SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist IncRNA in living cells


The 18-kb Xist long noncoding RNA (lncRNA) is essential for X-chromosome inactivation during female eutherian mammalian development. Global structural architecture, cell-induced conformational changes, and protein–RNA interactions within Xist are poorly understood. We used selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) to examine these features of Xist at single-nucleotide resolution both in living cells and ex vivo. The Xist RNA forms complex well-defined secondary structure domains and the cellular environment strongly modulates the RNA structure, via motifs spanning one-half of all Xist nucleotides. The Xist RNA structure modulates protein interactions in cells via multiple mechanisms. For example, repeat-containing elements adopt accessible and dynamic structures that function as landing pads for protein cofactors. Structured RNA motifs create interaction domains for specific proteins and also sequester other motifs, such that only a subset of potential binding sites forms stable interactions. This work creates a broad quantitative framework for understanding structure–function interrelationships for Xist and other IncRNAs in cells.

RNA structure | RNA–protein interaction | SHAPE-MaP | X-inactivation

Long noncoding RNAs (lncRNAs) play central roles in the regulation of gene expression through interactions with numerous protein partners (1) and are necessary for normal health and development (2, 3). The 18-kb Xist lncRNA is essential for X-chromosome inactivation during female eutherian mammalian development and is an archetype of gene-silencing lncRNAs. During the early stages of X inactivation, Xist accumulates in cis around the future inactive X chromosome and recruits protein complexes that apply repressive chromatin modifications, leading to stable gene silencing (3, 4).

Genetic deletion studies have demarcated several broad regions of function within Xist. Several tandem repeat regions (labeled A–F in the mouse) show moderate conservation (5–7), and at least two of these, repeat A and the rodent-specific repeat C, are implicated in silencing and localization to the inactive X. Deletion of the final 7.5-kb exon of Xist causes a defect in its localization (8), and the 1.5-kb region encompassing repeats F and B is required for accumulation of heterochromatin marks over the inactive X (4); however, beyond these initial characterizations, the mechanisms by which gene silencing, heterochromatinization, and localization of Xist on the X chromosome occur are not well understood. In particular, the role of RNA structure in orchestrating these distinct functions remains unclear.

Several previous studies have suggested the importance of RNA structures in specific regions of Xist (9–12), but overall, the locations and structures of functional domains within Xist are poorly defined. Detailed structural maps of other functional RNAs, such as ribosomal RNAs (13) and the HIV RNA genome (14–16), have been fundamental to understanding the mechanisms by which individual domains within large RNAs execute discrete cellular functions. A detailed and quantitative 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) provides a biophysically rigorous measurement of local nucleotide flexibility that is independent of base identity (15). SHAPE-MaP readily detects modifications in highly complex environments, including in the cell nucleus (17), and, unlike alternative RNA-probing methods with deep sequencing readout, is unaffected by biases introduced during complex ligation-based library preparation steps (15). SHAPE data are sufficient for distinguishing between structural models (18) and detecting distinct modes of protein binding in cells (17). SHAPE-informed structural models have consistently yielded rich insights into the biological functions of diverse RNAs (14, 15, 18–21) and, in many cases, uncovered novel functional elements (14–17, 20).

Using SHAPE-MaP, we examined full-length, authentic transcripts of mouse Xist at single-nucleotide resolution in mouse trophoblast stem cells (TSCs) and under protein-free conditions (ex vivo). TSCs demonstrate prototypical epigenetic patterns over the inactive X chromosome (22) and require Xist for continued silencing (23). SHAPE data identified 33 regions in Xist that form well-defined structures with complexities comparable to those of functional elements within RNA viruses and ribosomal RNAs (15, 21). We found extensive significant differences between in-cell SHAPE reactivities and those obtained ex vivo, indicating that many nucleotides of Xist interact with proteins or have different conformations in cells vs. a cell-free state. The perspective obtained

Significance

Long noncoding RNAs (lncRNAs) are important regulators of gene expression, but their structural features are largely unknown. We used structure-selective chemical probing to examine the structure of the Xist lncRNA in living cells and found that the RNA adopts well-defined and complex structures throughout its entire 18-kb length. By looking for changes in reactivity induced by the cellular environment, we were able to identify numerous previously unknown hubs of protein interaction. We also found that the Xist structure governs specific protein interactions in multiple distinct ways. Our results provide a detailed structural context for Xist function and lay a foundation for understanding structure–function relationships in all IncRNAs.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Raw sequencing data have been deposited in the Sequence Read Archive (accession no. SRP074108). Processed data are available in the SI Appendix and at www.chem.unc.edu/.

1To whom correspondence may be addressed. Email: weeks@unc.edu or jcalabre@emailmed.unc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.160008113/-/DCSupplemental.

Edited by Joan A. Steitz, Howard Hughes Medical Institute, New Haven, CT, and approved July 19, 2016 (received for review January 1, 2016)
here supports novel and specific models of the complex interrelationships among lncRNA sequence, structure, and function.

**Results**

**Ex Vivo Structure Probing.** We probed full-length Xist, after gentle and nondenaturing extraction from cells, using the SHAPE reagents 1-methyl-7-nitroisatoic anhydride (1M7), 1-methyl-6-nitroisatoic anhydride (1M6), and N-methyl-isatoic anhydride (NMIA) (24, 25) and obtained ex vivo SHAPE reactivities for 56% of nucleotides in Xist (Fig. L4). We also probed the Xist structure in living cells with 1M7. Full biological replicates of 1M7 probing, performed more than 1 y apart, showed good agreement over thousands of nucleotides under ex vivo conditions (Spearman’s ρ = 0.65; SI Appendix, Fig. S1A). In-cell replicates exhibited a more modest correlation (Spearman’s ρ = 0.50, SI Appendix, Fig. S1B). Critically, however, the in-cell replicates yielded highly similar outcomes in subsequent analyses (SI Appendix, Figs. S2 and S3 and Supporting Text).

We used the cell-free ex vivo data to guide initial RNA structure modeling. We searched for and identified 10 potential pseudoknots (26) and modeled the secondary structure of Xist using the three-reagent differential SHAPE strategy, which yields highly accurate RNA structure models (15, 25). The structure was also modeled without SHAPE data and with only 1M7 data (SI Appendix, Fig. S4 A and B).

To assess our models, we examined the structural context of 105 single-nucleotide variants (SNVs) within mouse Xist (27). For each structural model (no data, 1M7 only, or three-reagent differential), we counted the SNVs that disrupted structure by creating base pair mismatches. With increasing data quality, the probability that SNVs are structurally disruptive by chance decreases significantly (SI Appendix, Fig. S4C; P = 0.35, 0.15, and 0.027 for the no data, 1M7 only, and three-reagent models, respectively). Just as lack of selective pressure leads to increased SNV abundance in genetic elements with low functional potential (28), we infer that SNVs occur predominantly in unstructured regions in our model because many RNA structures within Xist are important for function.

We further assessed how well each secondary structure element was defined by its sequence and the experimental SHAPE data by calculating Shannon entropies at nucleotide resolution (15) (Fig. 1B). Previous work with large viral RNAs has shown that functional elements are overrepresented in regions with both low SHAPE reactivity (indicating a high degree of structure) and low Shannon entropy (indicating well-defined structure) (15, 21). We identified 33 regions with low SHAPE reactivity and Shannon entropy in the Xist RNA (Fig. 1 C and D, gray shading and SI Appendix, Fig. S5). Of the well-defined domains, three-fourths have not been described previously (SI Appendix, Supporting Text).

