Genome-wide mapping of conserved RNA Secondary Structures Reveals Evidence for Thousands of functional Non-Coding RNAs in Human

Stefan Washietl,[1] Ivo L. Hofacker, [1] Peter F. Stadler[2,3]

[1] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[2] *Department of Computer Science and Interdisciplinary Center of Bioinformatics, University of Leipzig Härtelstraße 16-18, D-04107, Leipzig, Germany*
[3] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM87501*

In contrast to the fairly reliable and complete annotation of the protein coding genes in the human genome, comparable information is lacking for non-coding RNAs. We present a comparative screen of vertebrate genomes for structural non-coding RNAs, which evaluates sequence conservation, secondary structure conservation, and thermodynamic stability of putative RNA structures. We predict more than 30 000 structured RNA elements in the human genome, almost 1000 of which are conserved across all vertebrates. Roughly a third is found in introns of known genes, a sixth are potential regulatory elements in untranslated regions, about half are located far away of any known gene. Only a small fraction of these sequences has been described previously. EST data demonstrate, however, that the majority of them is at least transcribed. The widespread conservation of secondary structure points to a large number of functional ncRNAs in the human genome, which we estimate to be comparable to the number of protein-coding genes.

The recent finishing of the human genome sequence emphasizes the "need for reliable experimental and computational methods for comprehensive identification of non-coding RNAs"[1]. A variety of experimental techniques have been used to uncover the human and mouse transcriptomes, in particular tiling arrays[2–4], cDNA sequencing[5,6], and unbiased mapping of transcription factor binding sites[7]. All these studies agree that a substantial fraction of the genome is transcribed and that a large fraction of the transcriptome consists of non-coding RNAs. It is unclear, however, which fraction are functional non-coding RNAs (ncRNAs), and which constitutes "transcriptional noise"[8].

Genome-wide computational surveys of ncRNAs, on the other hand, have been impossible until recently, because ncRNAs do not share common signals that could be detected at the sequence level. A large class of ncRNAs, however, has characteristic structures that are functional and hence are well conserved over evolutionary timescales: most of the "classical" ncRNAs, including rRNAs, tRNAs, snRNAs, snoRNAs, as well as the RNA components of RNAse P and the signal recognition particle, are of this type. The stabilizing selection acting on the secondary structure causes characteristic substitution patterns in the underlying sequences: Consistent and compensatory mutations replace one type of base-pair by another one in the paired regions (helices) of the molecule. In addition, loop regions are more variable than helices. These patterns can be exploited in comparative computational approaches[9–12] to discriminate functional RNAs from other types of conserved sequence. Recently, high levels of sequence conservation of non-coding DNA regions have been reported[13,14]. Here we screen the complete collection of conserved non-coding DNA sequences from mammalian genomes and provide a first annotation of the complement of structurally conserved RNAs in the human genome.

Results

Selection of conserved sequences and screening for structural RNAs We start from the genome-wide alignments of vertebrate genomes provided through the UCSC Genome Browser[15]. We limit our comparative screen to the *most conserved regions* as annotated by the PhastCons program, which constitute 4.81% of the 3,095 MB of the human genome. It has been estimated that about 5% of the human genome is under selective pressure[16,17]. Since we are interested in non-coding RNAs, we removed all annotated coding exons from the test set and retain only the 438,788 alignments of non-coding regions that are conserved at least in the four eutherian mammals (human, mouse, rat, dog). This amounts to 82.64 MB or 2.88% of the human genome (Tab. 1).

This dataset was screened for structural RNAs using RNAz [11], a program that combines a comparative approach (scoring conservation of secondary structure) with the observation[18,19] that ncRNAs are thermodynamically more stable than expected by chance. A structure conservation index (SCI) is computed by comparing the predicted minimum free energies of the sequences in an alignment with a consensus energy, which is computed by incorporating covariation terms into a free energy minimization computation[20]. Thermodynamic stability is quantified by means of a z-score that measures the folding energy relative to shuffled sequences (a regression approach replaces time-consuming shuffling methods). A support vector machine then classifies an alignment as “structured RNA” or “other” based on z-score and SCI. The significance of the classification is quantified as “RNA-class probability” p.

Fig.1 illustrates the strategy of our screen and shows an annotation for a 9 megabase region on chromosome 13. For details on the scanning procedure see the Methods section. In the complete genome, we detected 91,676 (15.1% of the conserved sequence) independent RNA structures on the $p = 0.5$ level and 35,985 (6.6%) structures on the $p = 0.9$ level (Tab. 1, Fig. 2a).

Estimating specificity The specificity of RNAz is generally high, $\approx 99\%$ for the $p = 0.9$ cutoff[11]. Due to the large number of input alignments, however, we have to expect a non-negligible number of false positives. We therefore repeated the complete screen with alignments randomized by shuffling[19]. We obtain a false positive rate of 28.9% ($p = 0.5$) and 19.2% ($p = 0.9$), respectively. As expected, the hits in the randomized dataset are on average smaller than the native ones, reducing the false positive rates to 25.7% ($p = 0.9$) and 16.3% ($p = 0.9$) in terms of sequence length. The estimate for the false positive rate implies lower bounds of 65000 ($p = 0.5$) and 29000 ($p = 0.9$) for the number of structural RNA elements in the human genome. On average, we predict 21 ($p = 0.5$) and 10 ($p = 0.9$) structural elements per megabase.

