Bioinformatics, 36(10), 2020, 3072–3076
doi: 10.1093/bioinformatics/btaa080
Advance Access Publication Date: 7 February 2020
Original Paper

Structural bioinformatics

Estimating the power of sequence covariation for detecting conserved RNA structure

Elena Rivas [1,*], Jody Clements[2] and Sean R. Eddy[1,3,4]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA, [2]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA, [3]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA and [4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on December 2, 2019; revised on January 22, 2020; editorial decision on January 28, 2020; accepted on January 29, 2020

Abstract

Pairwise sequence covariations are a signal of conserved RNA secondary structure. We describe a method for distinguishing when lack of covariation signal can be taken as evidence against a conserved RNA structure, as opposed to when a sequence alignment merely has insufficient variation to detect covariations. We find that alignments for several long non-coding RNAs previously shown to lack covariation support do have adequate covariation detection power, providing additional evidence against their proposed conserved structures.

Availability and implementation: The R-scape web server is at eddylab.org/R-scape, with a link to download the source code.

Contact: elenarivas@fas.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Comparative analyses of pairwise covariations in RNA sequence alignments have a successful history in consensus RNA secondary structure prediction, where the existence of a conserved structure is assumed *a priori* (Gutell *et al.*, 1985, 1992; Holley *et al.*, 1965; Michel *et al.*, 2000; Noller *et al.*, 1981; Pace *et al.*, 1989; Williams and Bartel, 1996). A statistically different question arises when covariation analysis is used to infer whether or not a genomic region is constrained by an evolutionarily conserved RNA secondary structure, as evidence for a structure-dependent function. For example, this question arises in analysis of long non-coding RNAs (lncRNAs) of uncertain mechanism or function. For this, one wants to determine if the covariation signal is distinguishable from a null hypothesis of primary sequence conservation patterns alone.

We previously introduced R-scape (RNA Structural Covariation Above Phylogenetic Expectation), a method for evaluating the statistical significance of covariation support for conserved RNA basepairs (Rivas *et al.*, 2017). R-scape analyses found that the covariation evidence for proposed conserved structures of several lncRNAs including HOTAIR (Somarowthu *et al.*, 2015), SRA (Novikova *et al.*, 2012) and the RepA region of Xist (Fang *et al.*, 2015; Maenner *et al.*, 2010) is not statistically significant (Rivas *et al.*, 2017).

Lack of significant covariation signal does not necessarily mean there is no conserved RNA structure. An alignment could merely have too little sequence variation to detect significant covariation (Fig. 1a). To know when an alignment has sufficient variation, we want to estimate the statistical power (the expected sensitivity) of detecting significant covariations. In a 'low-power' alignment, covariation analysis is inconclusive because a conserved RNA secondary structure could be present without inducing sufficient covariation signal. In a high-power alignment, observing no supporting covariations does provide evidence against a conserved structure.

2 Results and discussion

Many details of an alignment affect covariation analysis, but we hypothesized that detection power should depend primarily on the total number of single residue substitutions $s_{i,j}$ in two alignment columns i and j in a proposed consensus pair. We take the sequence phylogeny into account in inferring $s_{i,j}$ by inferring a maximum likelihood tree, using the Fitch parsimony algorithm (Fitch, 1971) to estimate a minimum number of substitutions s_i at each column independently, and taking $s_{i,j} = s_i + s_j$ (Section 3).

We tested this idea using synthetic RNA alignments evolved under simulated pairwise constraints. Figure 1b–d show simulations based on a cobalamin riboswitch alignment (Rfam RF00174) of 430 sequences and 42 annotated consensus basepairs. We choose a random sequence as the root and evolve it down a sub-sampled and rescaled phylogenetic tree, using an evolutionary model that includes basepair substitutions, insertions and deletions (Rivas and Eddy, 2015), to generate synthetic alignments with a desired number of taxa and average percentage identity (Section 3). We repeat this to create synthetic alignments over a wide range of sequence number

