extraction and modification

To obtain SHAPE-MaP data for the mouse 18S rRNA, trophoblast stem cells (TSCs) were cultured as described (*23*). Approximately 10^6 TSCs were washed and pelleted in ice-cold PBS, resuspended in 2.5 ml Lysis Buffer [40 mM Tris, pH 7.9, 25 mM NaCl, 6 mM $MgCl_2$, 1 mM $CaCl_2$, 256 mM sucrose, 0.5% Triton X-100, 1,000 U/ml RNasin (Promega), 450 U/ml DNase I (Roche)], and rotated at 4 °C for 5 minutes. Cells were then pelleted at 4 °C for 2 minutes at 2250 g, resuspended in 40 mM Tris pH 7.9, 200 mM NaCl, 1.5% SDS, and 500 μg/ml of Proteinase K, and rotated at 20 °C for 45 minutes. RNA was then extracted twice with phenol:chloroform:isoamyl alcohol (24:24:1) pre-equilibrated with 1× Folding Buffer (100 mM HEPES, pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$), followed by one extraction with chloroform. RNA was exchanged into 1.1× Folding Buffer using a desalting column (PD-10, GE Life Sciences) and incubated at 37 °C for 20 minutes. Approximately 3 μg RNA was then added to a one-ninth volume of 1M7 in neat DMSO (10 mM final concentration), and then incubated at 37 °C for 5 minutes. Modified RNA was purified (RNeasy Midi spin column, Qiagen) and eluted in approximately 50 μl H_2O. No-reagent negative control RNA was prepared in the same way except that neat DMSO was substituted for SHAPE reagent.

To prepare the denatured control, TSCs were grown as described (*23*), and total RNA was isolated using TRIzol (Ambion). Approximately 1.5 μg RNA was then resuspended in 150 μl 1.1× Denaturing Control Buffer [55 mM HEPES pH 8.0, 4.4 mM EDTA, 55% formamide (v/v)] and incubated at 95 °C for 1 minute. Aliquots of 45 μl of denatured RNA were then added to 5 μl of 100 mM 1M7, 1M6, or NMIA, and allowed to react at 95 °C for 1 minute. After modification, RNA was purified (RNeasy Mini spin column, Qiagen) and eluted in approximately 50 μl H_2O.

SHAPE-MaP

Mutational profiling reverse transcription reactions were primed with 2 pmol of an rRNA-specific primer (5′-ACCATCCAATCGGTAGTAGC-3′) (*22*). The resulting cDNA was purified (G-50 spin column, GE Healthcare) and eluted in 50 μl H_2O. cDNAs were then used as templates for for either by second strand synthesis (40 μl input, NEBNext Second Strand Synthesis Module, NEB) or PCR (0.5 μl input) with rRNA-specific primers (forward: 5′-GAGGTGAAATTCTTGGACCG-3′, reverse: 5′-ACCATCCAATCGGTAGTAGC-3′). PCR reactions were performed in 50 μl volumes (1× Q5 Reaction Buffer, 200 μM dNTPs, 0.5 μM each primer, 0.02 U/μl Q5 high-fidelity DNA polymerase) using a touchdown format: 98 °C for 30 s, 25 cycles of [98 °C for 10 s, 72 °C for 30 s (decreasing by 1 °C per cycle until 60 °C), 72 °C for 30 s], 72 °C for 2 min. The resulting dsDNA was purified (Agencourt XP beads, Beckman Coulter) before construction of high-throughput sequencing libraries (Nextera XT, Illumina). Libraries were purified (Agencourt XP beads, Beckman Coulter) prior to sequencing on an Illumina Miseq instrument, generating 2 × 250 paired-end reads.

SHAPE profile generation

Raw sequencing reads were quality-trimmed and aligned to the mouse 18S rRNA sequence (GenBank accession: NR_003278) using *ShapeMapper* (http://chem.unc.edu/rna/software.html). The resulting SHAPE reactivity profiles, corresponding to amplicon workflow and randomer workflow library preparations, were normalized to the same scale prior to comparison.

REFERENCES

1. P. A. Sharp, The Centrality of RNA. *Cell*. **136**, 577–580 (2009).

2. E. J. Merino, K. A. Wilkinson, J. L. Coughlan, K. M. Weeks, RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc*. **127**, 4223–4231 (2005).

3. C. M. Gherghe, Z. Shajani, K. A. Wilkinson, G. Varani, K. M. Weeks, Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S2) in RNA. *J. Am. Chem. Soc*. **130**, 12244–12245 (2008).

4. K. A. Wilkinson *et al*., Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *RNA*. **15**, 1314–1321 (2009).

5. J. L. McGinnis, J. A. Dunkle, J. H. D. Cate, K. M. Weeks, The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc*. **134**, 6617–6624 (2012).

6. K. A. Wilkinson *et al*., High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol*. **6**, e96 (2008).

7. C. Gherghe *et al*., Definition of a high-affinity Gag recognition structure mediating packaging of a retroviral RNA genome. *Proc. Natl. Acad. Sci. U.S.A*. **107**, 19248–19253 (2010).

8. E. J. Archer *et al*., Long-Range Architecture in a Viral RNA Genome. *Biochemistry*. **52**, 3182–3190 (2013).

9. J. Tyrrell, J. L. McGinnis, K. M. Weeks, G. J. Pielak, The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry*. **52**, 8777–8785 (2013).

10. J. L. McGinnis, K. M. Weeks, Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry*. **53**, 3237–3247 (2014).

11. S. A. Mortimer, K. M. Weeks, A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc*. **129**, 4144–4145 (2007).

12. K. E. Deigan, T. W. Li, D. H. Mathews, K. M. Weeks, Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A*. **106**, 97–102 (2009).

13. C. E. Hajdin *et al*., Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A*. **110**, 5498–5503 (2013).

14. G. M. Rice, C. W. Leonard, K. M. Weeks, RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA*. **20**, 846–854 (2014).

15. N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson, K. M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*. **11**, 959–965 (2014).

16. C. D. S. Duncan, K. M. Weeks, SHAPE Analysis of Long-Range Interactions Reveals Extensive and Thermodynamically Preferred Misfolding in a Fragile Group I Intron RNA. *Biochemistry*. **47**, 8504–8513 (2008).

17. J. L. McGinnis, C. D. S. Duncan, K. M. Weeks, High-throughput SHAPE and hydroxyl radical analysis of RNA structure and ribonucleoprotein assembly. *Meth. Enzymol.* **468**, 67–89 (2009).

18. J. M. Watts *et al.*, Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. **460**, 711–716 (2009).

19. R. C. Spitale *et al.*, RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* **9**, 18–20 (2012).

20. K.-A. Steen, G. M. Rice, K. M. Weeks, Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *J. Am. Chem. Soc.* **134**, 13160–13163 (2012).

21. S. A. Mortimer, K. M. Weeks, Time-resolved RNA SHAPE chemistry. *J. Am. Chem. Soc.* **130**, 16178–16180 (2008).

22. N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson, K. M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*. **11**, 959–965 (2014).

23. J. Quinn, T. Kunath, J. Rossant, in *Placenta and Trophoblast* (Humana Press, New Jersey, 2005), vol. 121, pp. 123–146.

CHAPTER 3: DETECTION OF RNA-PROTEIN INTERACTIONS IN LIVING CELLS WITH SHAPE

Introduction

Nearly all RNAs, regardless of function, interact with one or more protein partners in order to function properly (*1*, *2*). Characterizing ribonucleoprotein (RNP) complexes is thus an important step in understanding RNA function. Several well-validated approaches have been developed to explore RNP complexes (*3*). These methods provide many valuable insights but often have a limited scope due to affinity purification steps that require prior knowledge about the RNA or protein of interest. As RNA structure studies expand to 'omics scales, direct and accurate approaches for uncovering sites of interaction between the transcriptome and the proteome will become increasingly important.

SHAPE-MaP (selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling) combines well-validated SHAPE RNA structure probing chemistry (*4*, *5*) with massively-parallel sequencing to enable high-throughput interrogation of RNA flexibility at single-nucleotide resolution (*6*, *7*). When probed with SHAPE reagents, conformationally flexible nucleotides exhibit high reactivity. Conversely, nucleotides constrained by base pairing or by other interactions show low reactivities. The quantitative relationship between SHAPE reactivity and conformational flexibility is maintained even for nucleotides that are not solvent accessible as visualized in static RNPs (*5*), indicating that SHAPE can be used to probe the interiors of RNA-protein complexes. Previous work has shown that SHAPE reagents readily modify RNAs in living cells (*8-13*). Finally, SHAPE-MaP uniquely allows for thorough and quantitative analysis of specific individual RNAs within the contents of an entire transcriptome with the use of targeted primers (*6*, *14*). Thus, SHAPE-MaP offers a broadly useful strategy for probing the structure of the entire transcriptome, or elements thereof, under diverse experimental conditions.

A wide variety of RNA structure probing methods have been proposed (*15*, *16*), most of which depend on accurately identifying and quantifying cDNA ends created when reverse transcriptase enzymes encounter a chemical adduct or cleavage site. These methods all involve a critical adapter-ligation step. In principle, these methods make it straightforward to perform RNA structure probing on the entire contents of a given transcriptome; in practice, it is currently almost impossible to perform the adapter-ligation step quantitatively (*17*, *18*). Moreover, transcriptome-wide experiments are strongly subject to the classic multiple and sparse measurement problems such that many measurements are unlikely to be statistically significant (*6*) and do not survive follow-up validation (*19*). Thus, an important challenge in large-scale and in-cell RNA structure analyses is to robustly detect significant structural changes.

We hypothesized that most RNA-protein interactions would affect the flexibility of nucleotides at the binding site and that by comparing SHAPE reactivities of deproteinized RNA (*ex vivo*) with reactivities obtained by probing RNA in living cells (*in cellulo*), it would be possible to characterize sites of RNP interactions (**Fig. 3.1a**). We developed an analysis framework that enables detection of RNP interactions based upon three principles: (*i*) RNA-protein interactions strongly affect SHAPE reactivity, either positively or negatively; (*ii*) due to measurement errors and the large number of reactivity measurements made, not all apparent reactivity changes are significant; and (*iii*) most RNA-protein interaction sites (*20*) will span sites of five or more nucleotides in primary sequence.

To identify changes in SHAPE reactivity associated with protein interactions, we used the SHAPE reagent 1-methyl-7-nitroisatoic anhydride (1M7) to generate *in cellulo* and *ex vivo* SHAPE-MaP datasets for U1, 5S, and SRP RNAs (**Fig. 3.1a**). These RNAs enable evaluation of RNPs located both in the nucleus and in the cytoplasm and high-resolution structures of their complexes with proteins are available (*21-26*). Alternative SHAPE reagents have been proposed for *in cellulo* modification (*8*, *12*). We compared 1M7 SHAPE-MaP with recently published in-cell SHAPE (icSHAPE), which uses a clickable RNA acylation reagent (NAI-N3) to allow enrichment of RNAs modified with this relatively weakly reactive reagent. We found that icSHAPE measurements show very low correlation with those

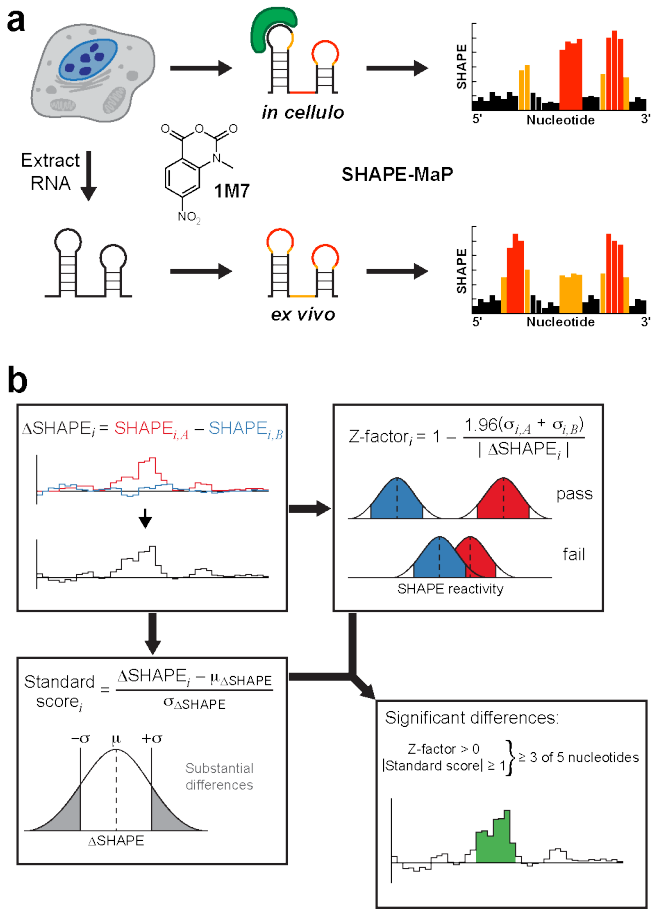


Figure 3.1 Experimental and analytical framework for detecting SHAPE-MaP reactivity differences. (a) Total cellular RNA is treated with 1M7 under native conditions in living cells (top) or following non-denaturing extraction into folding buffer (bottom). RNAs that interact stably with cellular proteins (green) exhibit different SHAPE reactivities under *in cellulo* versus *ex vivo* conditions. Black, orange, and red illustrate low, moderate, and high reactivities, respectively on the secondary structure diagram and in SHAPE-MaP profiles. (b) Calculation of differences in SHAPE reactivities (ΔSHAPE) between experimental conditions *A* and *B* (upper left). If (*i*) the Z-factor for a nucleotide is greater than zero, indicating that the 95% confidence intervals of measurements in the two conditions do not overlap, (*ii*) the standard score is greater than one standard deviation from the mean ΔSHAPE (lower left), and (*iii*) three of five nucleotides in a sliding window meet both Z-factor and standard score criteria (lower right), the reactivity difference is accepted as significant.

obtained with SHAPE-MaP. Thus, we chose 1M7 for its short half-life, ability to accurately report RNA secondary structure *ex vivo* (*4-7*, *27*) and in living cells (*9-11*), and because in-cell reactivity of 1M7 is sufficiently robust that downstream enrichment is not required.

Differences in SHAPE reactivities (ΔSHAPE) were calculated by subtracting *in cellulo* SHAPE reactivities from *ex vivo* reactivities (**Fig. 3.1b,** upper left) and averaging over a three-nucleotide sliding window to reduce local signal fluctuation. By this definition, positive ΔSHAPE values indicate protection from modification in the cellular environment, and negative ΔSHAPE reports enhanced reactivity in cells.

In a SHAPE-MaP experiment, discrete mutation events contribute to the overall reactivity at each nucleotide and are well modeled by a Poisson distribution (*6*). The standard error in the SHAPE reactivity measurement can therefore be estimated for every nucleotide (*6*). We used these error estimates to perform a modified Z-factor test (*6*, *28*) for all positions in a given RNA (**Fig. 3.1b,** upper right). This test compares the magnitude of ΔSHAPE with the associated *ex vivo* and *in cellulo* measurement errors, identifying nucleotides for which the magnitudes of the errors are too large for the ΔSHAPE values to be significant. We formulated the Z-factor test such that the underlying *ex vivo* and *in cellulo* SHAPE reactivities must differ by more than 1.96 standard deviations (Z-factor > 0), ensuring that the 95% confidence intervals of each measurement do not overlap.

For many nucleotides, SHAPE-MaP reactivity measurements have very small errors, allowing for the possibility that a trivially small ΔSHAPE could be considered significant according to the Z-factor test. We expected most stable protein-RNA interactions to have a strong effect on the reactivity of nucleotides at the binding site, so we calculated a standard score at each nucleotide to identify the largest ΔSHAPE values (**Fig. 3.1b,** lower left). This metric compares ΔSHAPE of a given nucleotide with the ΔSHAPE of all other nucleotides in the RNA, regardless of Z-factor. We required that the absolute value of each standard score be ≥ 1, meaning that individual ΔSHAPE values must be at least one standard deviation away from the mean ΔSHAPE. Thus, only the largest ΔSHAPE values are considered for further analysis. To determine final RNA-protein interaction sites, we filtered by Z-factor and standard score simultaneously (**Fig. 3.1b,** lower right). If, in a 5-nucleotide window, at least three nucleotides had

a Z-factor > 0 and an absolute standard score ≥ 1, those three (or more) nucleotides were considered to have significant cell-induced changes in SHAPE reactivity.

In this work, we show that biochemical RNA structure probing data generated with the well-validated SHAPE-MaP approach can be used to identify significant, meaningful changes in RNA structure between two states. Here, these states are the RNA in healthy mouse trophoblast stem cells and the same RNAs gently extracted from cells. We validate our approach with the abundant and well-characterized U1 small nuclear RNA (snRNA), 5S rRNA, and signal recognition particle (SRP) RNP complexes, illustrating that the statistical filters implemented in our analysis robustly identify sites of protein interactions. We then examine RNase MRP, an important RNP complex whose in-cell architecture is relatively poorly understood. Our analysis confirms several reported RNA-protein interactions within the complex, and also characterizes the underlying molecular phenotype of many disease-associated mutations.

