Chapter 2

Visualization of lncRNA and mRNA Structure Models Within the Integrative Genomics Viewer

Steven Busan and Kevin M. Weeks

Abstract

Every class of RNA forms base-paired structures that impact biological functions. Chemical probing of RNA structure, especially with the advent of strategies such as SHAPE-MaP, vastly expands the scale and quantitative accuracy over which RNA structure can be examined. These methods have enabled large-scale structural studies of mRNAs and lncRNAs, but the length and complexity of these RNAs makes interpretation of the data challenging. We have created modules available through the open-source *Integrative Genomics Viewer* (*IGV*) for straightforward visualization of RNA structures along with complementary experimental data. Here we present detailed and stepwise strategies for exploring and visualizing complex RNA structures in *IGV*. Individuals can use these instructions and supplied sample data to become adept at using *IGV* to visualize RNA structure models in conjunction with useful allied information.

Key words RNA structure, lncRNA, mRNA, Integrative Genomics Viewer, SHAPE-MaP, Base pairing

1 Introduction

RNA, like DNA, forms base-paired structures, as was first conclusively demonstrated for a simple synthetic helix in 1956 [1] and for tRNA in 1974 [2]. The secondary structures of RNAs of every class (including tRNA, rRNA, sRNA, miRNA, mRNA, and lncRNA) have been implicated in the biological functions of these RNAs [3–5]. Chemical or enzymatic probing of RNA in conjunction with thermodynamically-informed structure modeling has a long and successful history of defining structure models for RNAs not amenable to crystallization and for RNAs under biologically relevant or experimentally varied solution and cellular conditions [6]. Advances in RNA structure probing technologies such as SHAPE-MaP have facilitated studies of complex mRNAs [7], lncRNAs [8], and viral RNAs [9, 10]. These RNA molecules are often thousands of nucleotides in length, presenting notable challenges in data visualization and interpretability.

Haiming Cao (ed.), *Functional Analysis of Long Non-Coding RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 2254, https://doi.org/10.1007/978-1-0716-1158-6_2,

To enable efficient examination of the structures of long RNAs, we have created visualization modules for the *Integrative Genomics Viewer* (*IGV*). *IGV* is cross-platform and open-source software that supports visualization of diverse experimental data, especially from studies using arrays and high-throughput sequencing readout strategies [11]. *IGV* was developed to flexibly display genomic, clinical, and experimental data with an emphasis on integrative and interactive analyses. The software allows for visual comparison of many types of experiments and is quite responsive to user interaction.

We previously added several functionalities to *IGV* that enable exploration of RNA structure models and base-pairing probabilities, scaling easily from visualization of the entirety of long RNAs to focused examination of individual helices [12]. Base pairs are conveniently rendered as arcs. Here we present stepwise instructions and general recommendations for visualizing RNA structure models and associated data in *IGV* (*see* **Note 1**). We provide three example datasets derived from recent studies in the Supporting Information. Researchers can use these instructions and the sample data to quickly and efficiently become adept at interrogating RNA structural information in conjunction with a variety of complementary information. Due to the wide availability of next-generation sequencing, SHAPE-MaP probing strategies can be readily employed by diverse nonexpert laboratories. The visualization tools described here facilitate making RNA structure analysis a routine component of examining diverse biological systems.

2 Methods

2.1 Collect Files in Appropriate Formats

Download and extract provided "busan_rna_igv_vis_2019_SI.zip" to use sample data files with this tutorial *or* prepare files from your own sources.

1. Transcript sequence (required).
   (a) Text file with an .fa extension in FASTA format.
   (b) The first line of the file should begin with the ">" (greater-than) character, followed directly by a sequence name or ID with no special characters or spaces.
   (c) Remaining lines should contain the nucleotide sequence without spaces.
2. Chemical reactivity profiles (use either of the following two file formats).
   (a) .shape file: This is a tab-delimited text file with two columns. First column is nucleotide position, starting with 1. Second column is normalized SHAPE reactivity data values; no-data positions are set to −999.

(b) .map file:

- Same format as .shape, but with two additional columns. Third column is standard error, fourth column is nucleotide sequence.
- *ShapeMapper2* software outputs .map files containing chemical probing reactivities calculated using mutational profiling. *ShapeMapper2* and associated documentation is available at: https://github.com/Weeks-UNC/shapemapper2.