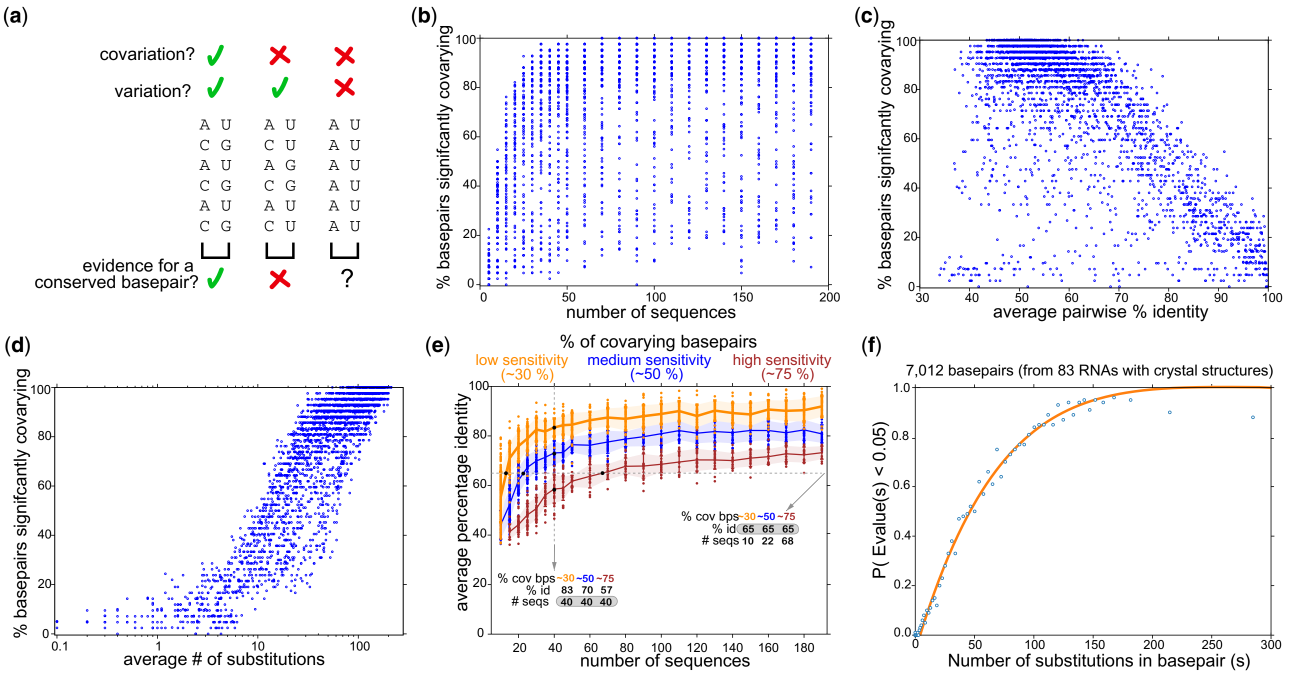


Fig. 1. (**a**) Three different patterns for two alignment columns proposed to form a consensus basepair. (**a, left**) **The two columns have variation and covariation (mutual information is 1 bit). This pattern is consistent with a basepair conserved throughout evolution.** (**a, middle**) The two columns have variation but not covariation (mutual information is 0.0). These two positions are unlikely to form a basepair. (**a, right**) Two columns with no covariation and no variation. This pattern is consistent with an A–U basepair, but there is no evolutionary evidence for it. (**b–d**) Scatter plots of power (% sensitivity) for detecting basepairs in simulated alignments. Each point represents the fraction of 42 consensus basepairs in a simulated Cobalamin alignment detected with an R-scape *E*-value <0.05, as a function of sequence number (b), average pairwise % identity (c) or inferred number of substitutions in two columns s_{ij} (d). (**e**) The same simulated alignments binned by low (yellow ~27 – 32%), medium (blue ~47 – 52%) or high (red ~74 – 76%) sensitivity and scatter plotted, illustrating how detection power increases either by increasing sequence number or sequence diversity. (**f**) Power of covariation as a function of the total number of substitutions. The orange line is the fitted power(*s*) curve. Each blue dot represents the empirical data we fit to: the mean fraction of significantly covarying basepairs and mean s_{ij} in a set of 100 annotated basepairs from Rfam seed alignments, out of 7012 total basepairs ordered by increasing number of substitutions

