Supporting Material

Mammalian Genomes Contain Thousands of Non-Coding RNAs with Conserved Secondary Structure

Stefan Washietl, Ivo L. Hofacker, Peter F. Stadler

Supporting Text: Methods

Alignments

Genome-wide alignments of vertebrates (“multiz8way”) were downloaded from the UCSC genome browser (*1*). The alignments included sequences of up to eight species: Human (hg17), chimp (panTro1), mouse (mm5), rat (rn3), dog (canFam1), chicken (galGal2), zebrafish (danRer1) and fugu (fr1). The chimp sequences were removed from the alignments because human and chimp are so similar that sequence differences between them provide essentially no information on RNA structure conservation.

Selection of the most conserved non-coding regions

We started from the “Most Conserved” track generated by the PhastCons program. This track was edited as follows:

1. Adjacent conserved regions that are separated by <50 nucleotides were joined because many known ncRNAs are not conserved over the full length but only contain shorter fragments of highly conserved regions (in microRNA precursors, for example, the two sides of the stems are detected as conserved while the loop region in between is not).
2. Conserved regions (after the joining step) with a length <50 nucleotides were removed because shorter RNA secondary structures are below the detection limit of RNAz.
3. All regions with any overlap with annotated coding exons according to the “Known Genes” and “RefSeq Genes” annotation tracks were removed.

The initial set of alignments consisted of all Multiz alignments corresponding to regions in the modified “Most Conserved” track. After the processing steps described below, we only considered alignments which were conserved at least in the four mammals (***“input alignments”***).

RNAz screen

The input alignments where screened for structural RNAs using RNAz (version 0.1.1) (*2*). Alignments with <200 columns were used as a single block. Alignments with length >200 were screened in sliding windows of length 120 and slide 40. This window size, on the one hand, appears long enough to detect local secondary within long ncRNAs and, on the other hand, is small enough to detect short ncRNAs (appr. 50–70 nucleotides) without loosing the signal in a much too big window.

The individual alignment block presented to RNAz were further processed in the following way:

1. We discarded alignments in which the human sequence contained masked positions by RepeatMasker. The vast majority of repeats was already filtered out in the input alignments: either they were not aligned by Multiz or not detected by PhastCons.

2. Some alignments in the input set contained a large fraction of gaps resulting from a documented problem of PhastCons when treating missing data. We therefore further edited the alignments and removed sequences with more than 25% gaps. The region was regarded as not conserved in this species. If the human reference sequence contained more than 25% gaps, the complete alignment was discarded.

3. The classification model of RNAz is currently only trained for up to six sequences. Therefore, we removed one sequence from alignments which were conserved in all seven species. One of the two sequences in the most similar pair of sequences in the alignment was removed because this pair provides the least comparative information. For the same reason only one representative was retained if two or more sequences in the alignment were 100% identical.

4. Columns of gaps were removed from the reduced alignments.

The resulting alignments were scored with RNAz using standard parameters. All alignments with classification score $p > 0.5$ were stored.

Finally, overlapping hits (resulting from hits in overlapping windows and/or hits in both the forward and reverse strand) were combined into clusters. The corresponding region in the human sequence was annotated as “structured RNA” with the maximum p value of the single hits in the cluster. It must be pointed out, however, that not each of these regions necessarily corresponds to a single RNA gene in the sense of a genetic unit. Long mRNA-like ncRNAs may contain several independent conserved structures. On the other hand, it is possible that multiple small ncRNAs with short intervening sequences are combined into a single cluster by our procedure. (For example, the six members of the *mir-17* cluster in Fig. 1B correspond to only four RNAz clusters). Therefore, and because of the limited data available on expression mechanisms and splicing patterns of ncRNAs, it is difficult to give more than an order-of-magnitude estimate for the number of ncRNAs in the humane genome.

