A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs

Elena Rivas[1], Jody Clements[2] & Sean R Eddy[1,3–5]

Many functional RNAs have an evolutionarily conserved secondary structure. Conservation of RNA base pairing induces pairwise covariations in sequence alignments. We developed a computational method, R-scape (RNA Structural Covariation Above Phylogenetic Expectation), that quantitatively tests whether covariation analysis supports the presence of a conserved RNA secondary structure. R-scape analysis finds no statistically significant support for proposed secondary structures of the long noncoding RNAs HOTAIR, SRA, and Xist.

Pairwise covariations in RNA alignments provide a means of deducing evolutionarily conserved RNA secondary structures[1–5]. In turn, a conserved secondary structure provides positive evidence that a noncoding RNA has a function. The first manual covariation analyses of small numbers of aligned RNA sequences used rules of thumb to infer conserved structures from a few compensatory base-pair substitutions[2,6], defined as two substitutions at a pair of positions of an RNA sequence alignment that preserve Watson–Crick or G:U base pairing. As the number of aligned sequences grows, apparent compensatory base-pair substitutions may be observed by chance. This is exacerbated by phylogenetic correlations; two independent single-residue substitutions that fortuitously look compensatory can propagate into several descendants and appear to be several compensatory base-pair substitutions (**Fig. 1**).

Because an RNA can function as an unstructured sequence[7], and because some RNAs, particularly some long noncoding RNAs (lncRNAs), may be unannotated coding mRNAs or various sorts of transcriptional noise[8], there is a need for computational tools that determine whether sequence alignment analysis provides statistical support for an evolutionarily conserved RNA secondary structure. There is extensive literature on RNA covariation analysis methods[9–12], but these methods have been underutilized, perhaps because no one computational tool has yet adequately combined covariation analysis with statistical significance testing, computational efficiency, and accessibility.

For example, the evidence for structure conservation in the lncRNA HOTAIR[13] consisted of use of the RNA drawing program R2R[14] to annotate an alignment of 33 sequences, using a proposed consensus structure based on chemical and enzymatic probing experiments. Examination of this HOTAIR alignment shows that in most cases only a single compensatory base-pair substitution supports each proposed covarying base pair, while many substitutions disrupt the proposed pair. R2R was intended for visualization of known RNA structures, not for quantitation of evidence for structure conservation[14]. It annotates a consensus base pair as covarying if any compensatory base-pair substitution, –even just one– is observed; and it does not consider substitutions that conflict with the proposed structure.

We have developed an accessible tool, R-scape, that analyzes a multiple-RNA sequence alignment and quantitates the statistical support for evolutionary conservation of an RNA structure. A pairwise covariation statistic is calculated for each alignment column pair, and statistically significant covariation is interpreted as evidence for a conserved RNA base pair. We compared several covariation statistics on a test set of annotated consensus structures for 104 RNA sequence alignments from Rfam[15] (see Online Methods), counting a false negative when an annotated base pair has a covariation statistic below threshold and a false positive when an unstructured pair scores above threshold. The G-test statistic[16] is more robust than other statistics tested, including mutual information (see Online Methods and **Supplementary Fig. 1**). A background correction[17] further improves covariation detection (**Supplementary Fig. 1a**). R-scape calculates the average product corrected (APC) G-test covariation statistic by default.

The APC G-test statistic can be calculated rapidly even for deep sequence alignments, but it does not explicitly deal with confounding covariation caused by phylogenetic correlation (**Fig. 1a**)[5]. Methods that do tend to be computationally expensive[9,11,12] (**Supplementary Fig. 2**). R-scape instead determines the statistical significance of the observed covariation scores by simulating alignments under a null hypothesis in which phylogenetic relationships are preserved but columns evolve independently. For example, a given toy alignment (**Fig. 1b**, top) has two independent substitutions in two different columns in the same tree branch, resulting in four apparent compensatory base pairs in the alignment. In R-scape-simulated null alignments (bottom), the same two substitutions are made on the same branch, but their sequence positions are randomized, and the apparent covariation remains. In another given toy alignment (**Fig. 1c**, top), a covarying base pair has five compensatory base-pair substitutions in five different branches. In R-scape-simulated null alignments (**Fig. 1c**, bottom), after randomizing the sequence position of

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA. [2]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia, USA. [3]Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts, USA. [4]FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts, USA. [5]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to E.R. (elenarivas@fas.harvard.edu).

