ARTICLE
PUBLISHED ONLINE: 9 JANUARY 2017 | DOI: 10.1038/NCHEMBIO.2272

nature chemical biology

Visualizing the secondary and tertiary architectural domains of lncRNA RepA

Fei Liu[1,2], Srinivas Somarowthu[1,4] & Anna Marie Pyle[1–3]*

Long noncoding RNAs (lncRNAs) are important for gene expression, but little is known about their structures. RepA is a 1.6-kb mouse lncRNA comprising the same sequence as the 5′ region of Xist, including A and F repeats. It has been proposed to facilitate the initiation and spread of X-chromosome inactivation, although its exact role is poorly understood. To gain insight into the molecular mechanism of RepA and Xist, we determined a complete phylogenetically validated secondary-structural map of RepA through SHAPE and DMS chemical probing of a homogeneously folded RNA *in vitro*. We combined UV-cross-linking experiments with RNA modeling methods to produce a three-dimensional model of RepA functional domains demonstrating that tertiary architecture exists within lncRNA molecules and occurs within specific functional modules. This work provides a foundation for understanding of the evolution and functional properties of RepA and Xist and offers a framework for exploring architectural features of other lncRNAs.

lncRNAs are a rapidly growing class of cellular transcripts that exceed 200 nt in length and lack protein-coding potential. During the past decade, lncRNAs have been increasingly recognized as essential components of the mammalian transcriptome[1]. They are involved in various cellular processes and diseases, including signaling, embryonic-stem-cell differentiation, brain function, chromatin remodeling, and cancer[2–7]. Genome-wide analysis has suggested that lncRNAs possess greater structural complexity than mRNAs[8–10]. Recent experimental characterizations of the lncRNAs SRA and HOTAIR have provided evidence that lncRNAs can adopt complex structures[11,12].

There is particular interest in the X-inactivation-specific transcript (Xist)[13,14], which functions in the dosage-compensation pathway in eutherian mammals. Females silence one of two X chromosomes, thereby balancing X-linked gene expression with that of monoallelic males[15]. The lncRNA Xist (>17 kb in mice) is indispensable for initiating, establishing and maintaining X-chromosome inactivation (XCI)[13,14]. The proposed functional roles of Xist include the recruitment of silencing factors and spreading over the inactive X chromosome (Xi) *in cis*, thereby causing transcriptional silencing of X-linked genes[16].

Unlike some lncRNA genes that are conserved only in primates, Xist has an overall gene structure that is conserved among all eutherian mammals. Its genetic architecture includes six regions composed of short tandem-repeat sequences, termed A–F[13,14,17] (**Fig. 1a**). The A repeat consists of 7.5 repeated A units in mice, and more than a decade ago this region was shown to be crucial for initiation of XCI[18]. Many protein factors that contribute to XCI have been proposed to directly interact with the A repeat, including Polycomb repressive complex 2 (PRC2), ATRX, and Spen (also known as Sharp)[19–24]. The F repeat is located ~0.7 kb downstream of the A repeat and consists of two repeated F units in mice[17]. This region has recently been demonstrated to directly interact with the Lamin B receptor (LBR) and to be required for Xist-mediated silencing of distal genes[25]. However, the precise mechanism by which the A repeat and F repeat recognize various protein partners and initiate X-chromosome silencing remains unclear.

The A- and F-repeat units within Xist are also embedded within a 1.6-kb mouse transcript known as RepA. The RepA RNA (not to be confused with the short A-repeat units themselves) is encoded by an internal promoter on the Xist-gene sense strand[19] (**Fig. 1a**) and has been proposed to recruit PRC2 histone methyltransferase to the future Xi before the expression of Xist. Additionally, RepA appears to upregulate the expression of Xist, which then initiates and spreads silencing across the inactive X chromosome[19].

Despite the availability of abundant functional data, structural work on this system has been limited. Early work focused on a single A-repeat unit or an isolated A-repeat-containing sequence (427 nt long). Most of these studies have described the entire A repeat as a set of tandem dual stem-loop structures[18], although some evidence has suggested that isolated A-repeat sequences can adopt more complicated configurations[26–28]. A later study using *in vivo* dimethyl sulfate (DMS) probing has provided the first in-cell secondary-structural information for the A-repeat sequence, as well as information on other novel structural elements within Xist[29]. Recently, another study has used *in vivo* RNA–RNA cross-linking to identify regions of human XIST involved in long-range interactions, at near-nucleotide resolution[30]. This study has revealed several conserved long-range interactions within Xist, but none involving the RepA region, thus indicating that A-repeat and F-repeat regions fold as isolated domains in the context of full-length Xist. This finding suggests that structural data obtained on the RepA transcript are relevant to the RepA region in Xist. Altogether, previous studies have provided secondary-structural information for fewer than 40% of the individual nucleotides in the RepA and Xist (RepA/Xist) region, owing to technological limitations. This incomplete data set, along with the use of truncated constructs, has probably contributed to the conflicting models of Xist structure and the limited understanding of Xist function. A complete secondary-structural map of RepA/Xist is needed to define the details of its architectural units, thereby setting the stage for phylogenetic analysis, and to understand the function of RepA/Xist at the molecular level.

In this work, we determined what is, to our knowledge, the first complete secondary-structural map of the RepA RNA by generating

[1]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, USA. [2]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [3]Department of Chemistry, Yale University, New Haven, Connecticut, USA. [4]Present address: Department of Biochemistry and Molecular Biology, Drexel University College of Medicine, Philadelphia, Pennsylvania, USA. *e-mail: anna.pyle@yale.edu

a homogenous, monodisperse RNA sample and then mapping its secondary structure by using selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) and DMS probing methodologies. We subjected the resultant map to a statistical and thermodynamic evaluation through jackknife[31] and Shannon entropy analyses[32]. Together, the data demonstrate that RepA forms a complicated structure composed of three independently folding modules. This map provides landmarks for the construction of meaningful sequence alignments, which enabled us to conduct phylogenetic analyses implicating regions of functional importance and providing functional validation for substructures in the map. We investigated the relative organization of RepA secondary structures in three-dimensional space by using UV-cross-linking experiments and computational 3D modeling, thereby providing tertiary-structural information on a lncRNA and revealing regions of sequence conservation. The resultant features provide structural insights into the evolution and function of RepA, and the study lays a groundwork for exploring the three-dimensional architecture of other lncRNAs.

RESULTS

RepA adopts a compact monodisperse state after folding

Functional RNAs, such as ribozymes, group II introns and riboswitch RNAs, can exert their function only as well-folded molecules[33–35]. Performing structural probing on less folded and less compact RNA samples, which contain a mixture of RNA conformations, prevents accurate structural predications. Therefore, the availability of a homogenous, monodisperse RNA sample is essential for meaningful structural characterization of any RNA. However, such samples are difficult to achieve for lncRNAs, owing to their length, because traditional RNA purification and refolding methods produce a mixture of lncRNA conformations[12]. To address this problem, we adopted an alternative native purification protocol that preserves the RNA secondary structure formed during *in vitro* transcription. Using this protocol (**Online Methods**), we obtained a homogeneous and monodisperse sample of RepA suitable for *in vitro* structural characterization.

To identify optimal ionic conditions for promoting the homogeneous compaction of monomeric RepA molecules, we studied RepA compaction as a function of Mg^{2+} concentration. We conducted a series of sedimentation-velocity experiments, using analytical ultracentrifugation (SV-AUC) at physiological K^+ concentration (150 mM) to directly monitor the degree of molecular compaction as a function of Mg^{2+} concentration. As Mg^{2+} concentration increased, the sedimentation coefficient (s) of RepA increased (**Fig. 1b**), and the hydrodynamic radius (R_H) decreased (**Fig. 1c**, **Supplementary Results** and **Supplementary Fig. 1**), thus indicating global compaction of the RepA molecule. Fitting the R_H values at each Mg^{2+} concentration to the Hill equation yielded a $K_{1/2\ Mg}$ of 4.8 ± 0.2 mM, a value smaller than those of several highly structured RNAs, such as the ai5γ group IIB intron (15 ± 2 mM)[36] and the lncRNA HOTAIR (8.6 ± 0.8 mM)[12]. This low $K_{1/2\ Mg}$ for RepA compaction signifies a relatively high degree of structural stability.