3. Base-pairing (secondary) structure model and/or estimated base-pairing probabilities.
   (a) A commonly used format for defining a base-pairing model has the extension .ct. These files are produced by the *Fold* module of *RNAstructure*.
   (b) Dot-bracket (.db, .dbn) files are also supported, most commonly used for small hand-edited structures or used alongside multiple sequence alignments in other software packages.
   (c) Pairing probabilities are calculated by the *partition* and *ProbabilityPlot* modules of *RNAstructure*. These files have the extension .dp.
   (d) For file format reference, *see* https://software.broadinstitute.org/software/igv/RNAsecStructure
   (e) Generation of RNA structure model files is not covered in detail here, since this method is focused on graphical exploration. For long RNAs, we recommend using *Superfold* (available at https://github.com/Weeks-UNC/Superfold), which automates the process of performing structure modeling over computationally manageable windows and merging the resulting structures. *Superfold* accepts a .map file as input and produces both a .ct file, containing a single predicted minimum free energy structure, and a .dp file, containing estimated base pairing probabilities.
4. Annotations such as gene coding regions, repeat sequences, or sites of known function (optional).
   (a) The .gff3 file format is convenient for most uses and is readily hand-edited. See the included examples in "busan_rna_igv_vis_2019_SI.zip" and further documentation at https://software.broadinstitute.org/software/igv/GFF.
   (b) Important: Ensure that the names listed in the first column of the .gff3 file match the name given in the first line

of the FASTA file and not the FASTA filename or other text.

5. One or more linear profiles from complementary experiments or computational analyses (this list is not exhaustive, *see* **Note 1**). These data are not required, but can substantially enrich RNA structure analyses.
   (a) Protein-binding enrichment data (e.g., CLIP or RIP data).
   (b) GC-content median over fixed windows.
   (c) SHAPE reactivity median over fixed windows.
   (d) Estimated per-nucleotide or median Shannon entropy over fixed windows.
   (e) Common file formats are .wig, .bedgraph, and .tdf (*see* https://software.broadinstitute.org/software/igv/FileFormats).
   (f) Important: Ensure that the sequence names listed in the first column of a .wig or .bedgraph file match the name given in the first line of the FASTA file (not including the ">" character) and not the FASTA filename.

2.2 Load and Import Files into IGV

1. Download *IGV* (available at https://software.broadinstitute.org/software/igv/download) and launch.
2. Load nucleotide sequence.
   (a) Click "Genomes" in menu bar and select "Load Genome from File".
   (b) Select FASTA file and click "Open".
   (c) Select "E_coli/sequence.fa" if using the example dataset or select your own .fa file.
3. Load SHAPE reactivity profiles, base-pairing probabilities, and annotations.
   (a) Click "File" in menu bar and select "Load from File".
   (b) Select "E_coli/SHAPE_reactivity.map", "E_coli/base_pairing_probability.dp", and "E_coli/gene_annotations.gff3" if following along with example dataset or select your own files, and click "Open".
   (c) Click "Continue" in any popup dialog boxes that appear (*see* **Note 2**).

2.3 Adjust Track Display

Steps listed here are optional and specific values given are suggestions. Users should adjust settings for comfortable display for their particular screen size, platform, and dataset.

1. Set view preferences.
   (a) Click "View," "Preferences," "General."

(b) Check "Display all tracks in a single panel."

(c) Uncheck "Show attributes panel."

2. Rename tracks.

(a) Right-click track, select "Rename track," and enter a descriptive name.

(b) For the example dataset, replace "SHAPE_reactivity.shape.wig" with "SHAPE reactivity" and "base_pairing_probability.dp.bp" with "Pairing probability."

3. If working with a higher resolution display, consider adjusting track name size settings.

(a) Shift-left-click and drag track names to select all tracks. Right click track or track names, select "Change font size," and increase the value to 16 (*see* **Note 3**).

(b) If track names are cut off or abbreviated: click "View," "Set Name Panel Width" and set to a larger value.