Furthermore, we observed that many of the hits in randomized alignments overlap with native predictions (Supplementary Table 1). This indicates that our shuffling process[19] does not effectively remove the signal in all cases. It follows that the above estimates are conservative; for details see Methods. We observed that the random hits are clearly enriched in highly conserved alignments. The false positive rate of RNAz is higher in this case, because these alignments contain little covariance information so that the classification is dominated by the thermodynamic stability alone. Since many known ncRNAs are contained in this set we decided against removing highly conserved alignments from our survey despite the increased false positive rate.

Performance on known ncRNAs A comprehensive annotation of ncRNAs in the human genome is not available, thus it is impossible to determine the overall sensitivity of our screen. For microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs), however, a comprehensive annotation is provided in the UCSC browser.

There are 207 annotated miRNA loci (see also Ref. 21) of which 45 loci are not in our set of input alignments for various reasons (see Supplementary Table 3). We detect 157 (96.9%) of the remaining 162 miRNAs. The effective sensitivity is 75.8% for miRNA precursors, which are among the most easy-to-find ncRNAs (Fig. 2c).

22 of the 86 annotated H/ACA-box snoRNAs are not included in the input set mostly because they are not detected by PhastCons , see Supplementary Table 4. We recover 55 of the remaining 64 sequences (85.9%). We can thus relatively accurately detect this class of ncRNAs which have resisted computational prediction so far. (Effective sensitivity: 64.0%)

Our screen performs poorly on C/D-Box snoRNAs, however. Out of the 256 known C/D snoRNAs about a half (129) are missing in the input alignments. Even though we detect 39.4% of C/D snoRNAs in our set, the effective sensitivity is only 19.5%. C/D-Box snoRNAs are hard to detect computationally even with specialized approaches[22].

From these examples we estimate that the overall sensitivity of the combination of the Multiz / PhastCons alignments and RNAz is on the order of 30%.

We then compared all hits with available databases of known ncRNAs (Tab. 2). Most of the "classical" structured ncRNAs are not contained in the input alignments because they are marked as repetitive DNA by RepeatMasker and were therefore excluded from the Multiz alignments. The RNA component of the signal recognition particle, the small nuclear RNAs (snRNAs) U1, U2, U4, U5, U6, and U7, as well as the four Y-RNAs belong to this class. On the other hand, we detect all snRNAs of the minor spliceosome (U4atac, U6atac, U11, and U12). We also detect very well conserved (although not very stable) structures within the RNAse P but miss RNAse MRP and telomerase RNA. Here the pseudoknotted structures[23, 24], which are not taken into account by RNAz , appear to be the problem.

We find local secondary structure motifs in various other documented ncRNAs which do not appear to have conserved global structures (Supplementary Table 2). The *Xist* gene, a 17kb ncRNA which plays a key role in dosage compensation and X chromosome inactivation[25] contains three independent conserved RNA secondary structures. Intriguingly, we find 8 RNAz hits in the human genome with significant sequence similarity to the *Air* RNA. This antisense transcript regulates imprinted gene expression in mouse[26] but is not conserved over its full length ($\approx$1000kb) in human. 7 of the 8 hits correspond to the same local secondary structure motif in *Air*. One of them can be found in an intron of *HERC2*, a locus located near the Prader-Willi imprinting center which in turn is regulated by antisense transcripts.

The RNAdb [27] compiles collections of expressed sequences without protein coding capacity. A comparison of our RNAz hits with the RNAdb identifies conserved structured elements in many of these transcripts, thereby supporting that they function as ncRNAs (Tab 2).

Candidates for novel snoRNAs and miRNAs A number of signals are novel ncRNAs that can be associated with known ncRNAs or ncRNA families through sequence similarity. Some of these are additional paralogs or orthologs of known RNA genes. For example, we found more than 100 hits with sequence similarity to snoRNAs. Some of these are most likely functional snoRNAs since they are human homologs of mouse snoRNAs described in reference 28.

Another class of signals are novel members of one of the large, well-described classes of ncRNAs.

A simple subscreen was performed to identify putative H/ACA box snoRNAs (Fig. 3b). We selected all RNAz hits with two stems at least 15 pairs in length and separated by an unpaired hinge, which in addition have the motif ACA in the consensus sequence in the last 20nt. We found 137 structures, of which 28 were known snoRNAs. Visual inspection shows that 30–40 additional clusters have typical H/ACA snoRNA-like secondary structure of which 15 also have the canonical H-box sequence ANANNA. In many known snoRNAs, only short parts of the stem are conserved in the predicted consensus sequence and/or only parts of the complete structure are detected as conserved structural element. As a consequence, this subscreen is not exhaustive and a more detailed analysis can be expected to bring up even more candidates.