We also confirmed the compaction of RepA after addition of Mg^{2+} by size-exclusion chromatography (SEC) (**Supplementary Fig. 2**). The SEC profiles demonstrated that the RepA population was stable and homogenous over a broad range of Mg^{2+} concentrations (0–20 mM). Both SV-AUC and SEC experiments indicated that a concentration of 15 mM Mg^{2+} (more than three times the $K_{1/2\ Mg}$) was sufficient to produce a completely compact and monodisperse form of RepA. We therefore chose this condition for subsequent *in vitro* structural characterization.

Determination of RepA secondary structure

Having obtained a well-folded RNA sample, we examined its secondary structure by using SHAPE and DMS probing. The SHAPE reagent 1-methyl-7-nitroisatoic anhydride (1M7) selectively acetylates the 2′-hydroxyl group of RNA nucleotides that have flexible backbones. DMS selectively methylates the heterocyclic nitrogen atoms on adenines and cytosines that are not base-paired. Because these two methods are orthogonal approaches for interrogating secondary structure, good agreement between them suggests a robust secondary-structural map.

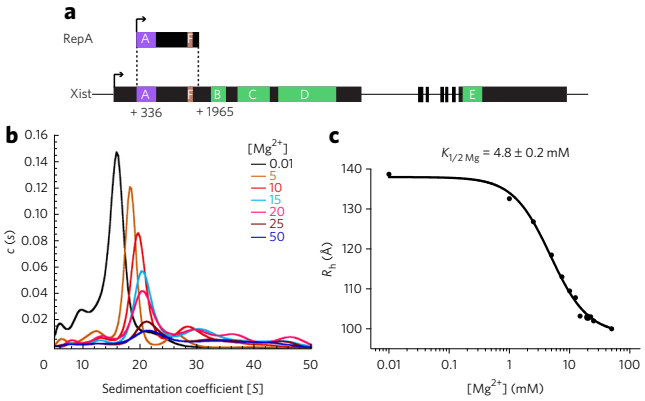

Figure 1 | RepA folds into a compact species after addition of Mg^{2+}. (**a**) RepA spans a region that extends from nucleotides 336 to 1965 in Xist and includes A-repeat and F-repeat sequences[19]. The transcription start sites in RepA and Xist are denoted with black arrows. All the repeat regions in RepA and Xist are highlighted and labeled. (**b**) SV-AUC profiles of RepA obtained under native conditions in the presence of increasing concentrations of Mg^{2+}. The graph was obtained with SedFit[43,44]. (**c**) Hill plot of the hydrodynamic radii (R_h, in angstroms) derived from the SV-AUC experiment described in **a**. The R_h values were derived from one independent measurement. $K_{1/2Mg}$ is 4.8 ± 0.2 mM, and the Hill coefficient (n_H) is 1.4 ± 0.1. Hill plots drawn for each set of measurements are shown in **Supplementary Figure 1**. The standard errors were estimated by least-square curve fitting.

We monitored SHAPE reactivity at single-nucleotide resolution and used normalized SHAPE reactivity values as pseudo–free energy constraints to guide RepA structure prediction (**Online Methods**). We evaluated the SHAPE-directed secondary-structural models by using jackknife resampling[31] to estimate the confidence of each base pair in the predicted model. The majority of the helices (70%) were described with high confidence, and most of the low-confidence helices were located adjacent to junctions (**Supplementary Fig. 3**). Next, we used the normalized DMS reactivities to examine the low-confidence regions. The low DMS reactivities in these helices supported the formation of direct base pairs; these results were consistent with the SHAPE-directed secondary-structural model (**Supplementary Fig. 3**). Overall, the resulting map (**Fig. 2**) indicates that RepA is a highly structured RNA molecule in which more than 60% of the nucleotides are base-paired.

RepA consists of three independently folding modules

Our experimentally derived map suggests that RepA contains a set of secondary-structural domains that radiate from a central junction (**Fig. 2**). To examine whether these individual sections are autonomous folding domains, we used the 3S shotgun analysis method[37] with two groups of fragments (**Fig. 3a**). One group contained fragments F1, F2, and F3, which together cover the full-length RepA. The boundaries of these fragments were designed to maintain all secondary-structural elements identified in full-length RepA. The other group contained fragments F4, F5, and F6, which overlap with fragments F1–F3 but are expected to disrupt secondary-structural domains predicted by the mapping and modeling procedure.

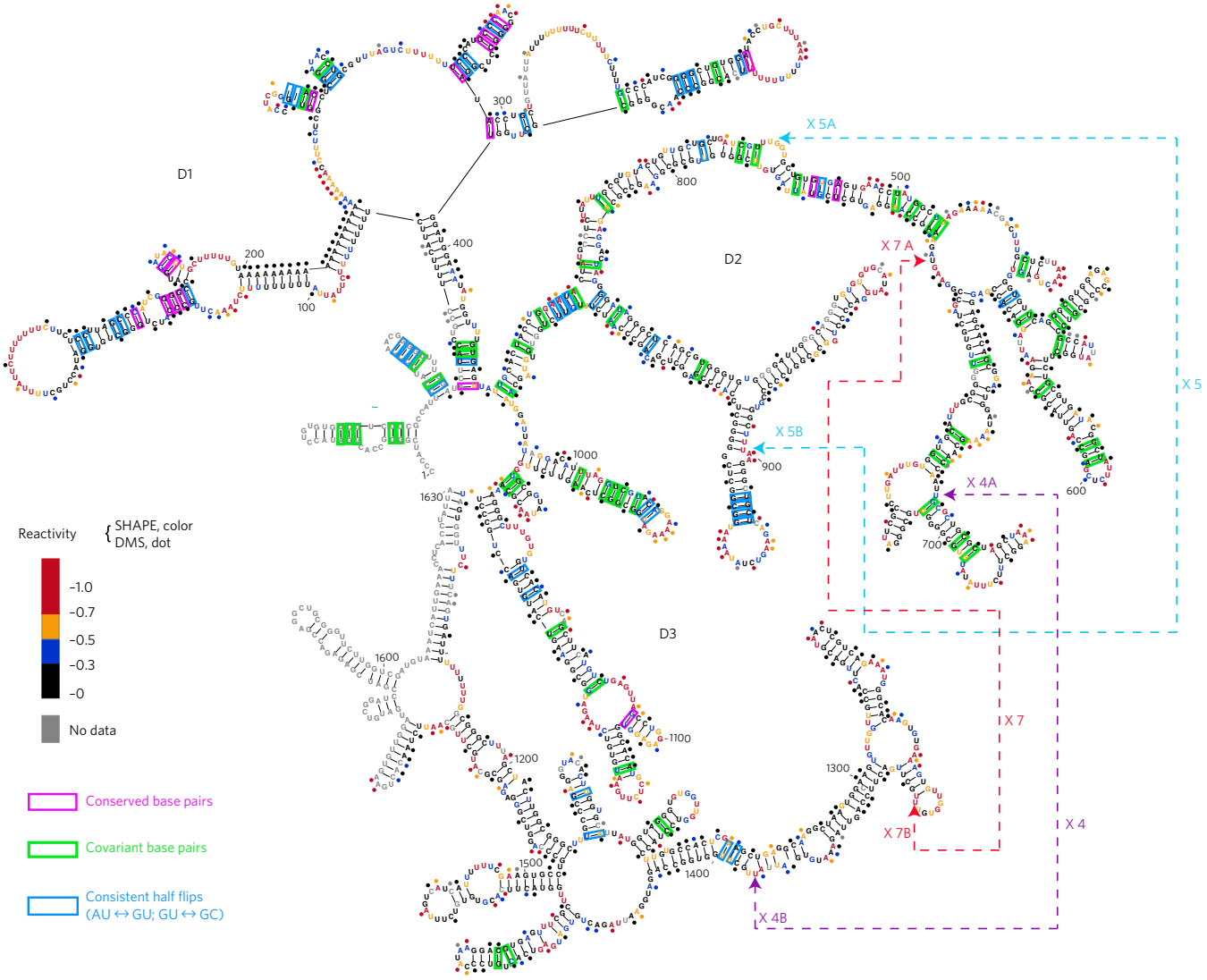

Figure 2 | Secondary structure of RepA derived from SHAPE and DMS probing. SHAPE reactivities are depicted by colored nucleotides. DMS reactivities are represented by colored dots over the nucleotides. SHAPE and DMS reactivities are denoted with the same color codes, as illustrated in the key. Highly reactive nucleotides are indicated with red and orange, and nucleotides with low reactivities are displayed in black or blue according to their reactivity values. Covariant base pairs in 56 mammalian sequences are highlighted in green, consistent half-flip pairs are highlighted in blue, and conserved base pairs are highlighted in pink. UV-cross-linked nucleotide positions are indicated by arrows with dashed lines. The secondary structure was drawn with VARNA (http://varna-gui.software.informer.com/).