(c) The default font size can be changed by clicking "View," "Preferences," "General," then clicking "Change" next to "Default font." This will only affect tracks or files loaded in the future.

4. Reorder tracks by left-clicking and dragging track names.

(a) Drag gene annotations track directly above other tracks.

(b) Drag SHAPE reactivity profile above base pairing probability arcs.

(c) If zoomed in far enough to see individual nucleotide identities, it can be useful to move the sequence track directly above base pairing arcs to visualize complementary pairs and the sequences of unpaired regions.

5. Adjust SHAPE profile track range.

(a) Right-click reactivity profile track, and select "Set Data Range."

(b) Set "Min" and "Mid" values to 0 and "Max" value to 3.

6. Widen pairing probability arc track.

(a) Right-click on the arc track, select "Change Track Height."

(b) Set to a larger value such as 100.

2.4 Examine Functional Sites in Example Biologically Important RNAs

2.4.1 E. coli *mRNA Gene Translation Start Sites*

The provided example of an *E. coli* transcript is notable in that it contains two nonribosomal protein genes *rimM* and *trmD* (encoding a ribosome maturation factor and a tRNA methyltransferase, respectively) that are located between two ribosomal protein-coding genes *rpsP* and *rplS* (encoding S16 and L19, respectively) [13] (Fig. 1a). The ribosomal proteins encoded by *rpsP* and *rplS* are translated at high levels; in contrast, the *rimM* and *trmD* gene products are translated at lower levels. In addition, the translation rates of *rimM* and *trmS* are largely uncoupled from those of the surrounding genes [14]. Examining the structures around the translation start sites of each gene provides clues to explain these differences.

1. Zoom in on the start codon regions of *rplS* and *rimM*. Use any of the following:
   (a) Click and drag to select range in ruler (as shown in Fig. 1a).
   (b) Click the "+" button in the upper right area of the toolbar several times and drag track window to scroll.
   (c) Double-click on an annotation graphic several times.
   (d) Enter a gene, annotation name, or numeric range in text box.

Note the differing structural contexts of the translation start sites of *rplS* and *rimM*. In particular, the region surrounding the AUG codon in *rplS* is unstructured, evidenced by high SHAPE reactivities and lack of highly probable base pair arcs in structure models (Fig. 1b). This lack of structure near the start codon likely provides a high ribosome accessibility, allowing translation initiation in the absence of a Shine–Dalgarno sequence [15].

2.4.2 *Murine LHR mRNA Sequence Motifs*

In mice, ZFP36L2, a zinc finger protein, regulates expression of the luteinizing hormone receptor (*LHR*) mRNA during oocyte maturation [16]. ZFP36L2 is a member of a class of zinc finger–containing proteins that bind RNA targets containing the sequence motif "AUUUA," termed adenine–uridine-rich elements (AREs) [17]. Surprisingly, gel-shift assays revealed that ZFP36L2 bound only one of the three AREs present within the *LHR* 3′ untranslated region [16], raising the intriguing possibility that the RNA structural context of these sequence motifs influences protein binding.

1. Examine the nucleotide sequence of a structure model (data from ref. 18).
   (a) Load the provided *LHR* sequence, SHAPE reactivity data, structure model, and annotations from supporting files folder "LHR", as in Subheading 2.2 (*see* Fig. 2a).