RECEIVED 1 APRIL; ACCEPTED 14 SEPTEMBER; PUBLISHED ONLINE 7 NOVEMBER 2016; DOI:10.1038/NMETH.4066

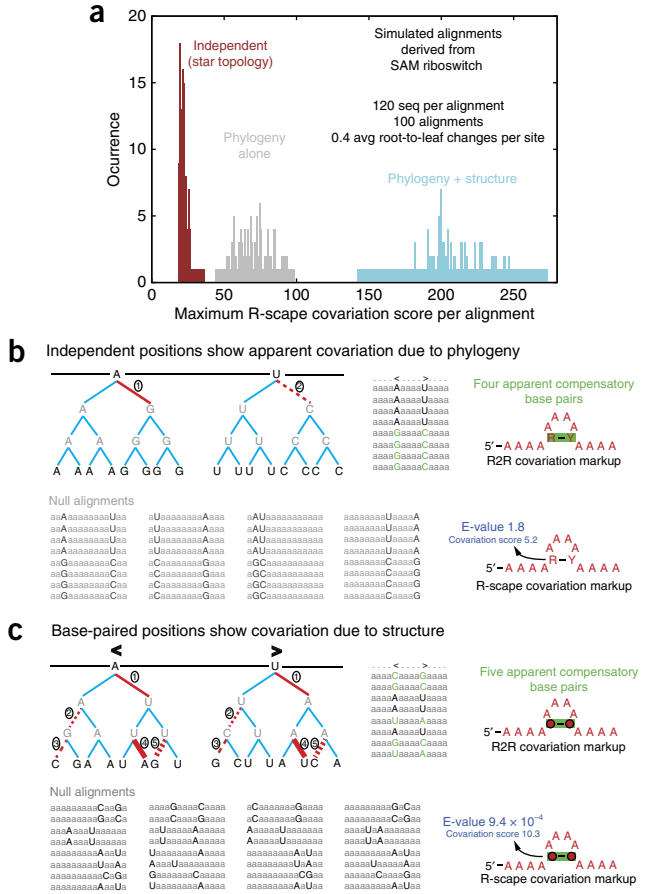

Figure 1 | Independent substitutions on a tree can create confounding covariations. (**a**) Histogram of maximum covariance score per alignment over 100 synthetic alignments simulated under three evolutionary models: no phylogeny and no structure constraint (red), phylogeny alone (gray), or phylogeny plus structure constraint (cyan). Avg, average; seq, sequences. (**b**) Toy alignment (top left) with two independent substitutions (marked 1 and 2 with distinct line styles) on the same branch, resulting in an apparent pairwise covariation annotated by R2R (top right). R-scape simulated null alignments (bottom left) retain this confounding covariation signal, and it is judged insignificant (bottom right). (**c**) Toy alignment with five compensatory base-pair substitutions (marked 1–5 with distinct line styles) showing a covariation pattern that is destroyed in the R-scape-simulated null alignments and thus judged significant.

each substitution on each branch, the five compensatory pair substitutions become ten uncorrelated substitutions, destroying the correlation seen in the alignment.

For a given alignment, R-scape produces many simulated null alignments (default 20) and calculates an APC G-test statistic for each alignment column pair, thus collecting an expected null distribution conditioned on the input alignment's characteristics, including its length in columns, sequence number, pairwise identity, base composition, substitution types, and phylogenetic correlation (see Online Methods). This empirical null distribution of the covariation statistic estimates the probability of obtaining a false positive on null data (the *P* value) at any threshold. We calculate an expectation value (E-value) by multiplying the *P* value by the total number of column pairs evaluated. An E-value $E(x)$ is the number of column pairs expected to give a covariation score of at least x when they are evolving independently, under no RNA structure constraint. A significant E-value is $\ll 1$.

R-scape analysis of the Rfam alignment for the structural 5S ribosomal RNA shows significant covariation support for 22/34 base pairs in the annotated consensus structure (**Fig. 2a**). R-scape also identifies eight significant pairs that are not present in the Rfam-annotated consensus structure, which shows that R-scape can not only support but also improve a structural annotation (**Fig. 2b**). Using an optional feature that predicts a new consensus secondary structure that includes the maximum number of significantly covarying pairs, R-scape proposes a modified 5S rRNA consensus structure in which 32/38 base pairs are significant. The R-scape structure is in agreement with the accepted 5S rRNA consensus structure[18], suggesting there are some errors in the curated Rfam structure. Three other examples in **Supplementary Figure 3** show R-scape support for consensus structures of two small noncoding RNAs and a *cis*-regulatory mRNA structure.

For alignments of typical, known structural RNAs (transfer RNA, bacterial RNAse P RNA, purine riboswitch), 70–100% of the annotated base pairs are supported by R-scape with E-values $< 10^{-5}$ (**Fig. 2c**, leftmost). For a recent study of ten γ-proteobacterial mRNA leader structures that autoregulate ribosomal protein synthesis[19], eight proposed structures have many significant covarying pairs, and only two show weaker support (one, the S7 leader, overlaps the ribosomal protein L5 coding region and therefore has restricted variation (**Fig. 2c**, center). For a recent screen identifying six small RNAs in α-proteobacteria[20], three (αr14, α15, and αr7) have good covariation support (**Fig. 2c**, center). Alignments of three human DNA repeat elements are negative controls with no known RNA structure constraint and no R-scape signal (**Fig. 2c**, rightmost).

For the proposed *HOTAIR* lncRNA structure, using the same alignment used by Somarowthu *et al.*[13], no significantly covarying base pairs are found for any of the four proposed domain structures. This result differs from the previous HOTAIR analysis[13] because R-scape accounts for the fact that the observed sequence variation is more frequently inconsistent than consistent with the proposed structure. Details of the analysis of proposed helices H7 and H10 are in **Supplementary Figure 4**.

For the proposed steroid receptor RNA activator noncoding RNA (ncSRA) lncRNA structure[21], R-scape does not find any significantly covarying pairs, only sequence variation as with HOTAIR. Details of the analysis of putative helices H3 and H4 and putative helices H9, H20, and H21 are provided in **Supplementary Figures 5** and **6**.

For the *Xist* lncRNA, consensus structures have been proposed for the repeat A (RepA) region, with compensatory base changes cited as support[22], but R-scape shows no significant covariation support for any base pair. The alignment in the published analysis of Xist RepA lncRNA has only ten sequences[22], which limits power in identifying covariation support.