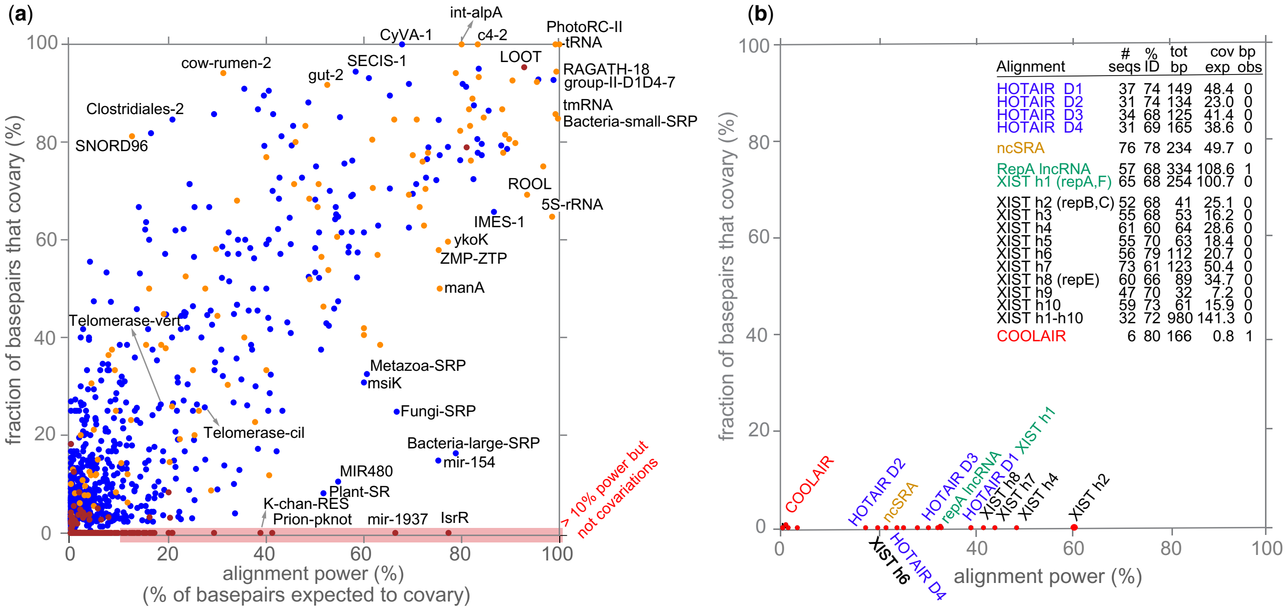
and diversity. We use R-scape on each alignment to determine the number of basepairs with significant covariation support (*E*-value < 0.05). The *E*-value of a given score is the expected number of pairs amongst the total set of tested pairs that are expected to get this score or better, under the null hypothesis of no covariation. Neither the number nor the diversity of sequences in the alignment alone suffices to estimate detection power (Fig. 1b and c), whereas $s_{i,j}$ does have a good relationship to power (Fig. 1d). Using either deeper or more diverse alignments increases $s_{i,j}$ and detection power (Fig. 1e).

We empirically fit the relationship between substitutions and detection power (at a significance threshold of $E < 0.05$) on a dataset of alignments of 87 RNA families in Rfam v14.0 with known 3D structures, consisting of 7012 annotated basepairs (Fig. 1f). These RNA families cover a wide range of different structures and functions. The fitted curve enables estimation of power(*s*), the expected sensitivity for detecting any proposed basepair with *s* total substitutions. For an alignment with *B* total proposed basepairs, the expected fraction of basepairs with significant covariation support is $\frac{1}{B}\sum_{b=1}^{B}\text{power}(s_b)$. We use this number, which we call *alignment power*, to compare covariation support across different alignments with different numbers of proposed conserved basepairs. We define an arbitrary threshold of 10% alignment power to distinguish *low-power* from *high-power* alignments.

In summary: given an input alignment with a proposed structure, first the minimum number of substitutions per column is calculated using the Fitch algorithm. Then, the number of substitutions per basepair (s) is calculated as the sum of substitutions for each position, and the basepair power(s) is estimated using the empirical power curve given in Figure 1f (see Section 3). Finally, the alignment power is defined as the average power for all basepairs in the proposed structure, and the expected number of detected covariations assuming the alignment is structural is given by the total power of all basepairs. This expected number of covariations can be then directly compared with the observed number detected by R-scape (Rivas *et al.*, 2017). Alignments with power but no covariation argue against the existence of a conserved RNA structure.

Only low-power alignments of conserved structural RNAs should lack significant covariation support. We analyzed all 3016 seed alignments for known conserved structural RNAs in Rfam v14.1 and compared the fraction of basepairs with significant covariation support versus estimated alignment power (Fig. 2a; Supplementary Table S1). Many Rfam alignments (66%, 1985/3016) have no statistically significant covariation support for any annotated consensus basepair, and almost all of these (98%, 1945/1985) are low-power alignments. Only 1% (40/3016) are high-power alignments with no significant detected covariations (shaded in red in Fig. 2a; Supplementary Table S2). Rfam, though curated, is a large compendium with a nonzero error rate. Upon examination, we believe these 40 families are enriched for inaccuracies. For example, the miR-1937 family (RF01942) (66% alignment power) is annotated in miRbase (Kozomara *et al.*, 2019) as a tRNA sequence fragment unlikely to be a bona fide miRNA.