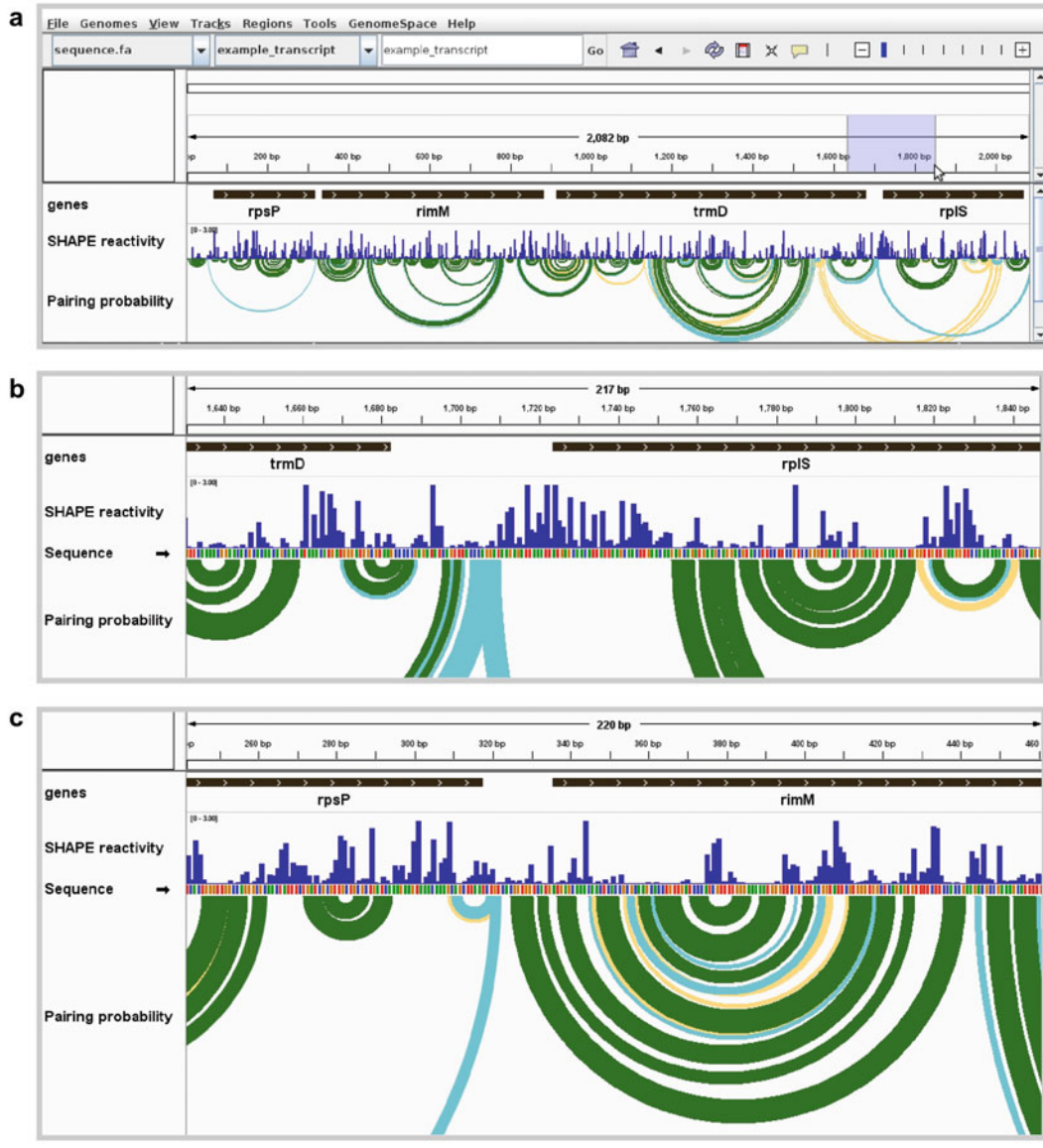

Fig. 1 Visualization of *E. coli* mRNA translation start sites. (**a**) Full view of an *E. coli* polycistronic transcript showing gene boundaries, SHAPE reactivity profile, and modeled base pairing probabilities. Pairing probability is indicated by arc color: green, >80% probability; blue, 30–80%; yellow, 10–30%. High SHAPE reactivities in blue shaded region are indicative of an unstructured region around the *rplS* translation start site. (**b**) Zoomed view of unstructured region surrounding the start codon of *rplS*, showing highly SHAPE-reactive positions modeled as unpaired (corresponding to no arcs indicative of base pairing). (**c**) Zoomed view of the start codon and surrounding region of *rimM*, showing positions with low SHAPE reactivities and corresponding well-determined base-pairing structure model. (Data from ref. 13)

