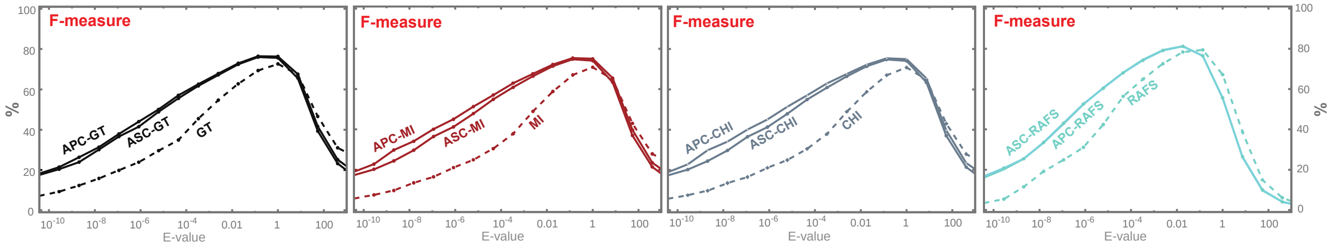
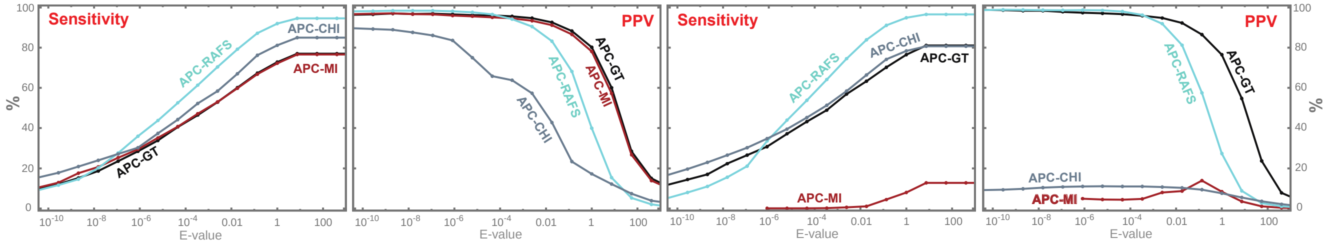


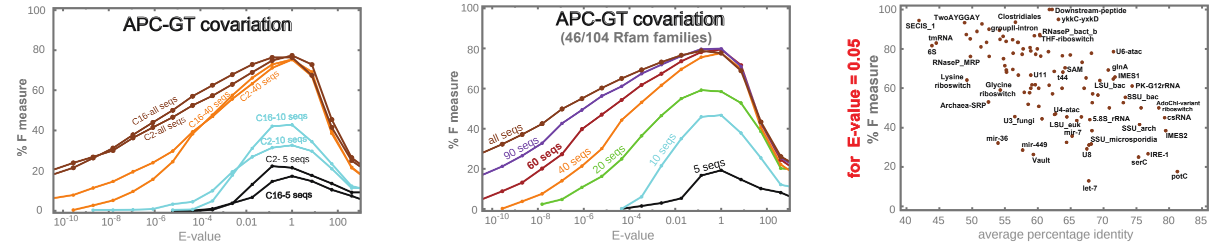
a. Corrected covariation statistics work better



b. GT is the most robust statistic: less affected by false positives (than RAFS) and gaps (than MI)