Previous analysis of several lncRNAs including HOTAIR (Somarowthu *et al.*, 2015), SRA (Novikova *et al.*, 2012) and the Xist RepA region (Fang *et al.*, 2015; Maenner *et al.*, 2010) found no significant covariation support for their proposed structures, but left open the possibility that the existing alignments lacked sufficient variation (Rivas *et al.*, 2017). We reanalyzed the four HOTAIR lncRNA domain alignments and consensus structures proposed by Somarowthu *et al.* (2015), and the SRA alignment and consensus structure in Novikova *et al.* (2012). All five alignments are high-power, estimated to be able to detect 23–50 significant basepair covariations each (Fig. 2b; Supplementary Table S3). Although the covariation analysis of these lncRNAs has been a subject of disagreement (Rivas and Eddy, 2018; Tavares *et al.*, 2019), these



Alignment	# seqs	% ID	tot bp	cov bp exp	cov bp obs
HOTAIR D1	37	74	149	48.4	0
HOTAIR D2	31	74	134	23.0	0
HOTAIR D3	34	68	125	41.4	0
HOTAIR D4	31	69	165	38.6	0
ncSRA	76	78	234	49.7	0
RepA lncRNA	57	68	334	108.6	1
XIST h1 (repA,F)	65	68	254	100.7	0
XIST h2 (repB,C)	52	68	41	25.1	0
XIST h3	55	68	53	16.2	0
XIST h4	61	60	64	28.6	0
XIST h5	55	70	63	18.4	0
XIST h6	56	79	112	20.7	0
XIST h7	73	61	123	50.4	0
XIST h8 (repE)	60	66	89	34.7	0
XIST h9	47	70	32	7.2	0
XIST h10	59	73	61	15.9	0
XIST h1-h10	32	72	980	141.3	0
COOLAIR	6	80	166	0.8	1

Fig. 2. (**a**) Power of covariation for structural RNAs. Each point represents one of 2209 Rfam families (seed alignments) with at least 10 annotated consensus basepairs, plotting the fraction of annotated basepairs that show a significant covariation signal (R-scape $E < 0.05$) versus 'alignment power', the fraction expected to show a significant signal. Points are color coded by positive predictive value (PPV; the fraction of significantly covarying basepairs that are 'true', i.e. in the annotated RNA consensus structure). Besides false positive base pairs from R-scape, low PPV can also occur because the RNA includes covarying pairs not in the annotated structure (such as pseudoknots, which are not reliably annotated in Rfam), or because of an incorrect annotated structure. PPV Blue are PPV >95%, yellow 50–95% and red <50%. Red shaded region along the bottom indicates alignments with sufficient power, defined as > 10%, but no significant detected covariations. (**b**) Results for HOTAIR, SRA, Xist and COOLAIR lncRNA alignments. Inset table shows details for each alignment, including the total number of annotated basepairs, the expected number that should show significant covariation (i.e. alignment power times total bp) and the number observed with significant covariation. Supplementary Table S3 describes all lncRNA alignments and proposed structures tested

results provide new evidence for the view that HOTAIR and SRA do not have evolutionarily conserved RNA structures.

Xist RNA is perhaps the best studied lncRNA, but it remains unclear whether Xist's role in X dosage compensation depends on any conserved RNA structure, as opposed to its sequence alone. Several different conserved structures have been proposed for the conserved 5′ RepA region of Xist (Fang *et al.*, 2015; Liu *et al.*, 2017; Maenner *et al.*, 2010), two of which are based on covariation analysis of alignments of 10–13 sequences (Fang *et al.*, 2015; Maenner *et al.*, 2010). Although R-scape finds no significant covariation support for the proposed RepA structures, our method finds that these are low-power alignments, so the R-scape covariation analyses are inconclusive (Fig. 2b; Supplementary Table S3).

A conserved structure for the ~1.3 kb Xist RepA lncRNA (including the conserved repeat A and F regions) has been proposed recently (Liu *et al.*, 2017) from a deeper and more diverse alignment of 57 sequences. Although the R2R visualization program used by Liu *et al.* (2017) highlighted many potential covariations, statistical analysis by R-scape identifies only one significant covarying basepair with an *E*-value of 0.005, out of 334 proposed pairs. Our method judges this alignment to be high-power, estimated to be sufficient to detect about 110/334 basepairs.

The repeat A + F region is the most conserved region of Xist, but Xist is a large RNA and it is possible that other Xist regions could show covariation support for conserved RNA structure (Fig. 3a). Starting with the human XIST genomic sequence, we used the *nhmmer* homology search program (Wheeler and Eddy, 2013) to identify 21 regions of significant sequence similarity with mouse Xist (*E*-value $< 10^{-5}$). Eleven regions correspond to insertions of well-studied ancient transposons according to Dfam analysis (Wheeler *et al.*, 2013). For the remaining 10 unique sequence conserved regions, we iteratively built up alignments of homologs from 47 to 65 vertebrate species. All of these are high-power alignments; none show significant covariation support for any basepair (Fig. 2b; Supplementary Table S3). In order to test for long range base pairing, we created a concatenated alignment of all ten XIST homology regions for 32 species. This concatenated alignment also has sufficient power but not covariations are observed.