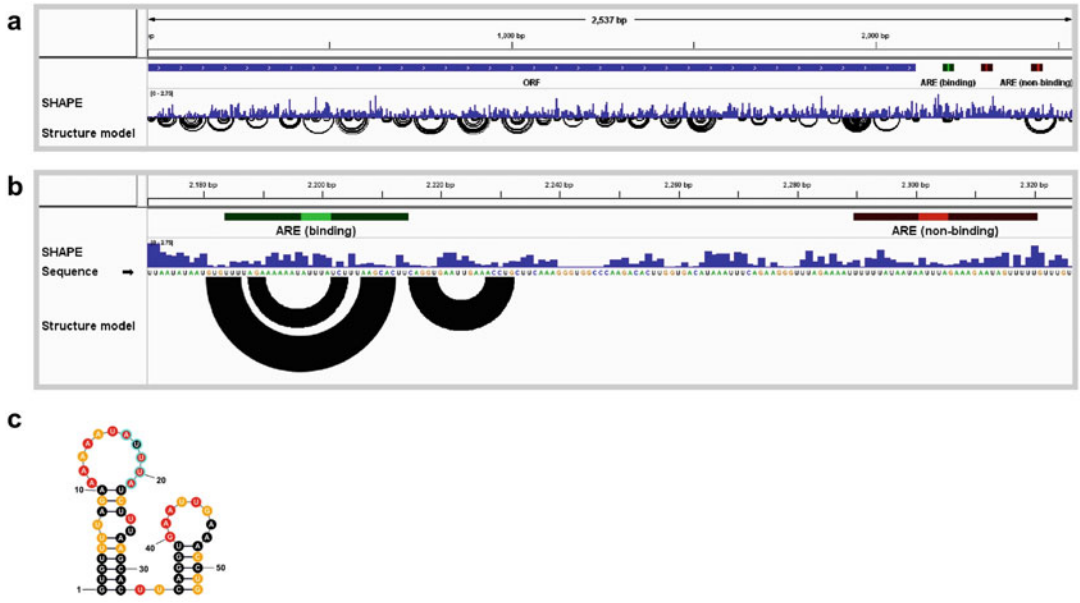

Fig. 2 Visualization of structures around AREs in *LHR* mRNA. (**a**) Full view of the *LHR* transcript showing SHAPE reactivity profile and base-pairing structure model. Annotations include an open reading frame (blue bar labeled ORF) and three AREs (highlighted in green and red). (**b**) Zoomed view of the region of two AREs, one functional and one nonfunctional based on binding assays with ZFP36L2 protein [16]. The arcs in the Structure model track, which indicate base pairs, suggest that the upstream ARE is highly structured, whereas the downstream ARE is not. (**c**) ARE structure rendered as a planar graph using the *StructureEditor* component of *RNAstructure*. SHAPE reactivities are indicated by color: red, reactivity >0.85; orange, 0.4–0.84; black, <0.4. Core ARE sequence highlighted with cyan outline. (Data from ref. 18)

(b) Zoom in on the region of the annotated AREs until the nucleotide sequence becomes visible (*see* Fig. 2b).

2. Visualize the highly structured element in Fig. 2b as a traditional planar graph (optional). This requires third-party software such as the *StructureEditor* component of *RNAstructure* available at https://rna.urmc.rochester.edu/RNAstructure.html (*see* **Note 4**).

SHAPE reactivities and structure modeling suggested that the functional motif is located in the context of a hairpin loop (Fig. 3c) and that RNA structure influences the binding affinity for ZFP36L2. These RNA structure-based hypotheses were supported through extensive mutagenesis studies [18].

2.4.3 *Murine Xist lncRNA Repeat Elements*

The *Xist* long noncoding RNA localizes to the nucleus and mediates the formation of a chromatin-modifying protein complex that silences gene expression on the female X chromosome [19]. *Xist* contains multiple repetitive sequence regions, some of which are implicated in *Xist* localization, interaction with members of the protein complex, and X-chromosome inactivation [20–22].

