Supplementary Materials:
The Tedious Task of RNA Gene Finding

Peter Menzel, Jan Gorodkin and Peter F. Stadler

1 Materials and Methods

1.1 Sources of genomic data

Genomic sequences of 27 species were downloaded from several sources, including ENSEMBL (release 47) [3] for most vertebrate genomes. Table 1.1 lists all genomes, version numbers and the download sources.

We collected alignments with structural annotation for seven RNA families from Rfam (version 8.1)[1] [1]. The SRP RNA alignment was taken from the SRP database[2] [8]. For each family we selected a certain set of seed sequences from the alignments. Except for Y-RNAs and vault RNAs, we chose only mammalian family members contained in the Rfam (or SRP DB) seed alignment of the family. To increase the phylogenetic diversity of the search, we added the *I. iguana*, *G. senegalensis* and *X. laevis* sequences to the Y-RNA and the *R. catesbeiana* to the vault RNA seed set. Note that these species are not included in the genome scans.

1.2 Software for homology search

For the descriptor search we used RNAMotif (version 3.0.4)[3]. RNAMotif takes a descriptor file and a (multi-) fasta file as command line arguments and runs the descriptor against the sequences in the fasta file. No prior database indexing is required. Erpin (version 5.5)[4] was used for automatic motif learning and searching. The accompanying script erpincommand.pl was used to generate the Erpin parameters for a given seed alignment. If we did not find hits with Erpin at all or only a few hits, we reduced the score cutoff in several steps to improve the sensitivity. In those species, in which both RNAMotif and Erpin did not report any true hits, we used RaveNnA [12] to screen these genomes, based on covariance models made of the seed alignments. Due to the high computational complexity of a RaveNnA search, we did this screening only on the SRP, RNAseMRP RNA, U3 RNAs and some of the vault RNAs. Sequence similarity search in the genomes was conducted using BLAST (NCBI, version 2.2.17) using all seed sequences as the query. We set the smallest possible word size of seven (-W 7) and scoring parameters -r 5 -q -4 -G 10 -E 6[5]. For the SRP RNA with a high number of repeats, we used MEGABLAST [13] with word sizes 100 and 24 in the vertebrates to reduce the number of hits. If we found no hits with MEGABLAST, we used the standard blast with word size 7.

1.3 Descriptor building and modifcation

Starting with the seed alignment, we manually construct an RNAMotif descriptor, reflecting the consensus structure of all sequences in the alignment. For stems and unpaired sequences we set minimum and maximum length constraints as they are observed from the seed sequences. After a scan of all genomes with RNAMotif, we modify the descriptor to be less restrictive or be more restrictive depending on the number of RNAMotif hits. Options for loosening the specifity of the descriptor are reducing the minimum length of a stem, allowing a certain amount of non-standard base-pairs (mispairs) and extending the length ranges of unpaired sequences in bulges and loop regions. Additionally we allowed for mismatches in primary sequence constraints. We did not allow for the loss whole stems and the explicit insertions of new stems, although regions described as single-stranded might be capable of folding into substructures. In certain cases, we allow for the loss of small bulges, but do not allow the introduction of bulges in stems, which would mean that we know the position of the new bulge a priori, since the RNAMotif notation does not support the insertion of bulges at random positions in a stem. Note that a decrease of descriptor's specifity does increase search time as well as the number of false positive hits.

[1] http://www.sanger.ac.uk/Software/Rfam/
[2] http://rnp.uthct.edu/rnp/SRPDB/SRPDB.html
[3] ftp://ftp.scripps.edu/case/macke/rnamotif-3.0.4.tar.gz
[4] http://rna.igmors.u-psud.fr/erpin/erpin5.5.serv1.tar
[5] http://stevemount.outfoxing.com/Posting0004.html

Table 1: Sources of genomic data

Species	Source
homo sapiens	ENSEBML release 47
pan troglodytes	ENSEBML release 47
macaca mulatta	ENSEBML release 47
ratturs norvegicus	ENSEBML release 47
mus musculus	ENSEBML release 47
bos taurus	ENSEBML release 47
canis familiaris	ENSEBML release 47
monodelphis domestica	ENSEBML release 47
ornithorhynchus anatinus	ENSEBML release 47
gallus gallus	ENSEBML release 47
anolis carolinensis	AnoCar1.0 assembly, released February 2007, http://www.broad.mit.edu/models/anole/
xenopus tropicalis	ENSEBML release 47
danio rerio	ENSEBML release 47
takifugu rubripes	ENSEBML release 47
branchiostoma floridae	Assembly 1.0, released March 2006 http://genome.jgi-psf.org/Brafl1/Brafl1.home.html
ciona intestinalis	ENSEBML release 47
strongylocentotus purpuratus	Spur 2.1, released September 2006 http://www.hgsc.bcm.tmc.edu/projects/seaurchin/
drosophila melanogaster	Flybase release 4.3, ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r4.3_20060303/
apis mellifera	Amel 4.0 assembly, March 2006, http://www.hgsc.bcm.tmc.edu/projects/honeybee/
caenorhabditis elegans	WormBase 180 assembly, released September 2007 , ftp://ftp.wormbase.org/pub/wormbase/genomes/c_elegans/sequences/dna/
trichinella spiralis	Trichinella spiralis 1.0 assembly, released March 2006, http://genome.wustl.edu/pub/organism/Invertebrates/Trichinella_spiralis/
schmidtea mediterannea	Schmidtea mediterannea 3.1 assembly, released October 2006 http://genome.wustl.edu/pub/organism/Invertebrates/Schmidtea_mediterannea/
capitella capitata	Capitella 1.0 assembly, released October 2006 http://genome.jgi-psf.org/Capca1/Capca1.home.html
hydra magnapapillata	TraceDB, retrieved Dezember 2007, ftp://ftp.ncbi.nih.gov/pub/TraceDB/hydra_magnipapillata/
nematostella vectensis	Nematostella vectensis 1.0 assembly, http://genome.jgi-psf.org/Nemve1/Nemve1.home.html
trichoplax adhaerens	Trichoplax adhaerens 1.0 assembly, released August 2006, http://genome.jgi-psf.org/Triad1/Triad1.home.html
monosiga brevicollis	Monosiga brevicollis 1.0 assembly, released July 2006, http://genome.jgi-psf.org/Monbr1/Monbr1.home.html

1.4 Filtering of hits

All putative RNAMotif hits were run against RepeatMasker[6] [9] to filter repetitive sequences. We set RepeatMasker to mask low complexity / simple repeats (-noint), e.g. $(AT)_n$ repeats which easily match any secondary structure. We did not filter RNA genes (-norna) to prevent masking of true positive hits.

To compare the remaining not masked hits and the hits from the BLAST and Erpin searches to the already known family members, we created a multiple sequence alignment with ClustalW [11].

For the E2 snoRNA, we run snoReport [2] on all hits, to count putative snoRNAs. SnoReport uses a Support Vector Machine to classify whether a given sequence belongs to the H/ACA or C/D snoRNA class or neither of both.

