Journal of Bioinformatics and Computational Biology
Vol. 6, No. 6 (2008) 1157–1175

Imperial College Press
www.icpress.co.uk

DUPLICATED RNA GENES IN TELEOST FISH GENOMES

DOMINIC ROSE[*,‖], JULIAN JÖRIS[*,**], JÖRG HACKERMÜLLER[*,†,††], KRISTIN REICHE[*,†,‡‡], QIANG LI[*,‡,§§] and PETER F. STADLER[*,†,§,¶,¶¶]

[*]*Bioinformatics Group, Department of Computer Science*
Interdisciplinary Center for Bioinformatics, University of Leipzig
Härtelstraße 16-18, D-04107 Leipzig, Germany

[†]*Fraunhofer Institute for Cell Therapy and Immunology*
Perlickstr. 1, 04103 Leipzig, Germany

[‡]*T-Life Research Center, Fudan University*
Shanghai 200433, China

[§]*Department of Theoretical Chemistry, University of Vienna*
Währingerstraße 17, A-1090 Wien, Austria

[¶]*Santa Fe Institute, 1399 Hyde Park Rd*
Santa Fe, NM 87501, USA
[‖]*dominic@bioinf.uni-leipzig.de*
[**]*julian@bioinf.uni-leipzig.de*
[††]*joerg.hackermueller@izi.fraunhofer.de*
[‡‡]*kristin.reiche@izi.fraunhofer.de*
[§§]*q.li@fudan.edu.cn*
[¶¶]*stadler@bioinf.uni-leipzig.de*

Received 1 October 2007
Revised 17 June 2008
Accepted 18 June 2008

Teleost fishes share a duplication of their entire genomes. We report here on a computational survey of structured non-coding RNAs (ncRNAs) in teleost genomes, focusing on the fate of fish-specific duplicates. As in other metazoan groups, we find evidence of a large number (11,543) of structured RNAs, most of which (∼86%) are clade-specific or evolve so fast that their tetrapod homologs cannot be detected. In surprising contrast to protein-coding genes, the fish-specific genome duplication did not lead to a large number of paralogous ncRNAs: only 188 candidates, mostly microRNAs, appear in a larger copy number in teleosts than in tetrapods, suggesting that large-scale gene duplications do not play a major role in the expansion of the vertebrate ncRNA inventory.

Keywords: Non-coding RNA; ncRNA; ncRNA evolution; ncRNA gene finding; teleost fish; teleosts; *Takifugu rubripes*; whole-genome duplication; comparative genomics; annotation.

1. Introduction

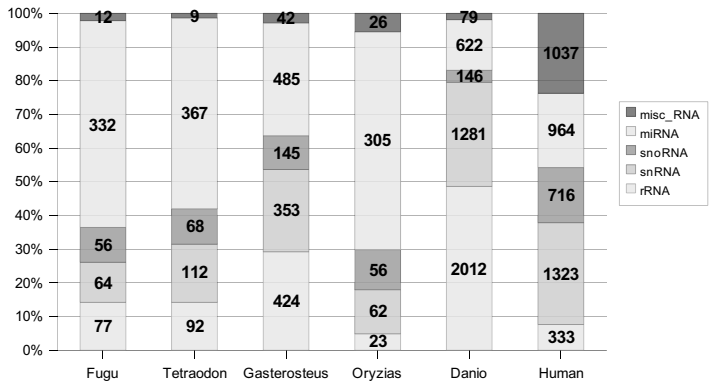
The heterogeneous class of non-protein-coding RNAs (ncRNAs) has in recent years moved from a biochemical curiosity to a main research topic in molecular biology. Recent results from high-throughput transcriptomics[1–4] have established that ncRNAs in fact dominate the transcriptomes of higher eukaryotes, even though the list of ncRNAs is still largely incomplete. With ncRNAs being implicated in a plethora of regulatory roles,[5,6] it becomes an interesting issue to understand their evolution in more detail and beyond the mostly anecdotal narratives available for individual ncRNA gene families (see Bompfünewerer *et al.*[7] for a review).

Compared to the annotation of protein-coding genes, non-coding RNA annotation of genomic sequences is still in its infancy. This is especially true beyond mammalian genomes, which mostly "inherit" the human and mouse annotation in a straightforward way. Particular problems for all large-scale studies of ncRNA evolution are massive differences in ncRNA coverage and biases in annotation, even between fairly closely related organisms. In addition to biased coverage, the currently available ncRNA annotation procedures cannot distinguish clearly between functional ncRNAs and the sometimes huge number of associated pseudogenes. Figure 1 summarizes the situation for the five available teleost fish genomes, making it obvious that a systematic study of ncRNA evolution cannot be meaningfully performed solely on the basis of the available annotation.