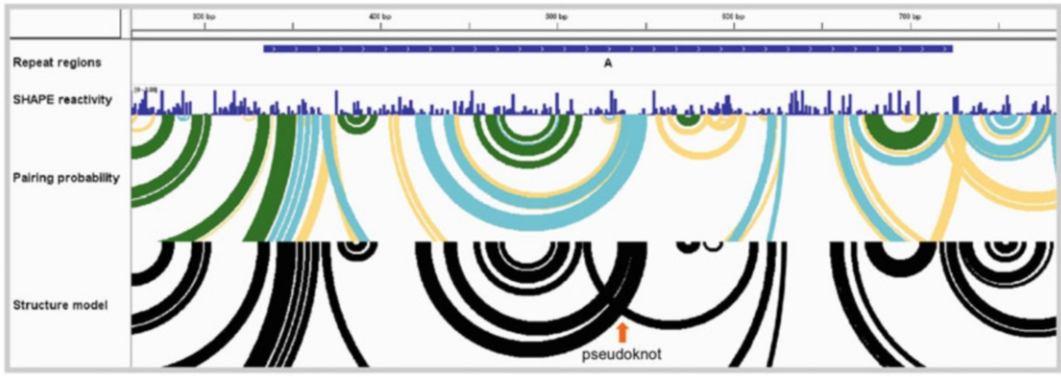

Fig. 3 Murine *Xist* long noncoding RNA repeat region. View of one *Xist* repeat region (repeat A) showing SHAPE reactivity profile, base-pairing probabilities, and base-pairing structure model. Pairing probability is indicated by arc color: green, >80% probability; blue, 30–80%; yellow, 10–30%. The abundance of medium- and low-probability base pairing arcs (in blue and yellow) suggests that no single structure (including the minimum free energy structure shown with black arcs) predominates in this region and that, instead, an ensemble of RNA structural states is present. (Data from ref. 8)

For functionally important regions of an RNA, there is often a single, thermodynamically stable secondary structure. Examples of such well-defined secondary structures include ligand-bound riboswitches, the bacterial 16S rRNA, or the stable base-paired structures overlapping the *rimM* gene (Fig. 1; green arcs). In contrast, some RNAs instead adopt a family (or ensemble) of structures. Modeling RNA structural ensembles remains an important experimental and computational frontier, but the visualization of estimated pairing probabilities is a useful approach that begins to address variability within populations of folded RNA molecules.

1. Examine base-pairing probabilities (data from ref. 8).
   (a) Load the provided murine *Xist* sequence, SHAPE reactivity data, structure model, base pairing probabilities, and repeat region annotations from supporting files folder "Xist," as in Subheading 2.2.
   (b) Zoom in on repeat A as in Fig. 3.

Although the minimum free energy structure model, by definition, displays a single secondary structure for *Xist* repeat region A (Fig. 3; black arcs), the pairing probability arcs show multiple overlapping low- and medium-probability helices (Fig. 3; blue and yellow arcs). These data support a model in which this region of the *Xist* RNA does not have a well-defined secondary structure overall. A possible pseudoknotted structure is evident in the "Structure model" track as overlapping arcs, highlighted with an orange arrow in Fig. 3.

3 Notes

1. This brief report is focused on the visualization of RNA structure probing data and structure models and is not a comprehensive guide to *IGV*. For general guides and documentation to *IGV*, see the following: https://software.broadinstitute.org/software/igv/UserGuide and https://software.broadinstitute.org/software/igv/FileFormats.
2. File conversion popup dialog
   (a) .ct, .map., and .shape file formats do not contain sequence name, strand, or nucleotide offset position. Therefore, upon import, *IGV* converts these files into file formats that contain this information, without overwriting the original input files.
   (b) The default settings in the file conversion popup dialog cover the most common RNA structure exploration scenario, that is, one transcript sequence and chemical reactivity profiles corresponding to this sequence. If examining a short gene-specific primer amplicon (*see* ref. 23) within a larger sequence, users may need to manually adjust the beginning and end positions of reactivity profiles or structures.
3. The *IGV* user interface may appear unusably small on some newer high-resolution displays. Recent Java 11 builds of *IGV* provide support for high-resolution displays (*see* https://software.broadinstitute.org/software/igv/download).
4. The *IGV* modules discussed here visualize secondary structures as arc diagrams. Rendering traditional RNA secondary structure figures (sometimes referred to as airport, planar, or tree diagrams) requires additional software. Commonly used packages include *VARNA*, *Ribosketch*, *RNAstructure StructureEditor*, and *XRNA*. See informal discussion at https://github.com/Weeks-UNC/shapemapper2.