Experimental evidence from chemical probing and crosslinking has been used in making structure predictions for the HOTAIR, SRA and Xist lncRNAs. However, essentially any RNA, even a random RNA sequence, has some secondary structure. Lack of covariation signal in high-power RNA sequence alignments for these lncRNAs suggests that whatever structure they adopt is not detectably constraining their evolution, and thus may not be relevant for their function.

An important caveat in covariation analysis is that the input sequence alignment is assumed to be reasonably correct. Spurious apparent covariations can be created artifactually by sliding conserved primary sequence regions under proposed stems. We identified an example of this artifact in a proposed conserved structure for COOLAIR, an *Arabidopsis* lncRNA (Hawkes *et al.*, 2016). The COOLAIR alignment is a low-power alignment of only six aligned sequences, yet R-scape identifies six significant covarying basepairs, four of them in one proposed helix (Fig. 3b). Upon inspection, it appears that misalignment introduced artifactual covariations (Fig. 3c). We realigned the COOLAIR sequences using Infernal (Nawrocki and Eddy, 2013) (Section 3), which brought regions of strong primary sequence identity back into alignment (Fig. 3d). The revised COOLAIR alignment is still low-power, and has only one significant supported basepair with a marginal *E*-value of 0.048.

The R-scape software now reports estimated statistical power calculations along with observed pairwise correlations. We expect that one important future use of covariation power analysis is to enable quantitative use of negative information by excluding pairs that are unlikely to be conserved basepairs because they have high-power and no significant covariation.

3 Materials and methods

3.1 Estimating substitution number s_{ij}

Given an input RNA sequence alignment, R-scape infers a maximum likelihood phylogenetic tree using FastTree (version 2.1.10) (Price *et al.*, 2010), then infers a maximum parsimony assignment