The SHAPE reactivities of fragments F1, F2, and F3 correlated with their corresponding regions in full-length RepA with Pearson's correlation coefficients (r_p) of 0.92 ± 0.01 for F1, 0.83 ± 0.08 for F2, and 0.89 ± 0.01 for F3 (**Fig. 3b**). These results indicate that fragments F1, F2, and F3 adopt the same structure in isolation (**Supplementary Fig. 4**) and in the full-length RepA, and are therefore independently folding domains. By contrast, the SHAPE reactivities of fragment F4, F5, and F6 showed relatively poor correlation with the corresponding full-length RepA profiles (F4, F5, and F6 had r_p values of 0.57 ± 0.06, 0.55 ± 0.1, and 0.72 ± 0.05, respectively) (**Fig. 3b**), thus indicating that they adopt structures different from those of the parent molecule (**Supplementary Fig. 4**). Our results suggest that RepA is composed of three independent structural modules corresponding to fragments F1 (domain 1, D1), F2 (domain 2, D2), and F3 (domain 3, D3), all of which are connected through a central junction and assemble into a complex structure (**Fig. 2**).

To evaluate potential local structural heterogeneity, we calculated the Shannon entropy (*S*) of each nucleotide by using SHAPE-directed base-pair probabilities[32] (**Supplementary Fig. 5** and Online Methods). Overall, the average *S* value of RepA was as low as 0.102, thus indicating an overall structural homogeneity of RepA.

D1 comprises 7.5 A-repeat units and forms an elaborate secondary structure different from that of the tandem dual stem-loop model[18] (**Fig. 4a**). Only repeat 5 adopts the dual stem-loop structure, whereas repeats 1, 2 and 6 form diverse types of stem-loop motifs. Repeats 3 and 4 form an extended stem-loop, probably because of the long consecutive AU stem that lies adjacent. Similarly, repeats 7 and 8 base-pair with each other, thereby forming a relatively long stem-loop. Additionally, one central six-way junction brings the A-repeat units closer in space and adds complexity to the overall structure of D1. D1 had the highest average Shannon entropy (0.122), as compared with D2 (0.098) and D3 (0.093), thus suggesting relatively higher dynamics of D1. This observation is consistent with recent RNA–RNA cross-linking results that have suggested that an A repeat exists in multiple confirmations *in vivo*[30]. Specifically, the nucleotides involved in the consecutive AU stem had the highest *S* values (>0.5), thus contributing to the formation of multiple conformations. Although our RNA sample for chemical probing was globally homogeneous (**Fig. 1b** and

Supplementary Fig. 1), our structural model for this region probably represents one of the A-repeat functional states.

D2 and D3, compared with D1 exhibit different structural features, and they appear to be much more stable. The *S* values of most nucleotides in these domains were very low. Specifically, only 0.9% and 2.6% of nucleotides in D2 and D3, respectively, were located in very dynamic structural elements (*S* values >0.4), compared with 10.5% in D1 (**Supplementary Fig. 5**). Moreover, D2 and D3 consist of more complicated structural motifs (**Supplementary Fig. 6**). D2 contains three three-way junctions and two four-way junctions. Three major helices, H12, H13, and H25, are linked by the central three-way junction, thereby promoting long-range contact. In D3, there is a 471-nt region where 14 helices assemble into a subdomain via three different junction segments. The two F-repeat units (highlighted in **Supplementary Fig. 6b**) are located in this region and participate in the base-pairing of the same helix, H34. Interestingly, internal loops and bulges are embedded within the majority of D2 and D3 helices, thus adding flexibility to these helices and potentially promoting tertiary contacts among structural motifs.

Covariation analysis of RepA secondary structure

Using our secondary-structure map, we improved the sequence alignment of the RepA region and generated a covariance model[38,39] (Online Methods), which allowed us to identify the conserved structural motifs that may contribute to RNA function. Compared with results from previous studies[29], our covariation analysis included a much broader range of species (56 mammalian species), thus greatly increasing its statistical power. Despite rapid changes in primary sequence through evolution, the RepA region displays a high degree of covariation in base-pairing (**Fig. 2**), including the base pairs adjacent to the central junction connecting the three domains.

Most of the newly proposed structural elements in D1 can form in all mammalian species examined (**Fig. 2**). Ten out of 11 helices in D1 show a high degree of conservation across species. Previous phylogenetic analysis and functional studies have focused on the GC-rich A-repeat units, whereas we found that the sequences of some uridine-rich linker regions are highly conserved even in distant species (**Fig. 4b**). Interestingly, not only the uridines but also the other nucleotides in these regions are conserved. These linkers are all located in the loops or bulges adjacent to extended stem-loops. Our results suggest that the conserved linkers may be important in mediating specific interactions between A-repeat units and in facilitating the formation of more complex structural elements. Additionally, the conservation of sequence and structure within these uridine-rich linkers indicates a potential binding site for polyuridine or single-stranded RNA–binding proteins.

We extended our phylogenetic analysis to the region downstream of the A repeat in Xist (D2 and D3 in RepA) and identified many new conserved structural motifs (**Fig. 2** and **Supplementary Fig. 6**). The nucleotides participating in the central three-way junction in D2, which connects three important helices (H12, H13, and H25), are highly conserved in rodents, primates, and more distant mammals (**Supplementary Fig. 7a**), thus indicating that the overall architecture of D2 is probably conserved across species. The structure of the region containing nucleotides 498 to 771 is also highly conserved in rodents, thereby suggesting its functional importance (**Supplementary Fig. 7b**). The corresponding regions in the rat and Chinese hamster sequences are capable of adopting similar intricate structures despite their modest sequence identity (85% and 74%, respectively). Similar structural motifs (H15–H19) have also been derived from *in vivo* DMS probing[29].

Unlike the D1 and D2 regions, D3 exhibits relatively poor sequence conservation in mammals (**Fig. 2** and **Supplementary Fig. 6b**). Among all the structural motifs, only helices H28–H30, H35, H37, and H41 show a modest degree of conservation across species. Therefore, the D3 region in other species may display substantially different structural features.

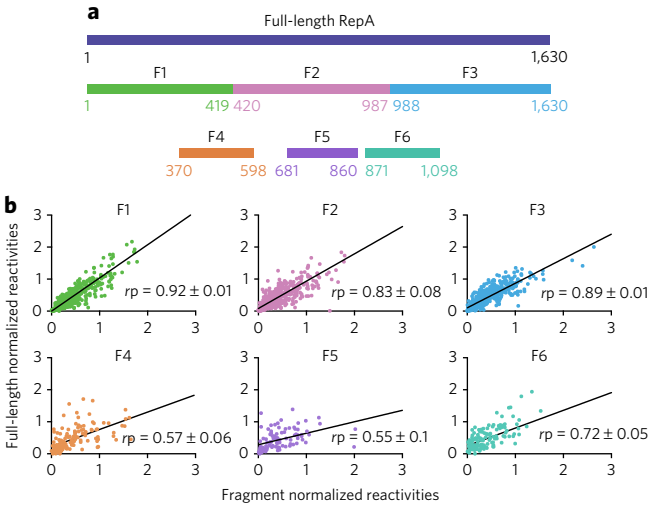

Figure 3 | Fragment analysis revealing independently folded domains in RepA. (**a**) Schematic representation of RepA fragments with respect to their positions along the sequence of full-length RepA. (**b**) Scatter plots comparing the SHAPE reactivity of each fragment with the corresponding region in full-length RepA. Pearson correlation values (r_p) between the SHAPE reactivities of each fragment and the corresponding regions in full-length RepA are indicated as mean ± s.e., with s.e. estimated by bootstrapping analysis. Three independent SHAPE experiments were performed on each fragment or full-length RepA.