Results

Comparison of SHAPE-MaP and icSHAPE

We compared the similarity of RNA structure probing data for SHAPE-MaP and icSHAPE (*12*) experiments using the SRP RNP complex and six mRNAs. When probing SRP *ex vivo*, we found that strong icSHAPE signals are generally indicative of flexible nucleotides (**Fig. 3.2a**), as expected for SHAPE reagents. However, in comparison to SHAPE-MaP, the icSHAPE results appear roughly binary, with relatively few intermediate reactivity values. In comparing *in cellulo* data, SHAPE-MaP and icSHAPE data show very poor correlations. The differences between *ex vivo* and *in cellulo* icSHAPE values exhibit strong, punctate positive values throughout the RNA and dramatically strong negative values near the 5′ end (**Fig. 3.2b**). The icSHAPE data would suggest that the SRP RNA undergoes extreme and widespread conformational changes in cells, which is not consistent with prior work on this RNA (*24*, *25*). Further *in cellulo* comparison of six mRNAs produced similar results; the correlation

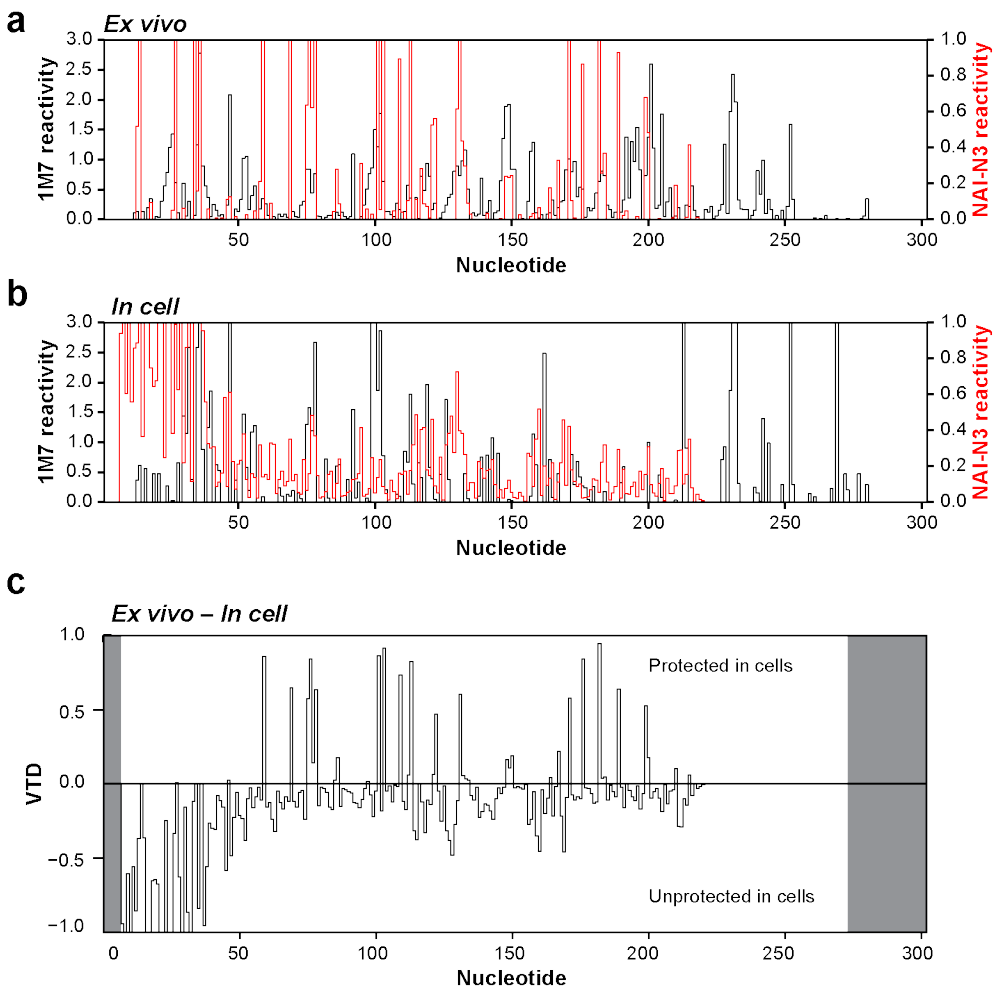


Figure 3.2 Comparison of SHAPE-MaP and icSHAPE reactivities. (a) *Ex vivo* SHAPE reactivity of the mouse SRP RNA derived from SHAPE-MaP (black) and icSHAPE (red) (*12*). Many nucleotides with significant SHAPE-MaP reactivities are scored as unreactive by icSHAPE and, conversely, icSHAPE reports multiple nucleotides as reactive that have little or no reactivity by SHAPE-MaP. (b) In-cell SHAPE reactivity of SRP RNA, colored as in (a). Note the very high icSHAPE reactivities at the 5′ end of the RNA, which likely reflects long modification times, adaptor ligation, or both. (c) Plot of the "*vitro–vivo* difference" (VTD) (*12*) for SRP RNA, in which the in-cell icSHAPE profile is subtracted from the *ex vivo* profile. The VTD profile exhibits strong, punctate positive values throughout the RNA and strong negative values near the 5′ end. These observations suggest that the SRP RNA undergoes extreme and large-scale conformational changes in cells, which is not consistent with accepted features of this RNA (*24*, *25*). Grey shading indicates regions for which no icSHAPE values were generated.

between icSHAPE and SHAPE-MaP was consistently poor, with correlation coefficients ranging from 0.1-0.3.

Validation of the ΔSHAPE approach

We used SHAPE-MaP to analyze three model RNAs *ex vivo* and *in cellulo*. The U1 snRNA is localized in the nucleus and forms the U1 snRNP complex upon binding several proteins: U1A, U1C, U1-70K, and the heteroheptameric Sm ring. Comparison of U1 snRNA *ex vivo* and *in cellulo* SHAPE reactivities revealed distinct qualitative reactivity differences throughout the RNA (**Fig. 3.3a**). Due to differences in the number of individual mutation events observed relative to the times a given nucleotide was sequenced, the estimated errors vary as a function of nucleotide position and are much greater for some reactivity measurements than others. This is a feature shared by all RNA structure probing experiments read out by massively parallel sequencing but is explicitly and uniquely measured using the MaP strategy. If a naïve approach had been taken that ignored these errors, multiple regions would have been (incorrectly) identified as having significant SHAPE reactivity differences (**Fig. 3.3b**, grey and green shading). Only a subset of these regions are involved in true RNA-protein interactions; the remainder are analysis artifacts caused by the measurement uncertainties that occur in any experiment, especially those read out by massively parallel sequencing. When we applied the complete analysis framework in which the Z-factor test is used to account for these errors, only three regions of significant ΔSHAPE were identified (**Fig. 3.3b**, green shading only). The locations of these positive ΔSHAPE values correspond precisely to known interactions sites of U1-70K, U1A, and the Sm ring proteins (**Fig. 3.3c**).

We next examined the differences in reactivities of the SRP RNA *ex vivo* versus *in cellulo* (**Fig. 3.4a**). The SRP RNA associates with six proteins and is comprised of an Alu domain at the 5′ end connected by a long central helix to the S domain. The Alu domain is bound by the SRP9-SRP14 (SRP9/14) heterodimer, and the larger S domain interacts with SRP19, SRP54, and the SRP68-SRP72 (SRP68/72) heterodimer. The SHAPE reactivity changes identified by our analysis were largely localized

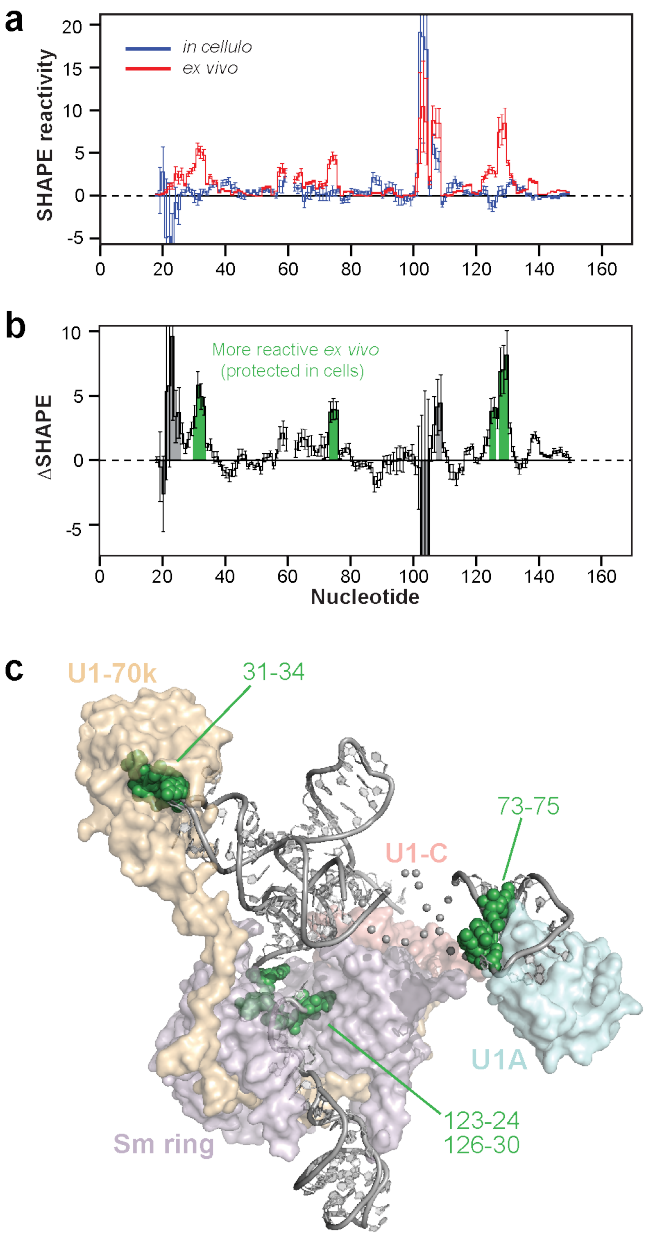


Figure 3.3 Identification of protein binding sites by ΔSHAPE analysis. (a) Smoothed SHAPE reactivities for U1 snRNA *in cellulo (*blue) and *ex vivo* (red). (b) ΔSHAPE values for the U1 snRNA. Significant reactivity changes established by the ΔSHAPE analysis framework are shaded green. If measurement errors were not taken into account, several off-target interaction sites would have been incorrectly identified as significant (grey shading). Primer-binding regions for which no data are available are shown with dashed lines. (c) Model of the human U1 snRNA complex including U1-70K (orange), U1-C (red), U1A (blue), and Sm ring proteins (purple; subunit D1 excluded for clarity). RNA is shown as a ribbon. Nucleotides that exhibit significant ΔSHAPE values are emphasized as spheres.

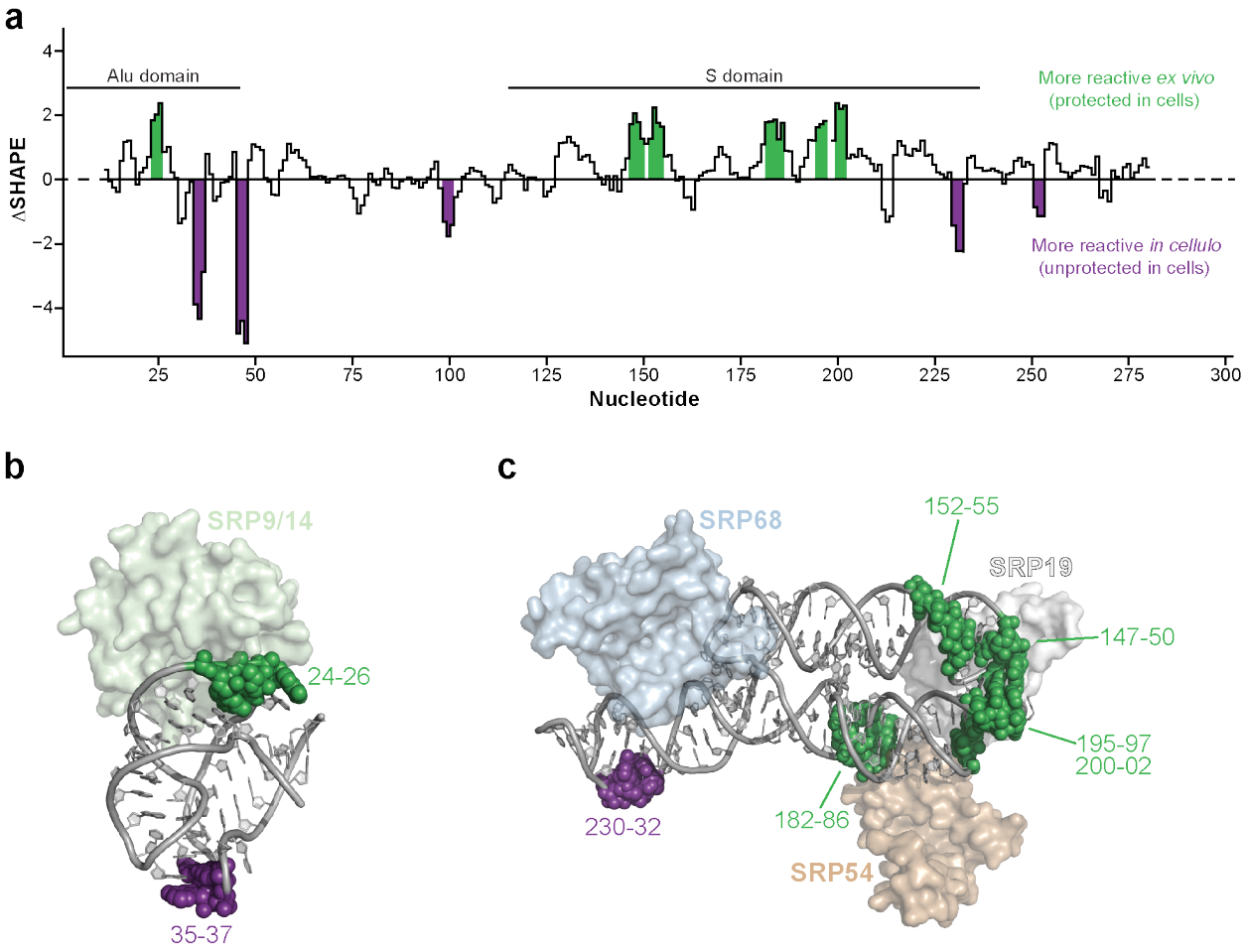


Figure 3.4 Summary of results obtained for the SRP RNA. (a) ΔSHAPE profile for the entire SRP RNA. *In cellulo* protections are shaded green, and *in cellulo* reactivity enhancements are purple. Locations of the Alu and S domains are indicated. (b) Crystal structure of the Alu domain bound to SRP9/14. Nucleotides with significant reactivity differences are labeled. (c) Model of the S domain bound to SRP68, SRP19 and SRP54 with significant reactivity differences indicated.

to these two domains, consistent with a lack of protein binding in the central helix.

In the Alu domain, we observed *in cellulo* protection at the SRP9/14 binding site (nts 24-26). We also detected enhanced *in cellulo* reactivity at nucleotides 35-37 and 46-48, consistent with protein-induced tertiary structure changes (**Fig. 3.4b**). In the S domain, we observed extensive *in cellulo* protection where SRP19 and SRP54 bind (**Fig. 3.4c**). Binding by SRP68/72 involves insertion of an α–helix into the major groove of the central helix, causing an adjacent asymmetric internal loop to open (*24*). Consistent with this observation, we detect enhanced *in cellulo* reactivity on the opened side of this loop at positions 230-232 (**Fig. 3.4c**). The interaction between SRP RNA and the complete SRP68/72 heterodimer has not been characterized at high resolution; however, cryo-electron microscopy data provide evidence that a portion of SRP68/72 interacts with the central helix at an internal "hinge" loop comprised of nucleotides 97-104 and 249-253 (*25*). In-cell SHAPE supports this observation, as enhanced *in cellulo* reactivity was noted on both sides of the loop at nucleotides 99-101 and 251-253, and suggests a local conformational change also occurs at nucleotides 230-232. Overall, every region of significant *in cellulo* protection in the SRP RNA identified by our analysis framework corresponds to sites of direct protein binding.

In examining the 5S rRNA, which forms a complex with ribosomal protein L5, we detected several regions of *in cellulo* protection (**Fig. 3.5a**). These sites correspond to previously identified contacts between 5S rRNA and L5 (**Fig. 3.5b**) (*8*, *29*). There were no other sites with significant ΔSHAPE values, although many ribosomal proteins are known to be located near the 5S particle in fully assembled ribosomes. These results are consistent with the observations that a significant fraction of cellular 5S RNPs are not ribosome-associated (*30*) and that 5S rRNA adopts multiple conformations even when associated with the ribosome (*31*). We infer that ΔSHAPE analysis primarily detects only the stable protein-RNA interactions in the 5S rRNA, and that these involve L5.