Berezikov and co-workers[29] identified 975 miRNA candidates in mouse/human and mouse/rat comparisons by means of a combination of phylogenetic shadowing and selection of stable stem-loop structures. Our set of input alignments contains 642 of these candidates, 472 overlap with our predictions ($p > 0.9$). Not all these stem-loops, which are stable as single sequences, are structurally conserved in all four mammals: some of them lack a stable consensus structure. A simple filter requiring a stem with at least 20 base pairs in the consensus structure, a mean z-score < -3.5 and a 22nt window with more than 0.95% pairwise sequence identity (the prospective mature miRNA sequence) retains 312 candidates, among them 109 known human miRNAs. Some of the unknown candidates show the typical mutation pattern of miRNA, see Figure 3a. Others exhibit clear structural conservation but show a very different mutation pattern. We speculate that these sequences are not miRNAs but belong to different, so far undescribed, classes of ncRNAs.

Structures conserved across all vertebrates The most highly conserved ncRNA candidates are of particular interest. We find 996 RNAz signals that are conserved in all 4 mammals, chicken and at least one of the two fish genomes (fugu, *Takifugu rubripes*, and zebrafish, *Danio rerio*). Of these, 152 can be at least partially annotated: 52 are miRNAs, 16 are snoRNAs, 28 are UTR elements, and 56 are similar to other described RNAs. 42 detected regions are contained within one of the different cDNA collections. 38 overlap with one of the 481 "ultraconserved elements" (segments longer than 200 base pairs that are identical between human, mouse and rat genomes) reported by Bejerano *et al.*[30]. A few of these can be unambiguously identified as structural RNAs because of the substitution pattern in the fish and chicken sequences. For most of them, however, we cannot give a definitive classification because there is too little sequence variation in this special set of extremely conserved sequences.

Genomic location of detected structures The majority of the 35989 structured RNA features detected with $p > 0.9$ is of completely unknown function (see Fig. 4 for a few selected examples). Approximately 70% of them are located in regions that are covered by ESTs. More than a third of the signals are found in introns and strongly support the notion that a plethora of functional ncRNAs are expressed from intronic DNA[31]. On the other hand, almost half of these RNAz signals are located at least 10kb away from, and hence are probably unrelated to, any known protein coding gene. Nevertheless, 50.3% of them are covered by ESTs. Roughly one sixth of the signals is located in UTRs. These are potential regulatory elements of the mRNA.

Clustering of RNAz hits with blastclust yields only small groups of RNAs or isolated sequences. This rules out the possibility that a substantial fraction of the RNAz hits are derived from pseudogenes or belong to repeat families that so far have not been annotated.

Discussion

The systematic use of comparative structural information has enabled us here to discover what we believe is a significant fraction of the Human "RNome". This computational screen crucially depends on both sequence conservation and evidence for the conservation of secondary structure. Both are strong indicators for stabilizing selection and indicate that the features detected by RNAz are indeed functional. Selection on secondary structure furthermore implies that these sequences function as RNAs rather than as (yet unknown) proteins or regulatory sequences at the DNA level.

Our data provides a strong basis for further theoretical and experimental studies. A systematic analysis and classification of all detected RNA structures, anticipated and dubbed "structural RNomics" a few years ago[32], together with rationally designed expression studies are promising strategies towards a better understanding of ncRNA function on a genome-wide scale.

Our study aimed at providing at least a preliminary *de novo* annotation of structural ncRNAs in the human genome. To our knowledge, this is the first attempt of this kind and the study would not be complete without addressing the heavily discussed question on the number of functional RNAs in the human genome[8]. The history of protein gene finding, however, demonstrates the problems of estimating gene numbers[33]. In the case of ncRNAs, the situation is even more difficult. It must be pointed out that not each of the detected structural regions necessarily corresponds to a single RNA gene in the sense of a genetic unit. Long ncRNAs (e.g. *Xist*) 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 structural cluster by our procedure (see Methods). 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 human genome. Assuming a sensitivity of 30%, we estimate about 100,000 structural features, of which about 30,000 are intronic and 50,000 are unrelated to protein-coding genes. Assuming that these "isolated" RNA genes have the same intron-exon structures as protein coding genes, which on average have 5 exons/gene, we estimate on the order 10,000 structured RNA genes. Considering the additional structures encoded in the introns, we estimate that the number of functional non-coding RNAs is comparable to the number of protein coding genes, which is consistent with previous estimates[7,31].

Methods

Alignments Genome-wide alignments of vertebrates ("multiz8way") were downloaded from the UCSC genome browser[15]. 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)[11]. 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.

Estimating specificity The specificity of RNAz was found to be $\approx 99\%$ and $\approx 96\%$, for $p = 0.9$ and $p = 0.5$, respectively[11]. 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 19. 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 (Supplementary Table 1).