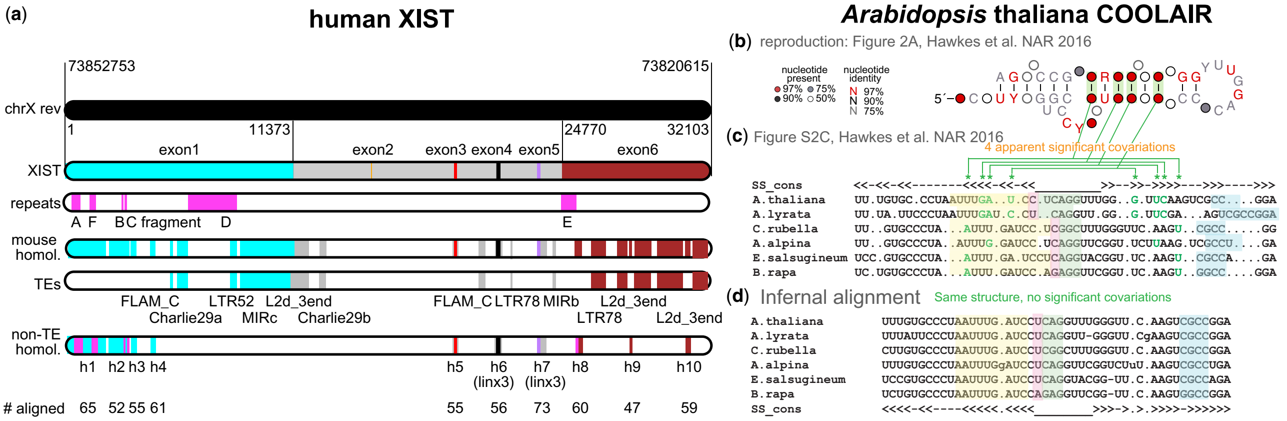

(a)
human XIST
73852753
73820615
chrX rev
1
11373
24770
32103
exon1
exon2
exon3 exon4 exon5
exon6
XIST
repeats
A F
BC fragment
D
E
mouse homol.
TEs
FLAM_C
LTR52
L2d_3end
Charlie29a
MIRc
Charlie29b
FLAM_C LTR78 MIRb
L2d_3end
LTR78
L2d_3end
non-TE homol.
h1 h2 h3 h4
h5
h6 (linx3)
h7 (linx3)
h8
h9
h10
# aligned 65 52 55 61
55 56 73 60 47 59
Arabidopsis thaliana COOLAIR
(b) reproduction: Figure 2A, Hawkes et al. NAR 2016
nucleotide present
● 97% ● 75%
● 90% ○ 50%
nucleotide identity
N 97%
N 90%
N 75%
5´–
(c) Figure S2C, Hawkes et al. NAR 2016
4 apparent significant covariations
SS_cons <<-<<-<<-------<<<<-<<-<<__________>>-->>->>>>--->>>---->>>
A.thaliana UU.UGUGC.CCUAAUUUGA..U.CC.UCAGGUUUGG..G.UUCAAGUCGCC....GGA
A.lyrata UU.UA.UUCCCUAAUUUGAU.C.CU..CAGGUU.GG..G.UUCGA...AGUCGCCGGA
C.rubella UU..GUGCCCUA...AUUU.GAUCC.UCGGCUUUGGGUUC.AAGU..CGCC.....GG
A.alpina UU..GUGCCCUA..AUUUG.GAUCC.UCAGGUUCGGU.UCUUAAG.UCGCCU....GA
E.salsugineum UCC.GUGCCCUA...AUUU.GA.UCCUCAGGUACGGU.UC.AAGU..CGCCA....GA
B.rapa UC.UGUGCCCUA...AUUU.GAUCC.AGAGGUUCGGU.UC.AAGU..GGCC....GGA
(d) Infernal alignment
Same structure, no significant covariations
A.thaliana UUUGUGCCCUAAUUUG.AUCCUCAGGUUUGGGUU.C.AAGUCGCCGGA
A.lyrata UUUAUUCCCUAAUUUG.AUCCUCAGGUU-GGGUU.CgAAGUCGCCGGA
C.rubella CUUGUGCCCUAAUUUG.AUCCUCGGCUUUGGGUU.C.AAGUCGCCGGA
A.alpina UUUGUGCCCUAAUUUGgAUCCUCAGGUUCGGUCUuU.AAGUCGCCUGA
E.salsugineum UCCGUGCCCUAAUUUG.AUCCUCAGGUACGG-UU.C.AAGUCGCCAGA
B.rapa UCUGUGCCCUAAUUUG.AUCCAGAGGUUCGG-UU.C.AAGUGGCCGGA
SS_cons <<<<-<<----<<<<<.<<<<_________>>>->.>.>>>>->>>>>>

Fig. 3. (**a**) Human XIST RNA conservation. Representation of the locations of ten human/mouse unique sequence conserved regions h1-h10 in XIST/Xist, relative to XIST repeat regions A-F and to other human/mouse conserved sequence corresponding to ancient transposable elements (TEs). (**b**) The proposed COOLAIR helix H10, redrawn from Hawkes *et al.* (2016) displaying R-scape's covariation annotation. (**c**) The proposed COOLAIR helix 10 alignment of six homologous sequences resulting in four apparent covariations. (**d**) A revised COOLAIR H10 alignment more consistent with primary sequence conservation. The apparent R-scape covariations from the original alignment in (b) disappear, while the proposed structure in (c) is maintained. Identical sets of residues in both alignments are shaded in the same color

of substitutions at each independent column i to each branch of the tree using our implementation of the Fitch algorithm (Fitch, 1971). For phylogeny-aware statistical significance testing as described in Rivas *et al.* (2017), R-scape then uses this information in tree-based simulations to construct synthetic negative control alignments that preserve average identity, composition, and phylogenetic relationships of the original alignment, while randomizing pairwise correlations. An empirical null distribution for a pairwise covariation statistic (default is the G-test statistic, related to mutual information) is then obtained from all pairs of columns i, j in these simulated alignments.

In the new method for statistical power estimation, the same tree and inferred maximum parsimony substitutions are used to obtain the total substitutions s_i (summed over branches) at each individual column i. For a proposed basepair involving two columns i, j, we use $s_{ij} = s_i + s_j$, the sum of the independent variation in both columns.

We also tested a more expensive variation where we parsimoniously infer pairwise substitutions jointly at all column pairs i, j to obtain s_{ij}, rather than assuming column independence, which gave similar results (data not shown).

3.2 Simulations

Simulated alignments were produced with the program R-scape-sim. Given an input sequence alignment with a consensus RNA secondary structure, R-scape-sim calculates a maximum likelihood phylogenetic tree with branch lengths. It sub-samples the original phylogenetic tree to a desired number of taxa, and linearly scales branch lengths to achieve a desired average percentage identity amongst the aligned sequences. One sequence is selected at random as a root and its evolution is simulated down the tree branches using a probabilistic evolutionary model. The evolutionary model consists of rate matrices for single (unpaired) residue substitution, pairwise (basepair) substitution, insertion, and deletion events (Rivas and Eddy, 2015; Rivas *et al.*, 2017). We used this simulation procedure on the Rfam Cobalamin riboswitch alignment (RF00174) to generate 29 976 synthetic alignments with sequence number ranging from 5 to 190, and average percentage identity ranging from 30% to 100%. The simulated alignments were randomly downsampled to 3000 in order to produce the scatter plots of Figure 1.