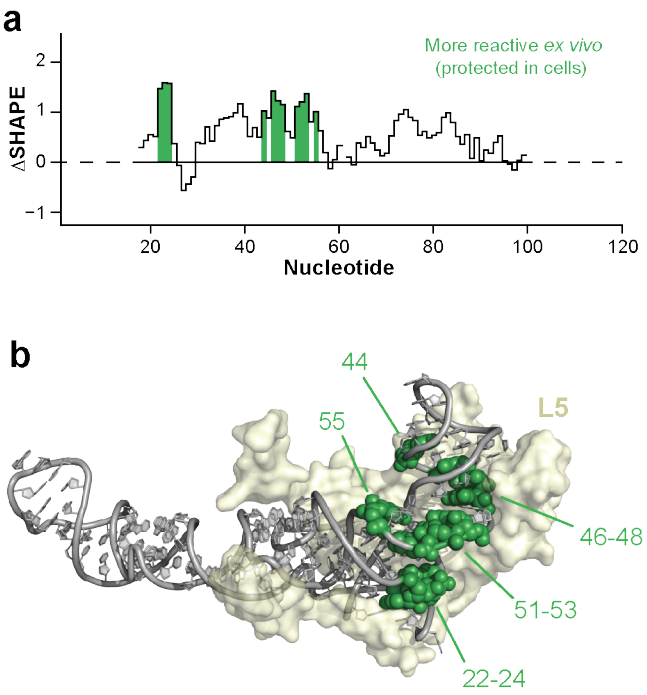


Figure 3.5 Summary of results obtained for the 5S rRNA. (a) ΔSHAPE profile of 5S rRNA with nucleotides protected *in cellulo* indicated in green. (b) Cryo-EM structure (*26*) of the 5S rRNA bound to ribosomal protein L5. Sites of significant ΔSHAPE are labeled.

Application of ΔSHAPE to RNase MRP

We next applied the ΔSHAPE analysis framework to in-cell analysis of the RNA component of mouse RNase MRP (RMRP). This RNA forms a complex with 10 proteins in eukaryotes that functions in rRNA processing and mitochondrial replication (*32*). In humans, numerous mutations within RMRP RNA cause a spectrum of autosomal recessive skeletal diseases ranging from cartilage-hair hypoplasia (CHH) to anauxetic dysplasia (AD) (*33*). The structure of and protein interactions with the RNA component of RMRP have been investigated *in vitro* using affinity selection, chemical probing, and crosslinking experiments (*32*, *34-36*). A recent cryo-EM study has revealed the overall three-dimensional architecture of the complex in yeast (*37*). However, the precise binding sites of proteins and interactions with substrate have not been examined natively in cells.

Multiple regions of the RMRP RNA have statistically significant enhanced reactivity or protection *in cellulo (***Fig. 3.6a**) and many of these can be attributed to interaction with protein components. These include in the P3 domain, a functionally critical element (**Fig. 3.6b**) (*38*), as well as nucleotides near the junction of helices P8, P9, and P12. Cryo-EM data suggest this latter region interacts with protein Pop4 and perhaps additional proteins (**Fig. 3.6c**). We also observed enhanced reactivity at internal loops in helix P12. Although the complete P12 helix is not present in the cryo-EM model, its proximity to the Pop3 protein suggests that the reactivity enhancements located in the P12 helix may be due to conformational changes induced by Pop3.

We also observed protections involving helices P2 and P19 that are not attributable to RNA-protein interactions. In the cryo-EM model of RMRP, these two regions are adjacent to the active site and are oriented such that they may stabilize or direct RMRP substrates to the catalytic center (**Fig. 3.6d**). Additional density in the cryo-EM map adjacent to these sites of protection may reflect RMRP substrates co-purified with the complex, and supports the hypothesis that P2 and P19 play roles in substrate recognition. There is notable overlap between ΔSHAPE-detected protection in P2 and P19 and sites of disease-associated mutations in RMRP (*39*). The substantial level of in-cell protection in the RMRP

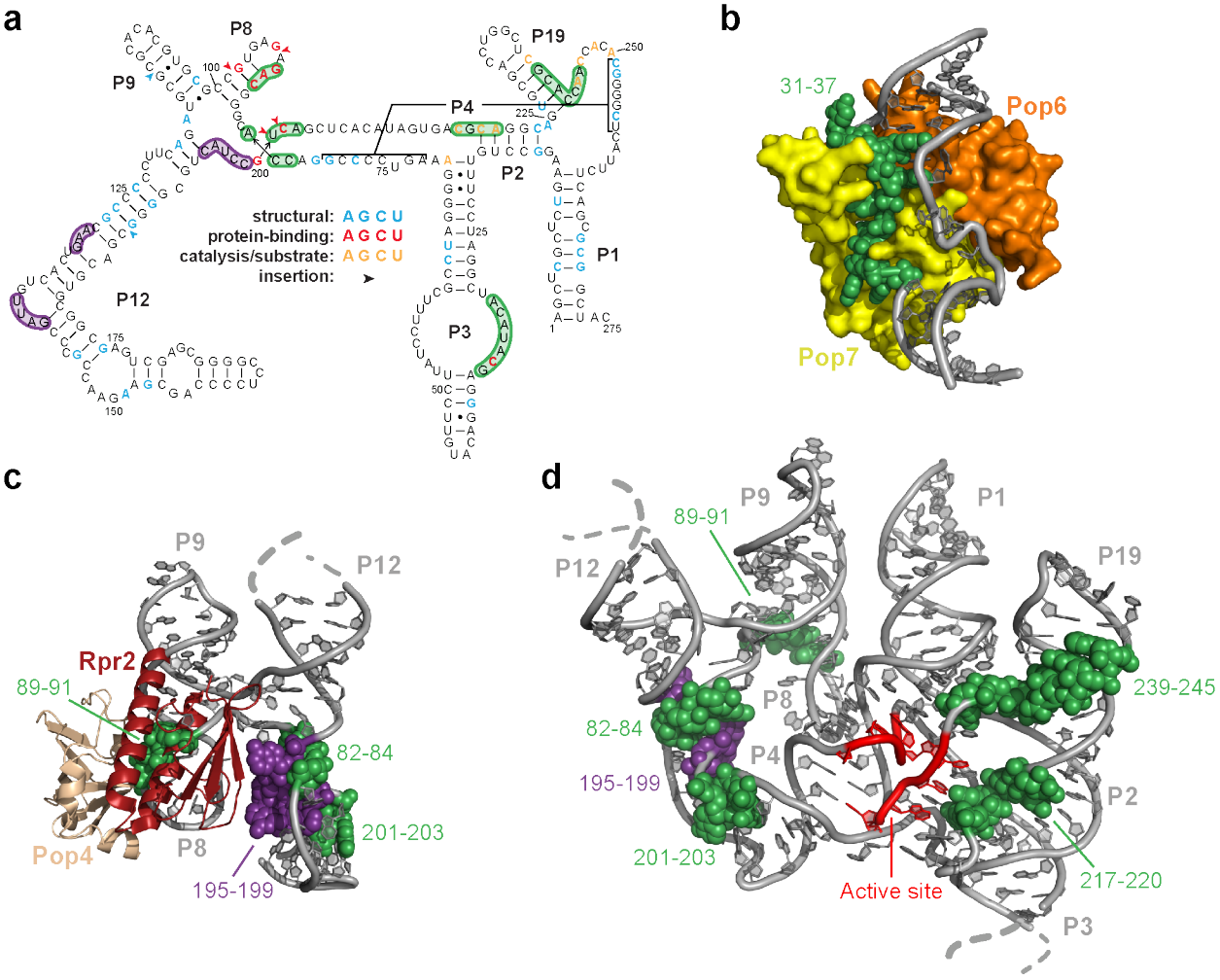


Figure 3.6 In-cell analysis of RNase MRP RNA interactions. (a) Secondary structure of the RNA component of RMRP (*40*), showing RNA-protein interactions detected by ΔSHAPE analysis. Nucleotides protected *in cellulo* are shaded green, and those with enhanced reactivity are purple. Nucleotide positions corresponding to disease-associated mutations that affect function due to inferred (based on ΔSHAPE analysis) RNA structure, protein interactions, or catalysis and substrate recognition are shown in blue, red, and yellow, respectively. (b) Crystal structure of eukaryotic Pop6 (orange) and Pop7 (yellow) proteins interacting with the P3 domain of RMRP (PDB 3IAB). Nucleotides 31-37 demonstrate ΔSHAPE protection *in cellulo* (green spheres) and interact tightly with Pop7. Nucleotides on the opposite side of the P3 internal loop are not tightly associated with Pop6/Pop7 and, correspondingly, do not exhibit strong interactions as assessed by ΔSHAPE. (c) Model of the junction between RMRP RNA helices P8, P9, and P12, showing interactions with Pop4 (tan). Nucleotides exhibiting significant ΔSHAPE values are shown as spheres and colored as in panel (a). In the cryo-EM model, yeast Rpr2 (a potential homolog of Snm1) also binds in this region (*37*) and this protein or an alternative mouse protein may interact with nucleotides 82-84, 195-199, and 201-203. (d) Model of core regions in the eukaryotic RNase P RNP, showing regions of protection and enhanced reactivity as in (a-c). Conserved active site nucleotides are colored red. Nucleotides 217-220 and 239-245 are protected *in cellulo* and form a path to the active site, supporting a role in substrate recognition.

active site cleft suggests that this RNP enzyme is saturated with its RNA substrates (*32*) in the cellular steady state.

Discussion

Our experiments with the well-characterized U1, SRP, and 5S RNPs validate the ability of the ΔSHAPE analytical framework (**Fig. 3.1**), enabled by SHAPE-MaP, to correctly and specifically identify regions of RNA protected by stably-associated proteins *in cellulo*, even in the context of a large number of individual measurements and variable level of confidence in each. In addition, this work illustrates the robust ability of the well-validated 1M7 reagent to react with RNP complexes located in both cytoplasmic and nuclear compartments in cells.

In comparing SHAPE-MaP with icSHAPE, we found poor agreement between the two approaches. SHAPE-MaP has previously been extensively validated against a large set of RNAs with complex structures (*6*), suggesting that icSHAPE does not provide a robust view of RNA structure *ex vivo* or *in cellulo (***Fig. 3.2**). icSHAPE also reports that the SRP RNA undergoes extensive internal conformational changes in cells, which is not consistent with prior studies of this RNA (*24*, *25*). icSHAPE differs from SHAPE-MaP in important ways. First, NAI-N3 reacts more slowly than 1M7 ($t_{1/2}$ = ~30 min vs. ~17 sec, respectively), which has important consequences. These include, first, that slow (but not faster) reagents are highly sensitive to specific ion and buffer choices (*41*) making it very difficult to compare in-cell and *ex vivo* experiments and, second, that long reaction times will reflect RNP assembly and disassembly, cellular turnover, and other events unrelated to the steady-state structure of an RNA. icSHAPE is also one of the many proposed strategies that require a complex purification procedure followed by multi-step adapter ligation-based sequencing library construction, steps that are difficult to perform quantitatively (*17*, *18*).

In addition to defining in-cell RNA-protein and RNA-substrate interactions, ΔSHAPE analysis makes it possible to categorize disease-associated mutations in terms of their likely phenotypic effects (**Fig. 3.6a**). Our analysis supports the interpretation that most mutations leading to CHH/AD spectrum

diseases in the RNase MRP complex result from misfolding of the RNA secondary or tertiary structure, as they are not located near protein or substrate interaction sites. These structural changes occur in helices P1, P3, P4, P9, and P12 (**Fig. 3.6a**, blue nucleotides). We also identified a subset of CHH/AD-related mutations located near protein interaction sites (**Fig. 3.6a**, in red). In individuals with these mutations, which are most concentrated within helix P8 and the P8-P9-P12 junction, improper assembly of the RNase MRP RNA-protein complex may be the root cause of disease. Finally, the remaining disease-related mutations are most consistent with compromising RNA-substrate interactions. These involve nucleotides that comprise the active site along with portions of P2 and P19 that are protected *in cellulo* due to putative substrate interactions (**Fig. 3.6a**, yellow).

The ΔSHAPE analysis framework is clearly a broadly useful tool for defining RNA-protein interactions. ΔSHAPE is also subject to limitations. Because ΔSHAPE requires a change in SHAPE reactivity between conditions, proteins that interact primarily with double-stranded RNA may be difficult to detect. For the RNAs studied here, in-cell protections almost always corresponded to direct protein-RNA interactions, while enhancements generally reported RNA conformational changes. In other cases, protein-induced conformational changes may lead to apparent protections in regions unrelated to protein binding. While the ΔSHAPE framework correctly identified sites of stable RNA-protein interaction, the stringency implemented here may lead to missing weaker protein binding sites. For example, nucleotides stably bound by Sm ring proteins are detected by ΔSHAPE (**Fig. 3.3**) but other nucleotides inside the Sm ring do not display protection. Finally, as with any chemical probing experiment, ΔSHAPE requires sufficient sequencing coverage of the RNA of interest in both tested conditions.

In sum, SHAPE-MaP efficiently and accurately detects RNA-protein interaction sites and occupancy in living cells. Using simple and intuitive statistical filtering, significant differences between *ex vivo* and *in cellulo* SHAPE reactivities were identified while avoiding false positive detection. The analysis framework developed here identified RNA binding sites for all stably bound protein factors for three model RNPs, found in both cytoplasmic and nuclear compartments, under native growth conditions without the need for specialized affinity purification. Application to the RNase MRP ribonucleoprotein

enzyme complex both identified sites of RNA-protein interaction and extensive substrate recognition in the active site cleft, and also enabled categorization of CHH/AD-related mutations by molecular phenotype.

This analysis framework works well for *de novo* identification of functionally essential regions in non-coding RNAs, and is complementary to RNA-protein crosslinking and immunoprecipitation (CLIP)[3] experiments. Critically, ΔSHAPE specifically detects the occupancy of a given site. As RNA structure studies increasingly shift towards in-cell and transcriptome-wide analyses, the robust analytical approach presented here will become an essential tool for rapid discovery and analysis of true RNA-protein interactions.

Methods

In cellulo modification

Mouse trophoblast stem cells (TSCs) were cultured as described (*42*). Live TSCs were washed once with PBS, and 900 μl of fresh growth media was added. For samples subjected to in-cell SHAPE probing, 100 μl of 100 mM 1M7 in neat DMSO (10 mM final concentration) were added and rapidly mixed by swirling the culture dish. Cells were then incubated at 37 °C for 5 minutes (although the 1M7 reagent is completely quenched by hydrolysis in ~2 minutes). Media was removed and the cells were washed once with PBS before isolation of total RNA (1 mL TRIzol; Ambion). The no-reagent negative control RNA was prepared similarly with the exception that neat DMSO was used instead of 1M7 in DMSO.

Ex vivo RNA extraction and modification

To preserve native secondary structures, RNA for ex vivo analysis was extracted using a gentle procedure, avoiding the use of harsh chemical denaturants. Approximately 10^6 TSCs were washed and pelleted in ice-cold PBS, resuspended in 2.5 ml Lysis Buffer [40 mM Tris, pH 7.9, 25 mM NaCl, 6 mM $MgCl_2$, 1 mM $CaCl_2$, 256 mM sucrose, 0.5% Triton X-100, 1,000 U/ml RNasin (Promega), 450 U/ml

DNase I (Roche)], and rotated at 4 °C for 5 minutes. Cells were then pelleted at 4 °C for 2 minutes at 2250 *g*, resuspended in 40 mM Tris pH 7.9, 200 mM NaCl, 1.5% SDS, and 500 μg/ml of Proteinase K, and rotated at 20 °C for 45 minutes. RNA was then extracted twice with phenol:chloroform:isoamyl alcohol (24:24:1) pre-equilibrated with 1× Folding Buffer (100 mM HEPES, pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$), followed by one extraction with chloroform. Note that use of TRIzol and similar reagents should be specifically avoided for native-like purification of RNA. RNA was exchanged into 1.1× Folding Buffer using a desalting column (PD-10, GE Life Sciences) and incubated at 37 °C for 20 minutes. Approximately 3 μg RNA was then added to a one-ninth volume of 100 mM 1M7 in neat DMSO (10 mM final concentration) and incubated at 37 °C for 5 minutes. Modified RNA was purified (RNeasy Midi spin column, Qiagen) and eluted in approximately 50 μl H_2O. No-reagent negative control RNA was prepared in the same way but substituting neat DMSO for 1M7.

Denaturing control

TSCs were grown as described (*42*) and total RNA isolated using TRIzol (Ambion). Approximately 500 ng RNA was then resuspended in 1.1× Denaturing Control Buffer [55 mM HEPES pH 8.0, 4.4 mM EDTA, 55% formamide (v/v)] and incubated at 95 °C for 1 minute. An aliquot of 45 μl of denatured RNA was added to 5 μl of 100 mM 1M7 and allowed to react at 95 °C for 1 minute. After modification, RNA was purified (RNeasy Mini spin column, Qiagen) and eluted in approximately 50 μl H_2O.

U1, SRP, and 5S SHAPE-MaP

Mutational profiling reverse transcription reactions were carried out using RNA-specific primers (*6*, *14*), which maximizes efficient use of sequencing reads. cDNA was purified using G-50 spin columns (GE Life Sciences). SHAPE-MaP sequencing libraries were created for each experimental condition (*ex vivo* +1M7, *ex vivo* DMSO, *in cellulo* +1M7, *in cellulo* DMSO, denaturing control +1M7) and RNA (U1 snRNA, 5S rRNA, SRP RNA) using the targeted specific-RNA approach (*6*) with minor changes. PCR 1 followed the touchdown format (*43*) and was performed as follows: 98 °C for 30 s, 20 cycles of [98 °C

for 10 s, 72 °C for 30 s (decreasing by 1 °C per cycle until 64 °C), 72 °C for 20 s], 72 °C for 2 min. PCR 2 was performed for 10 cycles using 2 μl unpurified PCR 1 product as template in a 50 μl reaction. Final libraries were then purified (PureLink PCR Micro spin columns; Life Technologies) prior to sequencing.