Many of the well-defined structural elements in Xist are located within the final 4,000 nucleotides (Fig. 1 C and D). Much of this region was missing from the original Xist annotation in mouse (7) and is dispensable for gene silencing in transgenic settings (29, 30). Nevertheless, the extent of defined structures within the region suggested functional roles. To test this possibility, we induced...
expression of full-length Xist or a 14.8-kb transcript lacking the 3' end from an isogenic site within the β-globin gene locus of a male mouse embryonic stem cell line (31). We found that the half-life of full-length Xist was threefold longer than that of the truncated version (SI Appendix, Fig. S6), consistent with a role for 3' structured elements in maintaining Xist stability in cells.

The 400-nt long repeat A region at the 5' end of Xist is one of the most clearly conserved regions of the RNA (5–7). Repeat A is required for stable accumulation of spliced Xist in cells and for gene silencing (3, 4). In the mouse, repeat A includes seven and one-half copies of a 24-nt repeat unit separated by U-rich spacers of variable lengths. Prior models of this region have emphasized self-contained structures consisting of either small intrarepeat stem-loops (30), large interrepeat structures (12), or a combination of both (10). In contrast, SHAPE data obtained in the context of full-length native Xist indicate that the repeat A region has high Shannon entropy and likely exhibits significant structural variability (Fig. 2A). A single hairpin with a GC-rich stem and AU-rich loop that bridges repeats three and four is the only well-defined element in repeat A in our model (Fig. 2A and B); these nucleotides exhibit high sequence conservation (Fig. 2C). Repeat A nucleotides also likely interact with adjacent segments of Xist in the full-length RNA (Fig. 2D), and the base of the repeat A stem loop may form a pseudoknot (Fig. 2B). Elements of prior repeat A models (10, 12, 30) occur among the structures generated by our ensemble analysis (SI Appendix, Supporting Text); however, high Shannon entropies support the model that this region is structurally dynamic, a feature that may facilitate accessible interaction with protein cofactors.

High probability pairing regions are predicted to exist in well-defined motifs just upstream (nucleotides 49–352) and downstream (nucleotides ~850–1,300) of repeat A. These regions have not been genetically disrupted in isolation of repeat A and their role in Xist function is not known. However, these regions bracket the essential repeat A element in Xist and may cooperate with the repeat to encode function in the 5' end of the IncRNA.

Repeat E, which has no known function, also forms a dynamic and flexible structure. This region spans roughly 1 kb at the beginning of exon 7 and consists of U-rich repeats of 20–25 nt (5). Repeat E exhibits low Shannon entropy and high SHAPE reactivity, indicating that this region is unstructured (Fig. 2D). Nucleotides in repeat E are accessible for unencumbered interaction with RNA binding proteins and we will show below that proteins extensively target this element.

Our model also provides structural context for previously characterized Xist mutant phenotypes. For example, a 16-nt insertion located 3' of repeat A causes a hypomorphic phenotype (32). The insertion falls in the middle of a well-defined hairpin structure with low Shannon entropy (Fig. 1D, filled arrowhead). The insertion likely leads to a realignment of the local structure that affects the biological activity of the repeat A region or attenuates a function of the hairpin itself. A 4-kb inversion of nucleotides 5,984–9,954 leads to a similar hypomorphic phenotype with incomplete silencing (33). This inversion overlaps 14 structural elements in the Xist RNA model (SI Appendix, Fig. S5).

**Fig. 2.** Ex vivo structural features of repeat regions A and E. (A) SHAPE reactivity (black), Shannon entropy (brown), and pairing probabilities (bottom) for repeat A and the surrounding region. Arcs are color coded as in Fig. 1. A single high-probability stem-loop structure is predicted to occur within repeat A (gray shading). Xist regions outside of repeat A may interact with this region. (B) Secondary structure of the repeat A stem-loop and predicted pseudoknot. (C) Comparative sequence alignment of the repeat A stem-loop element. Base pairs in mouse are indicated in dot-bracket notation. Nucleotides are color-coded to indicate conservation of base pairing. (D) SHAPE reactivity (black) and Shannon entropy (brown) for the repeat E region. SHAPE reactivities are high and Shannon entropies are low in this region, indicating a high probability of lack of defined structure as illustrated by base pairing probability arcs (bottom).
Localized Cellular Effects on Xist Structure. Each individual reactivity measurement in a SHAPE-MaP experiment includes an error estimate (15), thus allowing for statistically rigorous analysis of local structural changes. We hypothesized that sequences critical to Xist-protein interactions may be overrepresented among +ΔSHAPE sites (in which reactivity is lower in cells than ex vivo). We searched these sites for sequence motifs and identified two U-rich sequence motifs, E1 and E2 (SI Appendix, Fig. S2), located in repeat E. No other significant sequence motifs spanning ΔSHAPE sites were identified.

To identify sites in Xist where specific protein interactions occur, we searched for proteins both previously identified as Xist partners in TSCs (34) and present in the CLIPdb protein cross-linking and immunoprecipitation database (35) and identified CELF1, PTBP1, TARDBP, FUS, and RBFOX2 (34, 36, 37). We also performed digestion-optimized RIP-seq experiments in TSCs to identify binding sites for HuR, another Xist-interacting protein (34). We expected to find that 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 in each replicate. We found that, on average, 76% of ΔSHAPE sites overlapped with CLIP or RIP sites, whereas only 53% of the total reported CLIP sites coincided with ΔSHAPE signals (Fig. 4D and SI Appendix, Fig. S3D). This latter low number likely reflects differences between cell types, the high stringency used in the ΔSHAPE analysis (17), and the high background of CLIP experiments (38).

Given the low false-positive detection rate of protein binding when considering only +ΔSHAPE sites (17), we focused on CLIP sites corroborated by +ΔSHAPE values. We identified sites likely bound by CELF1, PTBP1, and HuR in repeat E, showed that sites for FUS are concentrated in the well-folded RNA domains spanning positions 13,900–15,000, and defined a single site strongly bound by FUS (SI Appendix, Fig. S3E). Owing to the relatively small number of proteins in our analysis, the data indicate that the 3’ end of Xist is extensively involved in in-cell interactions. This analysis also confirms that repeat E is a major protein-binding platform (Fig. 4B).

ΔSHAPE-confirmed CELF1 and PTBP1 CLIP sites are located almost exclusively in repeat E (Fig. 4B). These proteins function in RNA processing (39, 40) and may regulate Xist splicing or editing. We used sequence clustering to define consensus motifs from +ΔSHAPE-supported CLIP sites for CELF1 and PTBP1 and found that both overlap with motif E1 (Fig. 4C and SI Appendix, Fig. S3F). No strong consensus sequence was identified among non-ΔSHAPE-validated CLIP sites, although many fall within repeat E. Thus, CELF1 and PTBP1 likely interact with repeat E in a sequence-specific manner.

We identified HuR-binding sites throughout repeat E (Fig. 4B and SI Appendix, Fig. S3E). HuR promotes mRNA stability through interactions with AU-rich elements (AREs) (41). Consistent with an affinity for ARE motifs, HuR was widely detected throughout the U-rich repeat E (SI Appendix, Fig. S7A). Searching over subsequences corresponding to +ΔSHAPE in-cell protections returned a U-rich consensus containing elements from motifs E1 and E2 (SI Appendix, Fig. S7B). Repeat E may be particularly susceptible to ARE-mediated degradation, and coating this region with proteins, especially HuR, may inhibit RNA decay.