2 Results

2.1 SRP RNA

The SRP RNA seed set consists of 14 mammalian members taken from the SRP database, which we also used for comparing hits to annotated sequences. Scanning all genomes with the resulting RNAMotif descriptor, we found a few hits in all mammals and zebrafish. The descriptors in the second and third round additionally recovered SRP RNAs in lizard, frog, and the teleostei. Erpin found the annotated family members in the vertebrates (except fugu), but none of the known sequences the invertebrates, because the two stems chosen by erpincommand.pl for the search are too different from the vertebrate members. In this family, we also run RaveNnA on the small invertebrata genomes (except *H. magnipapillata*). All annotated sequences were returned among the highest scoring hits. Candidates found with BLAST were also found by RaveNnA. The SRP RNA, also named 7SL-RNA, is the basis for the common SINE repeat family *Alu*, and therefore a large amount of SRP related sequences is contained in the primate and rodent genomes [4]. Instead of the standard BLAST, we used megablast with word sizes of 24 and 200 to narrow down the number of BLAST hits. With word size 24 and the human SRP sequence as query, we find alone 147 BLAST hits with a length above 290 in the human genome, but only 2 hits were retained with word size 200. With word size 200, MEGABLAST returned only hits in the primates and rodents. Despite of all the pseudogene hits in the higher organisms we find SRP RNAs in all species down to trichoplax, in the lower taxa with the standard BLAST word size of 7, since the sequence similarity is no longer sufficient with MEGABLAST's large word sizes. In most species, the BLAST hit with lowest E-Value matched the known sequences except in rat, sea urchin, and schmidtea.

2.2 Y-RNA

The seed sequence set contains known Y-RNAs from human, bushbaby, mouse, rat as well as iguana and frog (*X. laevis*). The RNAMotif descriptor includes the closing stem and a long variable loop region, allowing for the four distinguishable Y-RNA classes Y1, Y3, Y4 and Y5 to fit. Since the structural constraints in this "universal" descriptor are pretty weak, we add sequence constraints in conserved regions of the 5' and the 3' part of the stem. The first genome scan showed hits in all vertebrates, with a lot of hits in human (111), chimp (104) and macaque (78). This is due to a high number of Y-RNA pseudogenes found in mammals. We compared our RNAMotif hits, with the Y-RNA sequences identified in [7]. For 9 of 14 vertebrate species, RNAMotif was able to recover all previously identified Y-RNAs, missing some of the Y classes in the remaining five species. In the next iterations we modified the descriptor to allow more variable bulge and loop regions. This increased the number of hits significantly in all species and previously missed Y-RNAs were detected in dog and opossum. Erpin finds all annotated Y-RNAs among the best scoring hits, except two members in xenopus and the zebrafish Y-RNA. The BLAST search recovered all sequences except the Y5 and Ya in xenopus and the single fugu Y-RNA, which was found by RNAMotif and Erpin. Due to the high number of Y pseudo genes in the mammals, we find not all of distinct Y-RNAs among the hits with lowest E-Values. All three methods failed to find the c. elegans sequence, but still report many additional hits in the invertebrates. An assessment whether those hits are putative Y-RNA genes is outstanding.

[6] http://www.repeatmasker.org

2.3 vault RNA

We used the three human vault RNAs (cluster on chromosome 5), the rat and mouse members and additional two sequences from the bullfrog *R. casbaiana* in our seed sequence set. The initial `RNAMotif` descriptor includes the consensus structure of all 7 sequences and we set sequence constraints for the conserved boxes A and B. The first iteration resulted in few `RNAMotif` hits in the seed genomes as well as in the other primates and cow. We compared the results to the sequences previously listed in [6]. In human, mouse, rat, and cow only the already annotated sequences were reported, in macaque and chimp we find two of the three known vault RNAs. The second iteration resulted in a single additional hit in chimp, but in no other species. In the third iteration we loose constraints on bulge lengths and allow mismatches in the conserved sequence regions. We now find a high number of hits in the whole species range, with only a small fraction filtered by repeat masker and many false positives. All mammalian vault RNAs are among the `RNAMotif` hits, but the opossum, lizard and chicken sequences were not found. The five *X. tropicalis* vault RNAs were retrieved, they feature high similarity to the bullfrog sequences in the seed set. The problem with all descriptors was the stem below the loop region, which has not the length of 4bp in all species, but was required by the descriptor in all three iterations. `Erpin` also missed the oppossum, chicken, frog and some anolis members, but found the annotated sequences of the other species with highest score. With the `BLAST` search we found the same hits as with `RNAMotif` and additionally the opossum and chicken members as well as a candidate vault RNA in lizard. Additionally the putative chicken vault RNA candidate identified in [6] was also identified with `BLAST`. We also compared the results to the candidate vault RNAs of *D. rerio*, *T. rubripes*, *B. floridae*, *C. intestinalis*, and *S. purpuratus* that have been predicted in [10]. Additionally we screened those genomes with `RaveNnA`. With `BlAST` we find only one of the *T. rubripes* candidate sequences. `RaveNnA` was able to identify all of the *D. rerio* and *B. floridae* sequences, one sequence in each of *C. intestinalis* and *S. purpuratus* and four of the five *T. rubripes* candidates sequences. Both `RNAMotif` and `Erpin` reported no true hits in those five species.

2.4 RNAse MRP

For the MRP RNA, we used the sequences from human, rat, mouse and cow in the seed set. All seven eutharia MRP RNAs were retrieved in the first `RNAMotif` run without false positive hits, but no more hits were found in other species with the less strict descriptors in the subsequent screens. `Erpin` was also not able to find MRP RNAs outside eutharia, also with greatly reduced cutoffs. This is probably due to an unfavorable selection of the substructure used for screening, which is only very short and not present in most species. `BLAST` also found almost all vertebrate members with high specifity as well as matches of the ciona, sea urchin, fly and worm MRP RNAs, however it misses the *A. mellifera* RNAseMRP. The `RaveNnA` screen also identified all known MRP RNAs, except the *A. mellifera* sequence.

2.5 E2/ACA6 snoRNA

We use only the human E2 snoRNA and its homolog ACA6 snoRNA in our seed set. In the `RNAMotif` descriptor we set sequence constraints in the regions of the conserved H and ACA boxes. E2/ACA6 annotations for several species were taken from snoRNAbase [5]. In the first scan, `RNAMotif` reported only hits in the primates and in the cow. The second iteration with a slightly looser descriptor we find a lot of hits in all species, but not many more of the known E2/ACA6 sequences. In the third iteration, we added an additional sequence constraint right after the first stem following the H box, and allowed a shortening of the 7bp stem by one base pair and loosened the constraints of several unpaired sequences. This resulted in a major reduction of the number of hits and additional E2 sequences were found in mouse and rat. Thus the third descriptor performed best both in specificity and sensitivity, but is still to restrictive to find E2/ACA6 RNAs in all vertebrates. `Erpin` also found most of the known vertebrate members, except in frog and the fish sequences. Using `BLAST` we find a handful of hits in all vertebrates, recovering both RNAs. In frog and the fish genomes we only find the ACA6 RNA. Candidates for E2 and ACA6 RNAs were returned in platypus and anolis, which show high sequence similarity with oppossum and frog respectively.