Whole-transcriptome SHAPE-MaP

Total RNA was modified as described above and then depleted of ribosomal RNA (mouse RiboZero; Epicentre). Mutational profiling reverse transcription reactions were primed with random DNA nonamers (*6*, *14*). cDNA was purified (Agencourt RNAClean XP beads, Beckman Coulter) and then converted to double-stranded DNA (NEBNext mRNA second-strand synthesis kit, New England Biolabs). The resulting DNA was purified (Ampure XP beads, Beckman Coulter) before construction of whole-transcriptome sequencing libraries (Nextera XT, Illumina).

Sequencing and SHAPE profile generation

Purified U1, 5S, SRP, or whole-transcriptome sequencing libraries were sequenced on an Illumina MiSeq (U1, 5S, and SRP) or NextSeq (transcriptome) instrument, generating 2 × 150 paired-end datasets. Initial SHAPE reactivity profiles, including error estimates, were created by aligning reads to U1 snRNA, 5S rRNA, SRP or RNase MRP RNA reference sequences (GenBank accession no. FM991912.1, M31319.1, HG323689.1, and NR_001460.1, respectively) using *ShapeMapper* (v1.0, http://chem.unc.edu/rna/software.html) (*6*, *14*). Median per-nucleotide read depth was greater than 10,000 for each of these RNAs.

From transcriptome-wide datasets, we identified the 50 most abundant transcripts using *Tophat* (*44*). SHAPE reactivity profiles were then generated for each of these RNAs by aligning to respective sequences with *ShapeMapper*. Transcripts with complete sequencing coverage and sufficient depth (median read depth > 5,000) were selected for comparison to icSHAPE profiles.

SHAPE reactivity normalization

SHAPE-MaP quantifies adduct formation based on the observed mutation rates of modified RNA relative to no-reagent and denaturing controls (*6*). We observed higher mutation rates in *ex vivo*-modified U1 snRNA than in the 5S and SRP RNAs. As a result, the SHAPE reactivities of U1 snRNA were generally elevated compared to the other RNAs. Independent normalization of U1 snRNA did not preserve the intrinsically high reactivity of this RNA relative to 5S and SRP. Thus, we normalized SHAPE reactivities with a common normalization factor to preserve the relative distribution of reactivities among the three simultaneously probed RNAs. RNase MRP SHAPE profiles were normalized independently, as they were derived from RNA probed separately from the three model RNAs. Initial SHAPE reactivities for both *in cellulo* and *ex vivo*-modified RNAs were first pooled together into a single distribution from which primer-binding sites were excluded. The first five nucleotides synthesized during reverse transcription were also excluded to eliminate spurious mutations caused by the suboptimal processivity of initiating retroviral reverse transcriptase (*45*). A normalization factor for the entire distribution was calculated by the boxplot method (*27*): the interquartile range (IQR) of the distribution was calculated; and reactivity values greater than 1.5 times the IQR were excluded as outliers with the number of outliers capped at 10%. The average of the 10% most reactive remaining nucleotides was then calculated, yielding the common normalization factor. Initial individual SHAPE profiles were then adjusted by dividing each reactivity and standard error by the common normalization factor.

icSHAPE profile generation

icSHAPE reads (*12*) were downloaded from the gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo, accession GSE60034). Reads corresponding to U1 snRNA, 5S rRNA, SRP and RNase MRP RNA were extracted by alignment to the respective sequence. Relevant reads were then converted to fastq format and analyzed using the published icSHAPE pipeline (https://github.com/qczhang/icSHAPE) (*12*). Limited reads for U1 snRNA, 5S rRNA, and RNase MRP

RNA resulted in icSHAPE profiles with very sparse data, so we restricted our comparison of RNP complexes to the SRP RNA.

Calculating ΔSHAPE, Z-factors, and standard scores to determine binding sites

The derivation of nucleotide-resolution standard error values associated with SHAPE reactivity measurements has been described fully (*6*), and is reviewed briefly here. Mutation rates for each experimental measurement (+1M7, no-reagent, denaturing control) are modeled as a Poisson distribution because discrete mutation events contribute to the overall reactivity at each nucleotide. The variance of a Poisson distribution equals the number of observations, and the standard error of a mutation rate (SE_{rate}) can be estimated as

$$SE_{rate} = \frac{\sqrt{\lambda}}{reads} = \frac{\sqrt{rate}}{\sqrt{reads}} \quad (1)$$

where λ is the number of mutations observed, *reads* is the read depth at a given nucleotide, and *rate* is the number of mutation events per read. The standard errors from each experimental measurement are then combined to yield SHAPE reactivity standard errors (*6*).

The change in SHAPE reactivity (ΔSHAPE) for each nucleotide *i* was calculated as

$$\Delta SHAPE_i = \frac{1}{3}\left[\left(\sum_{n=i-1}^{i+1} X_n\right) - \left(\sum_{n=i-1}^{i+1} C_n\right)\right] \quad (2)$$

where *X* and *C* are the *ex vivo* and *in cellulo* SHAPE reactivities, respectively. This produces ΔSHAPE values that reflect the difference in reactivity between *ex vivo* and *in cellulo* conditions averaged over a three-nucleotide sliding window. To account for smoothing, standard error values were averaged as

$$SE_i = \frac{1}{3}\sqrt{\sigma_{i-1}^2 + \sigma_i^2 + \sigma_{i+1}^2} \quad (3)$$

where σ_i and SE_i refer to the original error and smoothed error at nucleotide i, respectively. Z-factors (Z) (*28*) for each nucleotide i were calculated according to Eqn. 4, where the subscripts X and C indicate *ex vivo* and *in cellulo* conditions, respectively. Nucleotides for which $Z > 0$ were considered to undergo significant changes in SHAPE reactivity.

$$Z_i = 1 - \frac{1.96\left(SE_{X,i} + SE_{C,i}\right)}{|\Delta\text{SHAPE}_i|} \tag{4}$$

Standard scores (S) were calculated for each nucleotide i according to Eqn. 5, where $\mu_{\Delta SHAPE}$ and $\sigma_{\Delta SHAPE}$ represent the mean and standard deviation of the distribution of ΔSHAPE values, respectively.

$$S_i = \frac{\Delta SHAPE_i - \mu_{\Delta SHAPE}}{\sigma_{\Delta SHAPE}} \tag{5}$$

Putative binding sites were identified as regions within five-nucleotide sliding windows for which at least three nucleotides had $Z > 0$ and $|S| \geq 1$. Nucleotides that met these requirements were denoted as undergoing changes in SHAPE reactivity due to the influence of the cellular environment.

Modeling

The model of the complete U1 snRNP complex used in this study was generated from three individual models. Phosphorus atoms in a U1 snRNP model (omitting stem-loop 2 and the U1A protein and kindly provided by Kiyoshi Nagai) were first aligned to the phosphorus atoms in a 5.5-Å model of the complete complex (PDB: 3PGW) (*23*). To incorporate the U1A/stem-loop 2 interaction, we aligned the Cα atoms of U1A in a high-resolution model (PDB: 4PKD) (*46*) to the 5.5-Å model. The model of the SRP S domain bound to SRP68/72, SRP19, and SRP54 was generated by overlaying the SRP68- and SRP19-bound structure (PDB: 4P3E) (*24*) with the SRP19- and SRP54-bound structure (PDB: 1MFQ) (*22*) via alignment of the SRP19 atoms.

REFERENCES

1. D. D. Licatalosi, R. B. Darnell, RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11**, 75–87 (2010).

2. M. Guttman, J. L. Rinn, Modular regulatory principles of large non-coding RNAs. *Nature*. **482**, 339–346 (2012).

3. C. A. McHugh, P. Russell, M. Guttman, Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol.* **15**, 203 (2014).

4. E. J. Merino, K. A. Wilkinson, J. L. Coughlan, K. M. Weeks, RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127**, 4223–4231 (2005).

5. J. L. McGinnis, J. A. Dunkle, J. H. D. Cate, K. M. Weeks, The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc.* **134**, 6617–6624 (2012).

6. N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson, K. M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*. **11**, 959–965 (2014).

7. D. M. Mauger *et al.*, Functionally conserved architecture of hepatitis C virus RNA genomes. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3692–3697 (2015).

8. R. C. Spitale *et al.*, RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* **9**, 18–20 (2012).

9. J. L. McGinnis, K. M. Weeks, Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry*. **53**, 3237–3247 (2014).

10. J. Tyrrell, J. L. McGinnis, K. M. Weeks, G. J. Pielak, The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry*. **52**, 8777–8785 (2013).

11. J. L. McGinnis *et al.*, In-cell SHAPE reveals that free 30S ribosome subunits are in the inactive state. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2425–2430 (2015).

12. R. C. Spitale *et al.*, Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*. **519**, 486–490 (2015).

13. K. E. Watters, T. R. Abbott, J. B. Lucks, Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic Acids Res.* (2015), doi:10.1093/nar/gkv879.

14. M. J. Smola, G. M. Rice, S. Busan, N. A. Siegfried, K. M. Weeks, Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile, and accurate RNA structure analysis. *Nat. Protoc.* **10**, 1643–1669 (2015).

15. S. A. Mortimer, M. A. Kidwell, J. A. Doudna, Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **15**, 469–479 (2014).

16. C. K. Kwok, Y. Tang, S. M. Assmann, P. C. Bevilacqua, The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci.* **40**, 221–232 (2015).

17. C. A. Raabe, T.-H. Tang, J. Brosius, T. S. Rozhdestvensky, Biases in small RNA deep sequencing data. *Nucleic Acids Res*. **42**, 1414–1426 (2014).

18. K. M. Weeks, Toward all RNA structures, concisely. *Biopolymers*. **103**, 438–448 (2015).

19. M. Corley, A. Solem, K. Qu, H. Y. Chang, A. Laederach, Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. *Nucleic Acids Res*. **43**, 1859–1868 (2015).

20. D. E. Draper, Themes in RNA-protein recognition. *J. Mol. Biol*. **293**, 255–270 (1999).

21. O. Weichenrieder, K. Wild, K. Strub, S. Cusack, Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature*. **408**, 167–173 (2000).

22. A. Kuglstatter, C. Oubridge, K. Nagai, Induced structural changes of 7SL RNA during the assembly of human signal recognition particle. *Nat. Struct. Biol*. **9**, 740–744 (2002).

23. D. A. P. Krummel, C. Oubridge, A. K. W. Leung, J. Li, K. Nagai, Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*. **458**, 475–480 (2009).

24. J. T. Grotwinkel, K. Wild, B. Segnitz, I. Sinning, SRP RNA remodeling by SRP68 explains its role in protein translocation. *Science*. **344**, 101–104 (2014).

25. M. Halic *et al*., Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature*. **427**, 808–814 (2004).

26. R. M. Voorhees, I. S. Fernández, S. H. W. Scheres, R. S. Hegde, Structure of the mammalian ribosome-Sec61 complex to 3.4 Å resolution. *Cell*. **157**, 1632–1643 (2014).

27. C. E. Hajdin *et al*., Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A*. **110**, 5498–5503 (2013).

28. J. Zhang, T. Chung, K. Oldenburg, A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen*. **4**, 67–73 (1999).

29. J. B. Scripture, P. W. Huber, Binding site for Xenopus ribosomal protein L5 and accompanying structural changes in 5S rRNA. *Biochemistry*. **50**, 3827–3839 (2011).

30. J. A. Steitz *et al*., A 5S rRNA/L5 complex is a precursor to ribosome assembly in mammalian cells. *J. Cell Biol*. **106**, 545–556 (1988).

31. C. Leidig *et al*., 60S ribosome biogenesis requires rotation of the 5S ribonucleoprotein particle. *Nat. Commun*. **5**, 1–8 (2014).

32. O. Esakova, A. S. Krasilnikov, Of proteins and RNA: the RNase P/MRP family. *RNA*. **16**, 1725–1747 (2010).

33. S. Mattijssen, T. J. M. Welting, G. J. M. Pruijn, RNase MRP and disease. *WIREs RNA*. **1**, 102–116 (2010).

34. H. Pluk, H. van Eenennaam, S. A. Rutjes, G. J. Pruijn, W. J. van Venrooij, RNA-protein interactions in the human RNase MRP ribonucleoprotein complex. *RNA*. **5**, 512–524 (1999).

35. T. J. M. Welting, W. J. van Venrooij, G. J. M. Pruijn, Mutual interactions between subunits of the human RNase MRP ribonucleoprotein complex. *Nucleic Acids Res.* **32**, 2138–2146 (2004).

36. E. Khanova, O. Esakova, A. Perederina, I. Berezin, A. S. Krasilnikov, Structural organizations of yeast RNase P and RNase MRP holoenzymes as revealed by UV-crosslinking studies of RNA-protein interactions. *RNA.* **18**, 720–728 (2012).

37. K. Hipp, K. Galani, C. Batisse, S. Prinz, B. Bottcher, Modular architecture of eukaryotic RNase P and RNase MRP revealed by electron microscopy. *Nucleic Acids Res.* **40**, 3275–3288 (2012).

38. G. S. Shadel, G. A. Buckenmeyer, D. A. Clayton, M. E. Schmitt, Mutational analysis of the RNA component of Saccharomyces cerevisiae RNase MRP reveals distinct nuclear phenotypes. *Gene.* **245**, 175–184 (2000).

39. C. T. Thiel, G. Mortier, I. Kaitila, A. Reis, A. Rauch, Type and Level of RMRP Functional Impairment Predicts Phenotype in the Cartilage Hair Hypoplasia–Anauxetic Dysplasia Spectrum. *Am. J. Hum. Genet.* **81**, 519–529 (2007).

40. M. E. Schmitt, J. L. Bennett, D. J. Dairaghi, D. A. Clayton, Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison. *FASEB J.* **7**, 208–213 (1993).

41. S. A. Mortimer, K. M. Weeks, A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129**, 4144–4145 (2007).

42. J. Quinn, T. Kunath, J. Rossant, in *Placenta and Trophoblast* (Humana Press, New Jersey, 2005), vol. 121, pp. 123–146.

43. R. H. Don, P. T. Cox, B. J. Wainwright, K. Baker, J. S. Mattick, "Touchdown" PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.* **19**, 4008 (1991).

44. A. Roberts *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

45. C. Majumdar, J. Abbotts, S. Broder, S. H. Wilson, Studies on the mechanism of human immunodeficiency virus reverse transcriptase. Steady-state kinetics, processivity, and polynucleotide inhibition. *J. Biol. Chem.* **263**, 15657–15665 (1988).

46. Y. Kondo, C. Oubridge, A.-M. M. van Roon, K. Nagai, Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife.* **4** (2015), doi:10.7554/eLife.04986.

CHAPTER 4: SHAPE ANALYSIS REVEALS TRANSCRIPT-WIDE CELLULAR INTERACTIONS AND STABLE STRUCTURAL DOMAINS WITHIN THE *XIST* lncRNA

Introduction

Long noncoding RNAs (lncRNAs) play central roles in the regulation of gene expression through interactions with numerous protein partners (*1*, *2*). Several lncRNAs are necessary for normal development (*3*, *4*), and, as a result, their dysfunction is associated with diseases including cancer (*5*). Many more unstudied lncRNAs are located in disease-associated regions of the human genome, suggesting additional physiologically relevant examples of lncRNA-mediated gene regulation remain to be discovered (*6-9*). Despite the importance of lncRNAs and their cellular interactions, little is known about how lncRNA structure mediates function. Here we use quantitative in-cell and *ex vivo* nucleotide-resolution structure probing (SHAPE-MaP) to show that the X-inactive specific transcript (*Xist*) lncRNA forms multiple well-defined secondary structures with complexities comparable to those found in prominent viral RNAs, ribosomal RNA domains, and other regulatory RNAs. Nucleotide polymorphisms occur predominantly in predicted single-stranded regions, suggesting selective pressure maintains structured elements within the lncRNA. In contrast, repeated elements within *Xist*, which are common among lncRNAs (*10*), are distinctive in that they adopt dynamic structures with single-stranded regions that may function as landing pads for protein cofactors. Differences in chemical reactivity of the *Xist* lncRNA both in living cells and *ex vivo* suggest that the RNA binds proteins throughout its length and that some domains have dramatically different structures in cells. We identified a previously unknown interaction domain that bound numerous proteins and discovered examples of how lncRNA structure modulates specific protein interactions. Roughly half of the *Xist* RNA contains domains that either bind proteins or form well-defined structural elements, and these motifs span nearly the entire ~18 kilobase RNA, which rationalizes the conserved length of *Xist* among mammals. Our results create a framework

for further investigation of *Xist* and, more broadly, establish an experimental context for understanding complex relationships between lncRNA sequence, structure, and function.

Although it is clear that lncRNAs regulate gene expression at transcriptional, post-transcriptional and epigenetic levels, there are many unanswered questions about lncRNAs: Why are they so long? Are regulatory functions organized within these long RNAs? Do lncRNAs have well-defined structures? If so, to what extent does the cellular environment modulate structure? What are the specific structural features of the conserved repeat regions often found in lncRNAs? What features govern protein interactions? We sought to answer these questions as they pertain to the *Xist* lncRNA.