FUS is an abundant, nuclear-enriched protein involved in the regulation of transcription, RNA processing, and DNA damage repair. FUS binds to many RNAs, and its binding has been characterized as promiscuous (42). In contrast to this view, 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 in each replicate (Fig. 4B and SI Appendix, Fig. S3E). This region has a well-defined RNA structure (Fig. 1 and SI Appendix, Fig. S1) and a mixture of positive and negative ΔSHAPE sites (Fig. 3 C and D and SI Appendix, Fig. S5C). We analyzed the pairing probabilities over FUS-associated +ΔSHAPE sites within this region and identified a structural context for FUS binding: FUS-protected nucleotides occur in single-stranded motifs flanked by base-paired structures (Fig. 4D and SI Appendix, Figs. S6G and SI Appendix, Fig. S5C). An observation consistent with the complex structural rearrangements detected within the FUS interaction domain when comparing ex vivo and in-cell data. A local increase in FUS concentration via multimerization may lead to cooperative binding, which protects some regions from in-cell SHAPE modification while making others more accessible.
ΔSHAPE analyses support a single major CLIP-identified binding site for the TARDBP protein (Fig. 4E and SI Appendix, Fig. S3H). TARDBP is an RNA- and DNA-binding protein with a reported preference for UG-rich sequences; it is both a transcription repressor and a splicing regulator (44). The single TARDBP-binding site in Xist detected by our analyses of both replicate experiments is part of a UG-rich structural motif (positions 10,203–10,309) encompassing the splice junction between exons 6 and 7 (Fig. 4E). A threefold reduction in Xist transcript levels has been reported in adult mouse brains depleted of TARDBP via antisense knockdown (45); our analysis of these data further show that levels of incorrectly spliced Xist transcripts increased by twofold (SI Appendix, Fig. S9 A and B), suggesting that TARDBP controls the amount of Xist present in a cell. Although in principle many of the reported CLIP sites for TARDBP are detectable by ΔSHAPE (SI Appendix, Fig. S9), only a single site overlapped with a strong ΔSHAPE signal. The median SHAPE reactivity of this site was much higher than that of any other reported TARDBP CLIP site. These data suggest that 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. Most broadly, this analysis indicates that Xist RNA structure can specify a unique accessible protein-binding site.

It is intriguing that the 5' end of Xist lacks ΔSHAPE sites. The regions near and including repeat A are important for Xist silencing activity (30, 34). We hypothesize that RNA–protein interactions may be less stable here than in other regions, and reanalyzed the ΔSHAPE data with reduced stringency in an attempt to identify potential weaker sites. With these criteria, we identified only four to six additional interaction sites in the first 1,000 nucleotides of Xist (SI Appendix, Fig. S10), suggesting that proteins interact transiently with a dynamic 5' end or bind to double-stranded elements in such a way as to not exhibit SHAPE reactivity changes.

Conclusion

Comprehensive and quantitative nucleotide-resolution SHAPE-MaP structure probing revealed that Xist consists of multiple domains of well-defined secondary structure linked by structurally variable and dynamic regions (Fig. 1 and SI Appendix, Fig. S5), and supports existing domain-based models for IncRNA function (10, 46, 47). Fully one-half of the Xist IncRNA forms well-defined structure motifs, is significantly impacted by the cellular environment, or both. Structured elements at the 3' end of Xist appear to function in part by increasing the cellular stability of the transcript. Repeat-containing regions are generally unstructured and are extensively bound by protein cofactors (Figs. 1–4).

We identified three distinct structure-based mechanisms by which TARDBP mediates stable interactions with Xist. In each case, protein interactions corroborated by CLIP-seq or RIP-seq and ΔSHAPE data are focused within specific structural elements, and ΔSHAPE signals reveal specific details of these Xist–protein interactions. CELF1, PTBP1, and HuR exemplify widespread binding, likely with a degree of sequence specificity, to accessible, unstructured regions. FUS binding occurs in a region with a well-defined structure ex vivo that undergoes extensive rearrangement in cells. TARDBP appears to bind predominantly to a single site presented within a small structural domain. These findings highlight the impressive diversity of IncRNA–protein interactions and their distinct RNA structure-dependent interaction modes.

Cross-referencing of +ΔSHAPE sites with CLIP- and RIP-identified binding sites suggests that quantitative ΔSHAPE analysis is a rigorous approach for identifying stable RNA–protein interaction sites (Fig. 4 and SI Appendix, Fig. S3). Whereas CLIP studies often report binding across the entire transcript, our ΔSHAPE analysis revealed that stable binding sites tend to cluster within the RNA, as was observed for CELF1, PTBP1, and HuR within repeat E and for FUS within the FUS domain. In addition, only when our analysis was limited to +ΔSHAPE sites was a binding motif identified for HuR. ΔSHAPE can also detect site-specific interactions, as were observed for TARDBP, ΔSHAPE analyses of two independent replicates revealed similar overall patterns of protein interaction (SI Appendix, Fig. S3), despite relatively modest correlations for the in-cell experiments (SI Appendix, Fig. S1).

Protein-binding events may simply vary between individual Xist ribonucleoprotein complexes, perhaps due to limited access to the IncRNA, low stability of subsets of IncRNA–protein interactions, or limited availability of protein-binding partners (SI Appendix, Supporting Text). ΔSHAPE analysis clearly enables characterization of RNA–protein interactions and examination of RNA structure-mediated recognition in a way that will be broadly useful in future studies of Xist and other IncRNAs as additional protein partners are identified.

This work embraces numerous innovations in quantitative RNA structure probing to define RNA structure, RNA–protein interactions, and the effects of the cellular environment on RNA.
architecture. Our approach deemphasizes the global minimum free energy structure in regions where multiple structures are likely to be sampled simultaneously and uses experimentally derived metrics to define structural domains. For individual regions with a high propensity to form well-determined stable structures, we modeled Xist structures using the validated three-reagent differential SHAPE approach (15, 25). Differences in-cell and ex vivo states were interpreted in the context of robust analysis of measurement errors (17) (SI Appendix, Supporting Text). Limitations are that RNA interactions were constrained to 600-nat windows and canonical base pairing, and there are uncertainties in the thermodynamic parameters used in modeling. Nevertheless, in-cell SHAPE-MaP represents a major advance in converting RNA structure probing from a qualitative tool to a quantitative and predictive tool for understanding RNA biology.

The structured and unstructured domains identified here define maps that are expected to be useful in guiding investigations into the mechanisms by which Xist elements contribute to X chromosome inactivation. Xist and other lncRNA transcripts may span kilobases to coordinate long-range protein domain interactions (Figs. 3 and 4) that ultimately enable orchestration of epigenetic regulation on the kilobase to megabase scales (1, 3, 4). Many lncRNAs are likely to shape features identified here for Xist, including densely arrayed secondary structural features, multiple distinctive modes of protein interaction, and the ability to serve as multidomain organizers of cellular function.

Methods

In-cell modification was carried out by treating mouse TSCs in fresh growth medium with 1M7, 1M6, or NMIA. RNA was subjected to MaP reverse transcription (15) using Xist-specific primers, followed by Xist-specific PCR amplification and high-throughput sequencing library construction. SHAPE reactivities were calculated from raw sequencing reads using ShapeMapper, and secondary structures were modeled using SuperFold (15). Detailed descriptions of in-cell RNA probing, library construction, structure modeling, and bioinformatics analyses are provided in SI Appendix, Methods.