Characterization of tertiary interaction sites in RepA

Whereas secondary structure is informative, it is additionally valuable to obtain information on the arrangement of helices in three dimensions. To determine which RepA subdomains are located in proximity, we conducted UV-cross-linking experiments on the folded RepA molecule.

We first examined the UV-cross-linking pattern of RepA at different Mg^{2+} concentrations (0–15 mM; **Fig. 5a** and **Online Methods**). We observed that RepA formed cross-links under all conditions tested, and the addition of Mg^{2+} changed the intensity of several cross-linked bands (**Fig. 5a**). Specifically, the cross-linked band CL1 became more prominent at higher Mg^{2+} concentrations, whereas CL8 diminished. The cross-linked bands CL4 and CL6 appeared only when RepA was folded in the presence of Mg^{2+}. All cross-links became apparent at a relatively low Mg^{2+} concentration (5 mM); no additional cross-links appeared at higher concentrations. Comigration of cross-linked and unirradiated monomeric RNA on a native agarose gel confirmed that all cross-links were intramolecular. Altogether, these observations indicate that RepA undergoes a conformational change and tertiary-structure stabilization after Mg^{2+} addition, a conclusion consistent with the Mg^{2+}-dependent compaction of RepA observed in SV-AUC and SEC.

We used reverse transcription to identify the nucleotides involved in each cross-linked species. UV-cross-linked nucleotides generally interrupt primer extension one base before the cross-linked nucleotide residue[40]. Therefore, we assigned cross-linked nucleotides according to the position of strong reverse-transcription stops displaying intensities substantially higher than those in the unirradiated control (C) or the UV-treated but un-cross-linked parent RNA (P) (**Fig. 5b–d** and **Supplementary Fig. 8**). We identified three pairs of long-range cross-links—CL4, CL5, and CL7—within D2 and D3 (**Fig. 2**).

Specifically, cross-link CL5 between U480 (X 5A) and U899 (X 5B) formed between the internal loops of helices H14 and H27 in D2 (**Fig. 2** and **Supplementary Fig. 6a**). This cross-link reflects probable

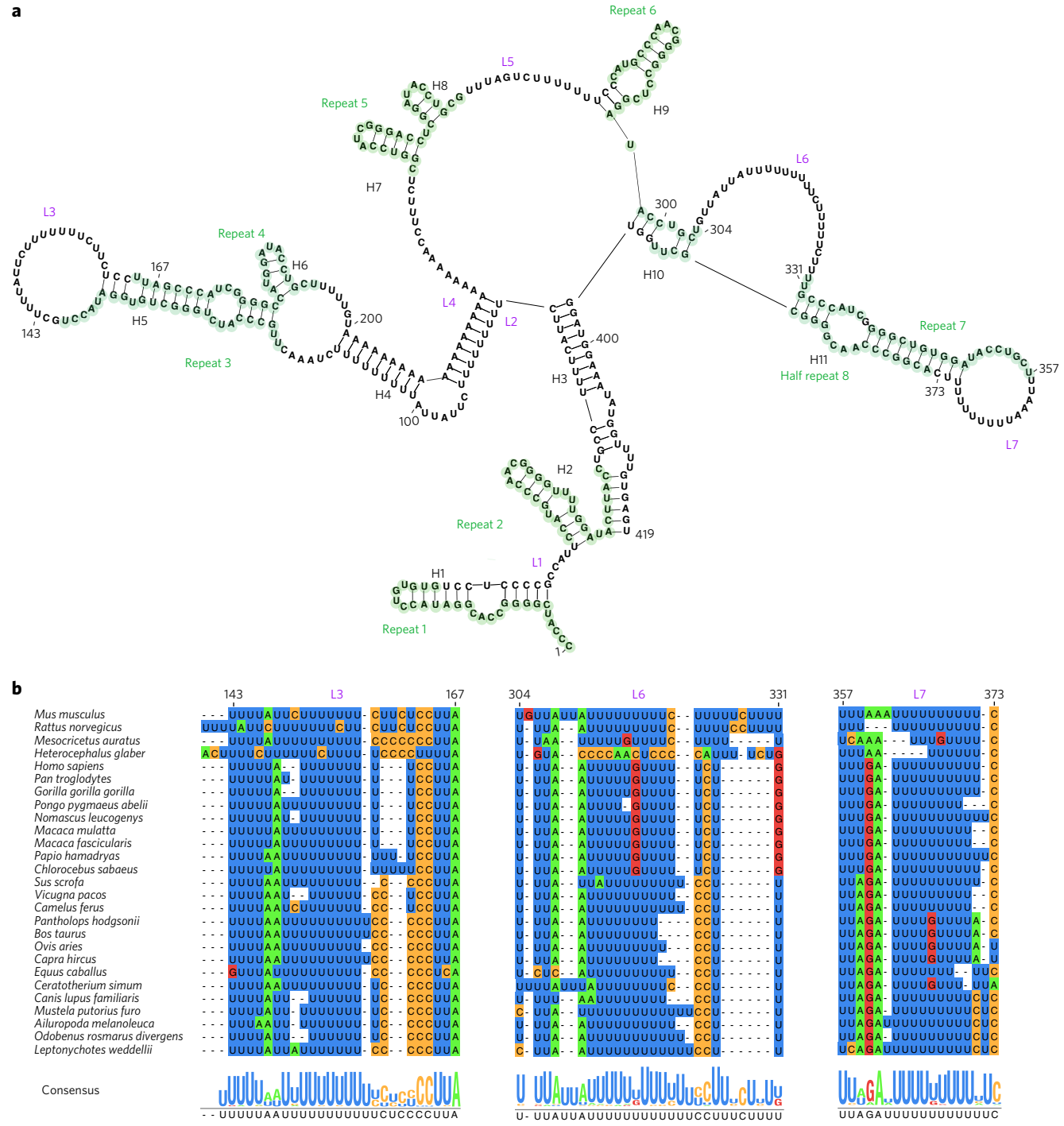

Figure 4 | D1 of RepA contains the functionally important A-repeat motif. (**a**) Secondary-structural map of D1. A-repeat units are highlighted in green and labeled accordingly. Linker regions are indicated by L1–L7. Helices in D1 are indicated by H1–H11. (**b**) Sequence conservation of the linker regions L3 (nucleotides 143–167), L6 (nucleotides 304–331) and L7 (nucleotides 357–373) across 27 mammalian species. The alignment of the sequences is presented and color-coded on the basis of nucleotide type, and the consensus sequence is shown on the bottom.

interactions between helices H14 and H27, which bring U480 and U899 into proximity. In this way, D2 adopts a compact configuration and may serve as a scaffold for structural elements in other regions.

The other two identified cross-links suggest long-range interactions between D2 and D3 (**Fig. 2** and **Supplementary Fig. 6**). U664 (X 4A) is part of a three-way junction in D2 and cross-links with U1394 (X 4B), a bulged nucleotide in helix H38 in D3. Importantly, U759 and U1364, which form CL7, are proximal to U664 and U1394 (CL4). These observations suggest that certain subdomains in D2 (nucleotides 494–508 and 632–776) and D3 (nucleotides 1272–1304 and 1349–1406) interact and form a specific tertiary architecture. The UV-cross-linking results are consistent with the low Shannon entropy in this region, thus suggesting that all the nucleotides involved in or adjacent to tertiary interactions are likely to remain in a single and stable conformation (**Supplementary Fig. 5**).