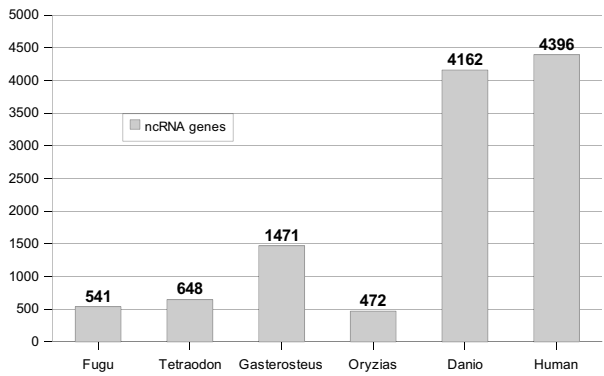
A further drawback of current annotation procedures is that they typically consider only a small subset of the ncRNA universe. Beyond the "housekeeping" RNAs (tRNA, snRNAs, RNAse P and MRP RNAs, and a few Pol III transcripts such as vault and Y RNAs), annotation is mostly restricted to microRNAs and the two classes of snoRNAs. In contrast, computational studies have provided convincing evidence for tens of thousands of RNAs whose secondary structure is under stabilizing selection,[1,8–10] while even more functional ncRNAs without evolutionarily conserved structure were found in the large-scale transcriptomics projects mentioned above.

Here, we study the overall characteristics of ncRNA evolution in teleost genomes. From an evolutionary perspective, teleost fish genomes are of particular interest because they have undergone a complete duplication of their genomes — usually called fish-specific genome duplication (FSGD) — just before the radiation of the crown-group teleosts.[11,12] This specifically raises the question to what extent duplicated ncRNAs are retained after such a large-scale duplication event.

Thus, the purpose of this contribution is twofold. First, we report an `RNAz`-based survey of teleost fish genomes. We demonstrate that, like many other groups of organisms (mammals,[8,9] urochordates,[13] nematodes,[14] insects,[15] and yeasts[16]), teleost genomes contain a large number of previously undescribed, clade-specific, non-coding RNAs. Second, we investigate from a global perspective the fate of ncRNAs in the wake of a genome duplication.



(a)



(b)

Fig. 1. `Ensembl-48` annotation of human and teleost fish ncRNAs. (a) Distribution of annotated ncRNA classes (without mitochondrial RNAs). (b) Absolute numbers of annotated ncRNA genes.

2. `RNAz` Screen: Unbiased Prediction of Structured RNAs

We prepared genome-wide alignments of the non-repetitive ncDNA of the five available teleost genomes (`Ensembl-48`) — fugu, *Takifugu rubripes*, Tr; pufferfish, *Tetraodon nigroviridis*, Tn; stickleback, *Gasterosteus aculeatus*, Ga; medaka, *Oryzias latipes*, Ol; and zebrafish, *Danio rerio*, Dr — using `NcDNAlign`[17] with fugu serving as the reference organism. Due to the larger evolutionary distances among teleosts compared to mammals, only a relatively small fraction of the genomes can be reliably aligned. In order to improve the performance, we consider only DNA that is alignable to the fugu genome after removing all "known coding sequences". Local alignments containing at least three species were then scored using the `RNAz` package (v1.0).[18] `RNAz` is a machine learning approach that allows the prediction of structural non-coding RNAs. It classifies given alignment slices as either "ncRNA"

or "other" by analyzing secondary structure conservation and thermodynamic stability of the fold. Alignments are scanned by moving a 120-nt window with a step width of 40 nt, so that consecutive slices overlap in 80 nt. This sliding-window mechanism is motivated by the fact that many structured RNAs are shorter than 100 nt. Such small RNA signals would "drown" in the noise of longer, mostly unstructured, alignments. On the other hand, there is no reliable signal for secondary structure conservation in too-short alignments. Previous RNAz-based studies discuss these issues in more detail.[8,15,16] A brief overview of our computational procedure is presented in Fig. 2.