ACKNOWLEDGMENTS

We thank Dirk Schübeler and Oliver Bell for generously sharing the HyTK embryonic stem cell line and Kathrin Plath for sharing the pSM33 line. This work was supported by National Institutes of Health (NIH) Grant GM064803 and National Science Foundation (NSF) Grant MCB-1121024 (to K.M.W.), NSF Grant MCB-0842621 and NIH Grant CA157268 (to J.M.C.), and laboratory start-up funding provided by the Lineberger Comprehensive Cancer Center (to J.M.C.). M.J.S. is an NSF Graduate Research Fellow (Grant DGE-1144081) and was supported in part by an NIH training grant in molecular and cellular biophysics (Grant T32 GM05870). T.W.C. was supported in part by an NIH training grant in bioinformatics and computational biology (Grant T32 GM07553). D.M.L. was supported in part by an NIH training grant in genetics and molecular biology (Grant T32 GM07092).

Supporting Information for:

SHAPE reveals transcript-wide interactions, complex structural domains, and principles of protein interaction across the Xist IncRNA in living cells

Matthew J. Smola¹, Thomas W. Christy¹, Kaoru Inoue², Cindo Nicholson³, Matt Friedersdorf³, Jack Keene³, David M. Lee², J. Mauro Calabrese²*, and Kevin M. Weeks¹*

¹Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599-3290
²Department of Pharmacology and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC
³Department of Molecular Genetics and Microbiology, Duke University, Durham, NC

*correspondence: weeks@unc.edu and jmcalabr@med.unc.edu
Supporting Text

Comparison with prior in-cell DMS probing of the Xist RNA

We compared the structures identified in our analysis with those proposed previously based on in-cell DMS probing (1). There are both local congruencies between the two approaches and also regions in which modeled structures differ substantially. Roughly three-fourths of the structural models proposed here are unique to this work (Fig. S5). Of the 16 regions for which models have been proposed based on both SHAPE and DMS modeling, seven show general agreement (Fig. S5, blue labels). Several factors likely contribute to the significant differences in models. DMS reactivity reports on two of the four RNA nucleotides (A and C) and development of well-validated methods for using DMS reactivity information to model large RNA structures remains ongoing. The SHAPE- and DMS-based works also differed in their approach for identifying regions with well-folded structures. The DMS-based work focused on folding of individual computationally pre-determined regions and emphasized minimal free energy (MFE) models (1). In contrast, the SHAPE-MaP approach folded all regions in the context of full-length Xist RNA and then emphasized regions with high levels of well-defined structure (low SHAPE and low Shannon entropy). In the context of full-length Xist, many regions appear to sample multiple conformations, which excluded individual domains proposed in the DMS-based work. There are likely also differences arising from effects of the cellular environment.

Comparison with prior models of the Repeat A region

Our model of repeat A structure exhibits both similarities and differences with previously proposed models. Wutz et al. proposed that this region forms simple short intra-repeat hairpin structures and identified loss-of-function mutations with the potential to disrupt hairpin formation (2). We examined these mutation sites in our model of repeat A and found that, whereas some intra-repeat hairpins do exist in our ensemble of possible structures, the repeat monomers are often involved in longer-range interactions that would also be impacted by mutagenesis. Thus, the observed phenotypes are consistent with both models. Elements of other repeat A models (1, 3) can be found among the possible structures generated by our approach, including the repeat A stem-loop, and similar levels of sequence conservation exist for portions of each model. In general, the Shannon entropy across repeat A is high, suggesting that a
minimum free energy structure may not be the best way to characterize this region and that, instead, multiple local and longer-range structures contribute to the overall ensemble in the repeat A region.

**Comparison of replicate ΔSHAPE data**

We note that, although prior studies have performed replicate experiments, replicate RNA chemical probing data in most studies have generally been combined for the final analysis, diminishing the value of the replicate information. In this work, we analyze each biological replicate fully independently. Overall, this analysis emphasizes that the transcript-wide and local interactions observed for Xist using SHAPE-MaP show excellent agreement. We performed full biological replicates of in-cell and *ex vivo* 1M7 structure probing. *Ex vivo* replicates show good agreement (Spearman R = 0.65; [Fig. S1](#fig-s1)), while in-cell replicates exhibit a more modest correlation (Spearman R = 0.50). This lower correlation may reflect both experimental variability and variability intrinsic to the cellular environment. We applied ΔSHAPE analysis to each replicate and identified 177 and 239 ΔSHAPE sites, respectively. The distributions of these sites along Xist are very similar, with few sites within the first 2.5 kb, many +ΔSHAPE sites within repeat E, and a mixture of positive and negative ΔSHAPE sites in the region surrounding the FUS domain ([Fig. S3C](#fig-s3c)). Analysis of sequences near +ΔSHAPE sites from each replicate identified similar consensus motifs E1 and E2 ([Fig. S3F](#fig-s3f)).

When we analyzed ΔSHAPE and CLIP sites, each replicate exhibited similar overlap, such that 79% and 73% of all ΔSHAPE sites overlapped CLIP sites ([Fig. S3D](#fig-s3d)). Interestingly, there are only 43 ΔSHAPE sites in common between to the two replicates (26 +ΔSHAPE, 17 –ΔSHAPE), overlapping 22 CLIP sites (51%). However, the global locations of ΔSHAPE-confirmed CLIP sites exhibit similar patterns between replicates ([Fig. S3C](#fig-s3c)). CELF1, PTBP1, and HuR bind extensively to repeat E; FUS interactions occur predominantly within the FUS domain; and TARDBP binds Xist near the exon 6-7 splice junction in both replicates.

The specific modes of interaction between Xist and the proteins examined here are also corroborated between replicates. Similar sequence motifs were identified for CELF1 and PTBP1, supporting sequencing-specific interactions with repeat E ([Fig. S3F](#fig-s3f)), while HuR interacts with
U-rich sequences in both replicates (Fig. S7B). The structural preference of Xist-FUS interactions for single-stranded regions flanked by base paired regions is observed in both replicates (Fig. S3G). The same Xist–TARDBP interaction near the exon 6-7 splice site is identified in both replicates (Fig. S3H). Thus, despite a relatively modest correlation between in-cell replicates, the overarching similarities in follow-up analyses suggest that in-cell and ex vivo SHAPE-MaP, in conjunction with the ΔSHAPE analysis framework, successfully identifies shared features of these datasets.
Methods

In-cell and ex vivo RNA structure probing

Mouse trophoblast stem cells (TSCs) were grown, and RNA was extracted and modified, both in cell and ex vivo, as described previously (4). Briefly, for in-cell SHAPE probing, live TSCs were washed once with phosphate buffered saline, 900 µl of fresh growth media were added, 100 µl of 100 mM 1M7 in neat DMSO were added (10 mM final concentration), and cells were 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). For ex vivo probing, RNA was obtained from cells using the non-denaturing approach described in (4). Briefly, 6 × 10^6 TSCs were washed with ice-cold PBS and resuspended in 6 mL ice-cold lysis buffer [40 mM Tris (pH 7.9), 25 mM NaCl, 6 mM MgCl₂, 1 mM CaCl₂, 256 mM sucrose, 0.5% Triton X-100, 1000 U/mL RNasin (Promega), and 450 U/mL DNase I (Roche)], and rotated for 5 minutes at 4 °C. Cells were pelleted at 4 °C for 2 minutes, resuspended in 6 mL digestion buffer [40 mM Tris (pH 7.9), 200 mM NaCl, 1.5% SDS, and 500 µg/mL Proteinase K], and rotated at 20 °C for one hour. RNA was extracted twice against an equal volume ice-cold phenol:chloroform:isoamyl alcohol (24:24:1) pre-equilibrated with 1× Folding Buffer (100 mM HEPES, pH 8.0, 100 mM NaCl, 10 mM MgCl₂), followed by two extractions with an equal volume of 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 were then added to a one-ninth volume of 1M7, 1M6, or NMIA, each at 100 mM in neat DMSO (10 mM final concentration), and incubated at 37 °C for 5 minutes. For both in-cell and ex vivo probing, background was assessed by performing no-reagent and denaturing controls (4).