2.6 let-7 microRNA

Most species have several different let-7 RNAs. The let-7 homolog mir-98 is found in mammals. Our seed set contains only the 11 human let-7 sequences from Rfam 8.1. For evaluation we compared the hits to annotated sequences in mirBase. In the first `RNAMotif` run, we find various hits in most species, covering most of the annotated let-7s, but also a few false positives. Since the descriptor did not include many sequence constraints, we probably found not only let-7 precursors, but other microRNA precursors too. Thus we set more sequence constraints covering the mature let-7 sequence. The number of hits with the modified descriptor then decreased and we find in at least hits in all vertebrates with only a few false positives. In a third iteration we allowed for two mismatches in the sequence constraint and allowed for the loss of the bulge in the stem, which resulted in few additional hits in *A. mellifera* and *C. elegans* with slightly more false positives. `Erpin` recovered the annotated let-7 homologs for most vertebrates. With `BLAST` search, the results are similar and we find addionally the fly sequence.

2.7 U5 snRNA

The seed sequence set contains known U5 family members from human, rat and mouse. We compared the hits in all species with the predicted U5 RNAs from (Marz et al. 2008, unpublished), and the macaque hits with the predicted Infernal hits. In the first `RNAMotif` run, we found almost all annotated sequences in the vertebrates except in fugu and zebrafish with only few false positives. In the next iteration we allow one mispairing in each stem and now find additional hits in the fish and some basal eukaryotes, but the number of false positive hits increased too. In the third iteration we remove the 3' hairpin from the descriptor, which results in several thousand hits and few more finds, but the specifity of the descriptor is worse. `Erpin` performs good in the chordata but did only find some the annotated sequences in the other taxa. Using the seed sequences as query for the `BLAST` search, we find almost all annotated sequences in all species usually among the HSPs with lowest E-values. Note that the sequences used for comparison were also produced by a blast search and filtered later by sequence-structure alignment.

2.8 U3 snoRNA

The seed set consists of the human, mouse, rat, and cow U3 RNAs. Our descriptor contains only the 3'-loop. In the third iteration the descriptor is sensitive enough to retrieve most vertebrate U3 members and some candidates in branchiostoma and sea urchin. `BLAST` finds all annotated sequences, except the *C. elegans* U3 RNA. Some candidates were found in *A. mellifera*, *T. spiralis*, *C. capitata*, and *N. vectensis*. `Erpin` missed the fugu and lancelet U3 as well as the *c. intestinalis* and *D. melanogaster* members. Both `BLAST` and `Erpin` detect the distant *T. adhaerens* U3 sequence. In this family, we also run `RaveNnA` on the small invertebrata genomes (except *H. magnipapillata*). All annotated sequences were returned among the highest scoring hits, including the *C. elegans* sequences from `RFam`. Candidates found with `BLAST` were also found by `RaveNnA`.

2.9 Summary

In this study we compared three approaches for homology search in whole genomes. The standard sequence-based approach uses `BLAST` to find sequences similar to a query sequence. The second approach utilizes a descriptor based search `RNAMotif`, that incorporates sequence and secondary structure motifs in the query. By manually generating a descriptor out of a sequence alignment with structural annotation and iteratively refining the descriptor depending on the outcome of the search, we try to maximize specificity and sensitivity of the descriptor for a certain family of ncRNA. The third method, `Erpin`, is based on automatic model learning from a structure-annotated alignment and then using that model for screening.

We explicitly did not try to put as much biological knowledge as possible in the input of the three programs, but to examine the prospects of a homology search, with a limited set of training data.

Comparing the results of the methods, we find that they all have strengths and weaknesses. Using a hand-built descriptor, we always have the trade-off between a high sensitivity, i.e. finding even distant homologs by using a loose descriptor, and a high specificity, i.e. having a low false positive rate usually achieved by a more strict descriptor. We showed, that by iteratively changing the descriptor, we can improve the search results both in sensitivity and specificity up to a certain extent, even if we conduct only three iterations. As the modifications we made to the

descriptor were not based on knowing all of the already known RNA family members in advance, but only on several seed members, our approach failed in identifying all known homologs. Especially if the RNA features a complex secondary structure, e.g. SRP or MRP, our modifications go not far enough to capture even family members outside the mammals. On the other hand, a structure-based search is much more specific if the target sequences belong to repeat classes – like the SRP RNA. `Erpin` on the other hand, automatically selects certain highly conserved regions from the seed alignment and only searches for those motifs. This approach typically results in a high specifity, but it is also likely to miss RNAs in phylogenetically distant RNA family members, that do not conserve the selected regions from the seed alignment. For the U3, SRP and RNAse MRP RNAs, we also run `RaveNnA` on the small invertebrate genomes. The sensitivity here was very good, and all of the annotated sequences were retrieved usually with the best score. Also `RaveNnA` performed best in the vault RNAs. The drawback of this method is the high computational effort necessary for whole genome scans. Using `BLAST`, we find annotated homologs in most species, although in some cases `BLAST` was not able to recover all family members, e.g. some vault RNAs and the *C. elegans* U3 RNA.

3 Tables and Figures

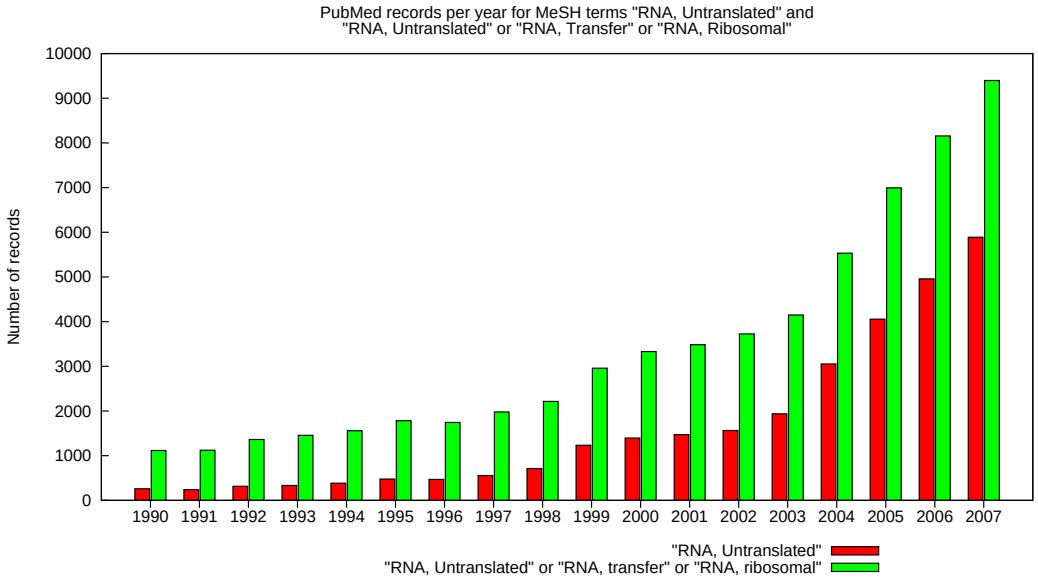

Figure 1: Number of publications per year in the PubMed database associated with MeSH terms for non-coding RNAs.