Estimating specificity

The specificity of RNAz was found to be $\approx 99\%$ and $\approx 96\%$, for $p = 0.9$ and $p = 0.5$, respectively (*2*). For benchmarking RNAz we used a defined set of high quality CLUSTAL W alignments of 2–4 sequences and 60%–80% mean pairwise identity.

In this screen, however, we used automatically generated genome-wide alignments essentially based on `Blast` hits. It was therefore not clear if the specificity is the same on these alignments and how other parameters (e.g. the sliding window) affects the false positive rate. We therefore estimated the false-positive rate for this particular special screen. To this end, we repeated the complete screen in exactly the same manner on randomized alignments. Alignments $<$200 columns were randomized as a whole, alignments $>$200 were randomized in non-overlapping windows of 200 before they were sliced in windows for scoring as described above for the true data.

For randomization, we used a slightly modified version of the program `shuffle-aln.pl` (available on request) which is described in detail in reference (*3*). This program shuffles the positions in an alignment in order to remove any correlations arising from a native secondary structure. It takes care not to introduce randomization artifacts and generates random alignments of the same length, the same base composition, the same overall conservation, the same local conservation and the same gap pattern.

This procedure is very conservative and we found that it cannot remove the signal in all cases. The number of possible permutations is reduced if all of the alignment characteristics mentioned above are strictly preserved. Furthermore, the typical mutation pattern of non-coding RNAs is not removed by shuffling of the columns. The number of "compatible" columns which can form a base pair in the consensus structure remains the same. This is one reason why we observe many random hits overlapping with native hits (Supporting Table 1).

In a screen of the urochordate *Ciona intestinalis* based on pairwise alignments (*4*), `RNAz` detected more than 300 tRNAs (about 55% of the `tRNAscan-SE` predictions) but found at $p > 0.5$ only 2 out of the more than 600 tRNA-pseudogenes predicted by `tRNAscan-SE`. This shows that `RNAz` very efficiently distinguishes between RNA secondary structures that are under stabilizing selection and similar sequences for which the selection pressure has been relaxed.

Sensitivity on microRNAs and snoRNAs

We used the "sno/miRNA" track created from the microRNA Registry (*5*) and the snoRNA-LBME-DB maintained at the *Laboratoire de Biologie Moléculaire Eucaryote*. The track contained 207 unique microRNA loci, 86 H/ACA snoRNA, and 256 C/D snoRNAs. We compared our predictions with the annotation tracks using the "Table browser" feature of the UCSC Genome Browser. Loci overlapping with our predictions were counted as detected. Loci that did not show any overlap with our input alignments were counted as "Not in input set" (Fig. 2C). We found that most of the microRNAs and snoRNAs are missed in our screen because they are not in our input set. To optimize future screens, and in particular sub-screens for miRNAs and H/ACA snoRNAs, we investigated in detail why miRNAs and H/ACA snoRNAs were missed in our selection of input alignments (Supporting Tables 3 and 4). MicroRNAs are mainly missed because they overlap with repeats or because they are not strictly conserved in all four mammals (It is more likely that the corresponding sequences are simply missing in one of the unfinished draft assemblies, in particular of the rat genome.) H/ACA snoRNAs are not well conserved on sequence level and `PhastCons` cannot detect conserved regions $>$50 nucleotides in many of them. In the case of C/D snoRNAs the problem is even more pronounced. Out of the 129 C/D snoRNAs not in our set, 63 are completely missed by PhastCons, in most of the other cases only short regions $<$50 are detected. Moreover, many snoRNAs which are contained in our set are not conserved over the full length. Given the fact the C/D snoRNAs in general do not exhibit very stable structures, the detection for RNAz is even more difficult if significant portions of the structure are missing in the input alignments.

Non-coding RNA annotation

We compared all hits to available databases of non-coding RNAs:

1. Rfam release 6.1 , August 2004 (*6*)
2. RNAdb, August 2004 (*7*)
3. NONCODE release 1.0, March 2004 (*8*)
4. microRNA registry release 5.0, September 2004 (*5*)
5. UTRdb, April 2004 (*9*)

We generated BLAST libraries for each of the databases[1] and matched the human sequence of all the detected RNAz clusters against them. Tab. 2 reports BLAST hits with E-values $E < 10^{-6}$.