Xist SHAPE-MaP

Modified RNA was subjected to mutational profiling (MaP) reverse transcription (5, 6). The resulting cDNAs were purified (Agencourt RNAClean XP beads, Beckman Coulter) and amplified by PCR (Q5 high-fidelity DNA polymerase, NEB) with Xist-specific primers (2 pmol each; primers are listed below). 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 before construction of high-throughput sequencing libraries (Nextera XT, Illumina) (6).

Primer sequences used to create amplicons that 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.

<table>
<thead>
<tr>
<th>Amplicon name</th>
<th>Forward primer</th>
<th>Reverse primer (RT primer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amp_22</td>
<td>TCCATCTAAGGAGCCTTGGGG</td>
<td>ATAGGTCACCTCACACAGCA</td>
</tr>
<tr>
<td>Amp_21</td>
<td>GCTTGGTGTGATGGAATATGG</td>
<td>CGTTATACCGCACAAGAAC</td>
</tr>
<tr>
<td>Amp_20</td>
<td>AGCGGACTGAGATAAAGCAAC</td>
<td>CATCACAGTCTAATTCACCTCG</td>
</tr>
<tr>
<td>Amp_19</td>
<td>TGTGGTGTGATGTCATCTCC</td>
<td>AAACTTAAAGGACTCAGGAAC</td>
</tr>
<tr>
<td>Amp_18</td>
<td>CGTCTGATAGTGCTTTCG</td>
<td>GGCTTGGAGTAGTGTGAAA</td>
</tr>
<tr>
<td>Amp_RepC</td>
<td>CCAGGCCCTAGAATTCCT</td>
<td>TGTGGCCTTTTCCTTAATAT</td>
</tr>
<tr>
<td>Amp_16</td>
<td>CCCATCTATACCCCTCCTAT</td>
<td>GCAAGGGTAGTATAGGACCTCAG</td>
</tr>
<tr>
<td>Amp_15</td>
<td>TCACATCTTTCTATTCACCC</td>
<td>AGTTAACACTGTGCACATTAC</td>
</tr>
<tr>
<td>Amp_14</td>
<td>GGTTCCTACCACATAGGCCC</td>
<td>AAAACCCCTCTCCTATGCAA</td>
</tr>
<tr>
<td>Amp_13</td>
<td>AAAAACTTTTGGCAGTCAGC</td>
<td>GCCCTGGTTCATAAAGGATAC</td>
</tr>
<tr>
<td>Amp_12</td>
<td>AGCGGAGGAGGGTTGTGCAG</td>
<td>TGAATGGAATGGAAAGGCCA</td>
</tr>
<tr>
<td>Amp_11</td>
<td>TCCATGGACCTTTTCTGAATCAC</td>
<td>AAGATACCTGTCTTAAACATTCCTG</td>
</tr>
<tr>
<td>Amp_10</td>
<td>TACTGAGGGTGATAGTCTGT</td>
<td>TCAGCAATGTCATATCAAAAC</td>
</tr>
<tr>
<td>Amp_9</td>
<td>CTCAAGCAACATGGAGGACT</td>
<td>GCATCTGATAGCTTCTTCT</td>
</tr>
<tr>
<td>Amp_8</td>
<td>AAAAAAGGTCTACAGGCAACA</td>
<td>AATAGACACAAAGAGCAAG</td>
</tr>
<tr>
<td>Amp_7</td>
<td>TGATGGTGTGATATGTCGCTG</td>
<td>ACACATGACAGAAAGACAG</td>
</tr>
<tr>
<td>Amp_6</td>
<td>GTGCTCTCCCTTCTTTACATTG</td>
<td>TCTGGACCTTATGGAAG</td>
</tr>
<tr>
<td>Amp_5</td>
<td>TTGTGTCTCTTTCTGTGACG</td>
<td>TTGCTTTATGGAATGGCAAT</td>
</tr>
<tr>
<td>Amp_4</td>
<td>CCCAGCATCCCCTTCCCTTATT</td>
<td>AATGGCAATGTCATATAGG</td>
</tr>
<tr>
<td>Amp_3</td>
<td>AGAGCCTACTAAACGGGCTTA</td>
<td>AGGGAATCATCAGCCG</td>
</tr>
<tr>
<td>Amp_2</td>
<td>TAAGGCGAGTGTTTCTTTC</td>
<td>ACACATGGAAGACATCTTAC</td>
</tr>
<tr>
<td>Amp_1</td>
<td>AACCTGGGTTGTAGTCAGAC</td>
<td>AAGTAGTCTACATAATGCTAG</td>
</tr>
<tr>
<td>Amp_0</td>
<td>TGGCTGAACTTACATATAATGTC</td>
<td>CTTGGCAACGTCCTTTC</td>
</tr>
<tr>
<td>Amp_-1</td>
<td>GAGCATGTTCTACAAGGCC</td>
<td>TGTGGAACGGAAG</td>
</tr>
<tr>
<td>Amp_-2</td>
<td>TGATGAAATGGAGTTCTAAGTA</td>
<td>TGGGCAAGGATAGTGC</td>
</tr>
<tr>
<td>Amp_-3</td>
<td>TAGGCGATTCCTAGTACTGT</td>
<td>TTTGAACTCCAGACTTAC</td>
</tr>
<tr>
<td>Amp_-4</td>
<td>AAGTGCCCTTACAGGCTGAGT</td>
<td>TGTCATGTTGGTAAGCA</td>
</tr>
</tbody>
</table>

**Sequencing and SHAPE profile generation**

Purified sequencing libraries were pooled and sequenced on an Illumina MiSeq or HiSeq.
instrument, generating $2 \times 150$ or $2 \times 100$ paired-end datasets. SHAPE reactivity profiles were created by aligning reads to the $Xist$ reference sequence (GenBank accession NR_001463.3) using ShapeMapper (v1.2, http://chem.unc.edu/rna/software.html). Default parameters were used with the following exceptions: windowSize = 3, maxInsertSize = 800, randomlyPrimed = off, minMapQual = 14 [for nucleotides spanning repeat E (10,251-11,017) only]. The folding module of ShapeMapper was not used. Final reactivity profiles were generated by excluding nucleotides 1-78, 2,451-2,599, and 17,801-17,918 and renormalizing the remaining nucleotides using the boxplot approach (7). SHAPE reactivity values for nucleotides 1-10,250 and 11,018-12,729 were, together, scaled to match values from nucleotides 10,251-11,017 and 12,730-17,918 because the two groups were amplified from cDNAs generated in separate reverse transcription reactions. 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 (1, 8) were computed using a Z-factor test (1, 5).

We confirmed that the enrichment by amplification approach faithfully reports RNA secondary structure by examining the 3’ domain of the mouse 18S rRNA (Fig. S1). SHAPE-MaP reactivity values obtained from randomly-primed MaP reverse transcription (2, 5, 6) show excellent agreement (Spearman R = 0.88) with those obtained using targeted primers (F: 5’-GAGGT GAAAT TCTTG GACCG-3’; R: 5’-ACCAT CCAAT CGGTA GTAGC-3’).