In a screen of the urochordate *Ciona intestinalis* based on pairwise alignments[34], 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[35] 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 subscreens for miRNAs and H/ACA snoRNAs, we investigated in detail why miRNAs and H/ACA snoRNAs were missed in our selection of input alignments (Supplementary 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: Rfam (release 6.1, August 2004)[21], RNAdb (August 2004)[27], NONCODE (release 1.0, March 2004)[36], microRNA registry (release 5.0, September 2004)[35], UTRdb (April 2004)[37].

We generated BLAST libraries for each of the databases and matched the human sequence of all the detected RNAz clusters against them. 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. 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.

The complete dataset of predicted structures can be downloaded and viewed under `www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/ncRNA.`

1. The Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

2. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).

3. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).

4. Johnson, J. M., Edwards, S., Shoemaker, D. & Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21**, 93–102 (2005).

5. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).

6. Imanishi, T. & *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biology* **2**, 0856–0875 (2004).

7. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).

8. Hüttenhofer, A., Schattner, P. & Polacek, N. Non-coding RNAs: hope or hype? *Trends Genet.* (2005). In press.

9. Hofacker, I. L. *et al.* Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.* **26**, 3825–3836 (1998).

10. Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11**, 1369–1373 (2001).

11. Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding rnas. *Proc. Natl. Acad. Sci. USA* **102**, 2454–2459. (2005).

12. Moulton, V. Tracking down noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102**, 2269–2270 (2005).

13. Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).

14. Dermitzakis, E. T. *et al.* Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033–1035 (2003).

15. Kent, W. J. *et al.* The human genome browser at ucsc. *Genome Res* **12**, 996–1006 (2002).

16. International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

17. Cooper, G. M. *et al.* Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res* **14**, 539–48 (2004).

18. Le, S. V., Chen, J. H., Currey, K. M. & Maizel Jr., J. V. A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.* **4**, 153–159 (1988).

19. Washietl, S. & Hofacker, I. L. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* **342**, 19–30 (2004).

20. Hofacker, I. L., Fekete, M. & Stadler, P. F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066 (2002).

21. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–D124 (2005).

22. Accardo, M. C. *et al.* A computational search for box C/D snoRNA genes in the *D. melanogaster genome*. *Bioinformatics* **20**, 3293–3301 (2004).

23. Childs, J. L., Poole, A. W. & Turner, D. H. Inhibition of *Escherichia coli* RNase P by oligonucleotide directed misfolding of RNA. *RNA* **9**, 1437–1445 (2003).

24. Lin, J. *et al.* A universal telomerase RNA core structure includes structured motifs required for binding the telomerase reverse transcriptase protein. *Proc. Natl. Acad. Sci. USA* **101**, 14713–14718 (2004).

25. Avner, P. & Heard, E. X-chromosome inactivation: counting, choice, and initiation. *Nat. Rev. Genet.* **2**, 59–67 (2001).

26. Rougeulle, C. & Heard, E. Antisense RNA in imprinting: spreading silence through Air. *Trends Genet* **18**, 434–7 (2002).

27. Pang, K. C. *et al.* RNAdb — comprehensive mammalian noncoding RNA database. *Nucl. Acids Res.* **33**, D125–D130 (2005). Database issue.

28. Hüttenhofer, A. *et al.* RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.* **20**, 2943–2953 (2001).

29. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–24 (2005).

30. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).

31. Mattick, J. S. RNA regulation: a new genetics? *Nature Rev. Genetics* **5**, 316–323 (2004).

32. Doudna, J. A. Structural genomics of RNA. *Nature Struct. Biol.* **7**, 954–956 (2000).

33. Pennisi, E. Bioinformatics. gene counters struggle to get the right answer. *Science* **301**, 1040–1 (2003).

34. Missal, K., Rose, D. & Stadler, P. F. Non-coding RNAs in the urochordate *Ciona intestinalis*. In *ECCB 2005* (2005). Under review.

35. Griffiths-Jones, S. The microRNA Registry. *Nucl. Acids Res.* **32**, D109–D111 (2004).

36. Liu, C. *et al.* NONCODE: an integrated knowledge database of non-coding RNAs. *Nucl. Acids Res.* **33**, D112–D115 (2005). Database issue.

37. Pesole, G. *et al.* UTRdb and UTRSite: specialized databases of sequences and functional elements of 5’ and 3’ untranslated regions of eukaryotic mRNAs. update 2002. *Nucl. Acids Res.* **30**, 335–340 (2002).

38. Scherer, S. W. & *et al.* Human chromosome 7: DNA sequence and biology. *Science* **300**, 767–772 (2003).

Acknowledgements This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893, by the German *DFG* Bioinformatics Initiative BIZ-6/1-2, and by the Austrian *Gen-AU bioinformatics integration network* sponsored by BM-BWK and BM-WA.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence should be addressed to P.F.S.
(email: studla@bioinf.uni-leipzig.de).

Table 1. Genomic coverage of filtering steps and phylogenetic conservation of ncRNA candidates.