columns with more than 50% gaps removed

all alignment columns included



c. A two component (WC/non-WC) covariation suitable for few sequences

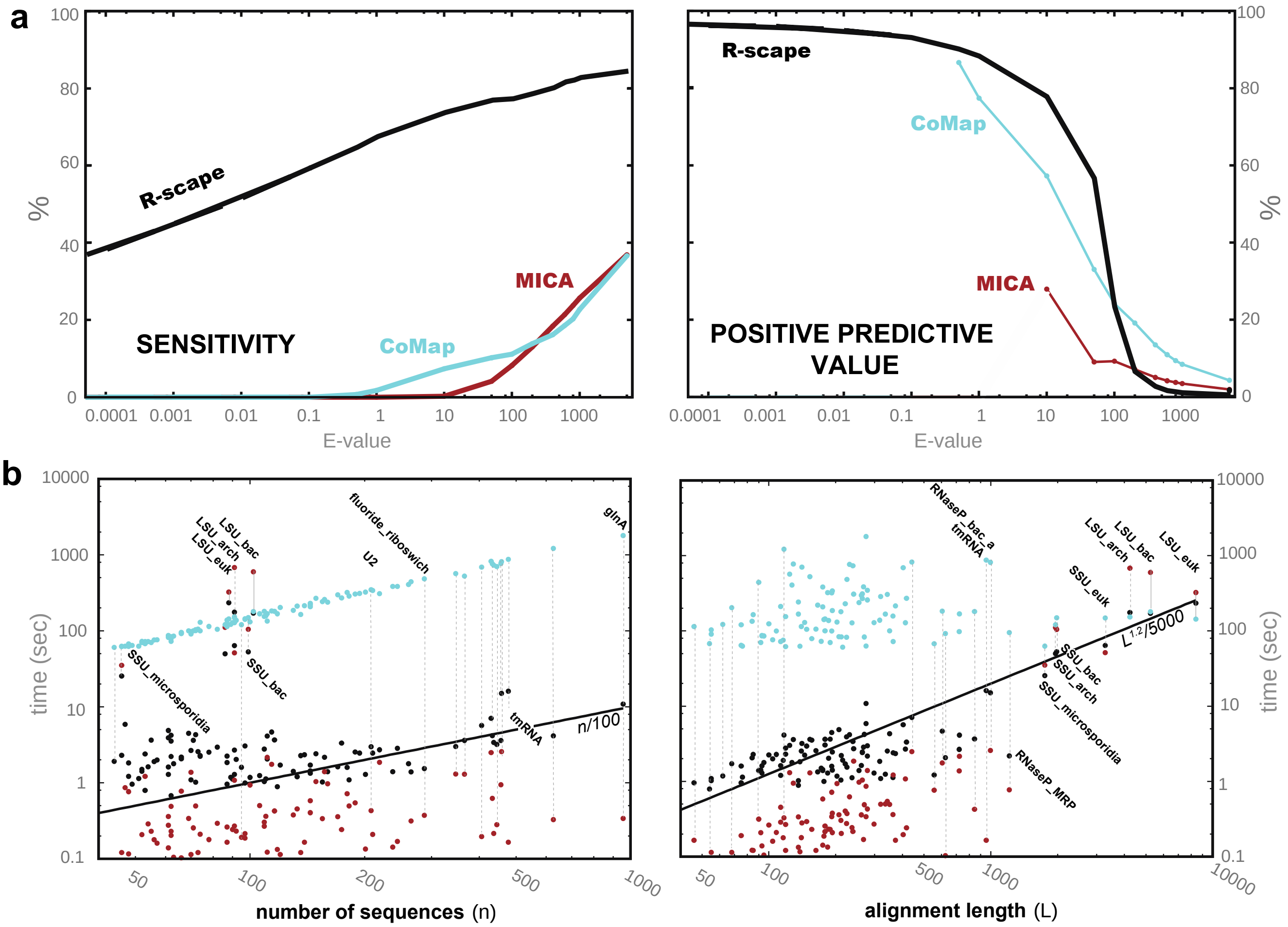
d. Covariation detection improves with more aligned sequences

e. Covariation detection improves with more divergent alignments

Supplementary Figure 1

Characterization of different covariation statistics on a positive testset of 104 RNAs.