A different RepA secondary structure has been proposed, and it is said to have covariation support for four base pairs in an alignment of 13 sequences[23]. Applying the same criteria used by Fang *et al.*[23] systematically to all column pairs shows 541 pairs with equivalent support (**Supplementary Fig. 7**), 538 of which are inconsistent with the proposed structure. This highlights another source of confounding signal. Independent G>A and U>C substitutions in conserved G+A and U+C columns (454/541 pairs in this RepA alignment) create an appearance of covariation support because of G:U wobble base pairing. R-scape null alignments

reproduce this sequence variation pattern, so R-scape finds no statistically significant covariation support.

Failure to identify significant covariation support for an evolutionarily conserved RNA secondary structure does not necessarily mean that a structure is not present. Deeper alignments or a more powerful statistical analysis might reveal a more subtle conserved secondary structure for any of these RNAs. For the two published RepA alignments the pattern of sequence conservation is consistent with a large number of possible base pairs, too many to deduce any single conserved structure; here especially, deeper alignments may help in revealing or disproving a conserved structure. In contrast, the published HOTAIR and ncSRA alignments

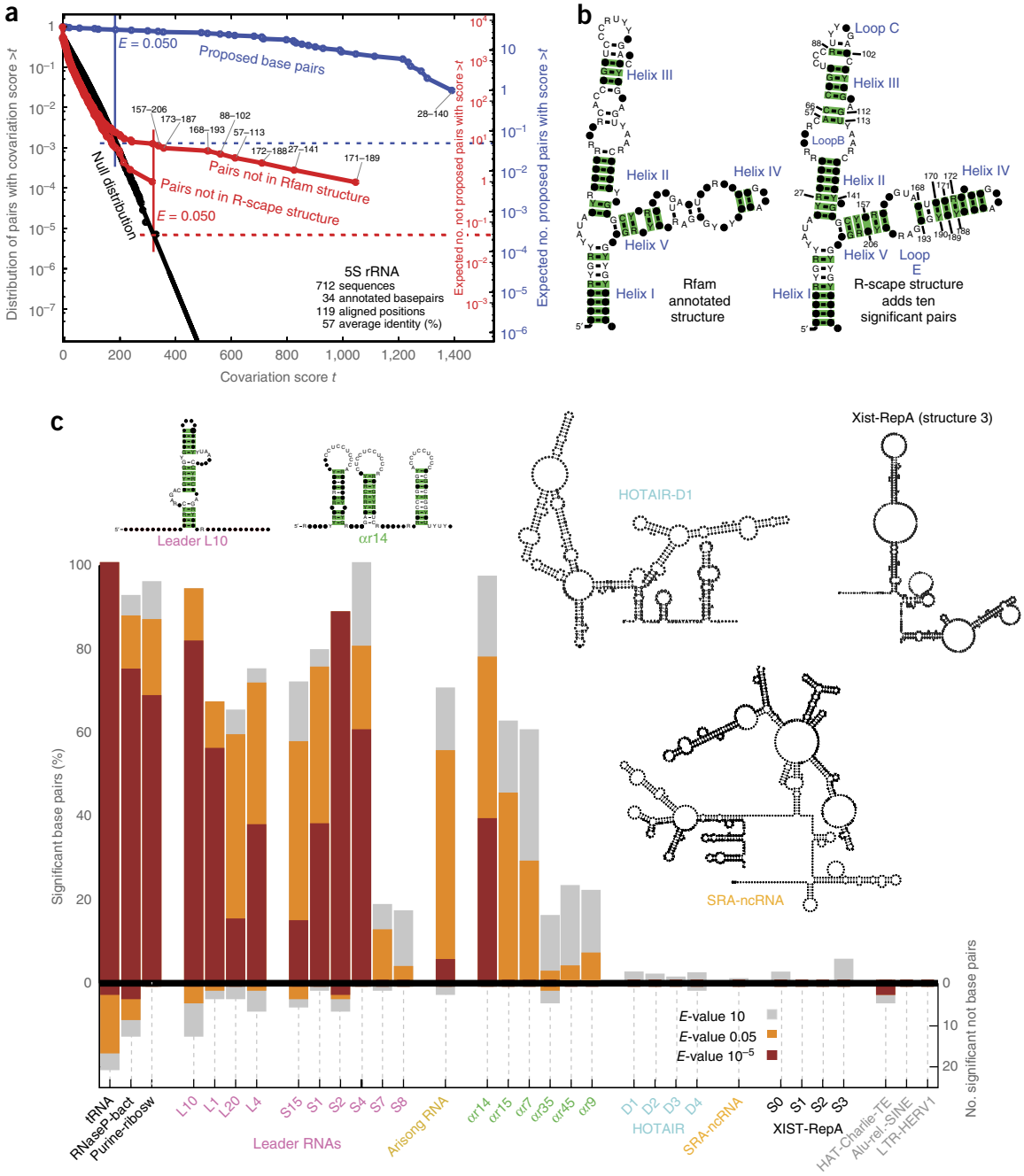

Figure 2 | Covariation analysis of known or proposed RNA secondary structures. (**a**) The plot shows the expected null distribution of covariance scores for 5S rRNA, a known structural RNA (fit: black lines; data: black circles), compared with the covariation scores observed for base pairs present (blue) or absent (red) in the Rfam structural alignment of 5S rRNA. Numbers indicate base pair coordinates in the alignment. (**b**) Covariation support for the Rfam-annotated 5S rRNA structure versus an alternative structure proposed by R-scape to include all significantly covarying pairs. Significant pairs (at $E < 0.05$) are highlighted in green. Coordinates are alignment column positions. Specific nucleotides are shown when their weighted frequency in the column exceeds 50%; black dots represent more variable positions. (**c**) Proposed RNA structures (named along *x*-axis) with R-scape-significant covariations in green ($E < 0.05$) for L10 leader[19], αr14 (ref. 20), HOTAIR[13], SRA-ncRNA[21], and RepA[22]. Plot shows in the positive *y*-axis the percentage of base pairs supported by covariation at three thresholds (red, $E < 10^{-5}$; orange, $E < 0.05$; gray, $E < 10$). Negative *y*-axis shows the number of additional significantly covarying pairs not in the proposed structure.