	Genome Coverage		Alignments	RNAz hits $p > 0.9$		
	Size	Fraction	Number	Size	Fraction of	Number
	(MB)	(%)		(MB)	input (%)	
Human genome	3,095.02	100.00	–			
PhastCons most conserved	137.85	4.81	1,601,903			
without coding regions	110.04	3.84	1,291,385			
without alignments $< 50nt$	103.83	3.33	564,455			
Set 1: 4 Mammals	82.64	2.88	438,788	5.46	6.62	35,985
Set 2: + Chicken	24.00	0.85	104,266	1.34	5.50	8,802
Set 3: + Fugu or zebrafish	6.86	0.24	30,896	0.14	2.03	996

Table 2. Comparison of RNAz predictions with ncRNAs from the literature. Apart from curated databases we compare our data with 4 collections of cDNAs. Fantom contains more than 15000 unique, putative ncRNAs from mouse that do not contain a significant coding sequence. H-invitational (hinv) contains more than 2000 transcripts with ORFs < 80aa that passed several additional filters designed to exclude likely protein-coding genes. The human chromosome 7 annotation project (chr7) has described over 350 putative ncRNAs derived from computer-based annotation in conjunction with extensive laboratory experimentation. Note that many ncRNAs may function e.g. as anti-sense regulators without exhibiting a conserved, functionally important, secondary structure. Thus one cannot use these databases to estimate the sensitivity of the RNAz screen. Nevertheless, a significant fraction of these transcribed putative ncRNAs exhibits evolutionarily conserved structures.

Database	Ref.	$p > 0.5$	$p > 0.9$
Rfam	21	267	189
NONCODE	36	273	177
RNAdb	27	446	327
miRNA Registry	21	176	168
UTRdb	37	388	159
Curated		984	563
hinv	6	478	205
Fantom	5	1908	781
chr7	38	180	90
antisense pipeline	27	149	59
cDNA collections		2539	1056
Total		3441	1585

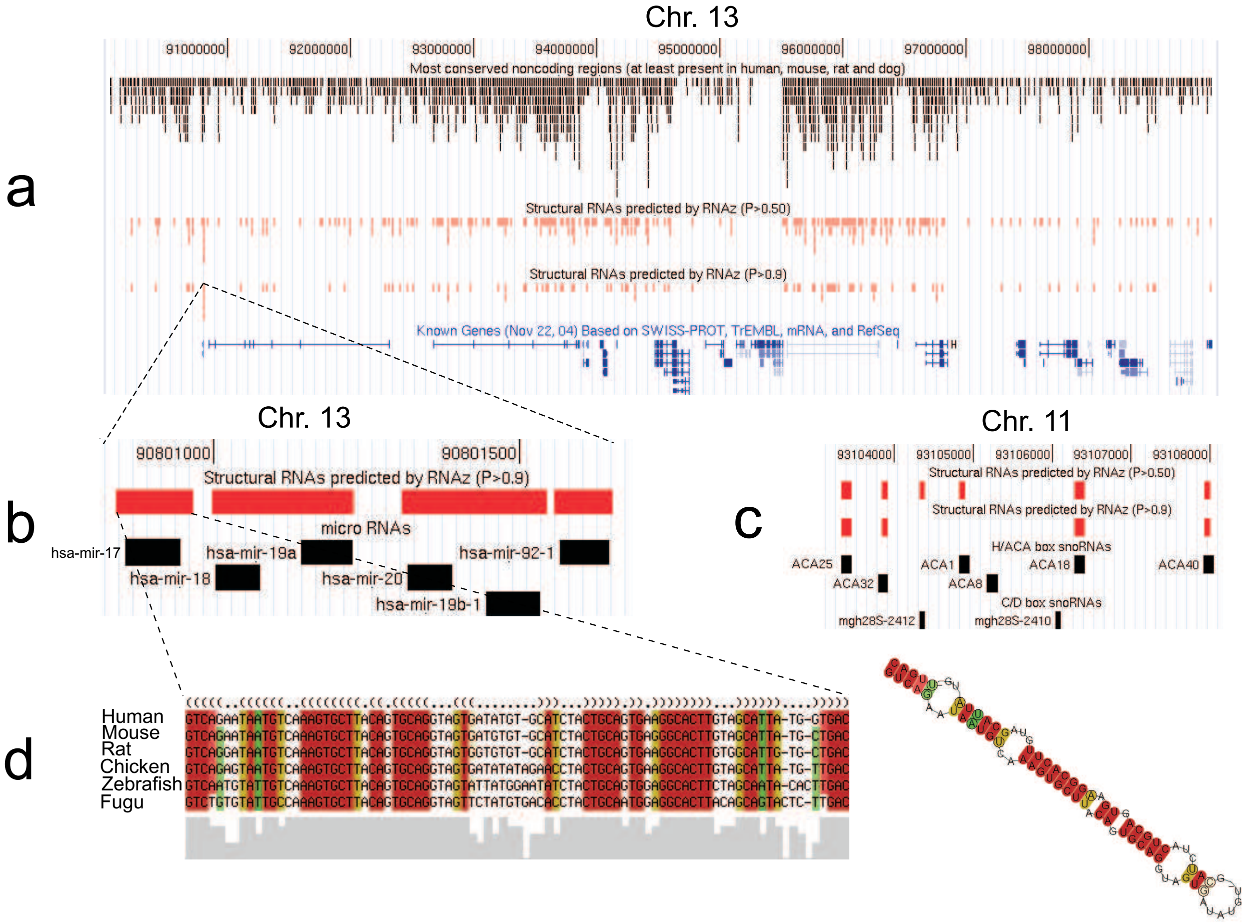
Chr. 13
a
91000000
92000000
93000000
94000000
95000000
96000000
97000000
98000000
Most conserved noncoding regions (at least present in human, mouse, rat and dog)
Structural RNAs predicted by RNAz (P>0.50)
Structural RNAs predicted by RNAz (P>0.9)
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Chr. 13
b
90801000
90801500
Structural RNAs predicted by RNAz (P>0.9)
micro RNAs
hsa-mir-17
hsa-mir-18
hsa-mir-19a
hsa-mir-20
hsa-mir-19b-1
hsa-mir-92-1
Chr. 11
c
93104000
93105000
93106000
93107000
93108000
Structural RNAs predicted by RNAz (P>0.50)
Structural RNAs predicted by RNAz (P>0.9)
H/ACA box snoRNAs
ACA25
ACA32
ACA1
ACA8
ACA18
ACA40
C/D box snoRNAs
mgh28S-2412
mgh28S-2410
d
(((((..((((((..((((((((.((.(((((...(((........)))...))))).)).))))))))...))))))....)))))
Human GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATGT-GCATCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-GTGAC
Mouse GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATGTGT-GCATCTACTGCAGTGAGGGCACTTGTAGCATTA-TG-CTGAC
Rat GTCAGGATAATGTCAAAGTGCTTACAGTGCAGGTAGTGGTGTGT-GCATCTACTGCAGTGAAGGCACTTGTGGCATTG-TG-CTGAC
Chicken GTCAGAGTAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATATAGAACCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-TTGAC
Zebrafish GTCAATGTATTGTCAAAGTGCTTACAGTGCAGGTAGTATTATGGAATATCTACTGCAGTGGAGGCACTTCTAGCAATA-CACTTGAC
Fugu GTCTGTGTATTGCCAAAGTGCTTACAGTGCAGGTAGTTCTATGTGACACCTACTGCAATGGAGGCACTTACAGCAGTACTC-TTGAC