At RNAz classification probabilities of $p > 0.5$ and $p > 0.9$, we obtained 19,916 and 6,690 "structured" sequence windows, which can be combined into 11,543 and 4,407 predicted structured RNA elements, respectively. In order to estimate the false discovery rate (FDR), the entire screen has been repeated using column-wise shuffled alignments as input. As described in Ref. 19 and tested in practice in Refs. 8, 9, and 13–15, the shuffling procedure of the RNAz-package generates randomized alignments that preserve the salient characteristics of multiple alignments: length, base composition, gap patterns, and conservation patterns. Comparing the amount of positively scored DNA from control and normal screen yields FDR estimates of 26% and 18% for low and high confidence levels, respectively. Alternatively, we can base FDR estimates on the numbers of positively scored sequence windows and observe similar FDRs: 26% (5,131/19,916, $p > 0.5$) and 18% (1,207/6,690, $p > 0.9$). These estimates might be somewhat optimistic, since the shuffling algorithm of the RNAz package does not account for dinucleotide content. Preserving dinucleotide content while generating randomized alignments is essential to accurately assess the significance of the prediction,[20] and previous RNAz-based studies have shown that taking this effect into account substantially increases the estimated FDR on mammalian sequences (FDR of ∼50% for the ENCODE regions[9]). In contrast, the impact of using dinucleotide instead of mononucleotide shuffling has

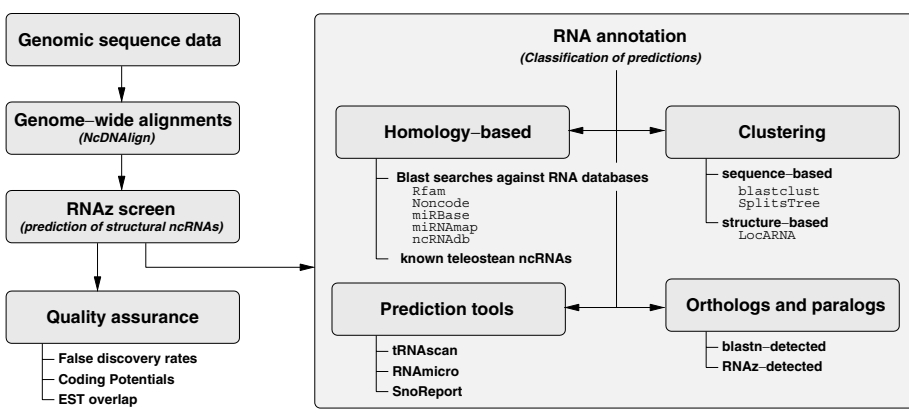


Fig. 2. Overview of the computational workflow.

Table 1. Summary of the RNAz screen.

Species	Tr	Tn	Ga	Ol	Dr
Genome size [Mb]	393	402	462	868	1,440
Without CDS [Mb]	361	211	429	838	1,410
Non-coding alignments	66,448	30,369	56,750	49,958	9,027
Aligned DNA [Mb]	8.45	4.9	8.8	5.44	0.96
Scored by RNAz [Mb]	10.7	5.84	5.41	7.7	1.3
RNAz ($p > 0.5$)	11,543	5,480	9,897	8,970	1,623
[kb]	1,469	692	1,229	1,116	201
RNAz ($p > 0.9$)	4,407	2,186	3,693	3,412	723
[kb]	602	260	431	398	85
FDR ($p > 0.5$) [%]	26	26	26	26	26
FDR ($p > 0.9$) [%]	18	18	19	18	16

CDS: coding sequence.

been quite small on drosophilid sequences.[15] As an additional control, we therefore applied SISSIz,[21] a novel approach to generate dinucleotide-controlled random alignments with the characteristics of a given input alignment. Not unexpectedly, the estimated FDR indeed increased up to 68%. The SISSIz-randomized control screen still contained 41 known ncRNAs, indicating that SISSIz leads to a conservative (pessimistic) FDR estimate. Table 1 summarizes the statistics of the RNAz screen.

We excluded currently annotated protein-coding sequences (Ensembl-48) from our analysis. The coding potential scores of CPC[22] further suggest that almost all RNAz hits are indeed in non-coding regions: only 100 (<1%) are predicted as "coding", a value that is less than CPC's false-positive rate, which its authors estimate at ∼2%.

In order to obtain at least a rough estimate of the sensitivity, we compared the RNAz output with existing annotation. Since annotated repetitive elements as defined by RepeatMasker have been removed from our input data, several classical RNA families have also been excluded, namely tRNAs, some of the snRNAs, and most of the "misc_RNAs". Only 321 of 541 known fugu ncRNAs passed the repeat masking step, and only 245 of these are sufficiently well conserved to be alignable with homologous sequences of the remaining teleosts; of these, 221 ncRNAs (90%) are recognized by RNAz. In order to obtain a more realistic sensitivity estimate, we retrieved all annotated ncRNAs from the fugu genome, added 100 nt of flanking sequence on both sides, and used NcDNAlign to retrieve their homologs and to construct multiple alignments. These were then scored with RNAz using the same parameters as the main screen. We obtained an overall sensitivity of about 85% (see Table 2). As in previous work, we note that in particular for snoRNAs the sensitivity is poor, while sensitivity values for miRNAs are encouraging. A considerable number of predicted ncRNAs overlap expressed sequence tags (ESTs) and hence show evidence of active transcription (see Table 3).