are already relatively deep (30–60 sequences), and their pattern of sequence variation is inconsistent with their proposed structures. Here it may be more likely that any functions of these RNAs may depend more on their linear sequences than on conserved secondary structure[7,24]. There are many other lncRNAs, and lncRNA function and structure remain controversial and difficult to study. Tools like R-scape will be useful for quantitative analysis of the covariation evidence supporting proposed structures of lncRNAs, or indeed of any RNA.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S.E.R. Egnor for suggesting the name R-scape and the Centro de Ciencias de Benasque Pedro Pascual in Spain, where part of this manuscript was drafted.

AUTHOR CONTRIBUTIONS

E.R. and S.R.E. designed the method and wrote the manuscript. E.R. wrote the code, and designed and carried out the experiments. J.C. wrote the R-scape web application.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Holley, R.W. *et al. Science* **147**, 1462–1465 (1965).
2. Noller, H.F. *et al. Nucleic Acids Res.* **9**, 6167–6189 (1981).
3. Pace, N.R., Smith, D.K., Olsen, G.J. & James, B.D. *Gene* **82**, 65–75 (1989).
4. Williams, K.P. & Bartel, D.P. *RNA* **2**, 1306–1310 (1996).
5. Michel, F., Costa, M., Massire, C. & Westhof, E. *Methods Enzymol.* **317**, 491–510 (2000).
6. Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J. & Stormo, G.D. *Nucleic Acids Res.* **20**, 5785–5795 (1992).
7. Davidovich, C. & Cech, T.R. *RNA* **21**, 2007–2022 (2015).
8. Ji, Z., Song, R., Regev, A. & Struhl, K. *eLife* **4**, e08890 (2015).
9. Akmaev, V.R., Kelley, S.T. & Stormo, G.D. *Bioinformatics* **16**, 501–512 (2000).
10. Lindgreen, S., Gardner, P.P. & Krogh, A. *Bioinformatics* **22**, 2988–2995 (2006).
11. Yeang, C.-H., Darot, J.F.J., Noller, H.F. & Haussler, D. *Mol. Biol. Evol.* **24**, 2119–2131 (2007).
12. Dutheil, J.Y. *Brief. Bioinform.* **13**, 228–243 (2012).
13. Somarowthu, S. *et al. Mol. Cell* **58**, 353–361 (2015).
14. Weinberg, Z. & Breaker, R.R. *BMC Bioinformatics* **12**, 3 (2011).
15. Nawrocki, E.P. *et al. Nucleic Acids Res.* **43**, D130–D137 (2015).
16. Woolf, B. *Ann. Hum. Genet.* **21**, 397–409 (1957).
17. Dunn, S.D., Wahl, L.M. & Gloor, G.B. *Bioinformatics* **24**, 333–340 (2008).
18. Szymanski, M., Barciszewska, M.Z., Erdmann, V.A. & Barciszewski, J. *Nucleic Acids Res.* **30**, 176–178 (2002).
19. Fu, Y., Deiorio-Haggar, K., Anthony, J. & Meyer, M.M. *Nucleic Acids Res.* **41**, 3491–3503 (2013).
20. del Val, C., Rivas, E., Torres-Quesada, O., Toro, N. & Jiménez-Zurdo, J.I. *Mol. Microbiol.* **66**, 1080–1091 (2007).
21. Novikova, I.V., Hennelly, S.P. & Sanbonmatsu, K.Y. *Nucleic Acids Res.* **40**, 5034–5051 (2012).
22. Maenner, S. *et al. PLoS Biol.* **8**, e1000276 (2010).
23. Fang, R., Moss, W.N., Rutenberg-Schoenberg, M. & Simon, M.D. *PLoS Genet.* **11**, e1005668 (2015).
24. Rinn, J.L. & Chang, H.Y. *Annu. Rev. Biochem.* **81**, 145–166 (2012).

ONLINE METHODS

R-scape: RNA structural covariation above phylogenetic expectation. An R-scape web server is available at http://www.eddylab.org/R-scape. The current version of the R-scape source code is freely downloadable from http://www.eddylab.org, and an archived tarball of the version used in this paper (v0.2.2) is included as part of the online **Supplementary Software**. The source code for the web server is freely available at http://www.github.com/EddyRivasLab/R-scape-web.