Phylogenetic analysis revealed that the cross-linked nucleotides and their surrounding primary sequences are well conserved, except

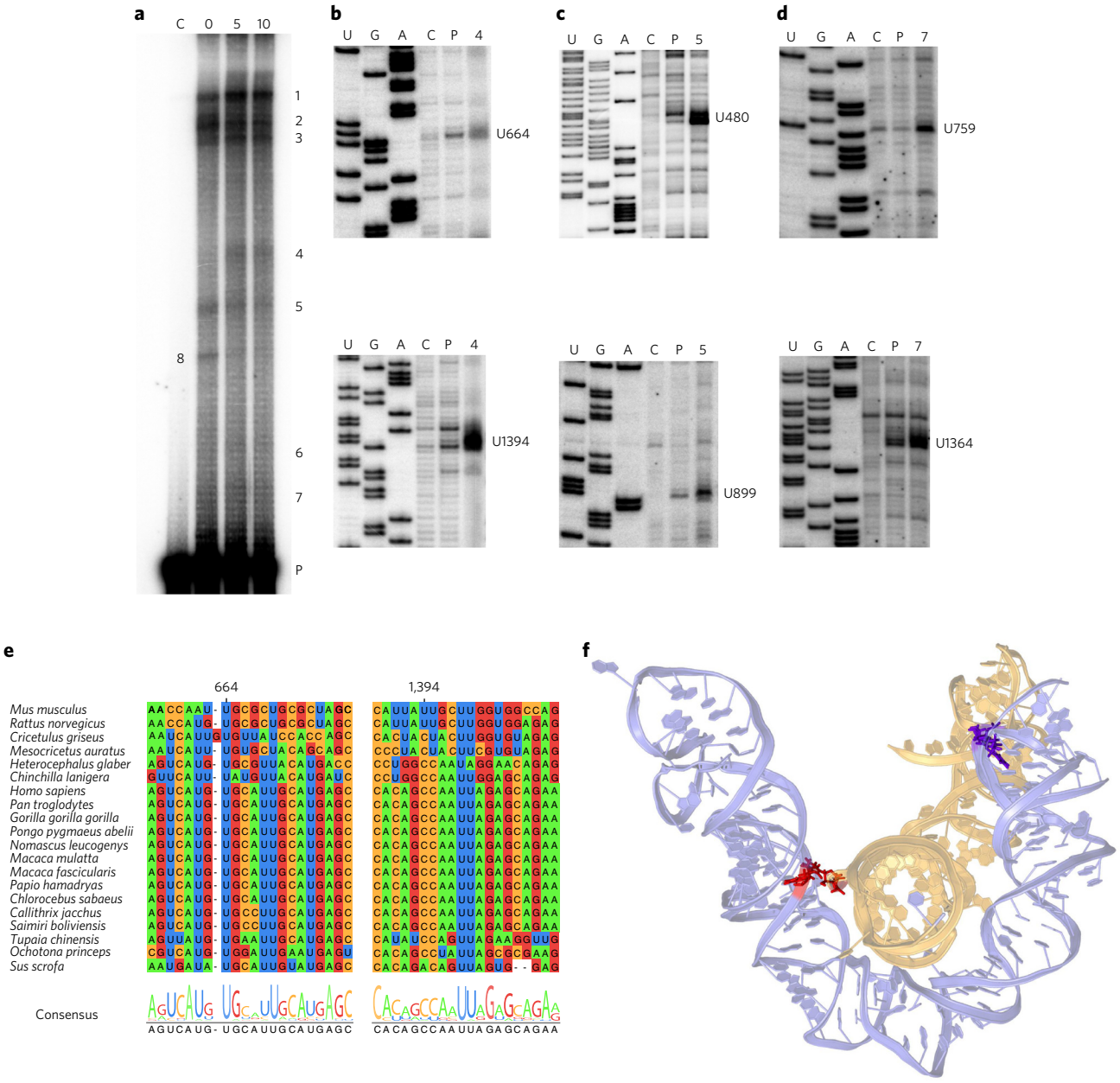

Figure 5 | Characterization of the UV-cross-linking sites within RepA. (**a**) UV-cross-linking patterns of RepA. Unirradiated control RNA (C) is shown in the leftmost lane. UV-treated RNA samples in the presence of 0, 5, or 10 mM Mg are shown in lanes 0, 5, and 10, respectively. The cross-linked RNA bands are labeled 1 to 8. The un-cross-linked parent RNA band is denoted by P. (**b**–**d**) Primer-extension mapping of the cross-linked nucleotides. Lanes U, G, and A show dideoxy-sequencing ladders. Lanes C, P, 4, 5, and 7 show reverse transcription (RT) products of corresponding RNA species. The nucleotide positions of identified RT stops are indicated to the right of the gels. Original gel images are shown in **Supplementary Figure 8**. (**e**) Primary sequence conservation of 20 mammalian species corresponding to the regions containing the cross-link CL4. The alignment of the sequences is presented and color-coded on the basis of nucleotide type, and the consensus sequence is shown on the bottom. (**f**) Three-dimensional model of the subdomains in D2 (nucleotides 494–508 and 632–776, purple) and D3 (nucleotides 1272–1304 and 1349–1406, yellow). Nucleotides participating in cross-links CL4 and CL7 are purple and red, respectively.

for U1364 (X 7B) and surrounding sequences, which are conserved only in mice and rats (**Fig. 5e** and **Supplementary Fig. 9**). Therefore, tertiary interactions suggested by the cross-links appear to be conserved in the species that we examined, although the specific arrangement of helices in three dimensions probably varies.

To visualize the tertiary contacts between D2 and D3, we modeled the spatially proximal subdomains with RNAComposer[41], using the identified cross-links as distance constraints (**Fig. 5f**). For comparison, we also modeled the subdomain of D2 alone. In the absence of D3, the corresponding subdomain in D2 was predicted to adopt a relatively extended and open conformation (**Supplementary Fig. 10**). In contrast, when it is constrained by tertiary interactions with D3, this subdomain forms a more compact structure (**Fig. 5f**). According to our 3D model, the D2 subdomain forms a platform for the docking of the distal stem-loop of D3, thereby placing U1364 in proximity to U759 and U664 in the vicinity of U1394.

Altogether, our UV-cross-linking and 3D modeling results suggest a specific structural model for RepA. In this model, D2 and D3 interact and establish a relatively rigid scaffold. This structure may facilitate the proper orientation of bound protein partners, thereby facilitating the formation of functional machinery.

DISCUSSION

In this study, we provide a complete and robust structural model of RepA, a functionally important lncRNA[19] that is identical to a 5′ region in Xist. This phylogenetically validated secondary structure of RepA should serve as a roadmap for the future study of XCI, thus allowing investigators to design targeted studies of specific structural elements and interactions and to improve understanding of the complex process of XCI. Furthermore, we defined specific tertiary contacts within RepA. This finding has important implications in RepA/Xist function and underscores the ability of lncRNAs to form complex and discrete secondary and tertiary structures.

In 2015, a secondary-structural model of the A repeat was derived from *in vivo* DMS probing data[29]; this model suggests an assembly of complex structural motifs within the A repeat that is more similar to our model than to the dual stem-loop model. In fact, repeats 5–8 are predicted to form the same structure in both our model and the *in vivo* model. However, repeats 1–4 are dissimilar, owing to different predictions of pairing of the U-rich linker (L2 or L6) with the A-rich region of L4. These discrepancies can be attributed to the use of only the most DMS-reactive nucleotides (11% of the total nucleotides) to constrain the RNA secondary-structure modeling in the 2015 study[29], whereas SHAPE and DMS probing data used in combination provided pseudo–free energy constraints on 85% of the nucleotides used for modeling of the secondary structure presented here. Owing to its inability to report on the pairing status of uridines, DMS probing provided information on only 29% and 12% of nucleotides in uridine-rich linkers L2 and L6, respectively. Hence, the DMS probing data alone are insufficient to determine which linker regions are base-paired. The SHAPE data presented here reflect the base-pairing status of all nucleotides, and the low SHAPE reactivity of linker L2 indicates that it is involved in base-pairing, whereas the high SHAPE reactivity of linker L6 indicates a single-stranded segment. This rich data set, provided by multiple chemical probes, combined with supporting phylogenetic covariation, provides a robust structural model of the A repeat.