Fig. 1. (a) Starting from the 5% best conserved non-coding DNA as defined by Multiz alignments and PhastCons in the USCS Genome Browser, RNAz employs a stringent filter for putative structured RNAs. These sequences are thermodynamically more stable than average and can fold into a common secondary structure. Two levels of confidence ($p > 0.5$) and ($p > 0.9$) are used. The RNAz hits, which, depending on the confidence level, cover 6–15% of the input alignments are highly enriched in known ncRNAs, such as the *mir17*-cluster of miRNAs (b) or a cluster of H/ACA and C/D box snoRNAs on chromosome 11 (c). The method not only detects signals for ncRNAs but in the process of classification constructs an explicit secondary structure model from the aligned sequences (d), see also Fig. 3 and Fig. 4.

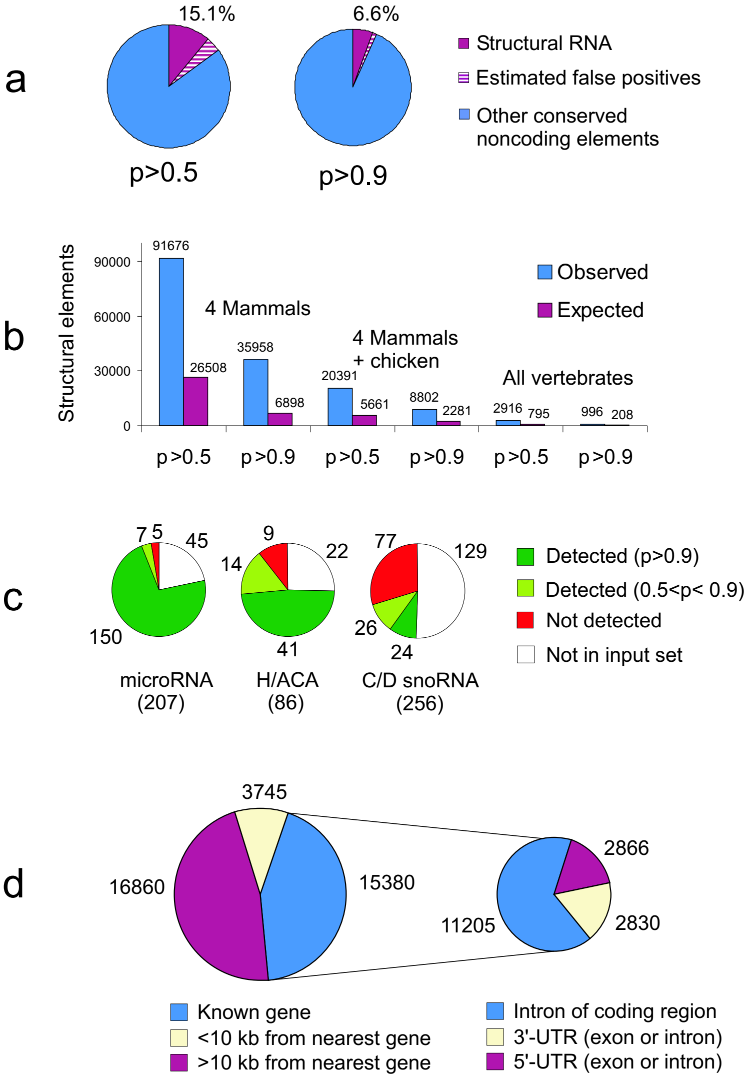