3.3 Empirical power(s) curve

For each of 7012 annotated consensus basepairs in 87 RNA families Rfam v14.0 with known 3D structures, we use R-scape to calculate the statistical significance (expectation value, E-value) and estimate s_{ij} for each proposed pair in Rfam 'seed' alignments. We binned proposed pairs with identical s_{ij} and calculated the frequency of pairs with significant support $E < 0.05$ in each bin. For 1, 653 such points for bins $s = 0$ to 1, 652, we fitted a polynomial of degree 10 by minimizing least square error. The choice of degree 10 was arbitrary, and simpler functions such as $1 - e^{-\lambda s}$ did not fit as well.

In binning the data by integer s, the number of basepairs per bin is variable. For large values of s, there are few basepairs per bin (often 0–2), leading to noisy data and possibly to a bad quality alignment, which is why we fit all point for $s \leq 150$, and only those with at least 80% power for $s > 150$. For Figure 1f, we plotted the Rfam data differently, in equal-size bins, by ranking all 7012 basepairs by increasing E-values, dividing them into 70 equal bins, and calculating the means s and power(s) in each bin. This plot is less noisy at high s. We did not re-estimate the fitted curve when we replotted. Fitted power(s) values starting from $s = 0.012$ are hard-coded in the R-scape source code; for $s > 226$ we set power(s) = 1.

Our approach treats power(s) as a function solely of s. This approximates away an important additional dependency on alignment length. R-scape E-values are multiple-test-corrected; the number of potential basepairs depends on the input alignment length. Detecting significant support for a basepair in a longer alignment requires more signal because the background of non-pairs is higher. We considered fitting power(s, p) to a range of different P-value thresholds p (i.e. before multiple test correction to E-values) but decided this was impractical. Instead, the fitted power curve treats all alignments as approximately the same length. The actual lengths of the Rfam seed alignments used in Figure 1f range from 40 consensus columns (HIV retroviral Psi packing element) to 3680 consensus columns (eukarya LSU rRNA).

3.4 Alignment power

The expected sensitivity for detecting a basepair b with s_b substitutions is power(s_b), from the empirical fitted curve shown in Figure 1f. In addition to reporting the pairs that significantly covary with their corresponding E-value, R-scape now also reports for each pair the inferred number of substitutions s_b and the estimated power(s_b).

As an overall summary statistic for an alignment with a proposed structure, R-scape reports the total number of basepairs expected to have significant covariation support,

$$cov - bp - \exp = \sum_{b=1}^{B} \text{power}(s_b),$$

and the *alignment power*, defined as the fraction of basepairs expected to have significant support,

$$\textit{alignment power} = \frac{\textit{cov} - \textit{bp} - \exp}{B} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{power}(s_b).$$

In this work, alignments with $> 10\%$ power are arbitrarily considered to have sufficient power.

3.5 R-scape statistical test modes

R-scape has two statistical modes to test the presence of a conserved RNA structure. By default, R-scape considers all pairs as equivalent and performs an statistical test as to which of the all possible $L(L-1)/2$ pairs (for an alignment of length L) are significantly covarying. This is R-scape's default *one-set test*. Alternatively, if a consensus secondary structure is provided, R-scape allows an optional *two-set test* consisting of two independent tests on two different sets. One test is on the proposed structure (the set of basepairs); the other parallel test is on all other possible pairs in the alignment (the set of non basepairs). On the set of basepairs, R-scape extracts the alignment's support for the annotated structure. On the set of non basepairs, R-scape identifies other possible covarying basepairs not present in the given structure.

Estimating the alignment power requires a proposed structure, thus the use of R-scape's *two-set* mode. Under the *one-set* mode, R-scape still reports the power for each of the significantly covarying basepairs, assuming that those could be part of a structure. The covariation and covariation power analyses provided in this manuscript for all lncRNAs have been obtained with R-scape's *two-set* mode on the proposed secondary structures.

3.6 lncRNA alignment sources

HOTAIR domain 1–4 alignments (D1–D4) and proposed consensus structure used in Somarowthu *et al.* (2015) were kindly provided to us by S. Somarowthu.

The SRA alignment and proposed consensus structure used in Novikova *et al.* (2012) were unavailable to us. The proposed secondary structure of the human ncSRA was reproduced by hand from Supplementary Figure S1 of Novikova *et al.* (2012). An SRA alignment was produced by imposing the human ncSRA proposed structure in the Multiz100way alignment of the ncSRA region obtained from the UCSC human genome browser (http://genome.ucsc.edu). This alignment includes 76 mammalian species.