Annotation relative to protein coding genes

For annotating the RNAz hits relative to known protein coding genes (Fig. 2D) , we used the "Known Genes" and "RefSeq Genes" annotation tables from UCSC genome browser. The UTR annotation is partly ambiguous. As a result, some hits in the second pie chart in Fig. 2D are classified both as intron of a coding region and UTR. Counting only unambiguous annotations, 9825, 2095 and 1987 hits are annotated as intron of coding region, 3'-UTR and 5'-UTR, respectively.

Supporting Text: Additional information for Figure 3

Fig. 3 shows 7 examples of conserved secondary structures predicted with $p > 0.9$. None of them has detectable sequence similarity to any described functional RNAs. The structures A,D,G are conserved across all vertebrates. B,C,F are conserved in all four mammals and chicken. Structure E is specific for mammals.

A Chr. 22, pos. 18,488,478, in intron of "RAN binding protein 1" (***D38076***)

B Chr. 5, pos. 169,242,548, in intron of "Dedicator of cytokinesis 2" (***D86964***)

C Chr. 11, pos. 116,734,202, in intron of KIAA1052 protein (***AB028975***)

[1] In case of the UTRdb we used the EMBL formatted files from ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR/data/ and extracted all annotated UTR elements >20 with flanking regions of 30 to build the BLAST library.

D Chr. 7, pos. 148,937,719, region without annotation

E Chr. 12, pos. 74,595,654, region without annotation. This hit has sequence similarity to a transcript in the Chr. 7 set of RNAdb. We found more than 50 conserved secondary structures throughout the genome with sequence similarity to this transcript. Within these hits, we could identify this structural motif 7 times by visual inspection.

F Chr. 12, pos. 77,343,753, region without annotation

G Chr. 10, pos. 32,739,292, region without annotation

References

1. W. J. Kent, *et al.*, *Genome Res* **12**, 996 (2002).
2. S. Washietl, I. L. Hofacker, P. F. Stadler, *Proc. Natl. Acad. Sci. USA* **102**, 2454 (2005).
3. S. Washietl, I. L. Hofacker, *J. Mol. Biol.* **342**, 19 (2004).
4. K. Missal, D. Rose, P. F. Stadler, *ECCB 2005* (2005). Under review.
5. S. Griffiths-Jones, *Nucl. Acids Res.* **32**, D109 (2004).
6. S. Griffiths-Jones, *et al.*, *Nucleic Acids Res* **33**, D121 (2005).
7. K. C. Pang, *et al.*, *Nucl. Acids Res.* **33**, D125 (2005). Database issue.
8. C. Liu, *et al.*, *Nucl. Acids Res.* **33**, D112 (2005). Database issue.
9. G. Pesole, *et al.*, *Nucl. Acids Res.* **30**, 335 (2002).

Supporting Table 1:
Detailed results of the native screen and the random control.

Native screen							
Set		clusters	size (MB)	% of input	% of genome	cluster length average	 maximum
A: Set 1	$p > 0.5$	91,676	12.47	15.09	0.44	136	1320
B: Set 1	$p > 0.9$	35,985	5.48	6.62	0.19	152	1320
C: Set 2	$p > 0.5$	20,391	2.80	11.52	0.10	137	665
D: Set 2	$p > 0.9$	8,802	1.34	5.50	0.05	152	665
E: Set 3	$p > 0.5$	2,916	0.38	5.57	0.01	131	488
F: Set 3	$p > 0.9$	996	0.14	2.03	0.00	139	488