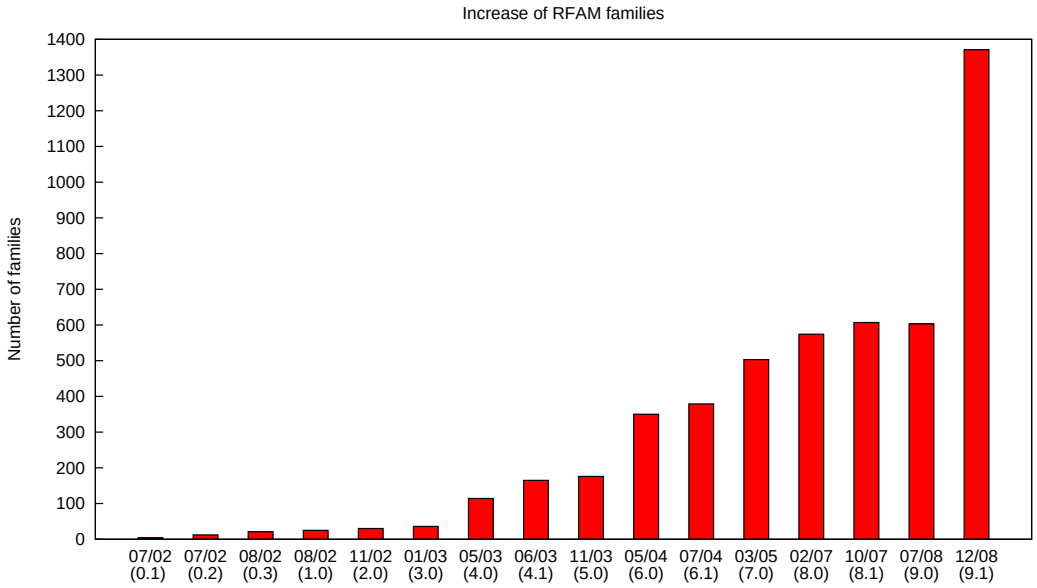

Figure 2: Number of RNA families in the `Rfam` releases.

Table 2: **E2/ACA6 snoRNA:** The column "Species" yields species names along with the number of known snoRNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2/#3 (#4), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of putative snoRNAs as reported by `snoReport`, #3 the number of all non repeat-masked hits and #4 is the total number of hits. For the numbers #1/#2/#3 in the "BLAST" and "Erpin" columns, #1 is the number of true positive hits or highly likely candidates,#2 is the number of putative snoRNAs as reported by `snoReport` and #3 is the total number of hits. If #1 is 2 than both E2 and ACA6 have been found, otherwise it is 1 if only one of both was found. All `BLAST` HSPs with length $\geq$ 120nt and e-value $\leq$ 5.0 and `Erpin` hits with cutoff 26.43 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RNAMotif 1**	**RNAMotif 2**	**RNAMotif 3**	**BLAST**	**Erpin**
homo sapiens (2) ‡	2 / 2 / 2 (6)	2 / 19 / 342 (1795)	2 / 2 / 2 (2)	2 / 2 / 6	2 / 2 / 4
pan troglodytes (2)	2 / 2 / 2 (12)	2 / 22 / 340 (1954)	2 / 2 / 2 (2)	2 / 2 / 6	2 / 2 / 4
macaca mulatta (2)	2 / 2 / 2 (3)	2 / 15 / 273 (711)	2 / 2 /2 (2)	2 / 3 / 6	2 / 2 / 4
rattus norvegicus (2)	0 / 0 / 0 (4)	0 / 18 / 194 (446)	1 / 2 /2 (2)	2 / 2 / 8	2 / 2 / 2
mus musculus (2)	0 / 0 / 0 (3)	0 / 20 / 181 (435)	1 / 2 / 2 (2)	2 / 2 / 4	2 / 2 / 2
bos taurus (2)	1 / 1 / 1 (1)	1 / 27 / 239 (362)	1 / 1 / 1 (1)	2 / 7 / 10	1 / 1 / 2
canis familiaris (2)	0	2 / 18 / 242 (391)	2 / 3 / 4 (4)	2 / 5 / 29	1 / 2 / 8
monodelphis domestica (2)	0 / 0 / 0 (3)	0 / 6 / 222 (397)	0	2 / 2 / 2	2 / 2 / 2[b]
ornithorhynchus anatinus	0	0 / 4 / 26 (33)	0	3 / 0 / 10	3 / 0 /3
gallus gallus (2)	0 / 0 / 0 (1)	0 / 7 / 107 (128)	0	2 / 2 / 2	2 / 2 / 2[b]
anolis carolinensis	0	0 / 12 / 176 (304)	0	2 / 2 / 2	1 / 1 / 1
xenopus tropicalis (2)	0 / 0 / 0 (1)	0 / 15 / 291 (355)	0 / 0 / 1 (1)	1 / 1 / 1	0 / 0 / 1[c]
danio rerio (1)	0 / 0 / 0 (2)	0 / 5 / 165 (361)	0 / 0 / 1 (2)	1 / 1 / 5	0 / 0 / 3[c]
takifugu rubripes (1)	0	0 / 0 / 24 (26)	0	1 / 1 / 1 [a]	0 [c]
branchiostoma floridae	0	0 / 7 / 101 (109)	0	0	0[c]
ciona intestinalis	0	0 / 1 / 20 (27)	0	0	0[c]
strongylocentrotus purpuratus	0	0 / 2 / 95 (95)	0	0 / 0 / 1	0 / 0 / 1[c]
drosophila melanogaster	0	0 / 1 / 10 (11)	0	0 / 0 / 1	0[c]
apis mellifera	0	0 / 1 / 24 (132)	0	0 / 0 / 1	0 / 0 / 1[c]
caenorhabditis elegans	0	0 / 0 / 36 (42)	0	0	0 / 0 / 1[c]
trichinella spiralis	0	0 / 0 / 14 (14)	0	0	0[c]
schmidtea mediterannea	0	0 / 2 / 283 (283)	0 / 0 / 11 (11)	0	0[c]
capitella capitata	0	0 / 2 / 30 (40)	0	0	0[c]
hydra magnipapillata	0 / 0 / 1 (124)	0 / 126 / 2074 (9164)	0 / 0 / 10 (11)	0 / 0 / 2	0 / 0 / 15[c]
nematostella vectensis	0	0 / 2 / 32 (36)	0 / 0 / 2 (2)	0	0[c]
trichoplax adhaerens	0	0 / 0 / 16 (19)	0	0	0[c]
monosiga brevicollis	0	0 / 0 / 2 (2)	0	0	0[c]