Although *in vivo* DMS probing is insufficient for generating a robust secondary structure, the raw data provided a valuable way to evaluate the biological relevance of our proposed structural map. We found that the majority of the structural elements in our model are consistent with the corresponding *in vivo* DMS probing data (**Supplementary Fig. 11**). One exception is helix H3, in which several nucleotides are highly DMS reactive *in vivo*, thus suggesting that H3 may be unzipped after protein binding in the cellular environment, whereas other structural elements within the A repeat are maintained.

Many studies have focused on the A repeat after a systematic deletion analysis demonstrated its indispensable role during XCI[18]. One of the deletion constructs, ΔSX, was created by deleting 0.9 kb of the 5′ end of Xist. ΔSX lacks not only the A repeat but also 172 nt immediately downstream (145 nt in D2 of RepA). Because the ΔSX deletion completely abolishes silencing activity, the A repeat of Xist has been deemed essential for silencing. ΔSX has been used in later studies[21,22,25,42], including one that has focused on identifying proteins that interact directly with Xist and are required for silencing[21]. In this case, because ΔSX did not recruit certain protein partners Spen (or Sharp), Rnf20, and Wtap *in vivo,* it has been deduced that the A repeat directly interacts with these proteins[21]. However, our model of RepA suggests that the aforementioned protein partners are equally likely to interact with D2, because the deletion creating ΔSX disrupts the highly conserved structural motifs in D2, which is located immediately downstream of the A repeat. Additionally, this deletion may trigger misfolding of the RNA and disrupt tertiary interactions. Therefore, ΔSX is a suboptimal control for investigating the functional role of the A repeat. Our structural model should facilitate the design of new constructs that can be used in studies of XCI.

Recent work on Spen has more conclusively identified a direct interaction with the A repeat, and our map suggests a precise structural motif that may be responsible. *In vitro* UV cross-linking between Spen and RepA (the same RNA construct used in this study) has suggested that the uridine-rich linkers in the A repeat are Spen-recognition sites[30]. However, Spen does not cross-link to the uridine-rich linkers equally. The dynamic inter-repeat duplex model previously proposed[30] does not provide a conclusive explanation for this observation. Interestingly, our structural map correlates with the differential cross-linking efficiency. The strongest protein cross-links are formed with the single-stranded uridine-rich linkers adjacent to helices H5 and H11 in our map (**Fig. 4a**). H5 and H11 are uniquely composed of two A-repeat units that form an extended stem-loop, thus suggesting that Spen may preferentially bind this structural motif. Moreover, these linker sequences are highly conserved in our phylogenetic analysis (**Fig. 4b**). In contrast, the linker regions between and upstream of the intrarepeat duplexes, which are inefficiently cross-linked to Spen[30], have a lower degree of sequence conservation. Overall, the correlation between the functional data and the structural elements proposed in our work suggests that our structural map represents a functional state of RepA/Xist.

The region downstream of the A repeat (D2 and D3 in RepA), including the F repeat, has long been considered to be a functionally dispensable part of Xist[18], and almost no functional or structural work has been done to characterize the role of this region. Very recently, an important player in XCI, the protein LBR, has been shown to directly interact with Xist through an LBR-binding site (LBS), which spans the regions that we designated D2 and D3 (ref. 25). However, deletion constructs created to test the LBS (ΔLBS, nucleotides 563–1347 in RepA) were designed without any knowledge of the potential importance of downstream substructures, and the ΔLBS construct is therefore predicted to disrupt some of the most important elements in D2 and D3, including the tertiary-interaction sites. As a result, the actual LBS-binding site remains undefined. Therefore, our structural model may serve as a guide in designing constructs for future mechanistic studies of interactions between RepA/Xist and protein partners.

Altogether, our findings show that RepA adopts specific functional secondary structures and assembles into a set of three distinct domains that can be validated biochemically and phylogenetically. Our findings have crucial implications for the interpretation of previous studies on XCI and the silencing machinery. Additionally, we demonstrated that RepA and the 5′ end of Xist possess a defined tertiary architecture, thereby suggesting that lncRNAs can adopt three-dimensional structures. These structures can form autonomously in the absence of protein partners, thus underscoring a central role of RNA folding in the formation of the lncRNA–protein machinery. Our results complement the valuable information obtained through global methods of monitoring RNA and protein complexes *in vivo* and provide precise biophysical insights into the functional roles of RepA/Xist during XCI at the molecular level.

Received 4 April 2016; accepted 16 November 2016;
published online 9 January 2017

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

References

1. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).
2. Wapinski, O. & Chang, H.Y. Long noncoding RNAs and human disease. *Trends Cell Biol.* **21**, 354–361 (2011).
3. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.* **9**, 703–719 (2012).

4. Lee, J.T. Epigenetic regulation by long noncoding RNAs. *Science* **338**, 1435–1439 (2012).
5. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2**, e01749 (2013).
6. Yang, L., Froberg, J.E. & Lee, J.T. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem. Sci.* **39**, 35–43 (2014).
7. Quinn, J.J. & Chang, H.Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**, 47–62 (2016).
8. Wan, Y. *et al.* Genome-wide measurement of RNA folding energies. *Mol. Cell* **48**, 169–181 (2012).
9. Clark, M.B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res.* **22**, 885–898 (2012).
10. Ding, Y. *et al. In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
11. Novikova, I.V., Hennelly, S.P. & Sanbonmatsu, K.Y. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.* **40**, 5034–5051 (2012).
12. Somarowthu, S. *et al.* HOTAIR forms an intricate and modular secondary structure. *Mol. Cell* **58**, 353–361 (2015).
13. Brown, C.J. *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. Brockdorff, N. *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. Lyon, M.F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
16. Galupa, R. & Heard, E. X-chromosome inactivation: new insights into cis and trans regulation. *Curr. Opin. Genet. Dev.* **31**, 57–66 (2015).
17. Nesterova, T.B. *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).
18. Wutz, A., Rasmussen, T.P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.* **30**, 167–174 (2002).
19. Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J. & Lee, J.T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750–756 (2008).
20. Sarma, K. *et al.* ATRX directs binding of PRC2 to Xist RNA and Polycomb targets. *Cell* **159**, 869–883 (2014).
21. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404–416 (2015).
22. McHugh, C.A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).
23. Moindrot, B. *et al.* A pooled shRNA screen identifies Rbm15, Spen, and Wtap as factors required for Xist RNA-mediated silencing. *Cell Rep.* **12**, 562–572 (2015).
24. Monfort, A. *et al.* Identification of Spen as a crucial factor for Xist function through forward genetic screening in haploid embryonic stem cells. *Cell Rep.* **12**, 554–561 (2015).
25. Chen, C.K. *et al.* Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science* **354**, 468–472 (2016).
26. Duszczyk, M.M., Zanier, K. & Sattler, M. 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).
27. Maenner, S. *et al.* 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol.* **8**, e1000276 (2010).
28. Duszczyk, M.M., Wutz, A., Rybin, V. & Sattler, M. The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization. *RNA* **17**, 1973–1982 (2011).
29. Fang, R., Moss, W.N., Rutenberg-Schoenberg, M. & Simon, M.D. Probing Xist RNA structure in cells using Targeted Structure-Seq. *PLoS Genet.* **11**, e1005668 (2015).
30. Lu, Z. *et al.* RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**, 1267–1279 (2016).
31. Ramachandran, S., Ding, F., Weeks, K.M. & Dokholyan, N.V. Statistical analysis of SHAPE-directed RNA secondary structure modeling. *Biochemistry* **52**, 596–599 (2013).
32. Mathews, D.H. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10**, 1178–1190 (2004).
33. Takamoto, K. *et al.* Principles of RNA compaction: insights from the equilibrium folding pathway of the P4-P6 RNA domain in monovalent cations. *J. Mol. Biol.* **343**, 1195–1206 (2004).
34. Pyle, A.M., Fedorova, O. & Waldsich, C. Folding of group II introns: a model system for large, multidomain RNAs? *Trends Biochem. Sci.* **32**, 138–145 (2007).
35. Fernández-Luna, M.T. & Miranda-Ríos, J. Riboswitch folding: one at a time and step by step. *RNA Biol.* **5**, 20–23 (2008).
36. Su, L.J., Brenowitz, M. & Pyle, A.M. An alternative route for the folding of large RNAs: apparent two-state folding by a group II intron ribozyme. *J. Mol. Biol.* **334**, 639–652 (2003).
37. Novikova, I.V., Dharap, A., Hennelly, S.P. & Sanbonmatsu, K.Y. 3S: shotgun secondary structure determination of long non-coding RNAs. *Methods* **63**, 170–177 (2013).
38. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
39. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
40. Harris, M.E. & Christian, E.L. RNA crosslinking methods. *Methods Enzymol.* **468**, 127–146 (2009).
41. Popenda, M. *et al.* Automated 3D structure composition for large RNAs. *Nucleic Acids Res.* **40**, e112 (2012).
42. da Rocha, S.T. *et al.* Jarid2 is implicated in the initial Xist-induced targeting of PRC2 to the inactive X chromosome. *Mol. Cell* **53**, 301–316 (2014).
43. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606–1619 (2000).
44. Brown, P.H. & Schuck, P. Macromolecular size-and-shape distributions by sedimentation velocity analytical ultracentrifugation. *Biophys. J.* **90**, 4651–4661 (2006).