Fig. 2. Statistical analysis of Human ncRNA candidates. (a) The RNAz method classifies, depending on the user-defined confidence cut-off, a small fraction of the conserved non-coding regions as ncRNA. (b) A second screen on shuffled data demonstrates that the effective false positive rate of entire screen is well below 25% and decreases with increasing confidence level and phylogenetic conservation. (c) The sensitivity is estimated for known miRNAs and snoRNAs. Between a quarter and two thirds of the known ncRNAs are not contained in the input alignments due to insufficient accuracy in the alignments, incomplete sequences, and removal of repeated DNA. This is the most severe limitation at present. (d) About half of the ncRNA candidates (shown here for the $p > 0.9$ level) are located far away from any known protein coding gene, the other half is associated with known genes. Two-thirds of the latter are located in introns.

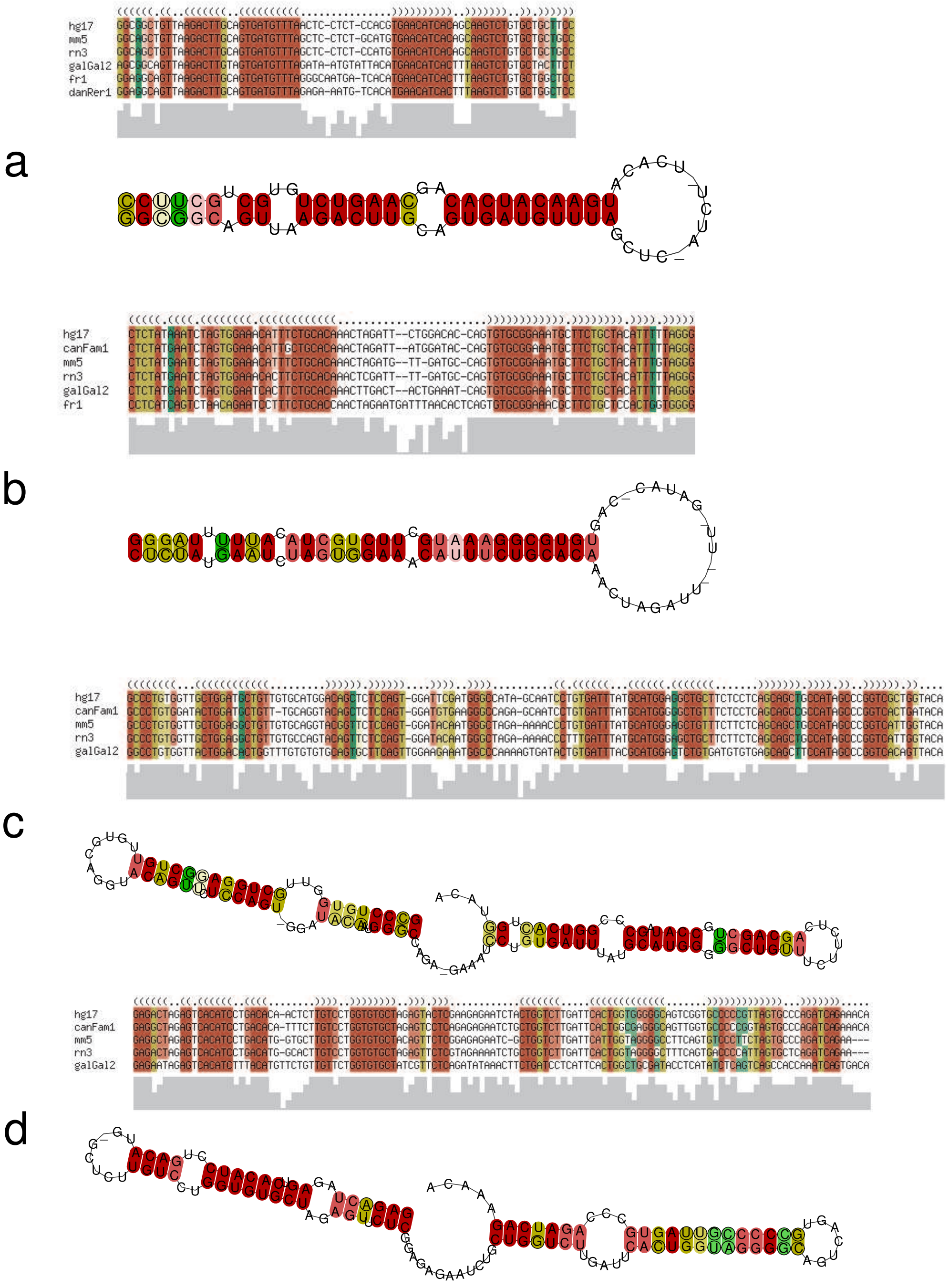

Fig. 3. (Caption see next page)

Fig. 3. 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.