Xist plays a central role in X-chromosome inactivation (XCI) during female eutherian mammalian development and is an archetype of gene-silencing lncRNAs. During the initiation of XCI, *Xist* spreads in *cis* along the future inactive X chromosome and recruits protein complexes, which in turn apply repressive chromatin markers to induce silencing (*11*, *12*). Despite its discovery more than 20 years ago (*13*, *14*), the sequence elements within Xist that contribute to its distinct function, and the mechanisms by which they do so, remain poorly defined. *Xist* is approximately 18 kilobases long, a length that is generally conserved (*13-15*); however, the primary sequence is less conserved than might be expected for an RNA of such importance. Several tandem repeat regions (labeled A-F in the mouse) exhibit moderate conservation (*13-15*), and at least two of these, Repeat A and the rodent-specific repeat C, have been implicated in silencing and localization to the inactive X, respectively (*16-21*). An additional 1.5 kilobase region encompassing repeats F and B has been shown to be required for the proper accumulation of heterochromatic marks over the inactive X (*22*). Beyond these three regions, the locations of additional functional domains within *Xist* are largely unexplored.

A roadblock to understanding the mechanisms by which Xist carries out its cellular function has been our near-complete lack of knowledge of the structures it adopts as a free RNA and in cells. Detailed structural maps of other functional RNAs, such as the ribosomal RNAs (*23*) and the HIV RNA genome (*24-26*), have been fundamental to understanding the mechanisms by which individual domains within

large RNAs execute discrete cellular functions. A structural map of Xist would be expected to have a similar transformative impact.

Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) (*25*, *27*) is a recently developed chemical strategy to analyze RNA structure that can be applied to any RNA regardless of its length. SHAPE-MaP provides a biophysically rigorous measurement of local nucleotide flexibility that is independent of base identity (*25*, *28*, *29*). SHAPE-MaP is sensitive enough to detect modifications in highly complex environments, including in the cell nucleus (*30*) and, unlike most RNA-probing methods with deep sequencing readout, is unaffected by biases introduced during ligation-based library preparation steps. SHAPE data are sufficient to distinguish between different structural models (*31*) and can detect various modes of protein binding in cells (*30*). SHAPE has been used extensively in recent years to characterize the structural properties of long RNAs (*24*, *25*, *32*). SHAPE-informed structural models have invariably yielded rich insights into the biological functions of diverse RNAs (*24*, *25*, *31-34*), and in many cases, uncovered novel functional elements (*24*, *25*, *30*, *34*, *35*).

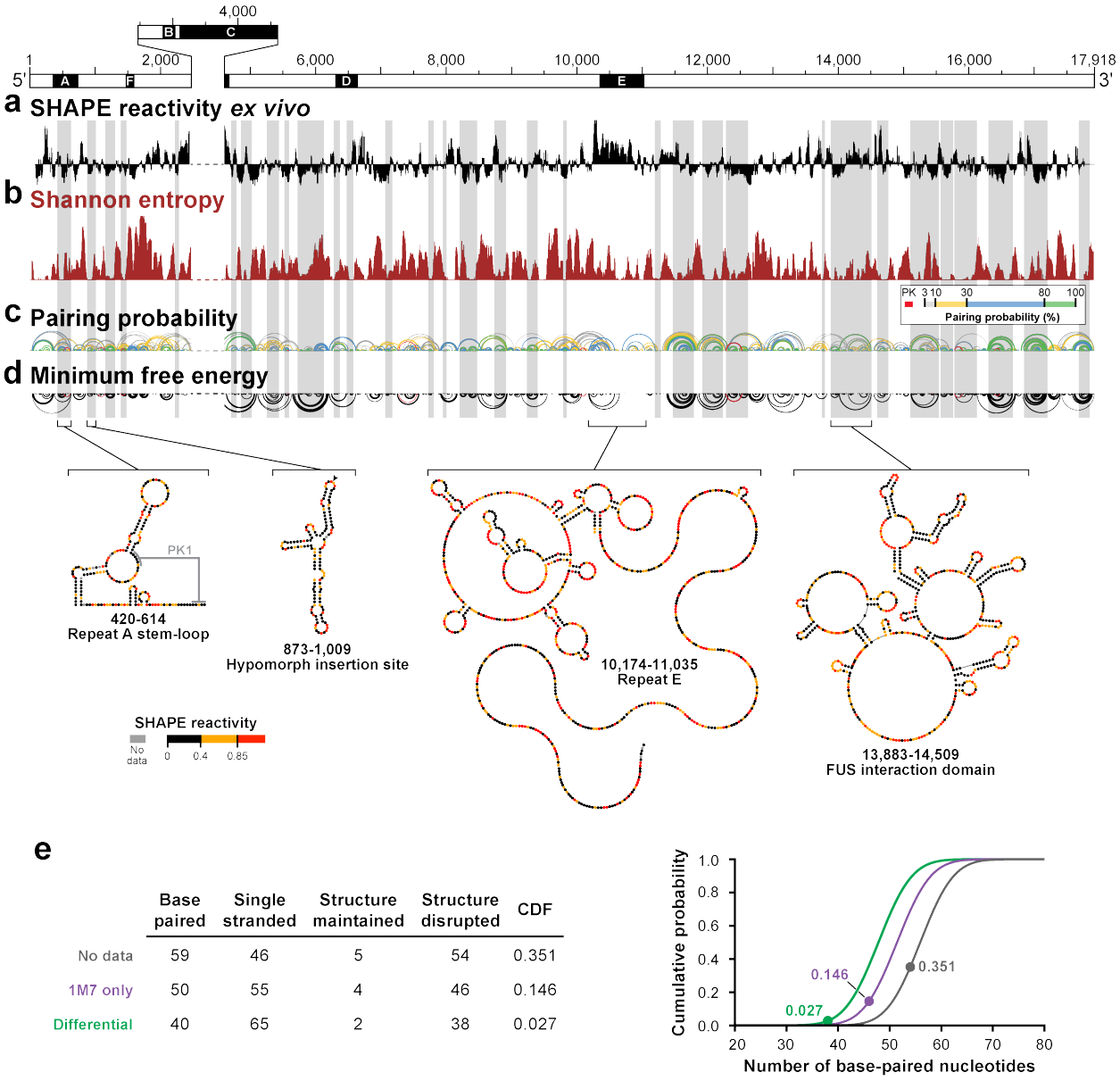
Using SHAPE-MaP, we examined full-length, authentic transcripts of mouse *Xist* at single-nucleotide resolution under protein-free conditions (*ex vivo*) and natively in mouse trophoblast stem cells (TSCs). These cells show prototypical epigenetic patterns over the inactive X chromosome (*36-38*) and depend on *Xist* for its continued silencing (*39*, *40*). The SHAPE data identified upwards of 30 regions in Xist that form complex, well-defined structures that resemble functional elements in viral and other RNAs (*25*, *27*, *32*). By comparison of *ex vivo* reactivities to those obtained in living TSCs, we found that large portions of *Xist* are bound by proteins and have different conformations *in vivo*, and found several domains in the 3′ half of *Xist* that appear to function as protein interaction platforms. Our data provide fresh insight into the mechanisms of *Xist*-mediated silencing and provide a broad structural foundation for understanding complex relationships between lncRNA sequence and function.

Results

Ex vivo structure probing

We first probed full-length *Xist* after gentle extraction and deproteination using the SHAPE reagents 1-methyl-7-nitroisatoic anhydride (1M7), 1-methyl-6-nitroisatoic anhydride (1M6), and *N*-methyl-isatoic anhydride (NMIA) (*41*, *42*) to obtain *ex vivo* SHAPE reactivities for 86% of all nucleotides in the *Xist* RNA (**Fig. 4.1a**). Due to the highly repetitive nature of repeats B and C, it is not possible to uniquely align sequencing reads to these regions, and we excluded them from our analysis (see Methods). We first searched for pseudoknots in the sequence guided by 1M7 reactivity data (*43*) and identified 10 potential pseudoknots. Incorporating these, along with data from all three SHAPE experiments, we modeled the secondary structure of *Xist* using the well-validated three-reagent differential SHAPE approach (*27*, *44*). We also modeled the structure without any SHAPE data and with only 1M7 data. As expected, data-driven SHAPE models are drastically different from the model generated without experimental data: the 1M7-only and differential models are only 49% and 46% similar to the no-data model, respectively.

In an independent assessment of our models, we examined the structural context of 105 known single-nucleotide polymorphisms (SNPs) within mouse *Xist*. Given the critical and conserved functions of *Xist*, and the presumed importance of structural elements within the RNA, we posited that SNPs that disrupt essential RNA structures would be strongly selected against. For each structural model (no data, 1M7 only, or three-reagent differential), we counted the number of SNPs that would lead to structural disruption by creating base pair mismatches. In the no-data model, 54 of the 105 SNPs were located in base-pairing positions; in the 1M7-only and three-reagent models, 46 and 38 SNPs would disrupt predicted structure respectively (**Fig. 4.1e,** left). The probabilities of selecting fewer disruptive positions by chance correspond to p-values of 0.35, 0.15 and 0.027 for the no-data, 1M7-only, and three-reagent models, respectively (**Fig. 4.1e,** right). Thus, with increasing data quality, the probability that SNPs are structurally disruptive by chance decreases significantly. This analysis suggests that the *Xist* secondary



e

	Base paired	Single stranded	Structure maintained	Structure disrupted	CDF
No data	59	46	5	54	0.351
1M7 only	50	55	4	46	0.146
Differential	40	65	2	38	0.027

Figure 4.1 Predicted structural architecture of the *Xist* lncRNA. (a) *Ex vivo* 1M7 reactivities are shown as the median reactivity over 55-nt sliding windows; values are plotted relative to the global median. (b) Shannon entropy values for the *ex vivo* secondary structure model, smoothed over 55-nt sliding windows, are plotted relative to the global median. High values indicate many possible structures, and vice-versa. Well-determined structures with low SHAPE reactivity and low Shannon entropy are emphasized with grey shading. (c) Base-pairing probabilities in the *Xist* RNA. Arcs represent base pairs and are colored by probability, with green arcs representing the most likely base pairs. (d) Minimum free energy secondary structure model of *Xist* obtained using SuperFold (*27*) with differential three-reagent data. Examples of well-formed domains are colored according to SHAPE reactivity. (e) Left, positions of known SNPs in the *Xist* RNA relative to secondary structure models. For each model, a bootstrapped Gaussian cumulative probability density function (CDF; right) was used to determine the likelihood of encountering fewer disruptive SNPs than reported in the Structure Disrupted column.

structure model informed by three-reagent differential SHAPE data is broadly accurate and, additionally, emphasizes the importance of high-quality RNA structure probing data. Just as lack of selective pressure leads to increased SNP abundance in introns, pseudogenes, and other genetic elements with low functional potential (*45*), the observation that SNPs predominantly occur in locally unstructured regions strongly suggests that the structure of *Xist* is important for function.

As part of our structure modeling approach, we assessed how well each secondary structure element was defined by both its sequence and the experimental SHAPE data by calculating Shannon entropies at nucleotide resolution (*25*, *27*, *46*) (**Fig. 4.1b**). RNA regions with high Shannon entropy likely sample multiple conformations, whereas those with low Shannon entropy likely adopt a single well-defined structure. Previous work with large viral RNAs has shown that functional elements can be identified *de novo* as regions with both low SHAPE reactivity (indicating a high degree of structure) and low Shannon entropy (indicating well-defined structure) (*25*, *32*). We identified 32 regions with low SHAPE reactivity and Shannon entropy in the *Xist* RNA (**Fig. 4.1a-b**). These individual structural domains resemble smaller lncRNAs such as steroid receptor RNA activator and HOTAIR, both of which exhibit extensive secondary structure (*47*, *48*). The structure of *Xist*, and by inference many as-yet unstudied lncRNAs, is best described as a series of well-formed domains interlinked by flexible regions (**Fig. 4.1c-d**). Many of the most well-defined structural elements in *Xist* are located in the 3′ end of the RNA and have not been considered in previous studies (*16*, *49*). The extent of defined structures in the 3′ end of *Xist* (**Fig. 4.1c-d**) suggests a functional basis for the conserved length of *Xist* transcripts.

The 400-nt repeat A region at the 5′ end of *Xist* is one of the most clearly conserved regions of the RNA (*13-15*). Repeat A is required for stable accumulation of spliced *Xist* in cells and for its function in gene silencing (*16*, *20*, *21*). When this 400-nt region is expressed as a short transcript in the absence of the remainder of *Xist*, it is sufficient to induce repression of neighboring genes (*17*). In the mouse, repeat A includes 7.5 copies of a 24-nt repeat unit separated by U-rich spacers of variable lengths. Current models of this region have emphasized self-contained structures consisting of stable stem-loops (*16*, *50-*

52). In contrast, in our model, based on data obtained in the context of full-length native *Xist*, the repeat A region generally lacks a defined global structure. A single small stem-loop, consisting of a GC-rich stem and AU-rich loop, is the only well-defined element in repeat A (**Figs. 4.2a-b**); these nucleotides exhibit high sequence conservation (**Fig. 4.2c**). This repeat A stem-loop motif is flanked by sequences with high Shannon entropies indicative of extensive structural variability (**Fig. 4.2a**). Furthermore, repeat A nucleotides likely interact, at least transiently, with adjacent segments of *Xist* (**Fig. 4.2a**). The inherently dynamic structure of repeat A may help the element interact with diverse subsets of proteins that accommodate its dual role in post-transcriptional processing of *Xist* and in gene silencing, respectively.

Repeat E, which has no known function, also forms a dynamic and highly flexible structure. This region spans roughly 1 kb at the beginning of exon 7 and consists of U-rich repeat units approximately 20-25-nts long (*15*). Repeat E exhibits low Shannon entropy and *high* SHAPE reactivity, indicating that it adopts an unstructured, mostly single-stranded conformation (**Fig. 4.2d**). Nucleotides in repeat E are clearly accessible for unencumbered interactions with RNA binding proteins.

Our model also provides structural context for previously characterized *Xist* mutant phenotypes. For example, a 7.7-kb inversion of nucleotides 5,947-13,670 leads to a hypomorphic phenotype with incomplete silencing capabilities (*53*). This large inversion disrupts 14 structural elements in the *Xist* RNA model. A 16-nucleotide insertion located directly 3´ of the Repeat A region in *Xist* causes a similar hypomorphic phenotype (*54*). The insertion site falls in the middle of a well-defined hairpin structure with particularly low Shannon entropy (**Fig. 4.1d**, see arrowhead). This likely leads to a rearrangement of local structure that may affect the biological activity of the Repeat A region or, alternatively, attenuate a function of the hairpin element itself.

The effects of the cellular environment on Xist structure

Xist interacts with many cellular proteins (*55-57*), and we hypothesized that such interactions would be mediated by distinct, independently-functioning domains, featuring both single-stranded and well-structured motifs. To identify regions of *Xist* most strongly affected by the cellular environment, we

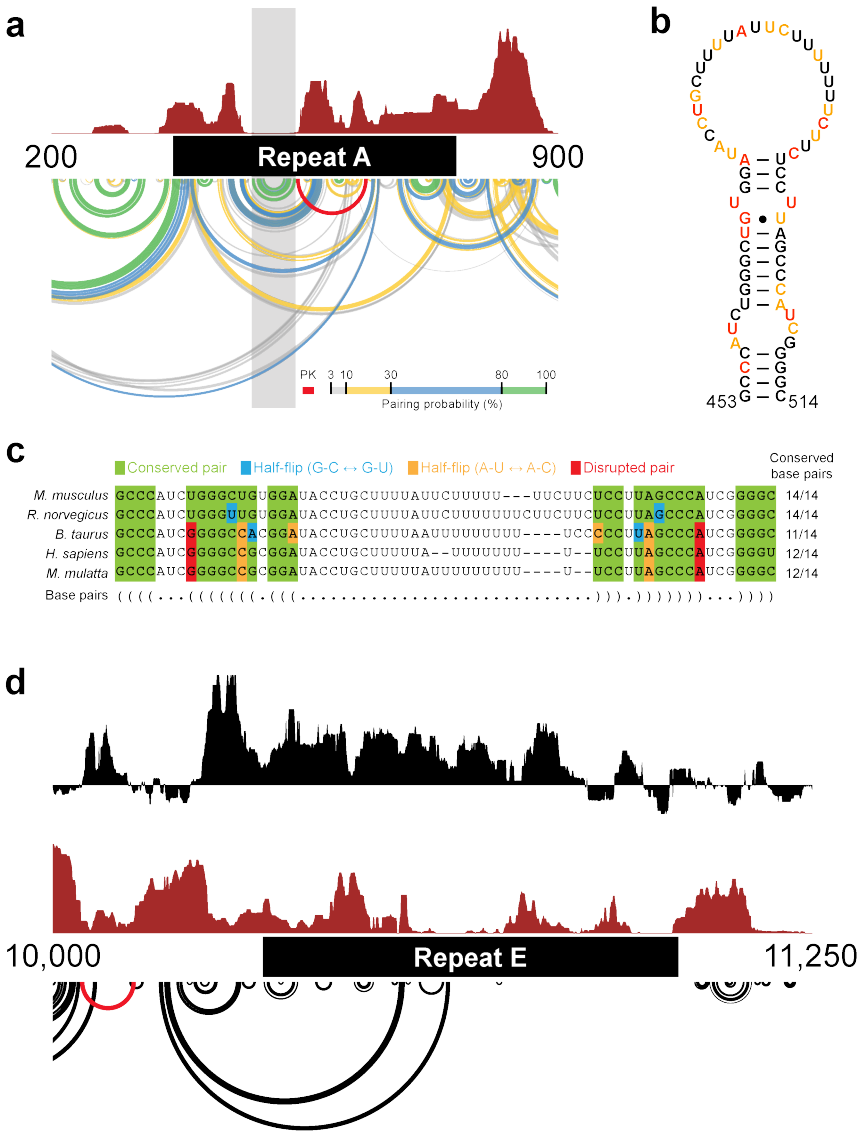


Figure 4.2 Structural features of repeat regions A and E. (a) Shannon entropy (top) and pairing probabilities for repeat A and the surrounding region (bottom). Probability arcs are colored based on probability with green likely (see scale). Only one high-probability stem-loop structure is predicted to occur within repeat A (grey shading). Other regions may interact with regions of *Xist* outside repeat A. (b) Secondary structure of the repeat A stem-loop. Nucleotides positions listed are relative to (Genbank accession no. NR_001463.3). (c) Sequence alignment of the stem-loop element for mouse, rat, cow, human, and monkey. Base pairs in the mouse structure are indicated in dot-bracket notation. Base-paired nucleotides are colored according to conservation. Nearly all base pairs are conserved (green), and most sequence variations maintain pairing potential through Hoogsteen (blue) or wobble interactions (yellow). (d) SHAPE reactivity (top) and Shannon entropy (middle) for the repeat E region. SHAPE reactivities are high and Shannon entropies are low throughout this region, indicating that it is likely largely unstructured as shown by the SHAPE-directed minimum free energy model of this region (bottom).

interrogated endogenous *Xist* structures in living cells using the 1M7 SHAPE reagent (**Fig. 4.3a**). We then evaluated absolute reactivity changes relative to *ex vivo* measurements. By searching for regions with an average absolute change greater than the global median, we identified 14 regions that are strongly affected by the cellular environment (**Fig. 4.3b**, purple shading). These regions overlap 16 well-defined RNA secondary structure domains and many dynamic regions (including repeats A and E). Prior work has shown that reduced in-cell SHAPE reactivities, relative to the *ex vivo* state, tend to report direct protein-RNA interactions, whereas increased reactivity in cells are often reflective of RNA conformational changes (*30*). By examining the relative contributions of positive and negative differences to the total absolute measured changes, we identified regions of *Xist* that likely interact with proteins and those that have different structures *ex vivo* and *in vivo* (**Fig. 4.3c-d**).