Acknowledgments

We acknowledge M. Simon (Yale University) for sharing the *in vivo* DMS probing data of Xist and E. Jagdmann for the synthesis of 1M7. We thank T. Dickey, C. Zhao, O. Fedorova, and all other members of the Pyle laboratory for constructive discussion and critical reading of the manuscript. This project was supported by the National Institutes of Health (R01GM50313). A.M.P. is supported as an Investigator, and F.L. is supported as a Postdoctoral Fellow, of the Howard Hughes Medical Institute.

Author contributions

F.L. and A.M.P. designed the project. F.L. performed the SV-AUC, SEC, SHAPE and DMS probing, and UV-cross-linking experiments and analyzed the data obtained from the aforementioned experiments; F.L. performed the phylogenetic studies; S.S. conducted jackknife resampling, Shannon entropy, and bootstrapping analyses, and performed 3D modeling experiments. F.L., S.S., and A.M.P. wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Any supplementary information, chemical compound information and source data are available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to A.M.P.

ONLINE METHODS

RNA synthesis and purification. *DNA templates.* Four different constructs of RepA were used in this work. A plasmid containing the full-length RepA sequence (GI 210076757) was purchased from Addgene (pCMV-Xist-PA, cat. no. 26760). The RepA sequence and fragments F1 (1–419), F2 (420–987), F3 (988–1630), F4 (370–598), F5 (681–860), and F6 (871–1098) were amplified with PCR and cloned into the pBlueScript vector immediately downstream of the T7 promoter and upstream of an XbaI (constructs RepA and F3) or BamHI (constructs F1, F2, F4, F5, and F6) restriction site.

In vitro transcription. Plasmids were linearized with the appropriate restriction enzyme (NEB, high fidelity), extracted twice with phenol–chloroform, precipitated with ethanol, washed with 70% ethanol, and dissolved in Tris-EDTA buffer (10 mM Tris-HCl, pH 8.0, and 1 mM Na-EDTA). *In vitro* transcription was carried out with 7 μg/mL template DNA in 15 mM $MgCl_2$, 40 mM Tris-HCl, pH 8.0, 2 mM spermidine, 10 mM NaCl, and 0.01% Triton X-100. Reactions were supplemented with 400 U/mL of RNase inhibitor (Roche) and T7 RNA polymerase, and incubated for 2 h at 37 °C. After transcription, 1.2 mM $CaCl_2$ and DNase (20 U/mL; Ambion) were added sequentially and incubated for 30 min at 37 °C. Finally, proteinase K (0.3 mg/mL; Ambion) was added and incubated for 30 min at 37 °C.

Purification by centrifugal filters and size-exclusion chromatography. Immediately after transcription, reactions were diluted five times with HEK buffer (25 mM K-HEPES, pH 7.0, 0.1 mM Na-EDTA, and 150 mM KCl) and applied to Amicon Ultra-0.5 (Millipore) centrifugal filters (30-, 50-, or 100-kDa molecular-weight cutoff were selected on the basis of the size of the RNA transcript). Four consecutive filtrations were performed at 5,400 r.p.m. for four minutes at room temperature.

After purification with the centrifugal filters, SEC was performed at room temperature with HEK buffer as the running buffer. Fragment constructs F4–F6 were purified with Superdex 200 10/300 GL (GE Healthcare). The other constructs were purified with Tricorn columns (GE Healthcare) that were self-packed with Sephacryl S400 (for constructs F1–F3), or S500 (for full-length RepA). Throughout the entire duration of these experiments, RNA transcripts were purified in the aforementioned manner and stored only at room temperature.

Sedimentation velocity analytical ultracentrifugation. SV-AUC experiments were performed with a Beckman XL-1 centrifuge with An-60 Ti rotor (Beckman Coulter). Before centrifugation, RNA was supplemented with 25 mM K-HEPES, pH 7.0, 150 mM KCl, 0.1 mM Na-EDTA, and appropriate $MgCl_2$ concentrations, as indicated in the results section, and incubated for 45 min at 37 °C. RNA concentrations were adjusted to obtain an initial absorption value of 0.4 at 260 nm on the instrument. All experiments were performed at 20 °C at 25,000 r.p.m. and independently repeated once. Data were analyzed with the continuous $c(s)$ distribution model, as implemented in Sedfit[43,44]. Hydrodynamic radii (R_h) were calculated under the assumption of a partial specific volume of 0.53 cm^3/g and a hydration of 0.59 g/g in Sedfit[36]. Hill plots were drawn for each set of data in Prism (GraphPad).

Notably, the homogeneity of RNA samples is evaluated not on the basis of the shape of the sedimentation curve itself but by mathematically fitting the calculated molecule radii to the Hill equation to obtain a value of $K_{1/2\,Mg}$. At the Mg^{2+} concentration indicated by a particular $K_{1/2\,Mg}$ value, 50% of the RNA molecules are compacted. Therefore, we chose 15 mM Mg^{2+} (more than three times the $K_{1/2\,Mg}$ value) to obtain a completely compact and monodisperse form of RepA.

Chemical probing. Before the chemical probing experiment, freshly purified RNA (20 pmol) was supplemented with HEMK buffer (25 mM K-HEPES, pH 7.0, 0.1 mM Na-EDTA, 150 mM KCl, and 15 mM $MgCl_2$), and incubated at 37 °C for 45 min.

Chemical-reagent titration. We performed reagent titration for both SHAPE and DMS probing before the actual structural characterization experiments. A range of final 1M7 or DMS concentrations (1–10 mM of 1M7 or 0.038–0.15% of DMS) was tested in the titration experiments. The lowest concentration of 1M7 or DMS that produced the saturated modification signals was chosen for the probing experiments.

Selective 2′-hydroxyl acylation analyzed by primer extension. The folded-RNA sample was divided equally into two tubes. The positive reaction was initiated by the addition of 1M7 (ref. 45), which was synthetized in house and dissolved in anhydrous DMSO to a final concentration of 2.5 mM, and the negative-control reaction was initiated by an equal amount of pure DMSO. Samples were incubated for 5 min at 37 °C, precipitated with pure ethanol, and washed twice with 70% ethanol. Pellets were resuspended in HEK buffer. Reverse transcription (RT) was performed with 1 pmol of RNA sample according to a previously described protocol[12]. SHAPE reactivities of nucleotides within full-length RepA and fragments F1–F6 were measured under the same reaction conditions.