**Structure modeling**

SHAPE-MaP provides a biophysically rigorous measurement of local nucleotide flexibility that is independent of base identity (1, 3, 5, 9, 10) and greatly improves secondary structure predictions (4, 8, 9). Potential pseudoknots in $Xist$ were identified using a sliding window approach (4, 5) in which full-length $Xist$ was folded in 600-nt windows offset by 100-nt increments using ShapeKnots (5-7). 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 for both strands of the potential helices. The model of ex vivo $Xist$ secondary structure was created by providing SuperFold (5, 6) with 1M7 reactivities in addition to differential SHAPE values and pseudoknotted helices (using the --differentialFile and --
pkRegion options). 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 interactions.

**Choice of Shannon entropy cutoff**
The Shannon entropy cutoff used in this work to identify well-defined structures has generally performed well on other large RNAs (5, 7, 11); however, it was unclear whether these same cutoffs would be appropriate for lncRNAs such as Xist. Thus, we examined the effect of raising and lowering the Shannon entropy cutoff by a factor of two (twice the global median or half the global median) to assess how the choice of cutoff would affect identification of well-defined regions. Unsurprisingly, we identified greater and fewer regions relative to our standard analysis when using higher and lower cutoffs, respectively. However, many of the new or expanded regions found when using the higher cutoff overlapped with portions of the RNA that appear capable of adopting a variety of structures (corresponding to many overlapping pairing probability arcs in Fig. 1C), suggesting that this cutoff was too relaxed. Conversely, using a lower cutoff resulted in many regions with high-probability, non-overlapping secondary structures being excluded from the analysis, suggesting the cutoff was too stringent. In general, a Shannon entropy cutoff equal to the global median appears to be an appropriate value for analyzing Xist structural models, although cutoffs used here in analyzing Xist may merit reevaluation for other lncRNAs.

**SNV 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 SNV locations to their equivalent position on the Xist transcript; a multi-sequence alignment was performed to ensure agreement of SNV locations. We then examined whether SNV locations corresponded to base-paired or single-stranded conformations in our structure models. A bootstrapping approach was used to calculate p-values indicating the likelihood of encountering fewer structure-disrupting SNVs 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 Gaussian cumulative distribution function from which we calculated the probability of finding fewer disruptive SNVs than observed 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 (12) and aligned with Clustal Omega (13). The repeat A stem-loop motif was then extracted from the alignment and analyzed for sequence conservation and covariation using R2R (14).

**Xist truncation mutant and half-life measurements**
Full-length *Xist* cDNA and its 14.8-kb counterpart from pCMV-*Xist*-PA (15) (Addgene plasmid 26760) were cloned via Gibson assembly (16) behind a doxycycline-inducible promoter in a plasmid containing inverted loxP sites (17) and a constitutively expressed G418 resistance gene. *Xist* plasmids were electroporated along with pIC-Cre (Addgene plasmid 19131) into Tc1 mouse embryonic stem cells (17), which contain inverted loxP sites and HSV thymidine kinase (*Tk*) inserted into the *Hbb-y* gene on chromosome 7. The Tc-1 embryonic stem cells were further engineered in this work to express the reverse-tetracycline transactivator (rtTA) (18). After electroporation, cells were selected in 200 µg/ml of G418 for 8 days, and in 3 µM ganciclovir and 200 µg/ml of G418 for an additional 4 days, to obtain cells that had both lost the HSV *Tk* gene and gained the *Xist* transgene. Successfully targeted colonies were confirmed by PCR, FISH, and RNA-seq of poly (A)-selected RNA (Fig. S6).

For half-life experiments, full-length and 14.8 kb *Xist* clonal cell lines were induced for 2 days with doxycycline prior to media wash out and replenishment in media containing actinomycin D (Sigma). pSM33 embryonic stem cells, which inducibly express Xist from its endogenous location on the X (19), were induced for 12 hours prior to actinomycin D treatment (to 5 µg/ml actinomycin D final). At specified time points post-actinomycin D treatment, RNA was collected, subjected to reverse transcription with random hexamers, and examined with qPCR using the following primer pairs: *Xist_p1* (F: 5’-ATGCA GGGCA TAGTG GTAGG-3’; R: 5’-
ACAGC TGCAT TTGTC CCTCT-3’) and Gapdh (F: 5’-TGTTC CTACC CCCAA TGTGT-3’; R: 5’-TGTGA GGGAG ATGCT CAGTG-3’). Xist expression levels were calculated relative to Gapdh from biological triplicate measurements for all embryonic stem cell lines and from biological duplicate measurements for TSCs. Error bars in Fig. S6C represent the standard error of the mean calculated from six total measurements at each data point (technical duplicates from biological triplicate preparations of embryonic stem cell RNA, and technical triplicate from biological duplicate preparations of TSC RNA).

To model half-life, RNA levels at each time point (measured by qRT-PCR) were represented as a percentage of RNA at 0 hr and transformed by their natural log. Linear models (time as a predictor of RNA percentage) were fit to the data for each biological replicate and the equations were used to find the time at which the percent RNA remaining relative to the 0 hr time point was 50% (half-life). Half-life comparisons between cell lines were made using an analysis of variance, followed by post-hoc pairwise T tests (pooled SD and Holm p-value adjustment) to compare each line directly to each other line.

**Computing regions of large absolute SHAPE-MaP 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 spans of at least 100 consecutive nucleotides in which absolute differences were greater than the global median. Use of these parameters emphasized that large regions of Xist that are impacted by the cellular environment. These parameters are sensible given that the structure of only a single lncRNA has been comprehensively mapped in cells. As more in-cell SHAPE data are obtained for additional lncRNAs, it is likely that more nuanced approached for identifying large-scale in-cell effects can be developed.

**Identifying protein binding sites and conformational changes with ΔSHAPE**
ΔSHAPE values were calculated by comparing the *ex vivo* and in-cell conditions, exactly as described (4) with the minor modification that differences greater than 20 [due to hyper-reactive nucleotides (10)] 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 overlap with 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. This implementation of ΔSHAPE analysis emphasizes strong changes in reactivity between the two conditions being compared, resulting in highly stringent identification of interaction sites. A lack of ΔSHAPE sites, therefore, does not necessarily exclude the possibility of protein interactions. For example, ΔSHAPE may fail to identify binding sites if interactions primarily involve double-stranded RNA, if interactions are transient or highly dynamic, or if proteins recognize an RNA structure that coexists in conjunction with other structures that reduce the impact of binding upon the overall SHAPE reactivity measurement.