Table 2. Sensitivity of RNAz on Ensembl-48-annotated fugu ncRNAs. Denoted percentages refer to the number of recovered ncRNAs using RNAz ($p > 0.5$) over the number of ncRNAs present in the input alignments.

Class	RNAz	Input	Ensembl-48	Sensitivity (%)
rRNAs	66	77	77	86
snRNAs	49	64	64	77
snoRNAs	22	55	56	40
miRNAs	261	276	332	95
Other	11	11	12	100
All	409	483	541	85

Table 3. A comparison of ncRNA candidates and EST sequences provides evidence of transcriptional activity. The table lists the number of fugu loci obtained by BLAST searches (E-value < 1e-5) between fugu sequences from (1) our teleostean ncRNA candidates ($p > 0.5$; overall 11,543 loci) and (2) the tetrapod-conserved subset (overall 1,581 loci; see Table 4) against EST sequences of four teleosts provided by the UCSC Table Browser. The small number of matches with fugu ESTs may be due to the fact that UCSC only offers access to sequences of a deprecated fugu assembly, which is only partially compatible with the Ensembl data our study is based on. For tetraodon, no ESTs were available. The last column lists the hits to the medaka EST database (http://medaka.lab.nig.ac.jp/est_index.html).

ESTs of	Tr	Ga	Dr	Ol	m-EST-DB
(1)	796	3,847	2,080	3,558	644
Annotated	292	1,436	768	2,229	63
Unknown	504	2,411	1,314	1,329	581
(2)	192	616	844	693	173
Annotated	33	159	194	159	42
Unknown	159	457	650	534	131

3. RNA Annotation

Our annotation procedure follows a recently published scheme.[23] Overall, 1,372 RNAz hits can be annotated by the following protocol. Firstly, known RNA genes and homologs of known ncRNAs were identified based on sequence comparison, here using blastn searches against all major ncRNA databases: Rfam 8.0,[24] NONCODE 2.0,[25] miRBase 10.0,[26] miRNAMap 1.0,[27] and ncRNAdb[28] (number of RNAz hits showing sequence conservation with entries of the respective ncRNA database: 115 Rfam, 104 NONCODE, 208 miRBase, 179 miRNAMap, 71 ncRNAdb).

In the second step, we used specialized programs to recognize novel members of three ncRNA classes. Since tRNAs were removed from the input set as multi-copy genes, we found only one tRNA and two tRNA pseudogenes with tRNAscan.[29] An experimental version of SnoReport[30] identified 885 snoRNAs (727 CD- and 136 HACA-box snoRNAs, 22 are classified as both), of which 8 match previously annotated miRNAs. This is within the expected FDR of SnoReport. We used RNAmicro[31] to determine putative miRNA precursors. At a confidence level of

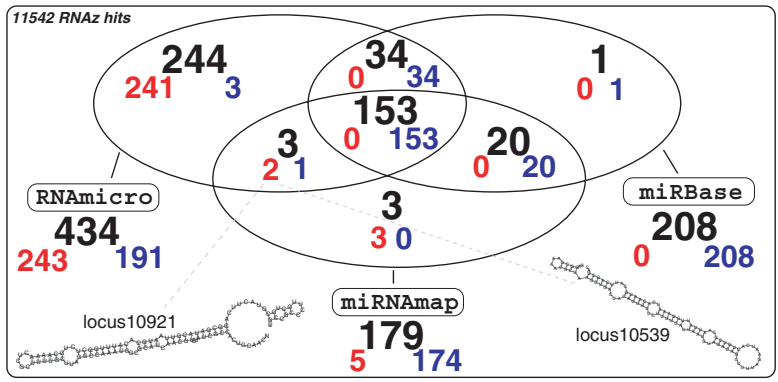


Fig. 3. MicroRNA annotation of `RNAz` hits. The Venn diagram contrasts the number of `RNAz` predictions that are positively classified by `RNAmicro` and hits that also have a `miRBase` and/or a `miRNAMap BLAST` hit. For each subset, we record three numbers in the form ${}_yx_z$ with $y + z = x$, where x is the number of microRNA precursors in the intersection of all three approaches, z is the number of corresponding miRNAs annotated in `Ensembl`, and y is the number of miRNAs not annotated in `Ensembl` (these structures are the most interesting ones because they constitute putatively novel miRNAs).