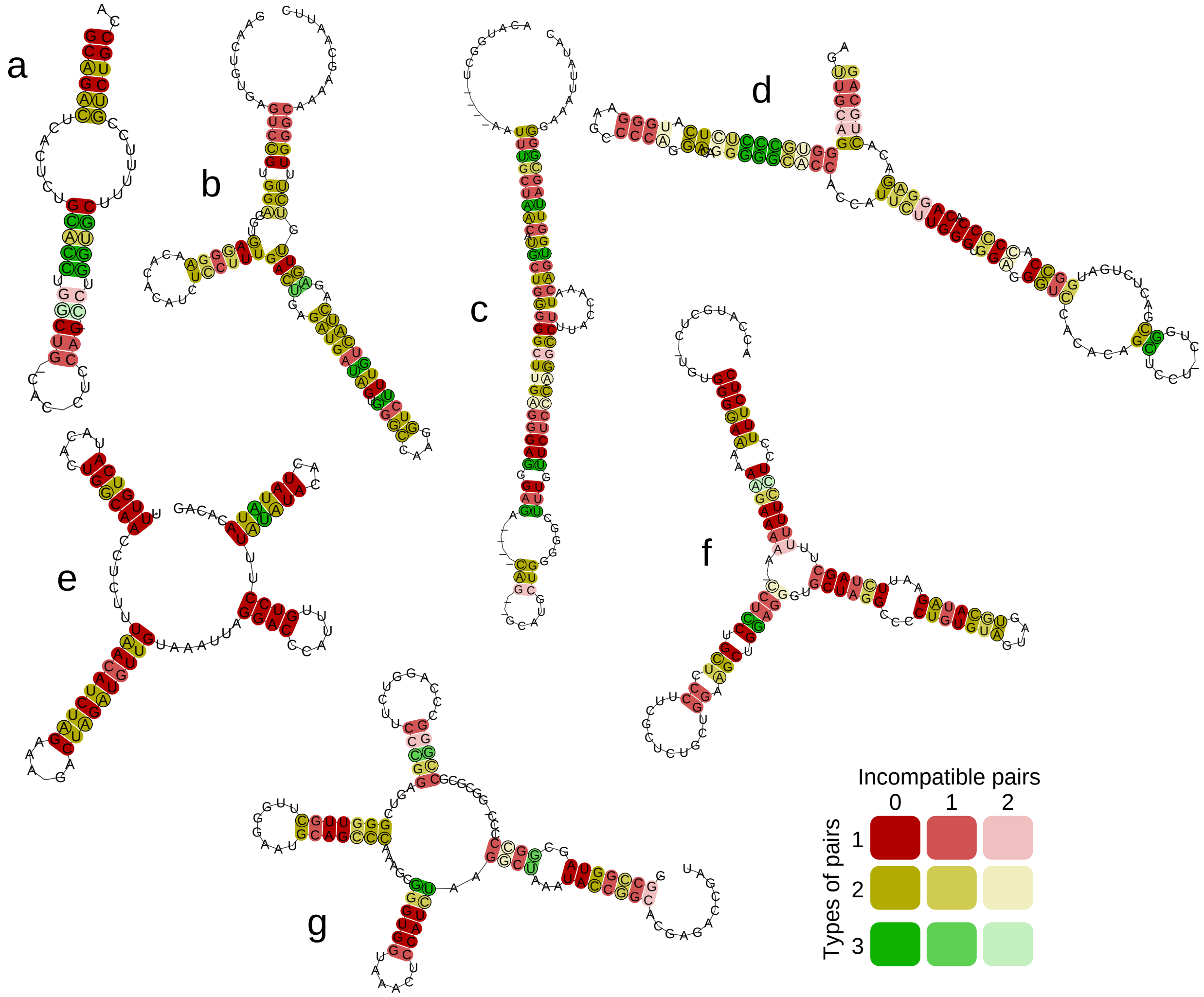

Fig. 4. Selected examples of novel structured RNAs. Structures a–c are located in introns of known protein coding genes, d–g were detected in intergenic regions. Detailed genomic positions are listed in the supplement. The predicted consensus structures with annotation of consistent and compensatory mutations are shown. Circles indicate variable positions in stems, colors indicate the number of different types of base pairs (red: conserved, ochre: two pairs, green: three pairs) that support stabilizing selection on the structure. Pale colors indicate that one or two sequences cannot form the pair in the consensus structure.