[a] HSP length $\leq$ 120nt
[b] `Erpin` cutoff 20.0
[c] `Erpin` cutoff 10.0

Table 3: **SRP RNA:** The column "Species" yields species names along with the number of known SRP RNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2 (#3), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of all non repeat-masked hits and #3 is the total number of hits. For the numbers #1/#2 in the "BLAST", "Erpin" and "Ravenna" columns, #1 is the number of true positive hits or highly likely candidates and #2 is the total number of hits. All `BLAST` HSPs with e-value ≤ 5.0, `Erpin` hits with score ≥ 20.17, and `RaveNnA` hits with score ≥ 20.0 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RnaMotif 1**	**RnaMotif 2**	**RnaMotif 3**	**Blast**	**Erpin**	**Ravenna**
homo sapiens (1) ‡	1 / 3 (3)	1 / 4 (4)	1 / 9 (9)	1 / 2 [a]	1 / 56	-
pan troglodytes (1) ‡	1 / 4 (4)	1 / 5 (5)	1 / 7 (7)	1 / 3 [a]	1 / 58	-
macaca mulatta (1)	1 / 3 (3)	1 / 3 (3)	1 / 8 (8)	1 / 3 [a]	1 / 56	-
rattus norvegicus (1) ‡	1 / 3 (3)	1 / 3 (3)	1 / 3 (3)	1 / 3 [a]	1 / 4	-
mus musculus (1) ‡	1 / 2 (2)	1 / 2 (2)	1 / 2 (2)	1 / 2 [a]	1 / 4	-
bos taurus (1) ‡	1 / 4 (4)	1 / 4 (4)	1 / 7 (7)	1 / 12 [b]	1 / 12	-
canis familiaris	1 / 3 (3)	1 / 3 (3)	1 / 3 (3)	1 / 3 [b]	1 / 9	
monodelphis domestica (1) ‡	1 / 3 (3)	1 / 3 (3)	1 / 3 (3)	1 / 5 [b]	1 / 4	-
ornithorhynchus anatinus (1) ‡	1 / 2 (2)	1 / 2 (2)	1 / 2 (2)	1 / 2 [b]	1 / 2	-
gallus gallus (1)	0	1 / 2 (2)	3 / 3 (3)	1 / 3 [b]	1 / 2	-
anolis carolinensis	0	0	1 / 2 [e] (2)	1 / 3 [c]	1 / 27 [d]	-
xenopus tropicalis	0	0	1 / 21 (21)	1 / 21 [b]	1 / 56 [d]	-
danio rerio (1)	1 / 1 (1)	1 / 2 (2)	1 / 41 (41)	1 / 53 [b]	1 / 88 [d]	-
takifugu rubripes	0	0	1 / 6 (6)	1 / 4 [b]	0 / 5 [d]	1 / 8
branchiostoma floridae (1)	0	0	0	1 / 15 [c]	0 / 51 [d]	1 / 70
ciona intestinalis (1)	0	0	0	1 / 21 [c]	0 / 1 [d]	1 / 12
strongylocentrotus purpuratus (1)	0	0	0	1 / 8 [c]	0 / 8 [d]	1 / 7
drosophila melanogaster (1)	0	0	0	1 / 2 [c]	0 / 3 [d]	1 / 2
apis mellifera	0	0	0	1 / 1 [c]	0 / 6 [d]	1 / 1
caenorhabditis elegans (4)	0	0	0	4 / 5 [c]	0 / 2 [d]	4 / 5
trichinella spiralis (1)	0	0	0	1 / 1 [c]	0 / 4 [d]	1 / 1
schmidtea mediterannea	0	0	0	1 / 12 [c]	0 / 11 [d]	1 / 14
capitella capitata	0	0	0	1 / 8 [c]	1 / 1	1 / 8
hydra magnipapillata	0	0	0	1 / 20 [c]	0 / 170 [d]	-
nematostella vectensis (1)	0	0	0	1 / 23 [c]	0 / 4 [d]	1 / 28
trichoplax adhaerens	0	0	0	1 / 1 [c]	0 / 3 [d]	1 / 1
monosiga brevicollis	0	0	0	0 [c]	0 [d]	0 / 1

[a] `MEGABLAST`, word size 200, e-value ≤ 5.0, HSP length ≥ 280
[b] `MEGABLAST`, word size 24, e-value ≤ 5.0, HSP length ≥ 280
[c] `BLAST` word size 7, e-value ≤ 5.0, HSP length ≥ 200
[d] `Erpin` cutoff 15.17
[e] Both hits were filtered by `RepeatMasker` and recovered manually.

Table 4: **RNAse MRP:** The column "Species" yields species names along with the number of known MRP RNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2/ (#3), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of all non repeat-masked hits and #3 is the total number of hits. For the numbers #1/#2 in the "BLAST" and "Erpin" columns, #1 is the number of true positive hits or highly likely candidates and #2 is the total number of hits. All `BLAST` HSPs with length $\geq$ 180nt and e-value $\leq$ 5.0, `Erpin` hits with score $\geq$ 33.46, and `RaveNnA` hits with score $\geq$ 10.0 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RNAMotif 1**	**RNAMotif 2**	**RNAMotif 3**	**BLAST**	**Erpin**	**Ravenna**
homo sapiens (1) ‡	1 / 1 (1)	1 / 1 (2)	1 / 4 (4)	1 / 3	1 / 7	-
pan troglodytes (1)	1 / 1 (1)	1 / 2 (3)	1 / 3 (3)	1 / 2	1 / 8	-
macaca mulatta (1)	1 / 1 (1)	1 / 1 (1)	1 / 2 (2)	1 / 3	1 / 7	-
rattus norvegicus (1) ‡	1 / 1 (1)	1 / 1 (1)	1 / 2 (2)	1 / 2	1 / 1	-
mus musculus (1) ‡	1 / 1 (1)	1 / 1 (2)	1 / 2 (2)	1 / 3	1 / 3	-
bos taurus (1) ‡	1 / 1 (1)	1 / 1 (1)	1 / 1 (1)	1 / 1	1 / 4	
canis familiaris (1)	1 / 1 (1)	1 / 1 (1)	1 / 1 (1)	1 / 1	0 / 112[b]	-
monodelphis domestica (1)	0	0 / 1 (1)	0 / 1 (1)	1 / 1	0 / 109[b]	-
ornithorhynchus anatinus (1)	0	0	0	1 / 5	0 / 19[b]	-
gallus gallus (1)	0	0	0 / 1 (1)	1 / 3	0 / 21[b]	-
anolis carolinensis (1)	0	0	0 (1)	1 / 1	0 / 28[b]	-
xenopus tropicalis (1)	0	0	0 / 1 (1)	1 / 1	0 / 31[b]	-
danio rerio (1)	0	0	0	1 / 10[a]	0 / 18[b]	1 / 1
takifugu rubripes (1)	0	0	0	1 / 4[a]	0 / 9[b]	1 / 1
branchiostoma floridae (1)	0	0	0 / 2 (2)	1 / 6 [a]	0 / 8[b]	1 / 2
ciona intestinalis (1)	0	0	0	1 / 3[a]	0 / 3[b]	1 / 1
strongylocentrotus purpuratus (1)	0	0	0	1 / 4 [a]	0 / 6[b]	1 / 4
drosophila melanogaster (1)	0	0	0	1 / 9[a]	0 / 5[b]	1 / 1[c]
apis mellifera (1)	0	0	0 (1)	0 / 37[e]	0 / 2[b]	0
caenorhabditis elegans (1)	0	0	0	1 / 7[e]	0 / 2[b]	1 / 1[d]
trichinella spiralis	0	0	0	0 / 1[a]	0 / 1[b]	0 / 1
schmidtea mediterannea	0	0 / 1 (1)	0 / 3 (3)	0	0 / 13[b]	0
capitella capitata	0	0	0	0 / 2[a]	0 / 6[b]	0 / 2
hydra magnipapillata	0	0 (9)	0 / 1 (1)	0 / 17[e]	0 / 80[b]	-
nematostella vectensis	0	0 (1)	0 / 1 (1)	0 / 3[e]	0 / 2[b]	0
trichoplax adhaerens	0	0	0 / 1 (1)	0	0 / 4[b]	0
monosiga brevicollis	0	0	0	0 / 7[a]	0 / 1[b]	0

[a] hits with HSP length $\geq$ 65
[e] hits with HSP length $\geq$ 35
[b] `Erpin` cutoff 28.46
[c] `RaveNnA` cutoff 0.0
[d] `RaveNnA` cutoff -10.0