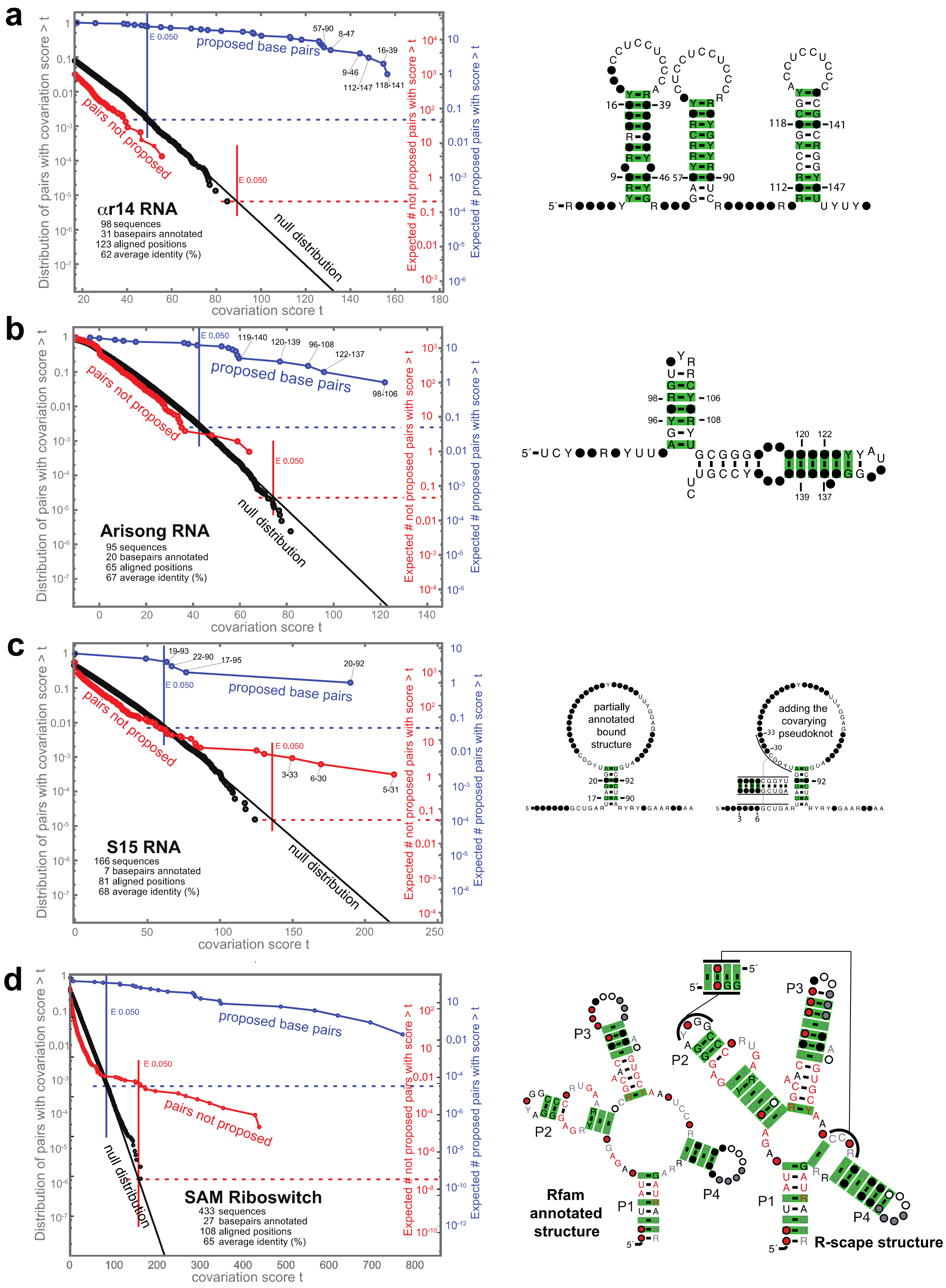
(a) Plots of the F measure---the harmonic mean of sensitivity (SEN) and positive predictive value (PPV), F=2*SEN*PPV / (SEN+PPV)---for four different covariation statistics as a function of the score's E-value, over all alignments, using R=scape with default parameters. **(b)** Effect of alignment gaps on the different covariation statistics, seen by including all alignment columns (right) as compared to the R-scape default (left). **(c)** Effect of measuring covariation using a binary classification (whether a pair is canonical Watson-Crick/G:U or not) versus using the full sixteen-way classification. **(d)** Covariation detection as a function of the number of sequences in the alignments. **(e)** The F measure for each of the 104 RNA Rfam alignments in the positive testset as a function of average percentage identity, at an E-value threshold of 0.05.



Supplementary Figure 2

Comparison of R-scape to related methods CoMap and MICA [12] on the testset of 104 RNAs.

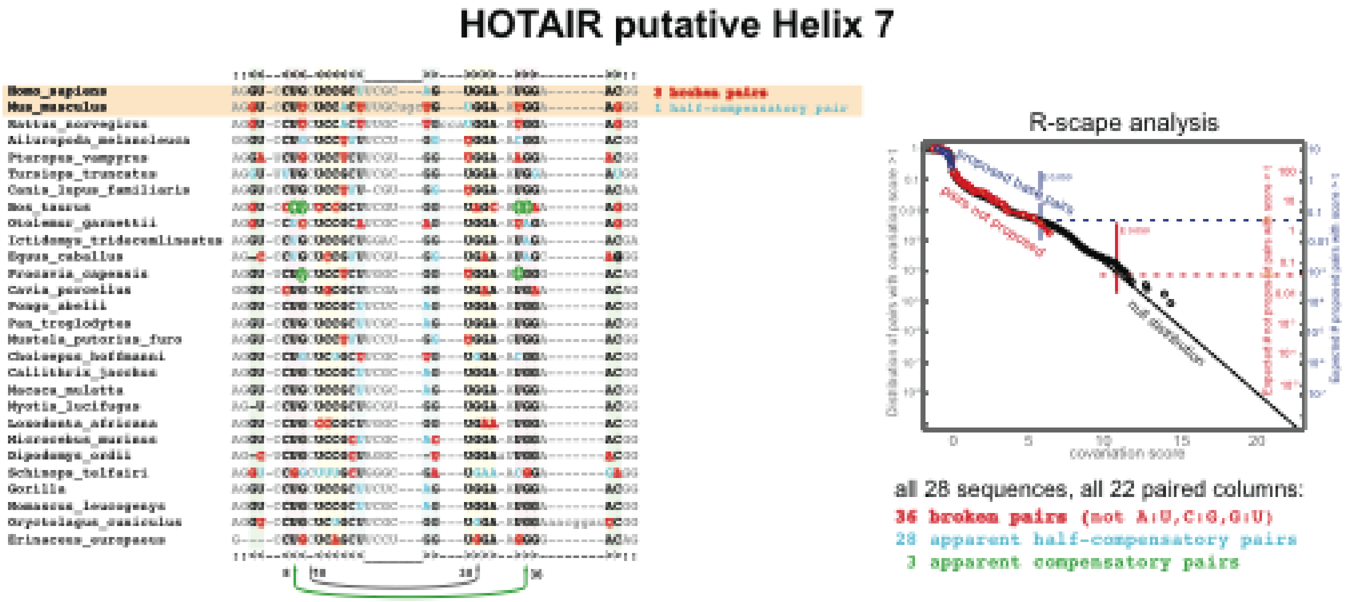
(a) Sensitivity (percentage of significant base pairs) and positive predictive value (percentage of significant pairs that are base pairs) as a function of the score's E-value. **(b)** Running times for the three methods (R-scape in black, CoMap cyan, MICA red) on a log-log plot as a function of the number of sequences in the alignment (left) and as a function of the alignment length (right). Running times are for a single 3GHz intel Core i7 with 8GB 1600GHz DDR3 RAM. Running times for R-scape and CoMap include the cost of generating a phylogenetic tree using FastTree [26].