Sites of CELF1, PTBP1, FUS, TARDBP, and RBFOX2 were downloaded from CLIPdb (20). 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. 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 as described (21) and RNA immunoprecipitation (RIP) of HuR-RNA complexes was performed similarly, except that micrococcal nuclease was used to partially digest RNA in the lysates before the RIP steps. RNA fragments from the HuR RIP and total input RNA (treated with micrococcal nuclease) were converted into cDNA libraries (NEBNext Small RNA Library Prep; NEB), and libraries were sequenced (Illumina Hi-Seq 2500). Libraries were constructed with custom adapters based on small RNA adapters but with the addition of six random nucleotides to the 5′ adapter. Raw reads from the digestion-optimized RIP libraries were preprocessed to remove adapter sequences up to the 6-nt random sequences (Cutadapt). PCR artifacts were removed by collapsing the reads, including the random portion of the adapter, to unique sequences. The random regions were then removed and the reads mapped to the mouse genome (mm9) using TopHat2 (22) using the following options: -N 1 --read-gap-length 1 -g10 --library-type fr-secondstrand. Uniquely mapped RIP reads were normalized to the
input reads across the genome in 5-nt intervals, and enrichment was determined using a mixture model approach (23). The posterior probabilities generated from the mixture model were used to calculate a log of odds ratio. HuR binding sites were then 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 +ΔSHAPE sites were extracted and expanded to 20 nt. These sequences were then searched for sequence motifs with MEME (24), allowing for up to two sites per motif, a minimum motif length of 4 nt, 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 two significant sequence motifs, E1 and E2, for each replicate. Expectation values for E1 were $1.4 \times 10^{-36}$ and $3.5 \times 10^{-14}$ for replicates 1 and 2, respectively, while E2 was reported with expectation values of $8.7 \times 10^{-29}$ and $3.3 \times 10^{-18}$ respectively. Sequence motifs within ΔSHAPE-confirmed CLIP sites were analyzed in the same way. This analysis identified motifs represented in the CELF1 and PTBP1 binding sites with expectation values of $9.3 \times 10^{-40}$ and $4.5 \times 10^{-17}$, respectively (replicate 1), or $9.9 \times 10^{-10}$ and $1.3 \times 10^{-15}$, respectively (replicate 2). Analysis of the 20 nt surrounding +ΔSHAPE sites within HuR RIP sites identified a U-rich consensus sequence with expectation values of $2.2 \times 10^{-30}$ and $1.2 \times 10^{-23}$ for each replicate, respectively. Similar analyses of the FUS and TARDBP ΔSHAPE sites revealed no significant consensus sequences.

**Clustering FUS-localized +ΔSHAPE sites by pairing probability**

ΔSHAPE-supported FUS binding sites did not exhibit strong sequence preferences (see above), so we investigated whether there was a structure-based preference for Xist-FUS interaction. Total pairing probabilities for each nucleotide in expanded +ΔSHAPE sites (see above) were extracted from the partition function dot plot (generated by SuperFold) 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 (25). Clusters were visualized using TreeView (26).

**Evaluation of data from TARDBP-deficient cells**
TARDBP protein levels were depleted by treatment of cells with antisense DNA oligonucleotides (27). RNA-seq reads for control and TARDBP-deficient samples (4 replicates each) were downloaded from the Gene Expression Omnibus (accession GSE27394). Adapter sequences were removed and filtered for base call quality before alignment to the mouse genome (mm10) with Bowtie2 (28). The number of reads overlapping introns was then computed and compared to exon-aligned reads.

Data availability
All processed SHAPE-MaP data and RNA structure models are freely available in the Supporting Dataset S1, and at the corresponding author's website. Raw sequencing data are available from the Sequence Read Archive, accession SRP074108.
References


**Supplemental Figure S1.** Correlation analysis and validation of Xist SHAPE-MaP probing. (A) Ex vivo SHAPE-MaP reactivities from two full biological replicates show good agreement (Spearman R = 0.65) over 15,210 nucleotides. (B) In-cell SHAPE-MaP reactivities exhibit a modest correlation (Spearman R = 0.50; 14,526 nt), possibly due to variations in cellular interactions encountered by Xist. (C) SHAPE-MaP reactivity correlations calculated for 100-nt sliding windows show that local correlations between replicates vary across the length of Xist both in ex vivo (top) and in-cell (bottom) experiments. Dashed lines represent the global median. Ex vivo reactivities show good agreement over hundreds of nucleotides, while in-cell reactivities exhibit regions of stronger and weaker correlation. In-cell correlations are weakest in the 3’ end of Xist (13,000-17,918), consistent with the possibility of non-saturating and experiment-specific interactions in this region. (D) Distributions of ex vivo (blue) and in-cell (red) windowed correlation values from (C). (E) Correlation between SHAPE reactivities obtained from mouse 18S rRNA using traditional random priming versus target-specific enrichment by amplification. Reactivities across 738 nucleotides show excellent agreement (Spearman R = 0.88). (F) SHAPE reactivity values plotted on the 3’ domain of mouse 18S rRNA show that the enrichment by amplification SHAPE-MaP approach accurately reports RNA secondary structure.

**Supplemental Figure S2.** Comparison of ex vivo and in-cell SHAPE reactivities. Panels A-C illustrate data from replicate 1. (A) Ex vivo reactivities smoothed over a 55-nt sliding window. Reactivities are 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) In-cell reactivities. (C) Comparison of ex vivo (blue) and in-cell (red) SHAPE reactivities, smoothed as described in (A). Overall, many regions of Xist exhibit similar structures in cells as ex vivo. Specific instances showing examples of large structural differences between in-cell and ex vivo SHAPE reactivities and also general structural agreement between in-cell and ex vivo SHAPE reactivities are highlighted with insets. (D) Ex vivo and in-cell SHAPE reactivities for replicate 2, shown as in (C). (E) Sequence motifs identified among ΔSHAPE sites, termed E1 and E2 for their location within repeat E, from two biological replicates. Information content was calculated using MEME (24).

**Supplemental Figure S3.** Side-by-side comparison of replicate ΔSHAPE analyses. Replicate 1
results are displayed above replicate 2 results. (A) Contributions of positive (blue) and negative (red) reactivity differences to the total absolute change in SHAPE reactivity between ex vivo and in-cell conditions. In each replicate, the contribution patterns are highly similar. (B) Ratio between positive and negative reactivity differences within regions of substantial reactivity change. Blue and red brackets indicate regions where in-cell protections or enhancements (or both) are most abundant. Protections dominate across most of Xist in each replicate. (C) Locations of positive (blue) and negative (red) ΔSHAPE sites, revealing similar distribution patterns between replicates. (D) Overlap between ΔSHAPE sites and CLIP- or RIP-identified sites are similar for each replicate. (E) The patterns of +ΔSHAPE-supported CLIP or RIP sites (filled circles) exhibit similar characteristics between replicates. (F) Sequence motifs derived from +ΔSHAPE-supported CELF1 and PTBP1 CLIP sites are similar to the E1 motif in each replicate. (G) Average pairing probabilities of +ΔSHAPE-confirmed CLIP sites reveal a structure-based preference for FUS binding, which appears as single-stranded binding sites flanked by base pairs in replicate 1 (solid line) and replicate 2 (dashed line). (H) ΔSHAPE sites from replicate 1 (left) and replicate 2 (right) support specific TARDBP-Xist interactions near the exon 6-7 splice junction.

**Supplemental Figure S4.** Inclusion of SHAPE structure probing data has a large impact on predicted Xist RNA structure. (A) Comparison of model obtained when guided by three-reagent differential SHAPE (8) data versus model obtained when no experimental data were used. Red arcs represent base pairs present in the three-reagent model only, and purple arcs represent base pairs present in the no-data model only. Green arcs illustrate base pairs shared between the two models and grey arcs represent regions of the RNA for which no SHAPE data are available. Base pairs common to both the differential and no-data models comprise 49.2% and 41.5% of the total base pairs present, respectively. As also indicated by the paucity of green arcs, the models predicted with and without three-reagent SHAPE data are dramatically different. (B) Comparison of models obtained when three-reagent differential data or only 1M7 data are used to model Xist secondary structure. The green arcs indicate agreement between the two structures. There are isolated regions of structural disagreement; extensive benchmarking experiments suggest that three-reagent data-driven models correctly resolve many subtle elements of RNA structure (8). (C) Left, positions of known SNVs in the Xist RNA relative to secondary structure models
supported by indicated chemical probing information. Numbers of SNVs found in base-paired and single-stranded positions in each model are listed, followed by the number of SNVs that maintain base pairing (for example, an A-U to G-U base pair) and the number of SNVs that disrupt base pairing. For each model, a bootstrapped Gaussian cumulative probability density function (CDF; right) was used to calculate p-values representing the likelihood of encountering fewer disruptive SNVs than reported in the Structure Disrupted column.