Table 5: **Y-RNA:** The column "Species" yields species names along with the number of known MRP RNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2/ (#3), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of all non repeat-masked hits and #3 is the total number of hits. For the numbers #1/#2 in the "BLAST" and "Erpin" columns, #1 is the number of true positive hits or highly likely candidates and #2 is the total number of hits. All `BLAST` HSPs with length $\geq$ 60nt and e-value $\leq$ 5.0 and `Erpin` hits with score $\geq$ 57.79 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RNAMotif 1**	**RNAMotif 2**	**RNAMotif 3**	**BLAST**	**Erpin**
homo sapiens (4) ‡	4 / 111 (111)	4 / 202 (202)	4 / 264 (265)	4 / 1971	4 / 846
pan troglodytes (4)	4 / 104 (104)	4 / 190 (190)	4 / 246 (246)	4 / 1974	4 / 820
macaca mulatta (4)	4 / 78 (78)	4 / 158 (159)	4 / 209 (211)	4 / 1772	4 / 710
rattus norvegicus (2) ‡	2 / 18 (18)	2 / 100 (100)	2 / 136 (137)	2 / 780	2 / 31
mus musculus (2) ‡	2 / 8 (8)	2 / 93 (94)	2 / 138 (141)	2 / 855	2 / 28
bos taurus (3)	3 / 13 (13)	3 / 88 (90)	3 / 136 (138)	3 / 1132	3 / 21
canis familiaris (4)	3 / 6 (6)	3 / 73 (74)	4 / 114 (117)	4 / 1949	3 / 33
monodelphis domestica (4)	2 / 5 (5)	2 / 31 (31)	3 / 45 (45)	4 / 579	3 / 249
ornithorhynchus anatinus (4)	3 / 6 (6)	3 / 61 (61)	3 / 97 (99)	4 / 677	3 / 16
gallus gallus (2)	1 / 2 (2)	1 / 97 (97)	1 / 117 (117)	2 / 867	2 / 8
anolis carolinensis (2)	2 / 5 (5)	2 / 48 (48)	2 / 68 (68)	2 / 72	2 / 6
xenopus tropicalis (4)	1 / 3 (3)	1 / 23 (23)	1 / 36 (36)	2 / 62	2 / 4
danio rerio (1)	1 / 4 (4)	1 / 178 (178)	1 / 206 (206)	1 / 891	1 / 2
takifugu rubripes (1)	1 / 3 (3)	1 / 31 (31)	1 / 51 (51)	0 / 65	1 / 1
branchiostoma floridae	0 / 2 (2)	0 / 53 (53)	0 / 81 (81)	0 / 25	0 / 3
ciona intestinalis	0	0 / 3 (3)	0 / 6 (6)	0 / 357	0 / 2
strongylocentrotus purpuratus	0	0 / 9 (9)	0 / 15 (15)	0 / 87	0 / 4
drosophila melanogaster	0	0 / 4 (4)	0 / 6 (6)	0 / 186	0 / 94 [a]
apis mellifera	0	0 / 1 (1)	0 / 1 (2)	0 / 430	0 / 6
caenorhabditis elegans	0	0	0 / 2 (2)	0 / 213	0 / 94 [a]
trichinella spiralis	0	0 / 2 (2)	0 / 2 (2)	0 / 76	0 / 1
schmidtea mediterannea	0	0 / 2 (2)	0 / 4 (4)	0 / 75	0 / 1
capitella capitata	0	0 / 3 (3)	0 / 7 (7)	0 / 48	0 / 169[a]
hydra magnipapillata	0	0 / 25 (25)	0 / 35 (35)	0 / 3553	0 / 53
nematostella vectensis	0	0 / 10 (10)	0 / 11 (11)	0 / 50	0 / 248[a]
trichoplax adhaerens	0 / 1 (1)	0 / 1 (1)	0 / 2 (2)	0 / 27	0 / 49[a]
monosiga brevicollis	0	0 / 7 (7)	0 / 7 (7)	0 / 9	0 / 21[a]

[a] `Erpin` cutoff 47.00

Table 6: **let-7 microRNA:** The column "Species" yields species names along with the number of known MRP RNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2/ (#3), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of all non repeat-masked hits and #3 is the total number of hits. For the numbers #1/#2 in the "BLAST" and "Erpin" columns, #1 is the number of true positive hits or highly likely candidates and #2 is the total number of hits. All `BLAST` HSPs with length $\geq$ 75nt and e-value $\leq$ 5.0 and `Erpin` hits with score $\geq$ 34.64 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RNAMotif 1**	**RNAMotif 2**	**RNAMotif 3**	**BLAST**	**Erpin**
homo sapiens (12) ‡	10 / 18 (22)	10 / 10 (10)	11 / 13 (15)	12 / 56	12 / 12
pan troglodytes (12)	10 / 23 (25)	10 / 10 (10)	11 / 13 (15)	12 / 43	12 / 12
macaca mulatta (12)	10 / 29 (37)	10 / 10 (10)	11 / 13 (13)	12 / 40	12 / 12
rattus norvegicus (11)	9 / 13 (25)	9 / 9 (9)	10 / 12 (13)	11 / 33	11 / 11
mus musculus (12)	10 / 16 (23)	10 / 10 (10)	11 / 12 (16)	12 / 45	12 / 12
bos taurus (12)	10 / 13 (13)	10 / 10 (10)	10 / 10 (10)	12 / 43	11 / 11
canis familiaris (7)	4 / 20 (27)	4 / 9 (9)	6 / 12 (12)	6 / 60	6 / 11
monodelphis domestica (9)	8 / 15 (18)	8 / 8 (8)	8 / 10 (10)	9 / 18	9 / 9
ornithorhynchus anatinus (7)	6 / 23 (23)	6 / 12 (12)	6/ 13 (13)	7 / 45	7 / 13
gallus gallus (11)	10 / 11 (14)	9 / 9 (9)	10 / 10 (10)	11 / 51	11 / 11
anolis carolinensis	11 / 25 (28)	11 / 13 (13)	11 / 18 (18)	12 / 14	12 / 14
xenopus tropicalis (9)	9 / 37 (37)	8 / 8 (8)	9 / 9 (9)	9 / 10	9 / 9
danio rerio (18)	12 / 15 (16)	12 / 14 (14)	13 / 150 (165)	16 / 43	16 / 19
takifugu rubripes (10)	9 / 16 (16)	9 / 14 (14)	9 / 16 (16)	10 / 22	10 / 17
branchiostoma floridae	0 / 14 (14)	0 / 4 (4)	0 / 6 (8)	0 / 9	0 / 4
ciona intestinalis (6)	0 / 1 (1)	0	0 / 1 (1)	0 / 18	3 / 15[b]
strongylocentrotus purpuratus	0 / 19 (19)	0	0 / 3 (3)	0 / 1	0 / 1
drosophila melanogaster (1)	0	0	0	1 / 12	0 / 3[b]
apis mellifera (1)	0 / 1 (2)	0	1 / 1 (1)	1 / 8	1 / 3[a]
caenorhabditis elegans (1)	1 / 2 (2)	1 / 1 (1)	1 / 1 (1)	1 / 3	1 / 1
trichinella spiralis	0 / 1 (1)	0	0	0	0[a]
schmidtea mediterannea (3)	0 / 23 (23)	0	0	0	0 / 4[a]
capitella capitata	0 / 1 (1)	0	0	1 / 2	1 / 1
hydra magnipapillata	0 / 44 (113)	0	0 / 2 (2)	0 / 45	0/423[a]
nematostella vectensis	0 / 13 (13)	0	0	0 / 1	0[a]
trichoplax adhaerens	0	0	0	0 / 1	0[a]
monosiga brevicollis	0	0	0	0 / 1	0[a]