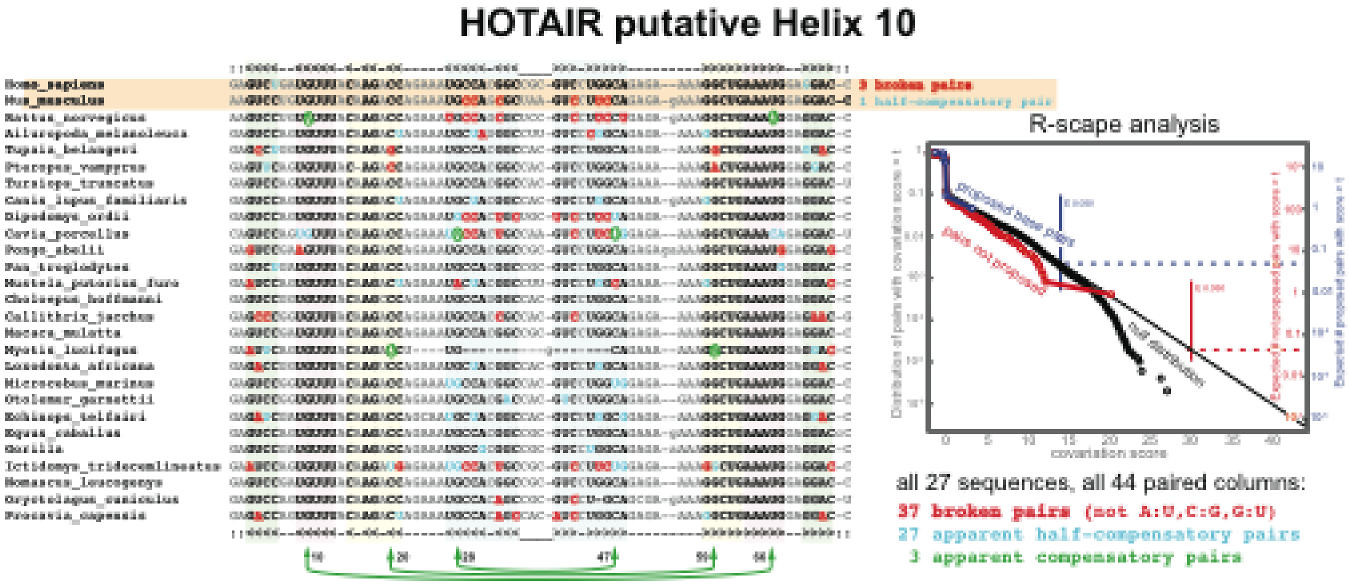

a
αr14 RNA
98 sequences
31 basepairs annotated
123 aligned positions
62 average identity (%)
proposed base pairs
pairs not proposed
null distribution
E 0.050
57-90
8-47
16-39
9-46
112-147
118-141
16
39
118
141
9
46
57
90
112
147
5′
b
Arisong RNA
95 sequences
20 basepairs annotated
65 aligned positions
67 average identity (%)
proposed base pairs
pairs not proposed
null distribution
E 0.050
119-140
120-139
96-108
122-137
98-106
98
106
96
108
120
122
139
137
5′
c
S15 RNA
166 sequences
7 basepairs annotated
81 aligned positions
68 average identity (%)
proposed base pairs
pairs not proposed
null distribution
E 0.050
19-93
22-90
17-95
20-92
3-33
6-30
5-31
partially annotated bound structure
adding the covarying pseudoknot
20
92
17
90
-33
-30
3
6
d
SAM Riboswitch
433 sequences
27 basepairs annotated
108 aligned positions
65 average identity (%)
proposed base pairs
pairs not proposed
null distribution
E 0.050
P1
P2
P3
P4
5′
Rfam annotated structure
R-scape structure
Distribution of pairs with covariation score > t
covariation score t
Expected # not proposed pairs with score > t
Expected # proposed pairs with score > t

Supplementary Figure 3

Examples of RNAs with significant covariation support for their proposed structures.