The input to the R-scape program is a multiple RNA sequence alignment, typically with a consensus structure, in Stockholm format. The output is a list of pairs of alignment columns that significantly covary, ranked by their E-value. The source code distribution includes examples of input and output files as well as documentation. The R-scape web server is a wrapper around the R-scape program.

If a consensus RNA structure is provided for the input alignment, R-scape evaluates the covariation support for that consensus structure. Optionally, R-scape can calculate an independent consensus structure, which is the maximum likelihood secondary structure constrained to use all significantly covarying pairs (for a given E-value cutoff). This constrained folding method uses the probabilistic 'basic grammar' model introduced with the RNA folding method Tornado[25]. This algorithm cannot predict pseudoknots, but the presence of nested significant pairs not in the structure but compatible with it (depicted in orange in the web application) are a good indication of their possible presence.

By default, R-scape (v0.2.2) uses two external programs: FastTree[26] (v2.1.8) to build a phylogenetic tree from the given alignment by approximate maximum likelihood, and a modified version of R2R[14] (v1.0.4) to draw the consensus structure annotated with the covarying base pairs with E-values smaller than a given E-value cutoff.

Different covariation statistics. We tested the following statistics for measuring pairwise covariations, out of many different statistics that have been suggested. Given two alignment columns i,j:

G-test[16]: $\mathrm{GT}(i,j) = 2\sum_{a,b} \mathrm{Obs}_{ij}^{ab} \log \frac{\mathrm{Obs}_{ij}^{ab}}{\mathrm{Exp}_{ij}^{ab}}$

Pearson's chi-square: $\mathrm{CHI}(i,j) = \sum_{a,b} \frac{(\mathrm{Obs}_{ij}^{ab} - \mathrm{Exp}_{ij}^{ab})^2}{\mathrm{Exp}_{ij}^{ab}}$

Mutual information (MI)[27,28]: $\mathrm{MI}(i,j) = \sum_{a,b} P_{ij}^{ab} \log \frac{P_{ij}^{ab}}{p_i^a p_j^b}$

MI normalized[29]: $\mathrm{MIr}(i,j) = \frac{\mathrm{MI}(i,j)}{H(i,j)} = \frac{\mathrm{MI}(i,j)}{-\sum_{ab} P_{ij}^{ab} \log P_{ij}^{ab}}$

MI with gap penalty[10]: $\mathrm{MIg}(i,j) = \mathrm{MI}(i,j) - \frac{N_{ij}^G}{N}$

Obs-Minus-Exp-Squared[30]: $\mathrm{OMES}(i,j) = \sum_{a,b} \frac{(\mathrm{Obs}_{ij}^{ab} - \mathrm{Exp}_{ij}^{ab})^2}{N_{ij}}$

RNAalifold (RAF)[31]: $\mathrm{RAF}(i,j) = B_{i,j}$,

RNAalifold Stacking (RAFS)[10]:

$$\mathrm{RAFS}(i,j) = \frac{1}{4}(B_{i-1,\,j+1} + 2B_{i,\,j} + B_{i+1,\,j-1})$$

where a,b are (nongap) residues; N is the total number of aligned sequences; $\mathrm{Obs}_{i,j}^{ab}$ is the observed count of $a{:}b$ pairs in columns i,j (only counting when both a,b are residues); N_{ij} is the total number of residue pairs in columns i,j (only counting when both a,b are residues); $P_{i,j}^{ab}$ is the observed frequency of pair $a{:}b$ in columns i,j, where

$$P_{ij}^{ab} = \frac{\mathrm{Obs}_{ij}^{ab}}{N_{ij}}$$

$\mathrm{Exp}_{ij}^{ab} = N_{ij} p_i^a p_j^b$ is the expected frequency of pair $a{:}b$ assuming i,j are independent, where $p_i{}^a$ are the marginal frequencies of a residues in column i (averaged to all other positions):

$$p_i^a = \frac{1}{L-1} \sum_{j \neq i} \sum_b P_{ij}^{ab}$$

$N_{ij}^G = N - N_{ij}$ is the number of pairs involving at least one gap symbol. The definition of B_{ij} used in the RAF and RAFS statistics is complicated and not shown here; their definition can be found elsewhere[10].

We also tested two background corrections that can be applied to any of the above covariation statistics[17]. Let $COV(i,j)$ be a covariation statistic; then, the average product correction (APC) is

$$\mathrm{COV}^{\mathrm{APC}}(i,j) = \mathrm{COV}(i,j) - \frac{\mathrm{COV}(i)\mathrm{COV}(j)}{\mathrm{COV}}$$

The average sum correction (ASC) is

$\mathrm{COV}^{\mathrm{ASC}}$ (i,j) = COV (i,j) − COV (i) + COV (j) − COV) is

where

$$\mathrm{COV}(i) = \frac{1}{L-1} \sum_{j \neq i} \mathrm{COV}(i,j)$$

is an average covariation for an individual column i; and

$$\mathrm{COV} = \frac{1}{L} \sum_i \mathrm{COV}(i)$$

is the average covariation overall.

By default, R-scape uses the APC G-test statistic, as this was the most robust statistic in our benchmark tests (**Supplementary Fig. 1**; also, see below). Compared to the related and more commonly used mutual information (MI) statistic, the G-test (based on observed counts) is different from MI (based on frequencies) for alignments with gaps. A column pair with many gaps could have similar MI to another column pair with no gaps, but the G-test score of the former will have smaller magnitude (fewer number of effective sequences) than the G-test score for the latter. This difference makes G-test a more robust statistic than MI on alignments with gaps (**Supplementary Fig. 1b**).