$p_{\texttt{RNAmicro}} > 0.5$, we obtained 434 candidates, of which 190 have `blastn` matches with `miRBase` or `miRNAMap` entries. Figure 3 summarizes the miRNA annotation in more detail.

An example of a novel fugu miRNA candidate not yet listed in `Ensembl-48` is a homolog of *xtr-miR-449* [Fig. 4(top)]. The structure shows all hallmarks of a microRNA precursor including the characteristic conservation pattern. Another novel miRNA precursor candidate found by `RNAz` and `RNAmicro` is the intronic locus2693 (scaffold_204, pos. 260585–260681), which contains the mature sequence of *dre-miR-728*. It is also not yet included in `Ensembl-48`'s annotation tracks.

As a third step, we employed the structural clustering procedure proposed by Will *et al.*[32] to find possible novel structural classes. All `RNAz` hits with either a classification probability exceeding 0.9 or a valid annotation form the input of the clustering procedure (2,293 loci). In brief, a modified Sankoff algorithm is used to compute local structural alignments and their consensus structure. The cluster tree (see Fig. 5) is then obtained by agglomerative clustering based on the alignment scores. It is further processed to identify an optimal partition by evaluating the squared error of the minimum free energies of the individual sequences located within the subtree rooted at an internal node relative to the minimum free energy of their common consensus secondary structure. If the increase of the squared error is unexpectedly high when traversing the cluster tree from the leaves to the root node, the merging of two internal nodes to one common cluster is stopped. In this way, we estimate that there are at least 106 relevant clusters (which depend on several thresholds, including the chosen cutoff to stop the merging procedure, a structure conservation index (SCI) > 0.7, and a consensus minimum free