(a) R-scape analysis of a multiple sequence alignment of αr14, a putative regulatory small RNA in α-proteobacteria [20,42]. **(b)** R-cape analysis of a multiple sequence alignment of Arisong RNA, a noncoding RNA identified in the ciliate *Oxytricha* [41]. **(c)** Example of detecting an underannotated structure, an S15 mRNA leader in γ-proteobacteria that autoregulates ribosomal protein synthesis [19]. Three out of the seven significantly covarying pairs are not in the proposed structure. These covarying pairs support the existence of a conserved pseudoknot, which was already known, but happened to not be annotated in the provided alignment [19]. **(d)** Example of using R-scape to improve a structural annotation for the Rfam seed alignment for SAM-I riboswitch. The R-scape modified structure has seven significant pairs not included in the Rfam-annotated SAM-I structure. The R-scape structure is in agreement with the secondary structure derived from the SAM-I riboswitch crystal structure (RK Montange & RT Batey, *Nature,* {\bf 441}**441**, 1172-1175, 2006). Notation is as in Figure 2.

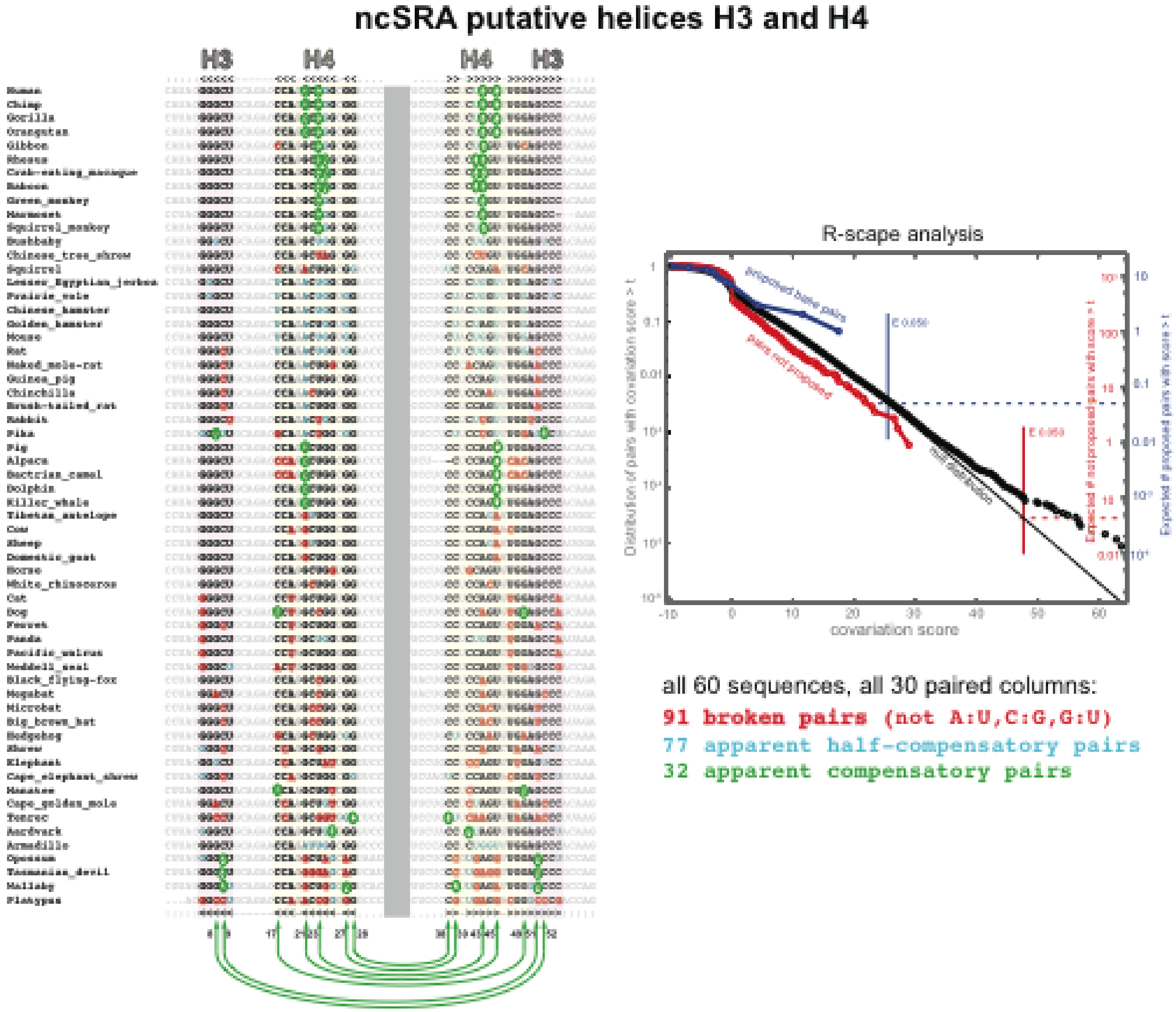




Supplementary Figure 4

Covariation analysis of HOTAIR putative helices H7 and H10.