Comparing the G-test with the RAFS statistic, although RAFS has better sensitivity (as has been previously reported)[10], RAFS is

more prone to report covariations between unpaired bases, especially for E-values larger than 10^{-4} (worse positive predictive value), and it is also more affected by alignment gaps. This could either be because RAFS is less specific (more false positives) or because it is more sensitive to true (tertiary structure) non-Watson–Crick interactions. We tested this by simulating alignments in which we preserved the base-paired columns but replaced all unpaired columns using the null model of phylogenetically dependent but position-independent changes. Results for these partially simulated alignments are similar to those for the original alignments in **Supplementary Figure 1b** (data not shown), suggesting that RAFS is less specific. The RAFS statistic is also slower to calculate, scaling with the square of the number of sequences in the alignment.

R-scape filters the input alignment before collecting observed counts such that columns with greater than 50% gap symbols are ignored. After this filtering, relative sequence weights are calculated using the GSC algorithm[32]. These steps help compensate for the fact that simple correlation statistics do not take the phylogenetic relationship of the sequences into account. Some covariation statistics have been developed that directly account for phylogeny[9,11] but at a cost of increased computational and model complexity.

The covariation statistics can either be applied as 16-class tests (taking the summation over *a*:*b* over the 16 possible base pairs) or as 2-class tests where pairs are separated into two groups according to whether they correspond to canonical Watson–Crick pairs (including G:U)[33]. A 2-class test is better on a small number of sequences (see **Supplementary Fig. 1**) and short alignments (not shown), because with fewer classes, the 2-class test is less susceptible to statistical fluctuations in small numbers of counts. The RAFS statistic is intrinsically of the 2-class type. A 2-class test looks specifically at whether covariation is consistent with Watson–Crick (or G:U) base pairing; a 16-class test detects any pairwise correlation, including non-Watson–Crick pairing seen in RNA tertiary contacts. For alignments with more than eight sequences, we use the 16-class covariation, as it performs similarly to the 2-class covariations for larger numbers of sequences while also allowing identification of possible non-Watson–Crick covariations (**Supplementary Fig. 1c**).

The same approach could be applied to other covariation statistics. A maximum entropy approach called direct coupling analysis (DCA), first introduced for protein sequence covariation analysis[34], has been applied to RNA[35,36]. DCA produces pairwise pseudoenergies based on pairwise correlation statistics observed in a multiple alignment. Strong pseudoenergies are thought to reflect direct structural interactions. Currently, applications of DCA methods to structure prediction are simply taking *N* top ranking pseudoenergies for an arbitrarily chosen *N*, without using a measure of statistical significance. The general approach used in R-scape could provide a means for assigning significance to DCA scores, allowing better discrimination of signal and noise and more meaningful comparison with other covariation statistics.

Benchmarking to choose the default covariation statistic. We evaluated the different covariation statistics on a test set of 104 alignments taken from seed alignments in the Rfam sequence family database[15], manually chosen to give wide representation of known structural RNAs with well-studied, more reliable consensus secondary RNA structures, and to have at least 40 sequences in the alignment. The test set includes: 2 tmRNA families, vault RNA, 6S RNA, U7 small nuclear RNA, 9 rRNA families (including 5.8S, 3 small-subunit, and 3 large-subunit rRNA families); 3 signal recognition particle RNAs; selenocysteine tRNA; 14 riboswitch RNAs; 1 leader peptide; 23 other *cis*-regulatory RNAs; 9 spliceosomal RNAs; 5 ribozymes; 13 sRNAs; 8 group-II introns; 5 miRNAs; 3 C/D-box snoRNAs; and 5 other RNA genes. The percentage identity of the alignments ranges from 41% to 80% (defined as the average pairwise % identity over all aligned sequence pairs, with pairwise identity calculated as the ratio of identical positions divided by the minimum length of the two sequences). The number of sequences varies from 44 (glycine riboswitch) to 956 (glnA, a bacterial regulatory sRNA).

In the analysis presented in **Supplementary Figure 1**, we aggregate all alignments together. The total number of consensus base pairs is 7,483; and the total number of alignment columns is 50,769. When columns with more than 50% gaps are removed, the average percentage identity of the alignments (ranging from 42% to 79%) remains similar to that of the original alignments. The number of base pairs remains unchanged, but the number of analyzed columns is reduced to 28,526. A complete list with summary statistics as well as the alignments themselves are provided in **Supplementary Figure 8** and **Supplementary Data Set 1**.

We tested the different covariation statistics on the 104 test alignments and their trusted consensus secondary structures, measuring the fraction of base pairs detected (and the fraction of detected pairs that are base pairs) at different E-value significance thresholds. The results, leading to the choice of the APC G-test statistic as the default, were summarized above.

The ability to detect significantly covarying base pairs depends on many factors, including the quality of the alignment and the number and diversity of the sequences in it. We tested the average effect of varying sequence number in the test alignments (**Supplementary Fig. 1d**). Typically, about 60% of base pairs are detected as significant at $E < 0.05$ when an input alignment contains 40 sequences, depending on other details of the alignment such as percent identity. More sequences have diminishing benefit. Fewer sequences compromise detection; with only ten sequences in the alignment, few base pairs are detected. **Supplementary Figure 1e** shows a scatter plot of the percent of detected base pairs (at an $E < 0.05$ threshold) versus average percent identity in the alignment, showing that there is substantial variation from alignment to alignment, some of which is accounted for by sequence diversity (unsurprisingly). These results emphasize that the failure to detect significant covariation for an individual base pair does not necessarily mean that the base pair is not present in a conserved structure.