References

1. Rich A, Davies DR (1956) A new two stranded helical structure: polyadenylic acid and polyuridylic acid. J Am Chem Soc 78:3548–3549. https://doi.org/10.1021/ja01595a086
2. Kim SH, Suddath FL, Quigley GJ et al (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. Science 185:435–440. https://doi.org/10.1126/science.185.4149.435
3. Eddy SR (2001) Non–coding RNA genes and the modern RNA world. Nat Rev Genet 2:919–929. https://doi.org/10.1038/35103511
4. Parker BJ, Moltke I, Roth A et al (2011) New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Res 21:1929–1943. https://doi.org/10.1101/gr.112516.110
5. Sonenberg N, Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell 136:731–745. https://doi.org/10.1016/j.cell.2009.01.042
6. Mailler E, Paillart J-C, Marquet R et al (2018) The evolution of RNA structural probing methods: from gels to next-generation

sequencing. Wiley Interdiscip Rev RNA 10: e1518. https://doi.org/10.1002/wrna.1518

7. Corley M, Solem A, Phillips G et al (2017) An RNA structure-mediated, posttranscriptional model of human α-1-antitrypsin expression. Proc Natl Acad Sci U S A 114: E10244–E10253
8. Smola MJ, Christy TW, Inoue K et al (2016) SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. Proc Natl Acad Sci U S A 113:10322–10327
9. Dethoff EA, Boerneke MA, Gokhale NS et al (2018) Pervasive tertiary structure in the dengue virus RNA genome. Proc Natl Acad Sci U S A 115:11513–11518. https://doi.org/10.1073/pnas.1716689115
10. Dadonaite B, Barilaite E, Fodor E et al (2017) The structure of the influenza a virus genome. Nat Microbiol 4(11):1781–1789. https://doi.org/10.1101/236620
11. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754
12. Busan S, Weeks KM (2017) Visualization of RNA structure models within the integrative genomics viewer. RNA 23:1012–1018. https://doi.org/10.1261/rna.060194.116
13. Mustoe AM, Busan S, Rice GM et al (2018) Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing. Cell 173:181–195.e18. https://doi.org/10.1016/j.cell.2018.02.034
14. Wikström PM, Björk GR (1988) Noncoordinate translation-level regulation of ribosomal and nonribosomal protein genes in the *Escherichia coli* trmD operon. J Bacteriol 170:3025–3031. https://doi.org/10.1128/jb.170.7.3025-3031.1988
15. Scharff LB, Childs L, Walther D, Bock R (2011) Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. PLoS Genet 7: e1002155. https://doi.org/10.1371/journal.pgen.1002155
16. Ball CB, Rodriguez KF, Stumpo DJ et al (2014) The RNA-binding protein, ZFP36L2, influences ovulation and oocyte maturation. PLoS One 9:e97324. https://doi.org/10.1371/journal.pone.0097324
17. Lai WS, Carballo E, Thorn JM et al (2000) Interactions of CCCH zinc finger proteins with mRNA. J Biol Chem 275:17827–17837. https://doi.org/10.1074/jbc.m001696200
18. Ball CB, Solem AC, Meganck RM et al (2017) Impact of RNA structure on ZFP36L2 interaction with luteinizing hormone receptor mRNA. RNA 23:1209–1223. https://doi.org/10.1261/rna.060467.116
19. Gendrel A-V, Heard E (2014) Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. Annu Rev Cell Dev Biol 30:561–580. https://doi.org/10.1146/annurev-cellbio-101512-122415
20. Chu C, Zhang QC, da Rocha ST et al (2015) Systematic discovery of Xist RNA binding proteins. Cell 161:404–416. https://doi.org/10.1016/j.cell.2015.03.025
21. Sunwoo H, Colognori D, Froberg JE et al (2017) Repeat E anchors Xist RNA to the inactive X chromosomal compartment through CDKN1A-interacting protein (CIZ1). Proc Natl Acad Sci U S A 114:10654–10659. https://doi.org/10.1073/pnas.1711206114
22. Sarma K, Levasseur P, Aristarkhov A, Lee JT (2010) Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome. Proc Natl Acad Sci U S A 107:22196–22201. https://doi.org/10.1073/pnas.1009785107
23. Smola MJ, Rice GM, Busan S et al (2015) Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat Protoc 10:1643–1669