**Supplemental Figure S5.** Structural models for well-defined elements of the *Xist* lncRNA. SHAPE reactivity, Shannon entropy, pairing probability arcs, and minimum free energy structure are reproduced from Figure 1. Grey shading indicates low SHAPE reactivity/low Shannon entropy regions. Structures of individual well-defined motifs are shown explicitly and highlight the complexity of structure throughout *Xist*, especially at the 3’ end. Structure labels are colored black, blue, and red to indicate whether they are new, agree, or disagree with previously proposed structures (1). Nucleotide positions are given relative to GenBank NR_001463.3.

**Supplemental Figure S6.** Generation and characterization of *Xist*-inducible cell lines. (A) Full-length or truncated *Xist* expression was driven by a doxycycline-inducible tetracycline response element (TRE) inserted into an isogenic site on chromosome 7 via recombinase-mediated cassette exchange in a modified Tc1 embryonic stem cell line (17). The truncation site is indicated with an unfilled arrowhead above Fig. 1A. (B) RNA FISH shows that both full-length and truncated *Xist* transcripts form expected nuclear foci upon induction with doxycycline. The additional single dot in each nucleus corresponds to the *Tsix* transcript from the single active X in these cells. (C) Stabilities of full-length 18-kb *Xist* (blue) and 14.8-kb truncated *Xist* (red) expressed in Tc1 embryonic stem cells, of spliced *Xist* inducibly expressed from its endogenous location on the X chromosome in male embryonic stem cells (19) (green), and of *Xist* endogenously expressed in TSCs (purple). Full-length *Xist* is more stable than its truncated counterpart when expressed in Tc1 cells, and *Xist* expressed in TSCs was more stable than *Xist* expressed in either embryonic stem cell line profiled. (D) RNA-seq read density at the *Xist* locus shows induction and *Xist* transcript length in each cell line. Full-length and truncated *Xist* exhibited 221-fold and 185-fold changes in expression levels, respectively, upon doxycycline induction.
Supplemental Figure S7. HuR interactions with *Xist*. (A) Read density comparing HuR RIP-seq versus input reads. Strong signals are observed at the 5’ end of *Xist* and within repeat E (inset). Yellow bars below the plot indicate signals in the 95th percentile of all HuR sites transcriptome-wide. (B) Sequences of +ΔSHAPE sites within HuR binding sites reveal U-rich consensus motifs in each replicate.

Supplemental Figure S8. Clustering of ΔSHAPE sites within the FUS interaction domain reveals a preferred structural context for FUS interactions. (A) Secondary structure model for the FUS interaction domain showing +ΔSHAPE and CLIP sites (blue and grey, respectively). CLIP sites that overlap with +ΔSHAPE sites are found in both internal and terminal loop regions, supporting a modest binding preference in these locations. (B) k-means clustering of all +ΔSHAPE sites within the FUS interaction domain shows those that overlap with FUS CLIP sites (red) fall predominantly in the largest cluster, characterized by single-stranded regions bracketed on both sides by base-paired nucleotides. Data shown are for replicate 1, but similar results are obtained using data from replicate 2 (Fig. S3G).

Supplemental Figure S9. TARDBP may regulate *Xist* post-transcriptional processing by stably interacting with a highly-accessible RNA site. (A) RNA-seq coverage of the *Xist* locus for control (top) and cells treated with an antisense oligonucleotide targeting *TARDBP* (bottom) samples (27). The *Xist* gene structure is shown with arrows and boxes indicating introns and exons, respectively. In TARDBP-deficient cells, a marked increase in intron-aligned reads was observed relative to the control untreated cells, especially within the first intron. Note that the y-axis indicates raw coverage values and is not normalized by the number of input reads. (B) Quantification of the increase in intron-aligned reads when TARDBP expression is inhibited. The ratio of exon reads to intron reads is plotted; bar heights represent the average of four replicates and error bars indicate the standard deviation. The increase in intron reads is highly significant, with *p* < 0.002. (C) Plot of median *ex vivo* SHAPE reactivities for all CLIP-identified TARDBP binding sites. The single site confirmed by ΔSHAPE analysis (red) shows a greater median reactivity than all other sites, consistent with a model that recognition requires an unstructured RNA site, available for protein interactions. (D) Histogram of median SHAPE
reactivities for all +ΔSHAPE sites, the majority of which exhibit a lower median reactivity than the single ΔSHAPE-supported TARDBP. This analysis rules out the possibility that the single TARDBP site was detected solely because of high \textit{ex vivo} reactivity.

**Supplemental Figure S10.** Impact of ΔSHAPE stringency on identification of potential protein interaction sites in the 5′ end of \textit{Xist}. (A) Standard ΔSHAPE analysis requires that the 95% confidence intervals of SHAPE reactivity measurements for a given nucleotide in each condition do not overlap and that the magnitude of the difference has a standard score greater than or equal to one (4). + and – ΔSHAPE sites are shown in blue and red, respectively. ΔSHAPE analysis of \textit{Xist} using these standard parameters identified only four interaction sites in the first 1,000 nucleotides of the transcript, none of which occur within repeat A (top). When the analysis stringency was decreased such that the confidence interval was reduced to 90% and standard score thresholds were incrementally lowered, a small number of additional ΔSHAPE sites appeared within repeat A (bottom). This may indicate that the 5′ end of \textit{Xist} interacts dynamically with protein partners in a way that is not easily observed without covalent crosslinking. (B) The same analysis as described in (A), performed on replicate 2, producing similar results. (C) We examined the effect of reduced analysis stringency on other regions in which we detect RNA-protein interactions using standard ΔSHAPE parameters and found that the conclusions outlined in the main text were unaffected. For example, interactions within the FUS domain detected using standard parameters (top) are largely unchanged even with reduced analysis stringency (bottom). (D) The same analysis as described in (C), performed on replicate 2. Throughout this figure, the scale is sufficiently detailed that vertical lines showing ΔSHAPE sites that fall in the same location indicate identical sites in each replicate.
Ex vivo Enrichment by amplification

Supplemental Fig. S1
Ex vivo vs. in-cell

Supplemental Fig. S2
A
Differential (three-reagent) model compared with no-data model

B
Differential (three-reagent) model compared with 1M7-only model

C

<table>
<thead>
<tr>
<th></th>
<th>Base paired</th>
<th>Single stranded</th>
<th>Structure maintained</th>
<th>Structure disrupted</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No data</td>
<td>59</td>
<td>46</td>
<td>5</td>
<td>54</td>
<td>0.351</td>
</tr>
<tr>
<td>1M7 only</td>
<td>50</td>
<td>55</td>
<td>4</td>
<td>46</td>
<td>0.146</td>
</tr>
<tr>
<td>Differential</td>
<td>40</td>
<td>65</td>
<td>2</td>
<td>38</td>
<td>0.027</td>
</tr>
</tbody>
</table>