Randomized screen								
Set		clusters	Overlap native A	size (MB)	% of input	% of genome	Cluster length average	 maximum
Set 1	$p > 0.5$	26,508	9039	3.20	3.87	0.11	121	496
Set 1	$p > 0.9$	6,898	2555	0.89	1.08	0.03	130	496
Set 2	$p > 0.5$	6,551	2158	0.81	3.35	0.03	124	394
Set 2	$p > 0.9$	2,281	881	0.31	1.26	0.01	134	394
Set 3	$p > 0.5$	795	179	0.096	1.40	0.00	121	338
Set 3	$p > 0.9$	208	63	0.026	0.38	0.00	127	279

Set 1: human/mouse/rat/dog, Set 2 = Set 1 + chicken, Set 3 = Set 2 + fugu or zebrafish
“cluster” refers to clustered regions of overlapping `RNAz` hits as described in “Methods”

Supporting Table 2:
Selected ncRNAs from literature with conserved RNA secondary structures detected in our screen.

Name	Type	$\max p$	hits	Comment
U11	snRNA	0.98	1	
U12	snRNA	0.94	2	
U4atac	snRNA	0.71	3	
U6atac	snRNA	0.98	12	
RNAseP	Ribozyme	0.57	1	
UM 9(5)	Transcript of unknown function	1.0	8	Transcript was found to be differentially expressed in the brain, 7 of the 8 hits match the same region of this long (1241nt) transcript
HUC-1	Other functional transcript	0.95	1	Tissue specific transcript that enhances H19 transcription (an antisense transcript for imprinting)
MALAT-1	transcript of unknown function	1.0	3	three independent hits along this 8kb transcript, which was identified in lung cancer cells as ncRNA
NCRMS	Other functional transcript	0.90	3	three independent hits in this 1.8 kb transcript; identified in rhabdomyosarcoma (RMS); host gene of mir-135a-2
BCMS	Other functional transcript	0.71	1	B-cell neoplasia associated transcript
aHIF	antisense transcript	0.98	1	aHIF is complementary to the 3' untranslated region of HIF1alpha mRNA, which encodes a protein known to stabilize p53 protein during hypoxia and to act as a transcription factor for hypoxia inducible genes
Air	Antisense transcript	0.96	8	Classical mouse model for imprinted antisense transcription.
CNS1	Other functional transcript	0.83	1	Expression of CNS1 accompanies the induction of the hyperacetylation of histone H3 on nucleosomes associated with the interleukin (IL)-4, IL-13 and IL-5 genes in developing Th2 cells
HOXA11 AS	Antisense transcript	0.53	1	
GA3824	Transcript of unknown function	0.74	1	Homo sapiens noncoding RNA GA3824 implicated in autism
XIST	Other functional transcript	1.0	3	Three independent hits in the long transcript responsible for X-inactivation in mammals
TTTY11	Transcript of unknown function	0.98	12	Identified in testis
TTTY3	Transcript of unknown function	0.86	1	Identified in testis
TTTY23	Transcript of unknown function	0.54	1	Identified in testis
His-1	Transcript of unknown function	1.0	2	Two independent hits on the same transcript; activation of this transcript leads to carcinogenesis