[a] `Erpin` cutoff 25.64
[b] `Erpin` cutoff 15.64

Table 7: **Vault RNA:** The column "Species" yields species names along with the number of known MRP RNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2/ (#3), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of all non repeat-masked hits and #3 is the total number of hits. For the numbers #1/#2 in the "BLAST" and "Erpin" columns, #1 is the number of true positive hits or highly likely candidates and #2 is the total number of hits. All `BLAST` HSPs with length $\geq$ 65nt and e-value $\leq$ 5.0, `Erpin` hits with score $\geq$ 21.94, and `RaveNnA` hits with score $\geq$ 10.0 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RNAMotif 1**	**RNAMotif 2**	**RNAMotif 3**	**BLAST**	**Erpin**	**Ravenna**
homo sapiens (3) ‡	3 / 3 (3)	3 / 3 (3)	3 / 917 (942)	3 / 31	3 / 473	-
pan troglodytes (3)	2 / 3 (3)	3 / 4 (4)	3 / 813 (830)	3 / 43	3 / 337	-
macaca mulatta (3)	2 / 2 (2)	2 / 2 (2)	3 / 791 (809)	3 / 32	3 / 281	-
rattus norvegicus (1) ‡	1 / 1 (1)	1 / 1 (1)	1 /1040 (1060)	1 / 61	1 / 104	-
mus musculus (1) ‡	1 / 1 (1)	1 / 1 (1)	1 / 857 (867)	1 / 44	1 / 177	-
bos taurus (1)	1 / 1 (1)	1 / 1 (1)	1 / 856 (877)	1 / 59	1 / 294	-
canis familiaris (1)	0	0	1 / 867 (897)	1 / 43	1 / 883	-
monodelphis domestica (1)	0	0	0 / 936 (946)	1 / 18	0 / 202	-
ornithorhynchus anatinus (2)	0	0	1 / 912 (939)	2 / 159	2 / 74	-
gallus gallus (1)	0	0	0 / 370 (379)	1 / 55	0 / 335	-
anolis carolinensis	0	0	0 / 1438 (1453)	1 / 1	0 / 55	-
xenopus tropicalis (5)	0	0	5 / 423 (425)	5 / 10	1 / 17	-
danio rerio (4)	0	0	0 / 282 (292)	0 / 47	0 / 111	4 / 828
takifugu rubripes (5)	0	0	0 / 273 (277)	1 / 12[a]	0 / 8	4 / 131
branchiostoma floridae (6)	0	0	0 / 430 (443)	0 / 1	0 / 5	6 / 319
ciona intestinalis (2)	0	0	0 / 66 (66)	0 / 20	0 / 2	1 / 58
strongylocentrotus purpuratus (3)	0	0	0 / 277 (277)	0 / 3	0 / 87	1 / 318
drosophila melanogaster	0	0	0 / 110 (113)	0 / 9	0 / 4	-
apis mellifera	0	0	0 / 182 (189)	0 / 20	0 / 96	-
caenorhabditis elegans	0	0	0 / 44 (44)	0 / 7	0 / 2	-
trichinella spiralis	0	0	0 / 19 (19)	0 / 2	0 / 15	-
schmidtea mediterannea	0	0	0 / 153 (153)	0 / 3	0 / 187	-
capitella capitata	0	0	0 / 201 (203)	0 / 6	0 / 1	-
hydra magnipapillata	0	0	0 / 674 (1090)	0 / 108	0 / 15558	-
nematostella vectensis	0	0	0 / 276 (278)	0 / 2	0 / 5	-
trichoplax adhaerens	0	0	0 / 15 (15)	0	0	-
monosiga brevicollis	0	0	0 / 96 (96)	0 / 5	0	-

[a] hits with HSP length $\geq$ 40

Table 8: **U5 snRNA:** The column "Species" yields species names along with the number of known MRP RNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2/ (#3), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of all non repeat-masked hits and #3 is the total number of hits. For the numbers #1/#2 in the "BLAST" and "Erpin" columns, #1 is the number of true positive hits or highly likely candidates and #2 is the total number of hits. All `BLAST` HSPs with length $\geq$ 100nt and e-value $\leq$ 5.0 and `Erpin` hits with score $\geq$ 18.05 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RNAMotif 1**	**RNAMotif 2**	**RNAMotif 3**	**BLAST**	**Erpin**
homo sapiens (5) ‡	5 / 5 (5)	5 / 64 (99)	5 / 6237 (6491)	5 / 115	5 / 66
pan troglodytes (7)	4 / 5 (5)	6 / 68 (114)	6 / 6060 (6330)	7 / 109	7 /70
macaca mulatta (15)*	1 / 1 (1)	5 / 53 (76)	6 / 5565 (5840)	7 / 99	10 / 66
rattus norvegicus (4) ‡	4 / 8 (8)	4 / 57 (63)	4 / 3270 (3512)	4 / 63	4 / 33
mus musculus (6) ‡	6 / 7 (7)	6 / 54 (76)	6 / 3548 (3811)	6 / 77	6 / 33
bos taurus (5)	4 / 7 (7)	5 / 66 (84)	5 / 3607 (3823)	5 / 95	5 / 36
canis familiaris (4)	3 / 5 (5)	3 / 72 (99)	3 / 4009 (5170)	4 / 136	4 / 50
monodelphis domestica (5)	5 / 7 (7)	5 / 99 (118)	5 / 5930 (6328)	5 / 48	5 / 39
ornithorhynchus anatinus (4)	3 / 7 (7)	4 / 34 (42)	4 / 2476 (2586)	4 / 34	4 / 40
gallus gallus (1)	1 / 6 (6)	1 / 26 (31)	1 / 1782 (1852)	1 / 55	1 / 18
anolis carolinensis (3)	3 / 3 (3)	3 / 23 (32)	3 / 5097 (5252)	3 / 31	3 / 60
xenopus tropicalis (1)	1 / 47 (47)	1 / 120 (124)	1 / 1572 (1649)	1 / 75	1 / 73
danio rerio (7)	0	7 / 180 (211)	7 / 1957 (2198)	7 / 169	6 / 11
takifugu rubripes (6)	0	6 / 18 (19)	6 / 390 (408)	6 / 22	4 / 23
branchiostoma floridae (9)	0	2 / 12 (13)	2 / 909 (926)	9 / 26	2 / 194[a]
ciona intestinalis (5)	0	0 / 3 (6)	1 / 148 (158)	5 / 93	5 / 69
strongylocentrotus purpuratus (8)	0	0 / 19 (19)	0 / 1250 (1250)	8 / 45	7 / 16
drosophila melanogaster (7)	0	0 / 1 (3)	0 / 111 (119)	6 / 18	0 / 16[a]
apis mellifera (3)	0	0 / 18 (58)	3 / 637 (831)	3 / 44	0 / 236[a]
caenorhabditis elegans (7)	0	0 / 17 (27)	6 / 226 (250)	7 / 28	0 / 51[a]
trichinella spiralis (2)	0	2 / 6 (6)	2 / 118 (118)	2 / 5	1 / 10[a]
schmidtea mediterannea (2)	0	0 / 140 (140)	0 / 1408 (1408)	2 / 45	0 / 62[a]
capitella capitata (3)	0	0 / 1 (1)	0 / 324 (339)	3 / 22	0 / 96[a]
hydra magnipapillata (7)	0	0 / 869 (2525)	0 / 13540 (16659)	7 / 1278	0 / 172
nematostella vectensis (6)	0	0 / 11 (12)	4 / 516 (547)	5 / 228	4 / 248[a]
trichoplax adhaerens (1)	0	1 / 4 (4)	1 / 104 (111)	1 / 7	1 / 3
monosiga brevicollis (0)	0	0	0 / 11 (11)	0	0 / 3[a]