Nucleotides in repeat E undergo striking changes in SHAPE reactivity; this region was reactive *ex vivo* but very unreactive in cells (**Fig. 4.3b-d**). There also appears to be extensive protein binding to Repeat D in cells. There are notable changes in absolute SHAPE reactivity in repeat A, but these are not as strong as within other regions in *Xist*. We infer that repeat A forms RNA-protein interactions in cells but that these interactions are less stable or distinct than in other *Xist* motifs. The clear lack of structure in these repeat regions *ex vivo* suggests that nucleotides in these regions of *Xist* are poised for relatively unhindered access to proteins.

We also observed large differences between *ex vivo* and in-cell data in regions that span many of the low SHAPE/low Shannon entropy domains in the *ex vivo* model (**Fig. 4.3b-c**; for example, positions 12,300-12,700, 13,800-15,900, and 16,700-17,300). These data emphasize the diversity of in-cell interaction motifs encoded by *Xist*. Regions that exhibit distinct changes in SHAPE reactivity in cells span nearly the entirety of the *Xist* RNA, are characterized by multiple distinct features, and are comprised of both dynamic elements (repeats A, D, and E), and large, structurally well-defined RNA domains.

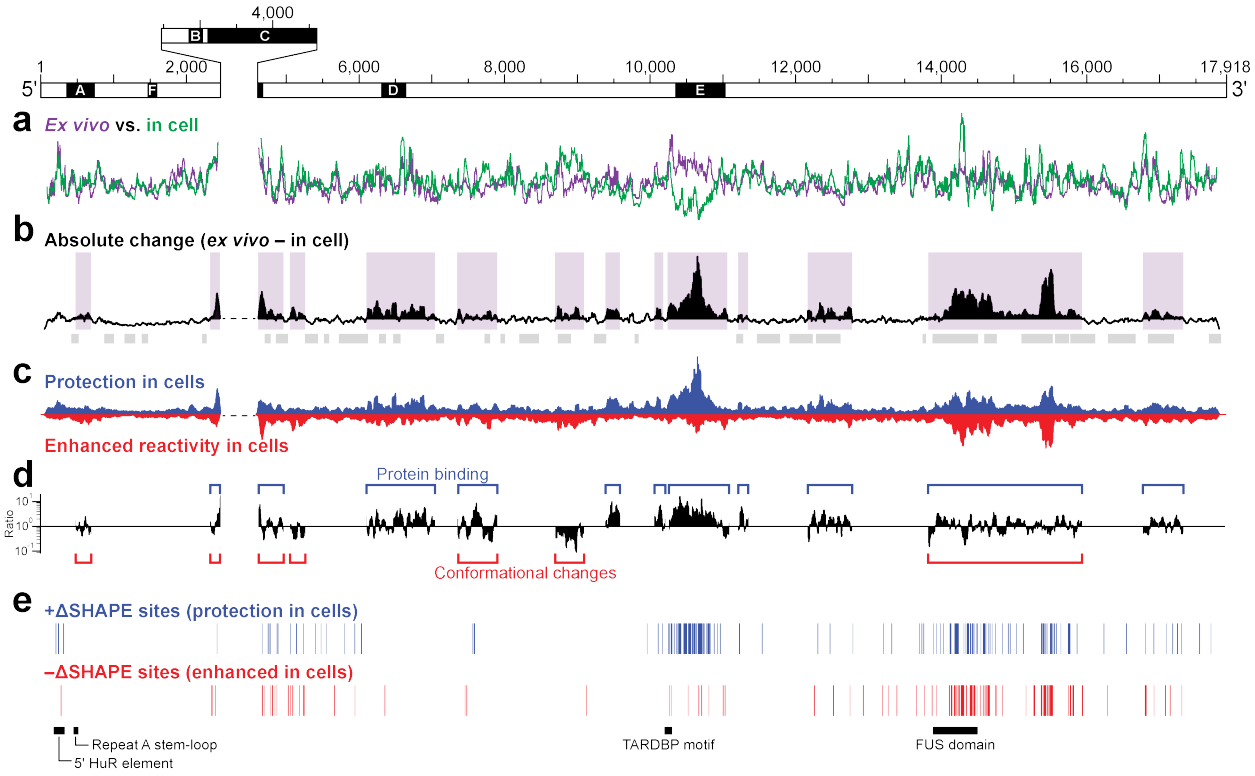


Figure 4.3 Effects of the cellular environment on *Xist* lncRNA structure. (a) Comparison of *ex vivo* (purple) and in-cell (green) SHAPE reactivities, plotted relative to the global median such that values above the line are more reactive than the median and those below the line are less reactive than the median. (b) The absolute change in SHAPE reactivity, computed over 50-nt sliding windows, is plotted over the length of *Xist*. Purple shading indicates regions that show the strongest differences in SHAPE reactivity when comparing *ex vivo* and in-cell data. Repeat E is characterized by a large absolute change, as are regions spanning 12,300-12,700, 13,800-15,900 and 16,700-17,300. Regions with low SHAPE reactivity and Shannon entropy are indicated with grey shading. (c) Contributions of positive (blue) and negative (red) reactivity differences to the total absolute change. In-cell values were subtracted from *ex vivo* values, such that positive differences represent reduced reactivity in cells, and vice versa. The sum of the blue and red areas equals the height of the black line in (b). (d) Ratio between positive and negative reactivity differences within regions of extensive reactivity change, highlighting the types of cellular effects that dominate each region. Values greater than 1 (above the line) indicate extensive in-cell protection, whereas those less than 1 (below the line) indicate enhanced reactivity in cells. Blue and red brackets indicate regions where protections or enhancements (or both) are most abundant. (e) Positive and negative ΔSHAPE sites. Blue sites are those which exhibit protection in cells while red sites indicate sites of enhanced reactivity, generally reporting protein binding and conformational changes, respectively.

Localized cellular effects on Xist structure

Each individual reactivity measurement in a SHAPE experiment includes an error estimate (*25*), thus allowing for statistically rigorous analysis of local changes in RNA structure. We have developed a ΔSHAPE comparison framework that incorporates these error estimates and accurately identifies small, specific sites within an RNA likely to be bound by protein or undergo conformational changes (*30*). Positive and negative ΔSHAPE values indicate protection from versus enhanced reactivity in cells, respectively (*30*). The ΔSHAPE analysis complements the large-scale structural changes identified above, and helps define sites of specific protein interactions or conformational changes in the *Xist* lncRNA in cells.

We identified ΔSHAPE sites throughout the length of *Xist* and found that the first 2.5 kb are largely lacking in ΔSHAPE sites, while in other regions they are abundant (**Fig. 4.3e**). Not surprisingly, regions with many ΔSHAPE sites are among those that exhibit strong absolute changes in SHAPE reactivity. We hypothesized that sequences critical to *Xist*-protein interactions may be over-represented among +ΔSHAPE sites and searched among them for sequence motifs. Our analysis revealed two highly abundant, U-rich sequence motifs, E1 and E2 (**Fig. 4.4**). These motifs are located exclusively within repeat E, and each motif contains a portion of the repeat unit. No additional significant sequence motifs, spanning ΔSHAPE sites, were identified outside of repeat E.

To determine the location of specific protein interactions along the *Xist* molecule, we searched the protein crosslinking and immunoprecipitation database (CLIPdb) (*58*) for proteins previously identified as *Xist* partners in TSCs: CELF1, PTBP1, TARDBP, FUS, and RBFOX2 (*55-57*). We also performed digestion-optimized RIP-seq experiments in TSCs to identify binding sites for the HuR protein, an *Xist*-interacting protein (*55*). We expected proteins that bound stably to *Xist* during our 2-min probing period would perturb the RNA structure and yield clear ΔSHAPE signals. For all proteins except RBFOX2, we identified CLIP or RIP sites that overlapped with positive and negative ΔSHAPE sites. We found that 79% of ΔSHAPE sites overlapped with CLIP or RIP sites, whereas only 47% of the total reported CLIP sites coincided with ΔSHAPE signals (**Fig. 4.5a**). This latter low number may reflect

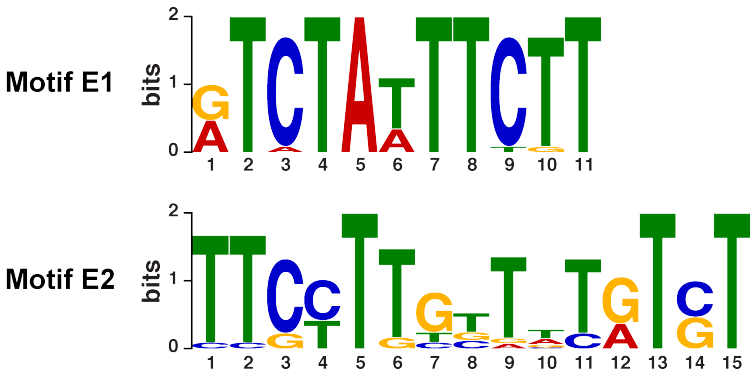


Figure 4.4 Sequence motifs identified among ΔSHAPE sites. These sites, termed E1 and E2 for their location within repeat E, were identified from sites protected in cells according to ΔSHAPE analysis.

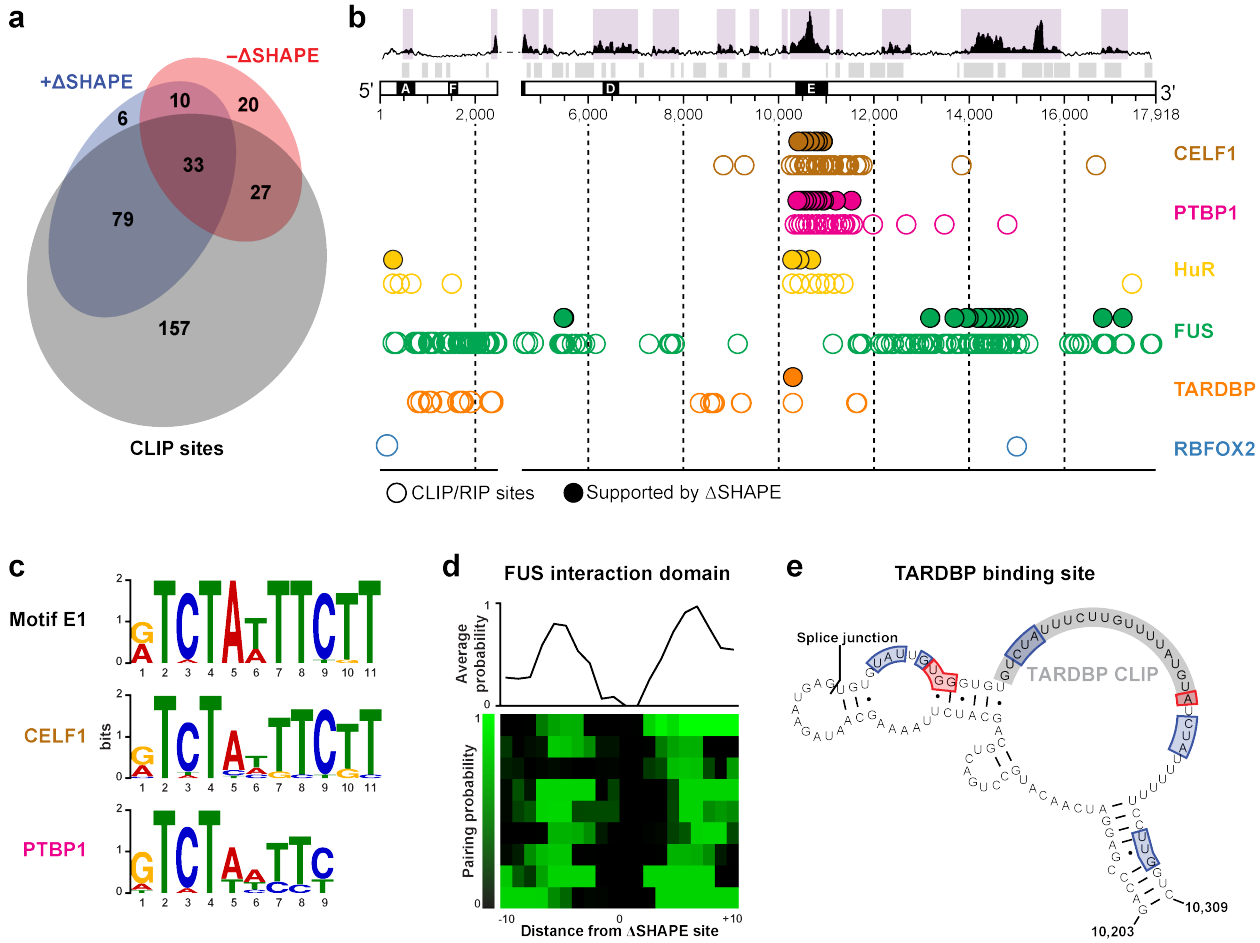


Figure 4.5 Correlation of ΔSHAPE sites with CLIP- and RIP-identified *Xist*-protein interactions. (a) Overlap between ΔSHAPE sites and CLIP- or RIP-identified sites. A subset of reported CLIP sites were confirmed as overlapping one or more ΔSHAPE sites for each protein with the exception of RBFOX2. Most ΔSHAPE-confirmed sites exhibit +ΔSHAPE changes (81% of total), while the remaining sites, confirmed only by –ΔSHAPE changes, may indicate proteins whose binding enhances SHAPE reactivity or correspond to non-specific sites that were incorrectly identified in CLIP experiments. (b) Locations of CLIP or RIP protein binding sites that overlap with positive ΔSHAPE sites. Regions of large absolute change (purple shading) and well-formed structure (low SHAPE/low Shannon entropy; grey shading) are shown at the top. CLIP- or RIP-defined protein binding sites are shown as open circles; sites specifically confirmed by ΔSHAPE measurements as filled circles. Data for CELF1, PTBP1, FUS, TARDBP and RBFOX2 were obtained from CLIP experiments (*58*); data for HuR were obtained using RIP (see Methods). (c) Sequence motifs identified among ΔSHAPE-confirmed CELF1 (middle) and PTBP1 (bottom) binding sites are similar to the E1 motif (top), indicating a preferred recognition motif in *Xist* for these proteins. (d) Clustering of pairing probabilities from CLIP-confirmed +ΔSHAPE sites reveal a structural preference for FUS binding. The average base-pairing probability is shown (top), derived from the major cluster of +ΔSHAPE sites within this region. (e) Structural context of the single ΔSHAPE-confirmed TARDBP binding site. The CLIP site is shaded grey. ΔSHAPE sites of in-cell protection and enhancement are boxed in blue and red, respectively. The splice junction between *Xist* exons 6 and 7 is highlighted.

differences between cell types, the high stringency used in the ΔSHAPE analysis (*30*), and the high background of CLIP experiments (*59*).

Given the low false positive detection rate of protein binding identified by +ΔSHAPE (*30*), we chose to focus on CLIP sites corroborated by positive ΔSHAPE values. This enabled us to identify sites likely bound by CELF1, PTBP1, and HuR in repeat E, to show that sites for FUS are concentrated in the well-folded RNA domains spanning positions 13,900-15,000, and to define a single site bound by TARDBP at position 10,285 (**Fig. 4.5b**, filled circles). Remarkably, despite the relatively small number of proteins in our analysis, it is clear that the 3′ end of *Xist* is extensively involved in in-cell interactions. Furthermore, this analysis confirms that repeat E is a major protein-binding platform (**Fig. 4.5b**).