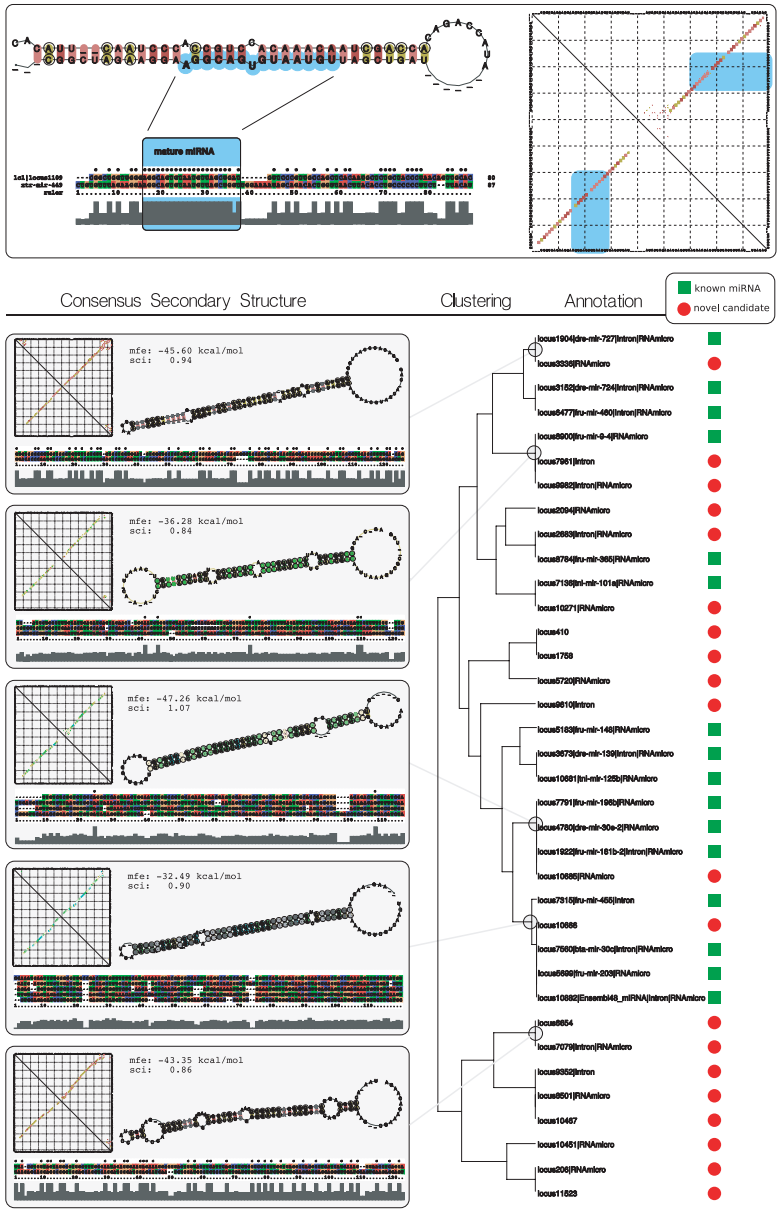


Fig. 4. Novel microRNAs in teleost genomes. Top: Fugu *miRNA-449* (locus 1285, scaffold_360, pos. 9131–9211) is an example of a homolog of a known miRNA not annotated in `Ensembl`. This `RNAz` prediction is classified as an miRNA precursor by `RNAmicro`, and a `BLAST` search against the mature miRNAs in `miRBase` reveals that it contains the sequence of the mature miRNA *xtr-miR-449*. Bottom: Structure-based clustering reveals several collections of hairpin structures. These clusters contain already-known miRNAs, but also closely related novel miRNA precursor candidates. Primary sequences of this example vary considerably, but secondary structure motifs are significantly well conserved. Novel ncRNA candidates (red circle) with a positive `RNAmicro` classification are prime candidates for novel miRNAs.

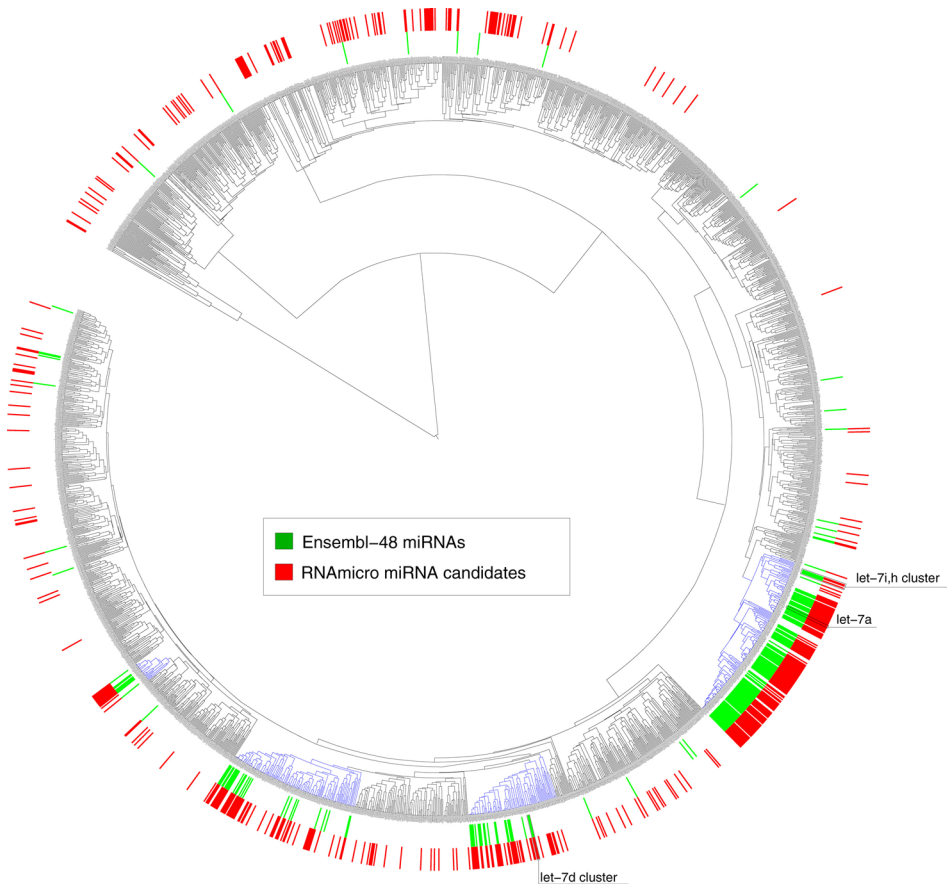


Fig. 5. Cluster tree of chosen high-scoring teleostean ncRNA candidates. The tree comprises 4,585 nodes, of which 2,293 are leaves. Known fugu miRNAs as provided by `Ensembl-48` are indicated in green (inner circle), while `RNAmicro`-predicted miRNA candidates are marked in red (outer circle). Certain subtrees containing typical miRNA clusters are drawn in blue. Encouragingly, the majority of known miRNA clusters is recovered (red+green). Overall, there is substantial evidence for novel miRNA(-like) classes of ncRNA structures in teleostean genomes (red only).

energy (MFE) < -20 kcal/mol) having an average cardinality of 2.9 sequences, an average SCI of 0.89, and an average MFE of -34.97 kcal/mol. Figure 5 displays the resulting cluster tree and Fig. 4(bottom) illustrates an example subtree in more detail. Besides known ncRNAs, both trees contain novel, closely related structures, demonstrating that our clustering approach indeed yields useful classifications. As in other recent work,[32,33] we predict several new miRNA candidates by means of clustering, including some candidates that are not recognized by `RNAmicro`.