Supporting Table 3: MicroRNAs missing from our input set

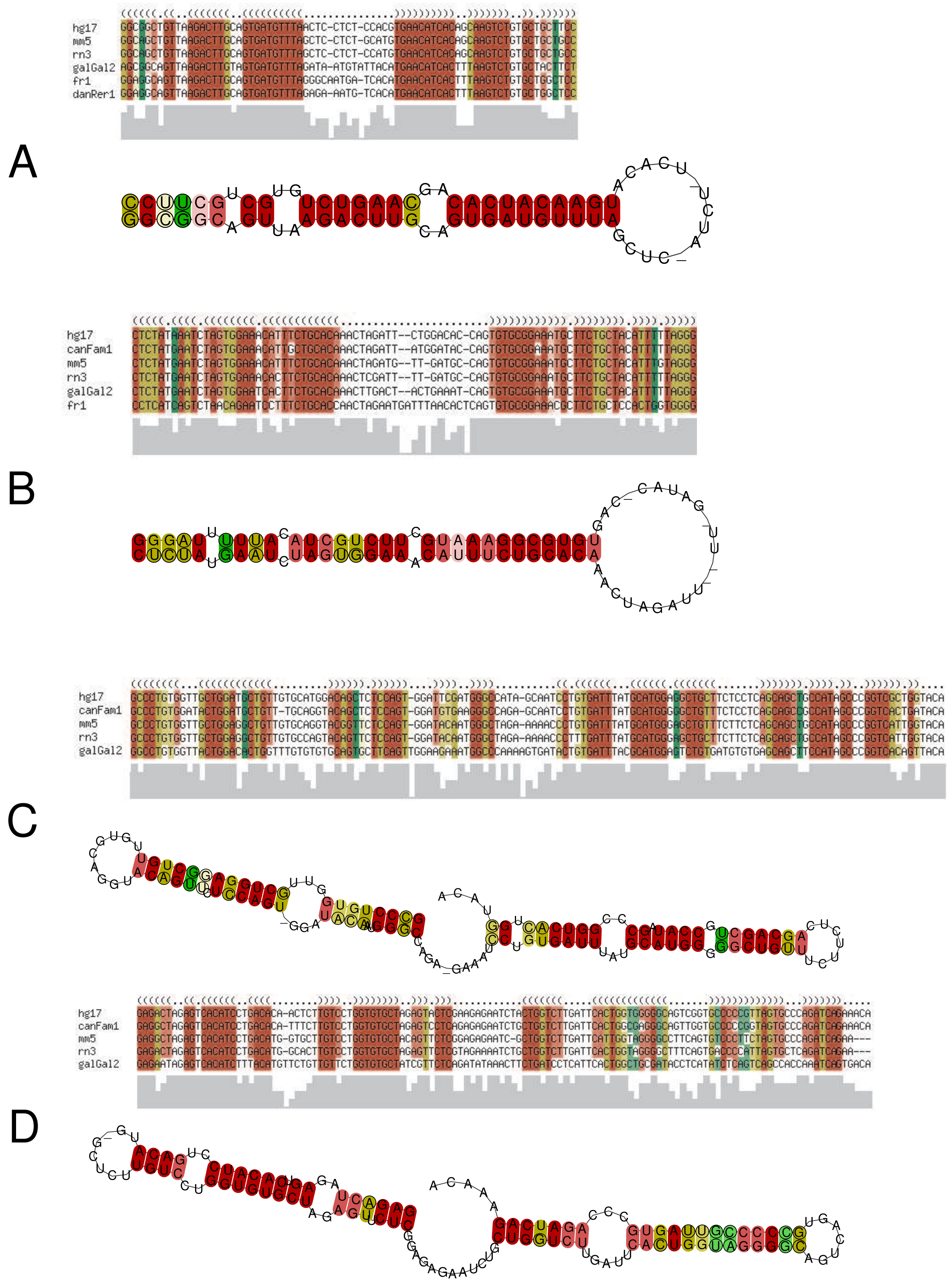
Name	Conservation	Repeat	Other
hsa-let-7g	rat missing		
hsa-let-7i	gap in dog		
hsa-mir-9-1		simple Repeat	
hsa-mir-15a	rat missing		
hsa-mir-16-1	rat missing		
hsa-mir-22			overlap with coding region
hsa-mir-23a			PhastCons artifact[1]
hsa-mir-28		LINE	
hsa-mir-95		LINE	
hsa-mir-130b		SINE	
hsa-mir-133a-2			overlap with coding region
hsa-mir-135a-1	part of rat sequence missing		
hsa-mir-138-1	mouse missing		
hsa-mir-147	PhastCons region <50		
hsa-mir-148a	rat missing		
hsa-mir-149	rat missing		
hsa-mir-150			overlap with coding region
hsa-mir-151		LINE	
hsa-mir-155	rat missing		
hsa-mir-182	part of rat sequence missing		
hsa-mir-197	long gap in mouse		
hsa-mir-198	rat missing		
hsa-mir-199b	rat missing		
hsa-mir-203			PhastCons artifact[1]
hsa-mir-205			overlap with coding region
hsa-mir-212	low complexity		
hsa-mir-302a	rat missing		
hsa-mir-302b	rat missing		
hsa-mir-302c	rat missing		
hsa-mir-302d	rat missing		
hsa-mir-321		tRNA	
hsa-mir-325		LINE	
hsa-mir-326		Arthur 1	
hsa-mir-328	PhastCons region <50		
hsa-mir-330		SINE	
hsa-mir-335	rat missing	SINE	
hsa-mir-337	dog missing		
hsa-mir-340	rat missing	MARNA	
hsa-mir-345		SINE	
hsa-mir-367	rat missing		
hsa-mir-370		SINE	
hsa-mir-371	PhastCons region <50		
hsa-mir-372	PhastCons region <50		
hsa-mir-373	rat and mouse missing		
hsa-mir-374	gaps in mouse and rat	LINE	