Chemical probing by DMS. Methylation with DMS (Sigma-Aldrich) was conducted at room temperature for 10 min. Before the reaction, 1 μl of DMS was diluted on ice with 129 μl ethanol and used as a 10× reaction-buffer stock. For control samples, corresponding amounts of pure ethanol were used. Reactions were stopped by the addition of 54.4 μl stop mix (5% 2-mercaptoethanol in ethanol). Samples were prepared for RT as previously mentioned.

Sequencing and structure mapping by capillary electrophoresis. Synthesis of fluorescent RT primers, preparation of sequencing ladders, and capillary electrophoresis were performed as previously described[12].

Data processing, normalization, and error assessment. All capillary data sets were analyzed with ShapeFinder, as previously described[46]. For secondary-structure prediction, the SHAPE reactivity profiles of all nucleotides were obtained by subtracting the peak areas of background (−) from the peak areas of the corresponding (+) reactions, and the data were then normalized, as previously described[47]. DMS probing data were processed in the same manner, except that the DMS reactivities of adenosines and cytosines were normalized separately, owing to their inherent differences in reactivity[48]. All the SHAPE and DMS experiments in this study were independently repeated three times, on three different RNA samples. Reproducibility among triplicates was measured with Pearson's correlation coefficient (r_p).

For the 3S shotgun approach, normalized SHAPE reactivity profiles were obtained with the peak areas of (+) reactions without subtraction of the background (−). Nucleotides with high peak areas in the background were identified as artificial reverse transcriptase stops and were eliminated from the analysis. The significance of structural similarities between each fragment and the corresponding region in full-length RepA was evaluated with Pearson's correlation coefficient. The standard error in correlation coefficients was calculated with the 'bootstrap' function in MATLAB (Mathworks), with 1,000 randomly generated data sets and removal of 10% of the nucleotides.

Structure determination. To generate the RepA secondary-structure map with the software RNAStructure (http://rna.urmc.rochester.edu/RNAstructureWeb/), SHAPE reactivity was used to provide pseudo energy constraints[49]. Resulting structures were manually evaluated to determine their match with DMS probing data. Resampling analysis and confidence estimation[31] were performed with MATLAB (Mathworks). A total of 1,000 'mock data sets' were generated by random removal of 10% of the nucleotides and setting them as 'no data'. All the mock data sets were then used as pseudo constraints to predict secondary structures with RNAStructure.

Conservation and covariance analysis. Multiple sequence alignments of 56 mammalian sequences comprising the full or partial RepA/Xist gene were downloaded from UCSC Genome Browser[38]. Covariance analysis was performed with Infernal 1.1 (ref. 39) in the following manner: first, a covariance model was built with *cmbuild* (Infernal 1.1), and a multiple sequence alignment of ten homologous sequences (sequence similarity range 95–75%) including murine Xist. Second, the covariance model was calibrated with *cmcalibrate*, and this was followed by a homolog search with *cmsearch* on all the downloaded sequences. Covariance in the resulting alignment was calculated with R2R[50] with 15% tolerance for noncanonical base pairs.

UV cross-linking. Before cross-linking, full-length RepA was internally labeled by *in vitro* transcription in the presence of [α-^{32}P]UTP, natively purified with Amicon filtration (100-kDa molecular-weight cutoff) and folded by incubation

doi:10.1038/nchembio.2272

in HEMK buffer (25 mM K-HEPES, pH 7.0, 0.1 mM Na-EDTA, 150 mM KCl and appropriate $MgCl_2$ concentrations, as indicated below) at 37 °C for 45 min. Analytical UV cross-linking was carried out with 20 nM RNA in 30 µl of HEMK buffer (0–15 mM $MgCl_2$ concentrations). At $MgCl_2$ concentrations higher than 10 mM, poor resolution of corresponding RNA bands was observed on the gel. In preparation for downstream primer extension, cross-linking reactions were conducted on samples of 200 nM RNA in 450 µl HEMK buffer (5 mM $MgCl_2$). Samples were distributed as drops of 30 µl each on a 96-well-plate lid (Corning Costar) and then exposed to short-wavelength UV light (UVP Handheld UV lamp, 6 W) at a distance of 15 cm for 10 min. Immediately after irradiation, samples were mixed with equal volumes of 2× urea loading buffer (22.6 M urea, 0.16% (w/v) xylene cyanol/bromophenol blue, 16% (w/v) sucrose, 80 mM Tris-HCl, pH 7.5, and 1.6 mM EDTA, pH 8.0). Cross-linked RNAs were analyzed on a 4% polyacrylamide gel (29:1 acrylamide/bisacrylamide ratio) containing 8.3 M urea and TBE (90 mM Tris, 90 mM boric acid, and 2 mM EDTA, pH 8.2).

Primer-extension mapping of intramolecular RNA cross-links. To map cross-linking sites, labeled cross-linked RNA was isolated from gel slices with the 'crush and soak' method and then precipitated with ethanol. The cross-linked nucleotide residues were mapped by primer extension with avian myeloblastosis virus (AMV) reverse transcriptase (Thermo Fisher Scientific), with 16 different DNA primers (5′-end-labeled with [γ-^{32}P]ATP) complementary to different regions of RepA (**Supplementary Table 1**). For the primer-extension assay, template RNA (i.e., RNA containing different RNA cross-links or controls) was heated to 95 °C for 1 min and then snap cooled on ice. DNA primers were mixed with the template, and the mixture was allowed to incubate on ice for 10 min. The template–primer complexes were then incubated with reverse transcriptase AMV and appropriate buffer and supplements that were provided along with the enzyme (Thermo Fisher Scientific) at 46 °C for 50 min. Dideoxy-sequencing ladders were obtained with a cycle sequencing kit (Affymetrix) and plasmids containing the full-length RepA sequence. RT products, along with corresponding sequencing ladders, were analyzed on an 8% polyacrylamide gel (29:1 acrylamide/bisacrylamide) containing 8.3 M urea and TBE (90 mM Tris, 90 mM boric acid, and 2 mM EDTA, pH 8.2).

RNA modeling. A tertiary-structure model of subdomains in D2 and D3 was constructed with RNAComposer[41] with the secondary structure mapped here and the cross-links that we identified as distance constraints (6 ± 2 Å). The input consisted of the following nucleotides: 494–508 and 632–776 in D2, and nucleotides 1272–1304 and 1349–1406 in D3. All the remaining nucleotides in D2 and D3 were removed, and the two subdomains were connected with a poly(A) linker that was 20 nt in length (of note, increasing the linker length to 30 nt or 40 nt did not affect the final model). Modeling was performed with 'batch mode', and the final model was chosen such that the distance between cross-linked nucleotides was <6 Å. Further, for comparison with the final model, a model of subdomain D2 was built in the absence of D3 with only secondary structure as the input (without any distance constraints).

Shannon entropy calculation. The SHAPE-directed base-pairing probabilities identified by RNAStructure's *Partition Function RNA* were used for calculations of Shannon entropy. The Shannon entropy of each nucleotide was calculated as previously described[32]:

$$S_i = -\sum_j P_{i,j} \log P_{i,j}$$

where S_i is the entropy of nucleotide I, and $P_{i,j}$ is the probability of nucleotides i and j base-pairing (which is the probability of nucleotide i being unpaired when $i = j$).

Data availability. All chemical probing data and MATLAB scripts for data analyses are available upon request.

45. Mortimer, S.A. & Weeks, K.M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129**, 4144–4145 (2007).
46. Vasa, S.M., Guex, N., Wilkinson, K.A., Weeks, K.M. & Giddings, M.C. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* **14**, 1979–1990 (2008).
47. McGinnis, J.L., Duncan, C.D. & Weeks, K.M. High-throughput SHAPE and hydroxyl radical analysis of RNA structure and ribonucleoprotein assembly. *Methods Enzymol.* **468**, 67–89 (2009).
48. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J.S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).
49. Low, J.T. & Weeks, K.M. SHAPE-directed RNA secondary structure prediction. *Methods* **52**, 150–158 (2010).
50. Weinberg, Z. & Breaker, R.R. R2R: software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics* **12**, 3 (2011).