Of particular interest are sets of unannotated `RNAz` hits that not only fold into similar structures and hence are identifiable in the cluster tree, but are also located

in close vicinity on the genome. Such arrangements are observed, for example, for polycistronic microRNA transcripts,[34] for multiple unrelated snoRNAs sharing the same host gene,[35] and for several Pol III transcripts including Y RNAs[36,37] and vault RNAs. Altuvia *et al.*[34] argued that miRNA precursors which are located within short chromosomal distances (<3,000 nt) from each other most likely arise from polycistronic transcripts. Interestingly, the 434 `RNAmicro` predictions contain 41 such clusters; however, there are only six such clusters within the 246 "novel" miRNA candidates, each comprising either two or three loci only. Overall, we observe 1,223 genomic clusters with a maximal distance of 1 kb consisting of two to eight loci. Increasing the cutoff distance to 10 kb yields 1,912 genomic clusters, the largest one comprising 110 sequences. Within the high-confidence $p_{\texttt{RNAz}} > 0.9$ predictions, we still find 234 clusters (730 at 10 kb), each with up to 5 (20) loci. Figure 6 shows the distribution of pairwise `LocARNA` distances used in structural clustering for `RNAz` hits that form genomic clusters. The distribution is clearly bimodal, consisting of a random bulk and a small set of signals with very similar structures. While fairly restrictive, the data indicate that there are at least a handful of novel structural classes of RNAs that tend to cluster in the same genomic location.

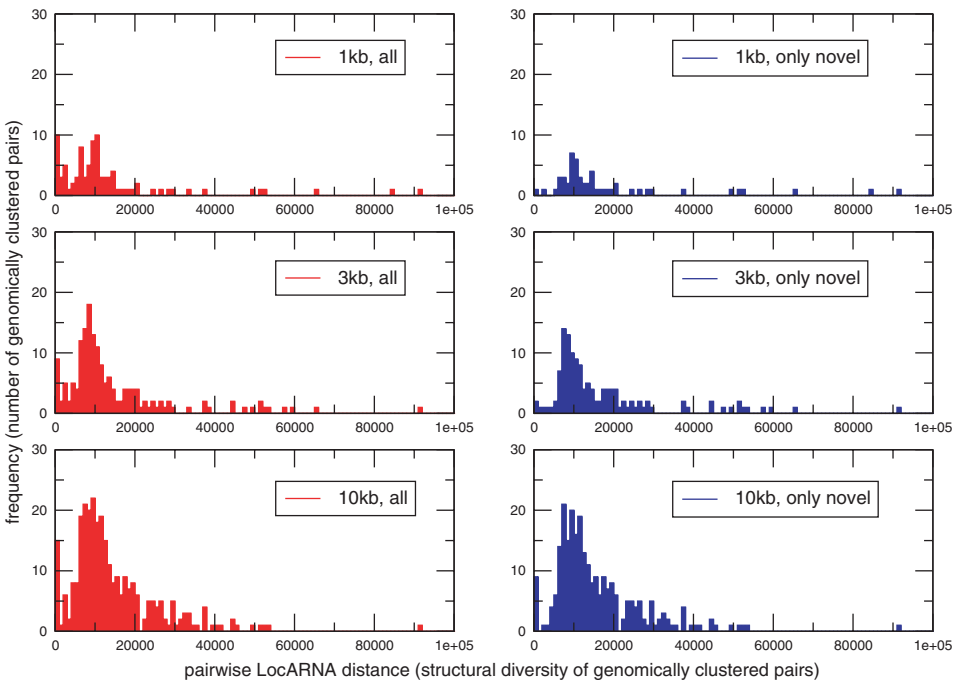


Fig. 6. Distribution of structure distances between pairs of adjacent high-confidence ($p_{\texttt{RNAz}} > 0.9$) `RNAz` hits with a maximum distance of 1 kb, 3 kb, and 10 kb, respectively. The left column shows histograms for all genomic clusters of `RNAz` predictions, while the right column is restricted to clusters containing putatively novel structures that are not included in `Ensembl-48`.