The Xist repeat A region alignment used in Maenner *et al.* (2010) and four alternative proposed consensus structures that we call RepA.S0 through RepA.S3 were reproduced from Supplementary Figure S5 in Maenner *et al.* (2010).

The Xist repeat A region alignment and proposed consensus structure in Fang *et al.* (2015) were kindly provided to us by W. Moss.

The RepA lncRNA alignment (spanning repeat A and repeat F regions) with a proposed consensus structure in Liu *et al.* (2017) was kindly provided to us by the authors.

We produced our own RepA lncRNA alignment of 65 sequences and got similar results: a high-power alignment sufficient to detect about 254 basepairs, but no significantly supported covarying pairs. The proposed structures for all ten XIST conserved regions were produced using R-scape.

As described in the main text, we used nhmmer (Wheeler and Eddy, 2013) to identify 21 significant local alignments (at $E < 10^{-5}$) between human XIST and mouse Xist, covering 79% of the XIST RNA sequence. We used nhmmer and Dfam (Wheeler *et al.*, 2013) to determine that 11 of these conserved regions correspond to known transposable elements including retroposons (L2d 3end), DNA transposons (Charlie29a, Charlie29b), SINEs (FLAM C) and retroviral LTRs (LTR78). For the remaining 10 conserved regions, we used an nhmmer profile of the mouse/human pairwise alignment to search a database of vertebrate genome sequences, resulting in 10 alignments consisting of 47–65 homologous sequences, which we name XIST h1 through XIST h10.

We used the *Arabidopsis thaliana* COOLAIR lncRNA sequence and the consensus structure proposed in Hawkes *et al.* (2016) to construct a single-sequence Infernal profile (Nawrocki and Eddy, 2013), then used Infernal to align all six COOLAIR homologs to this profile.

All alignments (with consensus structure annotation, where applicable) are included in Supplementary Material in Stockholm format.

Acknowledgements

We thank the Centro de Ciencias de Benasque Pedro Pascual, Spain, where ideas for this manuscript were developed.

Funding

This work has been partially supported by the Howard Hughes Medical Institute.

Conflict of Interest: none declared.

References

Fang,R. *et al.* (2015) Probing Xist RNA structure in cells using targeted structure-seq. *PLoS Genet.*, **11**, e1005668.

Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

Gutell,R.R. *et al.* (1985) Comparative anatomy of 16S-like ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.*, **32**, 155–216.

Gutell,R.R. *et al.* (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.

Hawkes,E.J. *et al.* (2016) COOLAIR antisense RNAs form evolutionarily conserved elaborate secondary structures. *Cell Rep.*, **16**, 3087–3309.

Holley,R.W. *et al.* (1965) Structure of a ribonucleic acid. *Science*, **14**, 1462–1465.

Kozomara,A. *et al.* (2019) miRBase: from microRNA sequences to function. *Annu. Rev. Cell Biol.*, **47**, D155–D162.

Liu,F. *et al.* (2017) Visualizing the secondary and tertiary architectural domains of lncRNA RepA. *Nat. Chem. Biol.*, **13**, 282–289.

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

Michel,F. *et al.* (2000) Modeling RNA tertiary structure from patterns of sequence variation. *Methods Enzymol.*, **317**, 491–510.

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

Noller,H.F. *et al.* (1981) Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.*, **9**, 6167–6189.

Novikova,I.V. *et al.* (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.*, **40**, 5034–5051.

Pace,N.R. *et al.* (1989) Phylogenetic comparative analysis and the secondary structure of Ribonuclease P RNA—a review. *Gene*, **82**, 65–75.

Price,M.N. *et al.* (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Rivas,E. and Eddy,S.R. (2015) Parameterizing sequence alignment with an explicit evolutionary model. *BMC Bioinformatics*, **16**, 406.

Rivas,E. and Eddy,S.R. (2018) Response to Tavares et al. Covariation analysis with improved parameters reveals conservation in lncRNA structures. *BioRxiv*. doi: 10.1101/2020.02.18.955047.

Rivas,E. *et al.* (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, **14**, 45–48.

Somarowthu,S. *et al.* (2015) HOTAIR forms an intricate and modular secondary structure. *Mol. Cell*, **58**, 353–361.

Tavares,R.C.A. *et al.* (2019) Phylogenetic analysis with improved parameters reveals conservation in lncRNA structures. *J. Mol. Biol.*, **431**, 1592–1603.

Wheeler,T.J. and Eddy,S.R. (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, **29**, 2487–2489.

Wheeler,T.J. *et al.* (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, **41**, D70–D82.

Williams,K.P. and Bartel,D.P. (1996) Phylogenetic analysis of tmRNA secondary structure. *RNA*, **2**, 1306–1310.