The structural alignments have been extracted from the HOTAIR Domain1 alignment (with 37 sequences) provided in [13]. The H7 and H10 alignments have 28 and 27 sequences respectively, after removing species for which the region does not include any residues. For any two base paired positions, changes are annotated in color relative to the most frequent Watson-Crick or G:U pair. Green arrows indicate the base pairs (one for H7 and 3 for H10) proposed as covarying in [13]. For putative helix H7, the proposed covarying pair (columns 8:36 marked in green) has covariation score -0.16 (E-value 7.74). Gray arrows indicate the best scoring putative Watson-Crick pair (columns 10:30, with a consensus C:G) which was not part of the proposed structure. This best scoring alternative pair would have one U:A compensatory and one U:G half-compensatory changes, and covariation score 3.66 (E-value 5.52). For both alignments, we also provide the R-scape analysis for all pairs. For putative helix H10, the one covariation above the null hypothesis corresponds to a G:G/U:C non-Watson-Crick covariation in a pair of adjacent columns that are not in the proposed structure and are too close to be a base pair.

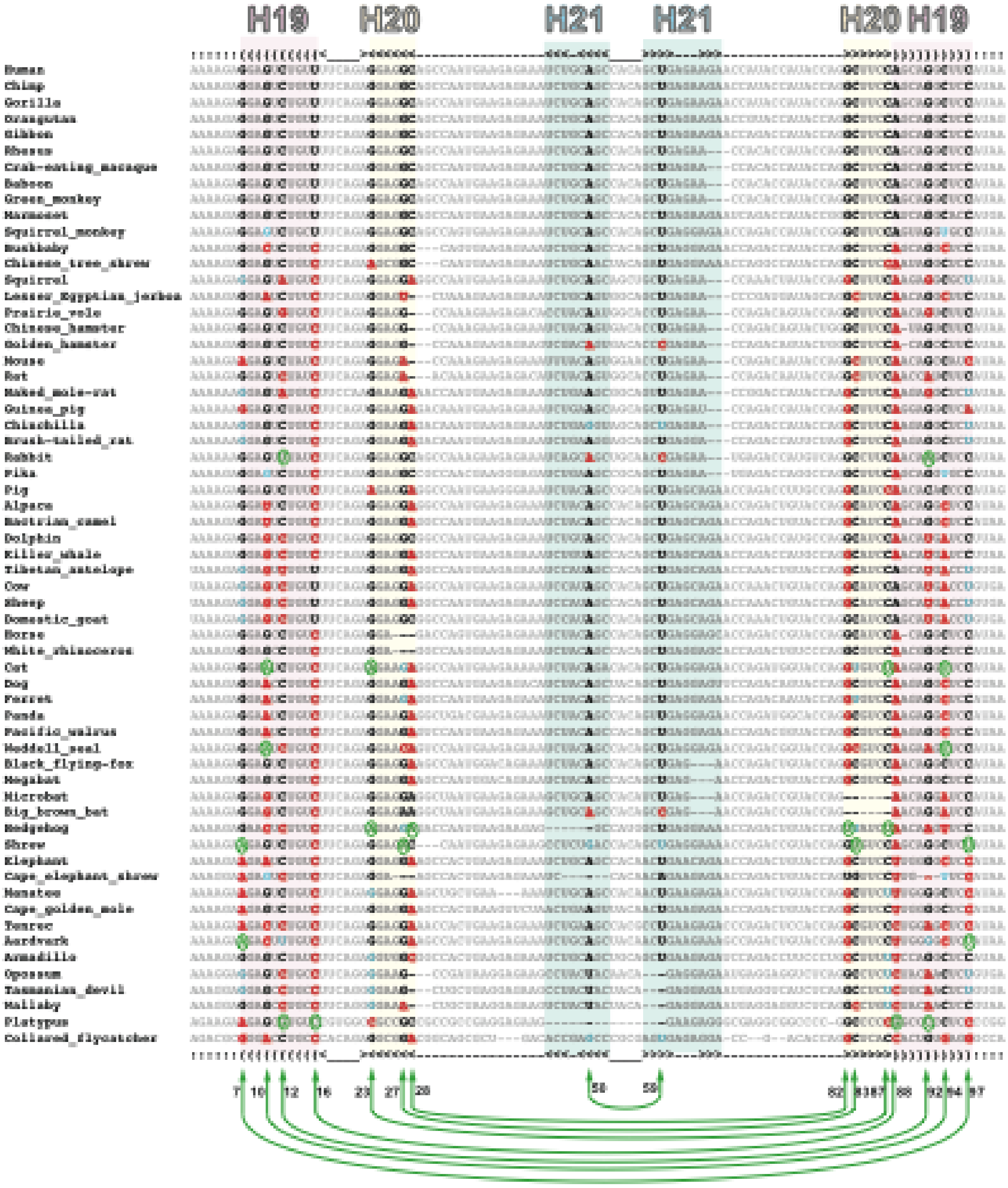


Supplementary Figure 5

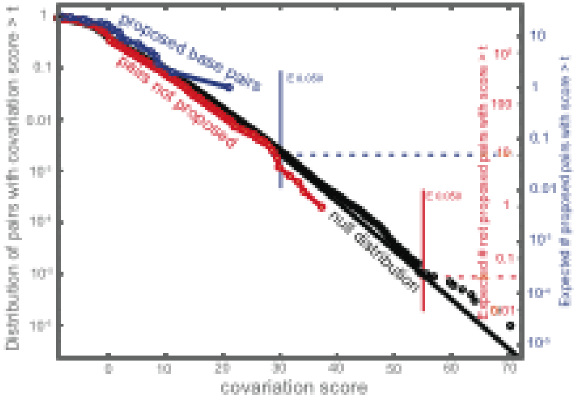
Covariation analysis of putative helices H3 and H4 of ncSRA.

Color annotation as in Supplementary Figure 4. Green arrows indicate the seven base pairs identified in [21] as significantly covarying. We also provide the R-scape analysis for all pairs in this partial ncSRA alignment.

ncSRA putative helices H19-H21



R-scape analysis



all 74 sequences, all 48 paired columns:

279 broken pairs (not A:U,C:G,G:U)

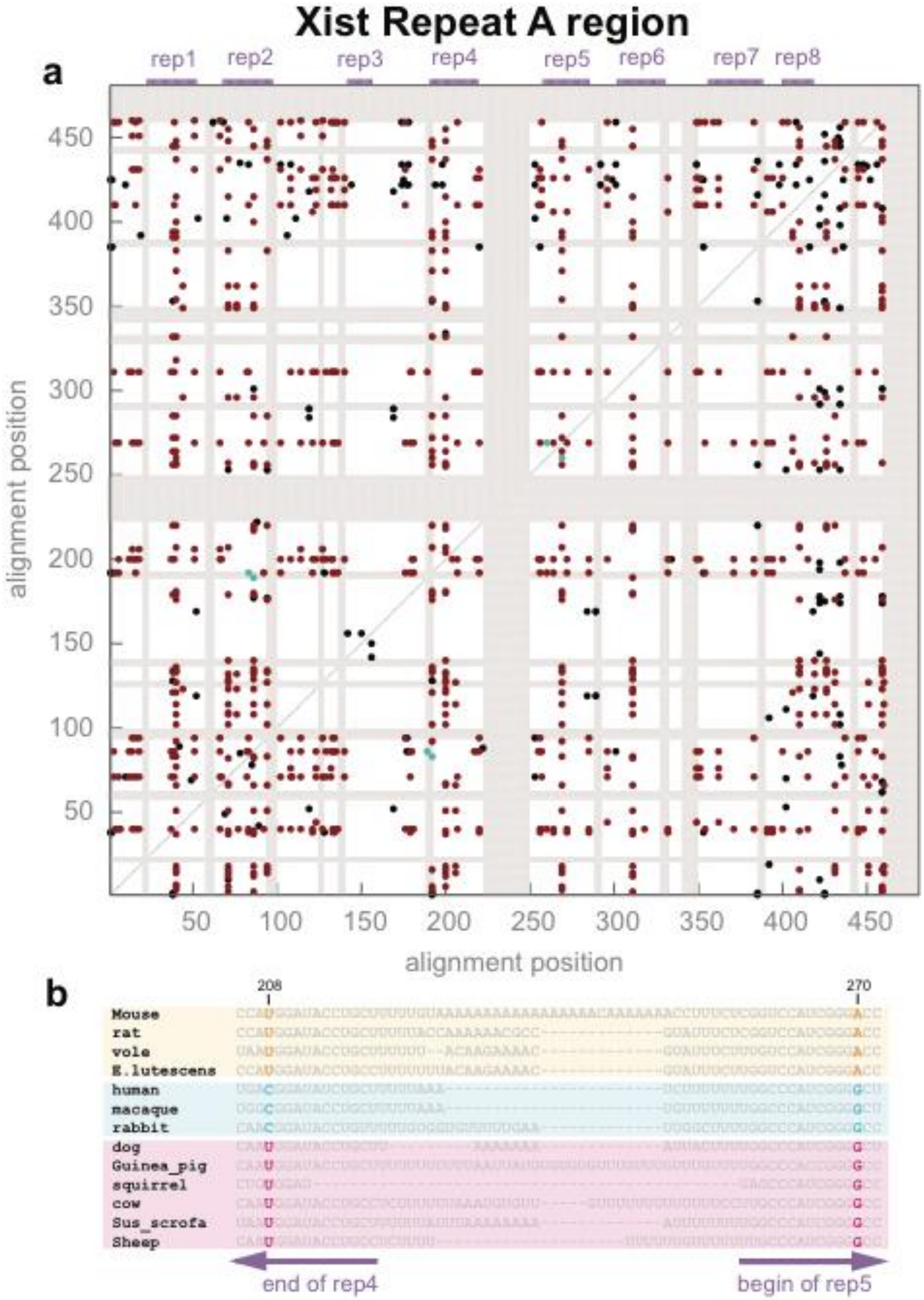
87 apparent half-compensatory pairs

19 apparent compensatory pairs

Supplementary Figure 6

Covariation analysis of putative helices H19, H20, and H21 of ncSRA.