Significance calculations on simulated null data. In R-scape, we calculate the significance of covariation scores by simulating phylogenetically related sequences under a null hypothesis of independently evolving columns. Given an alignment, we estimate a tree by approximate maximum likelihood using the FastTree method (v2.1.7 SSE3)[26], root the tree by midpoint rooting, and assign substitutions to branches by maximum parsimony (using the Fitch algorithm)[37]. We then simulate an alignment of the same depth (in sequences) and width (in columns), starting from a parsimoniously inferred root sequence. For each

doi:10.1038/nmeth.4066

ancestral node, we introduce the same set of single-nucleotide substitutions observed on each original descendant branch while randomizing their positions. For example, if we are introducing an A>G substitution on a branch, we choose a random A in the ancestral sequence and substitute it for a G in the descendant. The result is a sampled null alignment that has exactly the same base composition as the input alignment, exactly the same set of single-nucleotide substitutions, and similar pairwise percentage identities as the original sequences, while any correlated pairwise substitution has been scrambled.

E-value estimation. R-scape uses simulated null alignments to estimate the expected number of false positives (E-value) as a function of covariation score. Because simulated null alignments are generated by a resampling strategy on the input alignment, false-positive estimation takes into account the characteristics of the input alignment, including base composition, number of sequences, average pairwise identity, and phylogeny. By default, 20 synthetic null alignments are generated and scored, and the tail of the resulting survival distribution for covariation scores is fitted to a truncated gamma distribution by maximum likelihood[38] to estimate $P(\text{score} > x)$, the probability that one tested column pair will give a covariation score better than x.

The expected number of false positives is then $E = NP(\text{score} > x)$, where N is the number of column pairs tested. R-scape calculates two different E-values. One assumes that we are testing the support for a given secondary structure, in which case N is the number of proposed base pairs in that structure (and N scales with the length of the alignment L). The second assumes that we are testing for any other column pair that shows statistically significant covariation, in which case N is the total number of possible pairs in the alignment, less those in the proposed structure (and N scales with L^2).

Because E-values are based on stochastic simulations, there is some run-to-run variability. R-scape E-values are typically reproducible to an accuracy of about two-fold in different runs of the program.

R-scape uses the default E-value of 0.05 to define a significant covarying pair, which means that it would be expected to detect about 5 false positives overall in 100 different analyzed alignments.

Computational efficiency. R-scape is fast and memory efficient. RNA sequence alignments smaller than ~1,000 columns take under 10 s each and under 50 MB of memory (**Supplementary Fig. 2**). As an example of a large RNA, the Rfam bacterial small subunit ribosomal RNA seed alignment[15] of length 1,980 columns and 99 sequences takes 49 s and 800 MB of memory. The 956 sequence alignment of the glnA glutamine riboswitch RNA (274 alignment columns) takes 9 s and 16 MB of memory. Empirically, R-scape computation time scales approximately as $NL^{1.2}$ for sequence number N and alignment length L (**Supplementary Fig. 2**). Thus R-scape can be run systematically on any RNA alignment(s).

Simulations of sequences related by phylogeny alone, or by phylogeny and RNA structure. Figure 1a uses simulated data to illustrate the effect on covariation statistics for sequences related by phylogeny alone versus sequences related by phylogeny and with structurally constrained positions. From an alignment of a known structural RNA, we produce synthetic alignments guided by the consensus structure, alignment, and phylogeny of the given RNA according to one of the three scenarios: (i) simulated sequences are evolved following the phylogeny and the structure of the RNA; (ii) simulated sequences follow the phylogeny, but positions are independent from each other; and (iii) sequences and positions are independent from each other (alignment is not structural, and it follows a star topology).

As an example, **Figure 1a** starts with the Rfam seed alignment of the SAM riboswitch. From a subalignment of a random set of 120 sequences from the original Rfam SAM riboswitch alignment (433 sequences total), a phylogenetic tree was calculated using FastTree[26]. Starting from a random SAM riboswitch sequence as the root, three different evolutionary models are applied. The structural–phylogenetic model generates sequences along the branches of the SAM riboswitch tree, with base pairs in the ancestral sequence substituted under a 16 × 16 base-pair substitution process. The phylogenetic model generates sequences along the branches of the SAM riboswitch tree, but each ancestral position evolves independently. The independent model evolves extant sequences directly from the root, using the overall distance from root to leaf of the SAM riboswitch tree. In **Figure 1a**, for each simulated alignment, we collect the maximum R-scape covariation score from all possible pairs. We generated 100 synthetic alignments for each of the three methods.

The structural–phylogenetic model depends on a 16 × 16 rate matrix (a base pair evolving to another base pair) calculated from the base-paired positions in a collection of SSU and LSU ribosomal RNA alignments; and the model also depends on a 4 × 4 rate matrix calculated from the unpaired positions of the same alignments. The phylogenetic and independent models use a 4 × 4 rate matrix created from the same alignment but using all aligned columns. The rate matrices were obtained as the logarithm of the conditional probability (substitution) matrices obtained for the rRNA alignments normalized to one substitution per site. Insertions and deletions were created using the AIF evolutionary model[39] parameterized with these rate matrices. Individual insertions (possibly consisting of several residues) are assumed to be independent from each other, and once created they do not evolve.