ΔSHAPE-confirmed CELF1 and PTBP1 CLIP sites are located almost exclusively in repeat E (**Fig. 4.5b**). These proteins are associated with RNA processing (*60*, *61*) and may help regulate *Xist* splicing or editing. We used sequence clustering to define consensus motifs from +ΔSHAPE-validated CLIP sites for CELF1 and PTBP1 and found that both overlap with motif E1 (**Fig. 4.5c**). No strong consensus sequence was identified among non-ΔSHAPE-validated CLIP sites, indicating that repeat E interacts with CELF1 and PTBP1 in a sequence-specific manner.

We identified HuR binding sites throughout repeat E (**Fig. 4.5b**). HuR has complex roles including promoting mRNA stability through interactions with AU-rich elements (AREs) (*62*) and regulating nuclear RNA processing by repression splicing and increasing RNA stability to binding to intronic sequences (*63*). HuR was widely detected throughout the U-rich repeat E. Within repeat E, the HuR RIP signal is very strong, and the defined binding sites are much larger than the CLIP-informed binding sites of other proteins. A search for consensus motifs within these large sites returned the repetitive unit of repeat E. When subsequences corresponding to +ΔSHAPE in-cell protections were used, a more nuanced consensus was returned that contains elements from motifs E1 and E2. In addition to the strong association with repeat E, we also detect HuR binding at the 5′ end of *Xist*, upstream of repeat A. This element, which is predicted to fold into three helices with large internal loops, exhibits both positive

and negative ΔSHAPE signals. The pattern of ΔSHAPE binding sites fits a model in which the multiple RNA recognition motifs in HuR both disrupt a helical element and bind to elements of sequence separated by 100 nts of primary sequence. Although functionality of the *Xist* 5′ end is usually attributed to repeat A, our data emphasize the importance of the 300 nucleotides 5′ of the repeat region. The strong association of HuR in two defined regions may affect the stability of *Xist* transcripts or regulate *Xist* via combinatorial interactions with other RBPs that bind in the same regions.

Regions throughout *Xist* clearly serve as nucleation points for widespread interactions with proteins. FUS is an abundant, nuclear-enriched protein involved in the regulation of transcription, RNA processing, and DNA damage repair (*64-67*). Prior studies have shown that FUS binds to many RNAs (*68*). In contrast to the expectation of promiscuous binding, we find that in the context of full-length *Xist* RNA, +ΔSHAPE signals in CLIP sites indicative of FUS binding cluster strongly at nucleotides 13,000-15,000 (**Fig. 4.5b**). This region has a well-defined RNA structure (**Fig. 4.1**) and exhibits a complex mixture of positive and negative ΔSHAPE sites (**Figs. 4.3d-e**). We analyzed the pairing probabilities of FUS-associated +ΔSHAPE sites within this region and identified a structural context for FUS binding: FUS-protected nucleotides tend to occur in single-stranded motifs closely flanked by base paired structures (**Fig. 4.5d**). FUS can facilitate RNA-induced multimerization (*69*), and forms dynamic, liquid-like compartments *in vivo* (*70*), which may enable linkage of individual *Xist* molecules in the area that surrounds the inactive X (*71-73*). These observations are consistent with the abundant structural rearrangements we detect within the FUS-binding region, as the local increase in FUS concentration via multimerization may lead to widespread, likely cooperative, binding.

In contrast to the multiple-site binding observed for CELF1, PTBP1, HuR, and FUS, ΔSHAPE analysis supports a single CLIP-identified binding site for the TARDBP protein (**Fig. 4.5e**). TARDBP is an RNA/DNA binding protein with a reported preference for UG-rich sequences (*74*, *75*) and has been identified as both a transcription repressor (*76*, *77*) and splicing regulator (*74*, *75*, *78*). The single site detected by our analysis is part of a UG-rich structural motif (nts 10,200-10,320) encompassing the splice

junction between exons 6 and 7 (**Fig. 4.5e**). In adult mouse brains depleted of TARDBP via antisense knockdown, incorrectly spliced *Xist* transcripts increase 2-fold and overall *Xist* transcript levels are reduced 3-fold (*79*), suggesting that TARDBP controls the amount of *Xist* present in a cell. Several TARDBP binding sites were identified by CLIPdb, but only a single site overlapped with a region of positive ΔSHAPE. The median SHAPE reactivity of this site was much higher than any other reported TARDBP CLIP site. We examined the reactivity of all other positive ΔSHAPE sites and found that most exhibit lower *ex vivo* SHAPE reactivities than the confirmed TARDBP site, ruling out the possibility that the single TARDBP site was detected solely because of high *ex vivo* reactivity. These data suggest the remaining TARDBP sites are occluded by RNA structure or are not sufficiently stable to cause a detectable reduction in SHAPE reactivity when in-cell data and *ex vivo* data are compared. More broadly, this analysis of TARDBP binding emphasizes the role that *Xist* RNA structure plays in modulating the specificity of protein binding.

Conclusion

Here we used the SHAPE-MaP quantitative nucleotide-resolution structure probing strategy to show that the *Xist* lncRNA has domains of well-defined secondary structure linked by unstructured or dynamic regions (**Fig. 4.1**). Repeat regions are generally unstructured, which appears to facilitate binding by protein cofactors (**Figs. 4.1-4.5**). We identified three distinct modes of action by which protein cofactors form stable interactions with *Xist*. In each case, protein interactions corroborated by CLIP-/RIP-Seq and ΔSHAPE data are highly focused within specific structural elements. CELF1 and PTBP1 exemplify widespread binding to accessible, unstructured regions. FUS binding occurred in a region with a well-defined structure *ex vivo* that undergoes extensive rearrangement in cells. In contrast, TARDBP bound predominantly to a single site contained within a small, well-defined structural domain. Finally, HuR makes multiple contacts within the repeat E region and binds to a small region at the 5′ end of *Xist*.

In cross-referencing +ΔSHAPE sites with CLIP- and RIP-identified binding sites we have also shown ΔSHAPE to be a rigorous approach for identifying stable RNA-protein interaction sites (**Fig. 4.5**).

While CLIP studies have in several cases reported binding across the entire transcript, ΔSHAPE indicates that binding sites for a given protein tend to cluster, like those observed at repeat E and the FUS domain. In addition, only when +ΔSHAPE sites are analyzed is a distinct binding motif identified for HuR. ΔSHAPE also appears to detect specific interactions such as that observed at the TARDBP binding site. In general, ΔSHAPE analysis enables high confidence identification of RNA-protein interactions in a way that will be broadly useful in future studies of lncRNAs.

Direct structural interrogation clearly shows that the internal structure of a lncRNA can be complex and diverse. This is unsurprising, given the varied functions carried out by this class of RNAs. At 18 kb, *Xist* is much longer than most lncRNAs, including the SRA and HOTAIR RNAs, whose structures have been characterized recently (*47*, *48*). Like these other lncRNAs, the secondary structure of *Xist* is composed of distinct stable domains interspersed with regions that lack structure or that are structurally dynamic. These domains can now be used as a molecular roadmap, guiding future investigations into the mechanisms by which sequence elements embedded in *Xist* confer distinct biological activities, and how such elements contribute to XCI and lncRNA-induced gene silencing.

Diverse protein complexes are required to accomplish many of the functions carried out by lncRNAs, and our structural analyses revealed multiple mechanisms by which lncRNAs can interact with protein partners to accomplish biological tasks. This ability to coordinate proteins via distinct, domain-wise interactions may explain why certain lncRNAs are so long and why such RNAs are often capable of orchestrating epigenetic regulation on the kilobase to megabase scale (*80-83*). Our analysis of *Xist* supports the view that lncRNAs and other RNAs may have densely arrayed secondary structural features, exhibit multiple distinctive modes of protein interaction, and serve as multi-domain organizers of distinct cellular functions (*84*).

Methods

In-cell RNA modification

Mouse trophoblast stem cells were cultured as described (*85*). Live TSCs were washed once with PBS, and 900 µl of fresh growth media was added. For samples subjected to in-cell SHAPE probing, 100 µl of 100 mM 1M7 in neat DMSO were added (10 mM final concentration) and immediately mixed by swirling the culture dish. Cells were then incubated at 37 °C for 5 minutes (but note that the structure probing reaction is complete in ~2 min). Media was removed and the cells were washed once with PBS before isolation of total RNA (1 mL TRIzol; Ambion). The no-reagent negative control RNA was prepared similarly with the exception that neat DMSO was used instead of 1M7 in DMSO.

Ex vivo RNA extraction and modification

Approximately 10^6 TSCs were washed and pelleted in ice-cold PBS, resuspended in 2.5 ml Lysis Buffer [40 mM Tris, pH 7.9, 25 mM NaCl, 6 mM $MgCl_2$, 1 mM $CaCl_2$, 256 mM sucrose, 0.5% Triton X-100, 1,000 U/ml RNasin (Promega), 450 U/ml DNase I (Roche)], and rotated at 4 °C for 5 minutes. Cells were then pelleted at 4 °C for 2 minutes at 2250 g, resuspended in 40 mM Tris pH 7.9, 200 mM NaCl, 1.5% SDS, and 500 µg/ml of Proteinase K, and rotated at 20 °C for 45 minutes. RNA was then extracted twice with phenol:chloroform:isoamyl alcohol (24:24:1) pre-equilibrated with 1× Folding Buffer (100 mM HEPES, pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$), followed by one extraction with chloroform. RNA was exchanged into 1.1× Folding Buffer using a desalting column (PD-10, GE Life Sciences) and incubated at 37 °C for 20 minutes. Approximately 3 µg RNA was then added to a one-ninth volume of 1M7, 1M6, or NMIA, each at 100 mM in neat DMSO (10 mM final concentration), and then incubated at 37 °C for 5 minutes. Modified RNA was purified (RNeasy Midi spin column, Qiagen) and eluted in approximately 50 µl H_2O. No- reagent negative control RNA was prepared in the same way except that neat DMSO was substituted for SHAPE reagent.

Denaturing control

TSCs were grown as described (*85*), and total RNA was isolated using TRIzol (Ambion). Approximately 1.5 μg RNA was then resuspended in 150 μl 1.1× Denaturing Control Buffer [55 mM HEPES pH 8.0, 4.4 mM EDTA, 55% formamide (v/v)] and incubated at 95 °C for 1 minute.

Aliquots of 45 μl of denatured RNA were then added to 5 μl of 100 mM 1M7, 1M6, or NMIA, and allowed to react at 95 °C for 1 minute. After modification, RNA was purified (RNeasy Mini spin column, Qiagen) and eluted in approximately 50 µl H_2O.

Xist SHAPE-MaP

RNA was modified as described above. Mutational profiling reverse transcription reactions were primed with a mixture of *Xist*-specific primers (2 pmol each; **Table 4.1**) (*25*). The resulting cDNA was purified (Agencourt RNAClean XP beads, Beckman Coulter) and amplified by PCR (Q5 high-fidelity DNA polymerase, NEB) with *Xist*-specific primers (**Table 4.1**). These cDNAs (1.5 μL) were used as templates in individual 50 μl PCR reactions (1× Q5 Reaction Buffer, 200 μM dNTPs, 0.5 μM each primer, 0.02 U/μl Q5 high-fidelity DNA polymerase) using a touchdown format: 98 °C for 30 s, 25 cycles of [98 °C for 10 s, 72 °C for 30 s (decreasing by 1 °C per cycle until 60 °C), 72 °C for 30 s], 72 °C for 2 min. The resulting amplicons were purified (Agencourt RNAClean XP beads, Beckman Coulter) and pooled according to experimental treatment before construction of high-throughput sequencing libraries (Nextera XT, Illumina) (*27*).

Sequencing and SHAPE profile generation

Purified sequencing libraries were pooled and sequenced on an Illumina MiSeq or HiSeq instrument, generating 2 × 150 or 2 × 100 paired-end datasets. SHAPE reactivity profiles were created by aligning reads to the *Xist* reference sequence (GenBank accession NR_001463.3) using *ShapeMapper* (v1.0, http://chem.unc.edu/rna/software.html). Final reactivity profiles were generated by excluding nucleotides 1-78, 2451-2599, and 17801-17918 and renormalizing the remaining nucleotides using the boxplot approach (*43*). In-cell 1M7 reactivities were then scaled such that the median SHAPE reactivity

in the 95th percentile matched the *ex vivo* value. Differential SHAPE reactivities between 1M6 and NMIA (*44*) were computed using a Z-factor test (*25*).

Structure modeling

Potential pseudoknots in *Xist* were identified using a sliding window approach (*25*) in which full-length *Xist* was folded in 600-nt windows offset by 100-nt increments using ShapeKnots (*43*). Additional predictions were calculated at the 5′ and 3′ ends to increase sampling of terminal sequences and mitigate end effects. Predicted pseudoknots were inspected manually and retained if the structure was present in a majority of windows and if SHAPE reactivity was low on both sides of the potential helices. The model of *ex vivo Xist* secondary structure was created by providing SuperFold (*25*, *27*) with 1M7 reactivities in addition to differential SHAPE values and pseudoknotted helices. SHAPE reactivities and Shannon entropies were smoothed over centered 55-nt sliding windows. Regions in which the local median was less than the global median for at least 40 nts were flagged as well-structured regions. Regions separated by fewer than 10 nts were combined before expanding regions to include all secondary structure contacts.

SNP analysis

Sequence variation data were obtained from the Sanger Institute (http://www.sanger.ac.uk). Positions of *Xist* exons were obtained from Ensembl (http://www.ensembl.org). The genomic exon locations were used to convert genomic SNP locations to their equivalent position on the *Xist* transcript. A multi-sequence alignment was performed to ensure agreement of SNP locations. We then examined whether SNP locations corresponded to base-paired or single-stranded conformations in our structure models. A bootstrapping approach was used to determine the likelihood of encountering fewer structure-disrupting SNPs by chance. For 105 randomly chosen nucleotides, we recorded how many were base paired in a given structural model and iterated this process 100,000 times. We used the results to model a cumulative distribution function from which we calculated the probability of finding fewer disruptive SNPs than reported in the experimental data.

Conservation analysis

Sequences of *Xist* loci from mouse (*Mus musculus*), rat (*Rattus norvegicus*), cow (*Bos taurus*), human (*Homo sapiens*), and rhesus macaque (*Macaca mulatta*) were obtained from the UCSC Genome Browser (*86*) and aligned with Clustal Omega (*87*). The RASL motif was then extracted from the alignment and analyzed for sequence conservation and covariation using R2R (*88*).

Computing regions of large absolute reactivity changes

The absolute value of the SHAPE reactivity difference between *ex vivo* and in-cell conditions was summed over 50-nt sliding windows. Total positive and negative changes were calculated in the same way, except the absolute value was not used. Regions of differences were defined as those in which absolute differences were greater than the global median over at least 100 consecutive nucleotides.

Identifying protein binding sites and conformational changes with ΔSHAPE

ΔSHAPE values were calculated by comparing the *ex vivo* and in-cell conditions as described (*30*) with the slight modification that differences greater than 20 (due to hyper-reactive nucleotides (*29*)) were excluded from analysis. The 798 nucleotides that differed significantly in reactivity between *ex vivo* and in-cell conditions were then grouped into 175 interaction sites by a sliding window approach. Five-nucleotide windows were assessed for occupancy by at least three ΔSHAPE-identified nucleotides. Qualifying nucleotides within any adjacent windows meeting this criterion were pooled together as members of a single interaction site.

Sites of CELF1, PTBP1, FUS, TARDBP, and RBFOX2 were downloaded from CLIPdb. These data represent sites of cellular interactions in mouse brain tissue (FUS, TARDBP, RBFOX2) and cultured myoblasts (CELF1 and PTBP1) and are expected to provide a high-level view of *Xist*-protein binding. Any CLIP sites that overlapped with a +ΔSHAPE site were selected as confirmed sites of protein interaction.

HuR RNA immunoprecipitation and sequencing

Mouse trophoblast lysates were prepared according to the protocol described previously (*89*). The RNA immunoprecipitation (RIP) of HuR-RNA complexes were also performed similarly, with the addition of micrococcal nuclease to partially digest RNA in the lysates before RIP. RNA fragments from the HuR RIP and total input RNA (treated with micrococcal nuclease) were made into cDNA libraries (NEBNext Small RNA Library Prep, NEB), and then sequenced (Illumina Hi-Seq 2500). Raw reads from the digestion-optimized RIP libraries were preprocessed to remove adapter sequences and PCR artifacts. Preprocessed reads were mapped to the mouse genome (mm9) using *TopHat* (*90*). Uniquely mapped reads from the RIP were normalized to the input reads across the genome, and log of odds ratios calculated for each site using a mixture model approach as previously described (*91*). HuR binding sites were defined as regions with a log-odds score in the 95th percentile of all HuR sites transcriptome-wide.

Identification of sequence motifs among ΔSHAPE-identified interaction sites

Sequences corresponding to interaction sites associated with positive ΔSHAPE sites were extracted and expanded to 20 nts. These sequences were then searched for sequence motifs with *MEME* (*92*), allowing for at least two sites per motif, a minimum motif length of four nucleotides, and allowing any number of motifs per input sequence. The nucleotide distribution of the *Xist* transcript was used as the background when calculating significance. Searching in this manner yielded sequence motifs E1 and E2 with expectation values 1.4×10^{-36} and 8.7×10^{-29}, respectively. Sequence motifs within ΔSHAPE-confirmed CLIP sites were analyzed in the same way. This identified motifs represented in the CELF1 and PTBP1 binding sites with expectation values of 9.3×10^{-40} and 4.5×10^{-17}, respectively.