Color annotation as in Supplementary Figure 4. Green arrows indicate eight base pairs identified in [21] as significantly covarying. We also provide the R-scape analysis for all pairs in this partial ncSRA alignment.


Xist Repeat A region
a
rep1
rep2
rep3
rep4
rep5
rep6
rep7
rep8
alignment position
50
100
150
200
250
300
350
400
450
alignment position
b
208
270
Mouse
rat
vole
E.lutescens
human
macaque
rabbit
dog
Guinea_pig
squirrel
cow
Sus_scrofa
Sheep
end of rep4
begin of rep5

Supplementary Figure 7

Apparent covariations in 13 aligned Xist RepA region sequences [23].

(a) An alignment column pair was counted as covarying in [23] if it is entirely consistent with Watson-Crick or G:U base pairing, and at least one substitution and no more than two gaps are observed in each column. The dot plot shows 541 column pairs that satisfy these criteria in the RepA alignment used in [23], including (in blue) three of the four cited as support for the secondary structure in [30] (the other has a A:A non canonical pair, thus does not strictly satisfy the rule), 454 pairs that consist of a U+C column and a G+A column (red), and 84 other pairs (black). **(b)** Example of how single substitutions in conserved U+C and G+A columns can create apparent covariation

RNA	source	total bpairs in structure	number of sequences	average identity (%)	alignment length
HOTAIR Domain1	[13] (provided by authors)	149	37	74 (71)	526 (792)
HOTAIR Domain2	[13] (provided by authors)	143	31	74 (73)	515 (794)
HOTAIR Domain3	[13] (provided by authors)	125	34	68 (68)	468 (571)
HOTAIR Domain4	[13] (provided by authors)	165	31	69 (68)	637 (884)
SRA ncRNA	similar to [21]	234	76	78 (77)	887 (1181)
XIST RepA Structure0	derived from [22]	50 (53)	10	82 (77)	420 (560)
XIST RepA Structure1	derived from [22]	74 (90)	10	82 (77)	420 (560)
XIST RepA Structure2	derived from [22]	69 (72)	10	82 (77)	420 (560)
XIST RepA Structure3	derived from [22]	79 (83)	10	82 (77)	420 (560)
XIST RepA Targeted Structure-Seq	derived from [23]	88 (99)	13	76 (75)	442 (481)
Arisong RNA	[41] (updated from authors)	20	95	66 (65)	65 (150)
αr45	[20,42] (updated from authors)	52	31	87 (86)	180 (186)
αr35	[20,42] (updated from authors)	45	5	78	144 (146)
αr15	[20,42] (updated from authors)	29	51	73 (71)	112 (129)
αr14	[20,42] (updated from authors)	31	98	62 (61)	123 (153)
αr9	[20,42] (updated from authors)	40	26	78 (76)	146 (171)
αr7	[20,42] (updated from authors)	35	26	65 (64)	144 (168)
L1RNA	[19]	9 (15)	703	57 (54)	31 (72)
L10RNA	[19]	16 (59)	805	49 (46)	78 (319)
L20RNA	[19]	34	150	65 (63)	87 (127)
L4RNA	[19]	59	172	61 (58)	197 (328)
S15RNA	[19]	7	166	68 (67)	81 (109)
S1RNA	[19]	24	197	61 (60)	117 (179)
S2RNA	[19]	17 (45)	614	47 (43)	96 (279)
S4RNA	[19]	10	178	74 (73)	110 (129)
S7RNA	[19]	33 (34)	158	812 (79)	104 (179)
S8RNA	[19]	30	167	82	105 (108)
tRNA	RF00005 [15]	21	954	45 (44)	71 (118)
RNase P RNA Bacterial	RF00010 [15]	102	458	60 (58)	367 (996)
Purine Riboswitch	RF00167 [15]	22	133	55	102 (113)
SAM-I Riboswitch	RF00162 [15]	27	433	64 (63)	108 (231)
hAT-Charlie DNA transposon	DF0000021 [43]	0	2,000	28 (38)	181 (16,796)
Alu-related SINE/putative ncRNA	DF0000073 [43]	0	456	84 (83)	133 (403)
Long Terminal Repeat of retrovirus HERV1	DF0000167 [43]	0	13	85 (84)	520 (562)

Supplementary Figure 8

Properties of the structural alignments used in this study.

The alignments we analyzed are derived from the original alignments such that columns with less than 50% occupied positions are not considered. Information for the original alignments is given in parentheses if different from the analyzed alignment. Alignments are available as Stockholm files in the online Supplementary Information.