The code to produce these simulations (R-scape-sim) from a Stockholm alignment of a structural RNA as input is provided as part of the R-scape source code package, which also includes the rate matrices used.

Comparison with other methods. We identified two other methods (MICA and CoMap, from package CoMap v1.5.1) that calculate covariations and provide an estimation of statistical significance[12]. MICA (mutual information coevolution analysis) implements the mutual information statistic. We used MICA with a background-corrected MIp covariation statistic and Z-scores to estimate significance. CoMap (cosubstitution mapping) uses a phylogenetic tree to calculate its covariation statistic. We used parametric bootstrapping to estimate significance for CoMap scores (default for CoMap v.1.5.1).

In **Supplementary Figure 2**, we show that R-scape is at least as fast as the phylogenetic-free method MICA, while at the same time it performs favorably when compared with the fully phylogenetic method CoMap. The program CoMap was also run using

a phylogenetic tree created with PhyML[40] with similar results (data not shown).

Provenance of structural alignments. The alignment of the ciliate ncRNA Arisong with 69 sequences was provided by S. Jung as an updated version of the alignment given in the original manuscript[41]. The six α-proteobacteria ncRNA alignments (αr7, αr9, αr14, αr15, αr35, and αr45) were provided by C. del Val as updated versions of the alignments given in the original manuscript[42]. The alignments for the ten γ-proteobacteria ribosomal protein mRNA leader regions were obtained from the Supplementary Data of Fu *et al.*[19]. The alignments for the four HOTAIR domains D1–D4 and their proposed secondary structures were provided by S. Somarowthu[13].

We were unable to obtain alignments for the SRA ncRNA described in Novikova *et al.*[21] from the authors. Instead, a close approximation was produced by reproducing the proposed secondary structure of the human ncSRA by hand from **Supplementary Figure 1** of the above manuscript and imposing it as a consensus 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 alignment for the *Xist* repA region described in Maenner *et al.*[22] was obtained from their **Supplementary Figure 5**. Four alternative secondary structures (named here S0 to S3) were presented[22]. The consensus structures for the mouse sequence were reproduced by hand from **Figures 2–5**. The alignment includes ten vertebrate sequences with average length of 438 nucleotides and average percentage identity of 77%.

The alignment for the *Xist* RepA region described in Fang *et al.*[23] was provided by W. Moss. The alignment includes 13 vertebrate sequences with average length of 423 nucleotides and average percentage identity of 75%. In the alignment, we imposed by hand the proposed mouse secondary structure obtained by targeted structure-seq[23].

Rfam alignments where obtained from Rfam v12.0 seed alignments[15]. The human repeat alignments were obtained from Dfam v2.0 seed alignments[43].

All the alignments used in this analysis are provided in Stockholm format as part of the online **Supplementary Data Set 1**. Details of the properties of the alignments are provided in **Supplementary Figure 8**.

25. Rivas, E., Lang, R. & Eddy, S.R. *RNA* **18**, 193–212 (2012).
26. Price, M.N., Dehal, P.S. & Arkin, A.P. *PLoS One* **5**, e9490 (2010).
27. Shannon, C.E. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
28. Gutell, R.R., Larsen, N. & Woese, C.R. *Microbiol. Rev.* **58**, 10–26 (1994).
29. Martin, L.C., Gloor, G.B., Dunn, S.D. & Wahl, L.M. *Bioinformatics* **21**, 4116–4124 (2005).
30. Fodor, A.A. & Aldrich, R.W. *Proteins* **56**, 211–221 (2004).
31. Hofacker, I.L., Fekete, M. & Stadler, P.F. *J. Mol. Biol.* **319**, 1059–1066 (2002).
32. Gerstein, M., Sonnhammer, E.L.L. & Chothia, C. *J. Mol. Biol.* **236**, 1067–1078 (1994).
33. Gorodkin, J., Staerfeldt, H.H., Lund, O. & Brunak, S. *Bioinformatics* **15**, 769–770 (1999).
34. Weigt, M., White, R.A., Szurmant, H., Hoch, J.A. & Hwa, T. *Proc. Natl. Acad. Sci. USA* **106**, 67–72 (2009).
35. De Leonardis, E. *et al. Nucleic Acids Res.* **43**, 10444–10455 (2015).
36. Weinreb, C. *et al. Cell* **165**, 963–975 (2016).
37. Fitch, W.M. *Syst. Zool.* **20**, 406–416 (1971).
38. Goebel, B., Dawy, Z., Hagenauer, J. & Mueller, J.C. in *IEEE International Conference on Communications* Vol. 2, 1102–1106 (IEEE, 2005).
39. Rivas, E. & Eddy, S.R. *BMC Bioinformatics* **16**, 406 (2015).
40. Guindon, S. *et al. Syst. Biol.* **59**, 307–321 (2010).
41. Jung, S. *et al. Nucleic Acids Res.* **39**, 7529–7547 (2011).
42. del Val, C. *et al. RNA Biol.* **9**, 119–129 (2012).
43. Wheeler, T.J. *et al. Nucleic Acids Res.* **41**, D70–D82 (2013).

doi:10.1038/nmeth.4066