Clustering Fus-localized positive ΔSHAPE sites by pairing probability

Total pairing probabilities for each nucleotide in expanded +ΔSHAPE sites (see above) were extracted from the partition function using RNAtools (http://www.github.com/grice/RNAtools). Sites were then sorted by uncentered absolute correlation similarity into three clusters by *k*-means clustering implemented with Cluster 3.0 (*93*). Clusters were visualized using TreeView (*94*).

Evaluation of TARDBP antisense knock-down data

TARDBP protein levels were previously depleted by targeting RNase H to the TARDBP pre-mRNA with antisense DNA oligonucleotides (*79*). RNA-seq reads for control and TARDBP knock-down samples (4 replicates each) were downloaded from the Gene Expression Omnibus (accession no. GSE27394). Adapter sequences were removed and filtered for base call quality before alignment to the mouse genome (mm10) with Bowtie2 (*95*). The number of reads overlapping introns was then computed and compared to exon-aligned reads.

Amplicon name	Forward primer	Reverse primer (RT primer)
Amp_22	TCCATCTAAGGAGCTTTGGG	ATAGGTTCACTCACACAGCA
Amp_21	GCTTGGTGGATGGAAATATGG	CGTTATACCGCACCAAGAAC
Amp_20	AGCGGACTGGATAAAAGCAAC	CATCACAGTCTAATTCCATCCTG
Amp_19	TGTTGGTGTTTGCTTGACTTCC	AAACTTTAAGGACTCCAAAGTAAC
Amp_18	CGTCTGATAGTGTGCTTTGC	GGCTTGGGATAGGTCTGAAA
Amp_RepC	CCAGGCCCAGATACTTTCAG	TGTTTGCCCCTTTGCTAAAT
Amp_16	CCCATCTATACCCCCTCCAT	GCAAGGGTAGTATTAGGACCTTGAG
Amp_15	TCACATGCTTTCTTATTTCAGCC	AGTTAACACTGTGCACATTTAC
Amp_14	GGTTCCTACCACTATGCCCTG	AAAACCCCATCCTTTATGCAA
Amp_13	AAACCCTTTTGCAGTACAGC	GCCTCTGGTGTCAAAAGAGTAC
Amp_12	AGCAGAAAGAGGGTTGTACG	TGATGGAATTGAGAAAGGGCAC
Amp_11	TCCATTGACCACTTTTCTGAATCAC	AAGATACTTGTCTTAAACATTCTGC
Amp_10	TACTGAGGGTGATGAGTCTGT	TCAGCAATGTCATATCAAACAC
Amp_9	CTCAGACAACAATGGGAGCT	GCATTCTTTGAGCCTTTGTCT
Amp_8	ACAAAAAGCTTACAGGCCACA	AATAGACACAAAGCAAGGAAG
Amp_7	TGAGTGTGTATTGTGGGTGTGT	ACACTGCAGACAGAAAAGAC
Amp_6	GTCTCCTTGTGTTGTCTAATTCG	TTCTGGACCTATTGGGAAGGG
Amp_5	TTGTGTCTCTTTGCTATTGGTGG	TTCCTTTATGGGCAATGGCAAC
Amp_4	CCCAGCATCCCTTTCCATTTC	AATTGCCAATGTGCTATGAG
Amp_3	AGGACTACTTAACGGGCTTA	AGGGTAATCAATCACCTGCA
Amp_2	TAAGGCCAGTGTGTTTCTTC	ACACATGGGAGACAATATTTAGC
Amp_1	AACCTGGGTGTGTTAGCATG	AAGTGCTACATAATGCTAGAAAG
Amp_0	TGGCTGAAACTTAATCATAATGC	CTGGGCATGGGTTAAGCA
Amp_-1	GAGACATGGTCTCATAAAGCC	TGTGTGGAACCGAGGAAATA
Amp_-2	TGATGAAAGTGCAGTTCTAAGTA	TTGCCCCAGGATAATGCAAA
Amp_-3	TAGGCCATTTTAGCTATGACTGT	TTTGAACTCCCAGACCTCTTC
Amp_-4	AGTTGCCTTAGAGCTGAAGT	TTGTGACATGTTGGTAAGCA

Table 4.1. Primer sequences used to create *Xist* amplicons. Amplicons generated with these primers enable specific, transcript-wide structural interrogation of the *Xist* RNA in the context of total mouse cellular RNA. Sequences are written in the 5'→3' direction.

REFERENCES

1. J. L. Rinn, H. Y. Chang, Genome Regulation by Long Noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).

2. L. Yang, J. E. Froberg, J. T. Lee, Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem. Sci.* **39**, 35–43 (2014).

3. A. Fatica, I. Bozzoni, Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* **15**, 7–21 (2014).

4. B. K. Dey, A. C. Mueller, A. Dutta, Long non-coding RNAs as emerging regulators of differentiation, development, and disease. *Transcription.* **5**, e944014 (2014).

5. J. R. Prensner, A. M. Chinnaiyan, The emergence of lncRNAs in cancer biology. *Cancer Discov.* **1**, 391–407 (2011).

6. M. K. Iyer *et al.*, The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).

7. T. R. Mercer *et al.*, Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**, 99–104 (2012).

8. J. Harrow *et al.*, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

9. J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, A. Hamosh, OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–98 (2015).

10. T. R. Mercer, J. S. Mattick, Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* **20**, 300–307 (2013).

11. A.-V. Gendrel, E. Heard, Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. *Annu. Rev. Cell Dev. Biol.* **30**, 561–580 (2014).

12. J. T. Lee, M. S. Bartolomei, X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell.* **152**, 1308–1323 (2013).

13. C. J. Brown *et al.*, The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell.* **71**, 527–542 (1992).

14. N. Brockdorff *et al.*, The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell.* **71**, 515–526 (1992).

15. T. B. Nesterova *et al.*, Characterization of the Genomic Xist Locus in Rodents Reveals Conservation of Overall Gene Structure and Tandem Repeats but Rapid Evolution of Unique Sequence. *Genome Res.* **11**, 833–849 (2001).

16. A. Wutz, T. P. Rasmussen, R. Jaenisch, Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.* **30**, 167–174 (2002).

17. J. Minks, S. E. Baldry, C. Yang, A. M. Cotton, C. J. Brown, XIST-induced silencing of flanking genes is achieved by additive action of repeat A monomers in human somatic cells. *Epigenetics Chromatin*. **6**, 1–1 (2013).

18. K. Sarma, P. Levasseur, A. Aristarkhov, J. T. Lee, Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 22196–22201 (2010).

19. J. Zhao, B. K. Sun, J. A. Erwin, J.-J. Song, J. T. Lee, Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*. **322**, 750–756 (2008).

20. Y. Hoki *et al*., A proximal conserved repeat in the Xist gene is essential as a genomic element for X-inactivation in mouse. *Development*. **136**, 139–146 (2008).

21. M. E. Royce-Tolland *et al*., The A-repeat links ASF/SF2-dependent Xist RNA processing with random choice during X inactivation. *Nat. Struct. Mol. Biol*. **17**, 948–954 (2010).

22. S. T. da Rocha *et al*., Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. *Mol. Cell*. **53**, 301–316 (2014).

23. H. F. Noller, C. R. Woese, Secondary structure of 16S ribosomal RNA. *Science*. **212**, 403–411 (1981).

24. J. M. Watts *et al*., Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. **460**, 711–716 (2009).

25. N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson, K. M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*. **11**, 959–965 (2014).

26. C. A. Lavender, R. J. Gorelick, K. M. Weeks, Structure-Based Alignment and Consensus Secondary Structures for Three HIV-Related RNA Genomes. *PLoS Comput. Biol*. **11**, e1004230 (2015).

27. M. J. Smola, G. M. Rice, S. Busan, N. A. Siegfried, K. M. Weeks, Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile, and accurate RNA structure analysis. *Nat. Protoc*. **10**, 1643–1669 (2015).

28. K. E. Deigan, T. W. Li, D. H. Mathews, K. M. Weeks, Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 97–102 (2009).

29. J. L. McGinnis, J. A. Dunkle, J. H. D. Cate, K. M. Weeks, The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc*. **134**, 6617–6624 (2012).

30. M. J. Smola, J. M. Calabrese, K. M. Weeks, Detection of RNA-protein interactions in living cells with SHAPE. *Biochemistry*. in revision.

31. J. L. McGinnis *et al*., In-cell SHAPE reveals that free 30S ribosome subunits are in the inactive state. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2425–2430 (2015).

32. D. M. Mauger *et al*., Functionally conserved architecture of hepatitis C virus RNA genomes. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3692–3697 (2015).

33. J. Tyrrell, J. L. McGinnis, K. M. Weeks, G. J. Pielak, The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry*. **52**, 8777–8785 (2013).

34. J. L. McGinnis, K. M. Weeks, Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry*. **53**, 3237–3247 (2014).

35. C. A. Lavender *et al*., Model-Free RNA Sequence and Structure Alignment Informed by SHAPE Probing Reveals a Conserved Alternate Secondary Structure for 16S rRNA. *PLoS Comput. Biol.* **11**, e1004126 (2015).

36. J. M. Calabrese *et al*., Site-Specific Silencing of Regulatory Elements as a Mechanism of X Inactivation. *Cell*. **151**, 951–963 (2012).

37. S. Kalantry *et al*., The Polycomb group protein Eed protects the inactive X-chromosome from differentiation-induced reactivation. *Nature Cell Biology*. **8**, 195–202 (2006).

38. W. Mak, J. Baxter, J. Silva, A. E. Newall, A. P. Otte, Mitotically stable association of polycomb group proteins eed and enx1 with the inactive x chromosome in trophoblast stem cells. *Current Biology* (2002).

39. T. Ohhata, C. E. Senner, M. Hemberger, A. Wutz, Lineage-specific function of the noncoding Tsix RNA for Xist repression and Xi reactivation in mice. *Genes Dev*. **25**, 1702–1715 (2011).

40. J. W. Mugford, D. Yee, T. Magnuson, Failure of extra-embryonic progenitor maintenance in the absence of dosage compensation. *Development*. **139**, 2130–2138 (2012).

41. S. A. Mortimer, K. M. Weeks, A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc*. **129**, 4144–4145 (2007).

42. K.-A. Steen, G. M. Rice, K. M. Weeks, Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *J. Am. Chem. Soc*. **134**, 13160–13163 (2012).

43. C. E. Hajdin *et al*., Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A*. **110**, 5498–5503 (2013).

44. G. M. Rice, C. W. Leonard, K. M. Weeks, RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA*. **20**, 846–854 (2014).

45. Y. Guo, D. C. Jamison, The distribution of SNPs in human gene regulatory regions. *BMC Genomics*. **6**, 140 (2005).

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

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

48. S. Somarowthu *et al*., HOTAIR Forms an Intricate and Modular Secondary Structure. *Mol. Cell*. **58**, 353–361 (2015).

49. A. Wutz, R. Jaenisch, A shift from reversible to irreversible X inactivation is triggered during ES

cell differentiation. *Mol. Cell.* **5**, 695–705 (2000).

50. M. M. Duszczyk, K. Zanier, M. Sattler, A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat. *Nucleic Acids Res.* **36**, 7068–7077 (2008).

51. S. Maenner *et al.*, 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol.* **8**, e1000276 (2010).

52. M. M. Duszczyk, A. Wutz, V. Rybin, M. Sattler, The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization. *RNA.* **17**, 1973–1982 (2011).

53. C. E. Senner *et al.*, Disruption of a conserved region of Xist exon 1 impairs Xist RNA localisation and X-linked gene silencing during random and imprinted X chromosome inactivation. *Development.* **138**, 1541–1550 (2011).

54. Y. Hoki *et al.*, Incomplete X-inactivation initiated by a hypomorphic Xist allele in the mouse. *Development.* **138**, 2649–2659 (2011).

55. C. Chu *et al.*, Systematic Discovery of Xist RNA Binding Proteins. *Cell.* **161**, 404–416 (2015).

56. C. A. McHugh *et al.*, The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature.* **521**, 232–236 (2015).

57. A. Minajigi *et al.*, A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* (2015), doi:10.1126/science.aab2276.

58. Y.-C. T. Yang, CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 1–8 (2015).

59. M. B. Friedersdorf, J. D. Keene, Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol.* **15**, R2 (2014).

60. C. Barreau, L. Paillard, A. Méreau, H. B. Osborne, Mammalian CELF/Bruno-like RNA-binding proteins: molecular characteristics and biological functions. *Biochimie.* **88**, 515–525 (2006).

61. E. J. Wagner, M. A. Garcia-Blanco, Polypyrimidine tract binding protein antagonizes exon definition. *Mol. Cell. Biol.* **21**, 3281–3288 (2001).

62. M. N. Hinman, H. Lou, Diverse molecular functions of Hu proteins. *Cell. Mol. Life Sci.* **65**, 3168–3181 (2008).

63. N. Mukherjee *et al.*, Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell.* **43**, 327–339 (2011).

64. X. Wang *et al.*, Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature.* **454**, 126–130 (2008).

65. M. Polymenidou *et al.*, Misregulated RNA processing in amyotrophic lateral sclerosis. *Brain Res.* **1462**, 3–15 (2012).

66. W.-Y. Wang *et al*., Interaction of FUS and HDAC1 regulates DNA damage response and repair in neurons. *Nature Neuroscience*. **16**, 1383–1391 (2013).

67. J. C. Schwartz *et al*., FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes Dev*. **26**, 2690–2695 (2012).

68. X. Wang, J. C. Schwartz, T. R. Cech, Nucleic acid-binding specificity of human FUS protein. *Nucleic Acids Res*., gkv679 (2015).

69. J. C. Schwartz, X. Wang, E. R. Podell, T. R. Cech, RNA Seeds Higher-Order Assembly of FUS Protein. *Cell Rep*. **5**, 918–925 (2013).

70. A. Patel *et al*., A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*. **162**, 1066–1077 (2015).

71. D. Smeets *et al*., Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. *Epigenetics Chromatin*. **7**, 8 (2014).

72. A. Cerase *et al*., Spatial separation of Xist RNA and polycomb proteins revealed by superresolution microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2235–2240 (2014).

73. H. Sunwoo, J. Y. Wu, J. T. Lee, The Xist RNA-PRC2 complex at 20-nm resolution reveals a low Xist stoichiometry and suggests a hit-and-run mechanism in mouse cells. *Proc. Natl. Acad. Sci. U.S.A.*, 201503690 (2015).

74. E. Buratti, F. E. Baralle, Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of *CFTR* exon 9. *J. Biol. Chem.* **276**, 36337–36343 (2001).

75. E. Buratti, A. Brindisi, F. Pagani, F. E. Baralle, Nuclear factor TDP-43 binds to the polymorphic TG repeats in CFTR intron 8 and causes skipping of exon 9: a functional link with disease penetrance. *Am. J. Hum. Genet*. **74**, 1322–1325 (2004).

76. S. H. Ou, F. Wu, D. Harrich, L. F. García-Martínez, R. B. Gaynor, Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. *J. Virol*. **69**, 3584–3596 (1995).

77. A. S. Lalmansingh, C. J. Urekar, P. P. Reddi, TDP-43 is a transcriptional repressor: the testis-specific mouse *acrv1* gene is a TDP-43 target in vivo. *J. Biol. Chem*. **286**, 10970–10982 (2011).

78. C. Lagier-Tourenne, M. Polymenidou, D. W. Cleveland, TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Hum. Mol. Genet*. **19**, R46–R64 (2010).

79. M. Polymenidou *et al*., Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nature Neuroscience*. **14**, 459–468 (2011).

80. G. D. Penny, G. F. Kay, S. A. Sheardown, S. Rastan, N. Brockdorff, Requirement for Xist in X chromosome inactivation. *Nature*. **379**, 131–137 (1996).

81. F. Sleutels, R. Zwart, D. P. Barlow, The non-coding Air RNA is required for silencing autosomal

imprinted genes. *Nature*. **415**, 810–813 (2002).

82. D. Mancini-Dinardo, S. J. S. Steele, J. M. Levorse, R. S. Ingram, S. M. Tilghman, Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev*. **20**, 1268–1282 (2006).

83. J. Grant *et al*., Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature*. **487**, 254–258 (2012).

84. M. Guttman, J. L. Rinn, Modular regulatory principles of large non-coding RNAs. *Nature*. **482**, 339–346 (2012).

85. J. Quinn, T. Kunath, J. Rossant, in *Placenta and Trophoblast* (Humana Press, New Jersey, 2005), vol. 121, pp. 123–146.

86. W. J. Kent *et al*., The human genome browser at UCSC. *Genome Res*. **12**, 996–1006 (2002).

87. F. Sievers *et al*., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. **7**, 1–6 (2011).

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

89. J. D. Keene, J. M. Komisarow, M. B. Friedersdorf, RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat. Protoc*. **1**, 302–307 (2006).

90. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. **25**, 1105–1111 (2009).

91. N. Mukherjee, P. J. Lager, M. B. Friedersdorf, M. A. Thompson, J. D. Keene, Coordinated posttranscriptional mRNA population dynamics during T-cell activation. *Molecular Systems Biology*. **5**, 288 (2009).

92. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **2**, 28–36 (1994).

93. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A*. **95**, 14863–14868 (1998).

94. A. J. Saldanha, Java Treeview--extensible visualization of microarray data. *Bioinformatics*. **20**, 3246–3248 (2004).

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