[1] PhastCons region extends into the very gap-rich surrounding of the miRNA. Alignment discarded because it contains too many gaps.

Supporting Table 4:
H/ACA snoRNAs missing from our input set

Name	Conservation	Repeat	Other
ACA2A	gap in mouse and rat		
ACA5	`PhastCons` region <50		
ACA5b	`PhastCons` region <50		
ACA10	`PhastCons` region <50		
ACA11	gap in mouse		
ACA29			alignment artifact[1]
ACA33	`PhastCons` region <50		
ACA39	`PhastCons` region <50		
ACA42	not detected by `PhastCons`		
ACA48	not detected by `PhastCons`		
ACA56	rat missing		
ACA59 (Chr. 1)		SINE	
ACA59 (Chr. 17)		SINE	
ACA67	`PhastCons` region <50		
U17a		other	
U17b		other	
U64			alignment artifact[1]
U66	`PhastCons` region <50		
U71a	`PhastCons` region <50		
U71b	rat missing		
U98b	`PhastCons` region <50		

[1] The sequence in chicken is much longer and opens up long gaps in the other sequences, which are thus discarded.



Supporting Figure 1. (Caption see next page)

Supporting Figure 1. Examples of microRNA and H/ACA snoRNA candidates detected with $p > 0.9$.

The miRNA candidates (A,B) exhibit several characteristic features: (i) a stable hairpin consensus structure; (ii) the sequence of one arm of the stem is highly conserved over 22 nt (the putative mature miRNA); (iii) the opposite stem is also conserved but not that strictly; (iv) the loop sequence is diverged due to the absence of functional constraints in this region; (v) compensatory, or at least consistent, mutations are found in the outer parts of the stem where only structure but not sequence is important for function. The sequence in A is located on human chr.20 (pos. 33,041,857) in an intron of a mysine protein gene (***AB040945***). The position of candidate B is chr.15:43,512,536, in the UTR region of FOAP-11 (***AF228422***).

The H/ACA snoRNA candidates (C,D) fold into the typical bipartite hairpin secondary structure. We observe a H-box motifs ANANNA in the hinge regions and ACA motifs in the tail regions. Both candidates can be found in introns of genes implicated in translation. Candidate C is located at chr.9:92,134,300 in an intron of Isoleucine-tRNA synthetase (***D28473***). Candidate D is located at chr.11:8,663,564 in an intron of the ribosomal protein L27a. Primary sequence, secondary structure, and genetic context all strongly suggest a role as classical pseudouridylation guides for these RNAz hits.

Species abbreviations: hg17 human, mm5 mouse, rn3 rat, canFam1 dog, galGal2 chicken, fr1 fugu, danRer1 zebrafish.