[a] `Erpin` score $\geq$ 13.0

Table 9: **U3 snoRNA:** The column "Species" yields species names along with the number of known SRP RNAs in the organism. Species marked with ‡ are contained in the seed set. The columns "RNAMotif1-3" provides numbers #1/#2 (#3), where #1 is the number of true positive hits or highly likely candidates, #2 is the number of all non repeat-masked hits and #3 is the total number of hits. For the numbers #1/#2 in the "BLAST", "Erpin" and "Ravenna" columns, #1 is the number of true positive hits or highly likely candidates and #2 is the total number of hits. All `BLAST` HSPs with length $\geq$ 200nt and e-value $\leq$ 5.0, `Erpin` hits with score $\geq$ 20.0, and `RaveNnA` hits with score $\geq$ 10.0 are counted. Alternative cutoffs are indicated by a footnote.

Species	**RnaMotif 1**	**RnaMotif 2**	**RnaMotif 3**	**Blast**	**Erpin**	**Ravenna**
homo sapiens (1) ‡	1 / 2 (5)	1 / 2 (3)	1 / 47 (63)	1 / 58	1 / 30	-
pan troglodytes (1)	1 / 1 (3)	1 / 1 (1)	1 / 36 (52)	1 / 50	1 / 27	-
macaca mulatta (1)	0 (1)	1 / 2 (2)	1 / 27 (39)	1 / 47	1 / 17	-
rattus norvegicus (1) ‡	0 (20)	1 / 2 (5)	1 / 37 (75)	1 / 15	1 / 24	-
mus musculus (1) ‡	1 / 5 (22)	1 / 5 (5)	1 / 33 (68)	1 / 17	1 / 20	-
bos taurus (1) ‡	1 / 4 (4)	1 / 5 (5)	1 / 39 (44)	1 / 22	1 / 21	-
canis familiaris (1)	1 / 2 (2)	1 / 2 (2)	1 / 26 (34)	1 / 11	1 / 18	-
monodelphis domestica (1)	1 / 8 (15)	1 / 8 (8)	1 / 41 (52)	1 / 10	1 / 17	-
ornithorhynchus anatinus	0	0 (1)	1 / 25 (31)	1 / 7	1 / 6	-
gallus gallus (1)	0	0	1 / 11 (14)	1 / 1	1 / 2	-
anolis carolinensis	1 / 1 (3)	1 / 1 (1)	1 / 29 (34)	1 / 1	1 / 3	-
xenopus tropicalis (1)	0	0 / 1 (1)	0 / 13 (15)	1 / 30	1 / 94[a]	-
danio rerio (1)	0 (22)	0 (2)	1 / 21 (43)	1 / 29	0 / 110[a]	1 / 2847
takifugu rubripes (1)	0 (3)	0	0 / 11 (15)	1 / 4	0 / 50[a]	1 / 24
branchiostoma floridae	0 (2)	0	4 / 10 (14)	7 / 7	2 / 2	7 / 49
ciona intestinalis (1)	0	0 / 1 (1)	0 / 3 (3)	1 / 24	0 / 4[a]	1 / 64
strongylocentrotus purpuratus	0 / 1 (1)	0 / 1 (1)	6 / 12 (12)	9 / 9	1 / 3	9 / 171
drosophila melanogaster (2)	0	0	0 / 1 (1)	2 / 3	0 / 9[a]	2 / 17
apis mellifera	0	0	0 (4)	3 / 3	0 / 8	3 / 2291
caenorhabditis elegans (2)	0	0	0	0	0 / 5[a]	2 / 75
trichinella spiralis	0	0	0	1 / 1	1 / 3[a]	1 / 34
schmidtea mediterannea	0	0	0 / 7 (7)	0	0 / 50[a]	0 / 3858
capitella capitata	0	0	0 / 4 (4)	12 / 12	0 / 24[a]	1 / 96
hydra magnipapillata	0 (697)	0 / 21 (32)	1 / 31 (74)	18 / 78	0 / 27	-
nematostella vectensis	0	0	0	14 / 17	0 / 12[a]	1 / 164
trichoplax adhaerens (1)	0	0	0	1 / 1	1 / 1	1 / 14
monosiga brevicollis	0	0	0	1 / 1[b]	0 / 7[a]	1 / 1

[a] `Erpin` score $\geq$ 15.0
[b] HSP length $\geq$ 170nt

References

[1] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucl. Acids Res.*, 33(1):D121–124, 2005.

[2] J. Hertel, I. L. Hofacker, and P. F. Stadler. SnoReport: Computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, 2007.

[3] T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–7, Jan. 2007.

[4] J. O. Kriegs, G. Churakov, J. Jurka, J. Brosius, and J. Schmitz. Evolutionary history of 7sl rna-derived sines in supraprimates. *Trends in Genetics*, 23(4):158–161, Apr. 2007.

[5] L. Lestrade and M. J. Weber. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, 34:D158–162, 2006.

[6] A. Mosig, J. L. Chen, and P. F. Stadler. Homology search with fragmented nucleic acid sequence patterns. In *WABI 2007 (R. Giancarlo & S. Hannehalli, eds.)*, pages 335–345, 2007.

[7] A. Mosig, M. Guofeng, B. M. R. Stadler, and P. F. Stadler. Evolution of the Vertebrate Y RNA Cluster. *Th Biosci.*, 126:9–14, 2007.

[8] M. A. Rosenblad, J. Gorodkin, B. Knudsen, C. Zwieb, and T. Samuelsson. SRPDB (Signal Recognition Particle Database). *Nucleic Acids Res. (database issue)*, 31:363–364, 2003.

[9] A. F. A. Smit and R. Hubley. RepeatMasker Open-3.0. 1996-2008. http://www.repeatmasker.org.

[10] P. F. Stadler, J. J.-L. Chen, J. Hackermüller, S. Hoffmann, F. Horn, P. Khaitovich, A. K. Kretzschmar, A. Mosig, S. J. Prohaska, X. Qi, K. Schutt, and K. Ullmann. Evolution of vault rnas. *J. Mol. Biol.*, page 2008. submitted.

[11] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22(22):4673–4680, 1994.

[12] Z. Weinberg and W. L. Ruzzo. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, page bti743, 2005.

[13] L. W. Zheng Zhang, Scott Schwartz and W. Miller. A greedy algorithm for aligning dna sequences. *J Comput Biol*, 7:203–14, 2000.