4. Orthologs and Paralogs

A BLAST search (E-value $< 10^{-5}$) shows that 944 (8%) of the fugu RNAz predictions have sequence similarity with highly conserved non-coding (vertebrate) elements (HCNEs),[38] of which 246 candidates have some annotation. A total of 541 loci are not only alignable with vertebrates (human, mouse, dog, chicken, or shark using NcDNAlign); the extended alignments are still classified as structured RNAs by RNAz (see Table 4). Within this set of highly conserved structures, the estimated FDR significantly reduces to only 7%–11%, and only two sequences are classified as coding by CPC. Up to 164 of these sequences have homologs within the set of 997 high-confidence, highly conserved predictions of the prior mammalian RNAz screen.[8] Our data thus provide at least 377 additional well-conserved ncRNA candidates that were not detected by the earlier survey.

According to the UCSC Table Browser, at least 250 (2%) teleostean ncRNA candidates have sequence similarity with untranslated regions (UTRs) of human protein-coding genes (5′ UTRs: 136; 3′ UTRs: 162; some match both types of UTRs). Regarding the set of tetrapod-conserved RNAz predictions, 140 (26%) out of 541 candidates cover human intronic sequences and 181 match human UTRs (5′ UTRs: 112; 3′ UTRs: 116). Due to the lack of a reliable fugu UTR annotation, we analyzed the 1-kb flanking region of the 5′ and 3′ boundaries of Ensembl's fugu protein-coding genes. Interestingly, the 5′ flanking regions (3′ flanking regions) contain 1,505 (1,448) RNAz hits, of which 94% (93%) are potential *cis*-regulatory signals of unknown function. Among the tetrapod-conserved RNAz hits, 114 candidates reside within 1 kb upstream or downstream of protein-coding genes, 61 of which are not annotatable.

In contrast, only a tiny fraction of our structured RNA candidates seems to be conserved within any of the available invertebrate genomes. Again, note

Table 4. Phylogenetic conservation of fugu RNAz predictions. For all RNAz hits ($p > 0.5$), we list the number of elements with blastn-detected homologs (left part) and the number of hits that again were classified as structured (RNAz) when aligned with the sequences from human, mouse, dog, chicken, or shark (right part). By considering both approaches, we overall obtain 1,581 tetrapod-conserved ncRNA candidates.

Genome	blastn hit			Positive RNAz prediction		
	All	Annotation		All	Annotation	
		Yes (%)	No (%)		Yes (%)	No (%)
Any non-teleost	1,487	278 (19)	1,209 (81)	541	230 (43)	311 (57)
Homo sapiens	989	222 (22)	767 (78)	478	208 (44)	270 (56)
Gallus gallus	1,050	217 (21)	833 (79)	455	204 (45)	251 (55)
Canis familiaris	1,010	225 (22)	785 (78)	467	203 (44)	264 (56)
Mus musculus	912	211 (23)	701 (77)	429	194 (45)	235 (55)
Callorhinchus milii	610	153 (25)	457 (75)	357	190 (53)	167 (47)
Petromyzon marinus	188	40 (21)	148 (79)	—	—	—
Drosophila melanogaster	32	21 (66)	11 (34)	—	—	—
Caenorhabditis elegans	8	5 (63)	3 (37)	—	—	—

that this observation is misleading because the best-conserved "housekeeping" ncRNAs, in particular tRNAs and snRNAs, were removed from the input data.

From an evolutionary perspective, an important issue is the fate of duplicated non-coding RNA signals. Since the ancestral vertebrate genome already went through two rounds of whole-genome duplication, we have to expect up to eight paralogs in teleost genomes. For the case of microRNAs, we demonstrated in previous work that duplicated ncRNAs frequently have survived the genome duplication(s).[39] We therefore compared the `RNAz` predictions at sequence level and list them by their copy number in Table 5. The repeat-masked RNAs are not informative for this purpose because their copy numbers are highly variable between relatively closely related species (see supplemental data).

The fraction of predictions with more than eight copies seems to contain repetitive or pseudogenic elements rather than correctly identified, evolutionary duplicated ncRNAs. The fraction of unannotated hits grows with a larger number of copies. One explanation for this observation might be that these sequences diverged by accumulating mutations after they were duplicated, but simultaneously preserve their secondary structure, so that it becomes more and more difficult to reliably annotate them comparatively at the sequence level, while they still remain detectable by algorithms that incorporate structural features (e.g. `RNAz` or `LocARNA`). Interestingly, the mutual distance obtained by structure-based clustering between duplicated ncRNA candidates increases with the number of loci that belong

Table 5. Distribution of paralogous fugu `RNAz` hits. Paralogs are obtained by `BLAST` searches (`E-value` < 1e-3) of the fugu `RNAz` candidate sequences against themselves. They are compared with each other to exclude likely pseudogenes. We observe 1,403 duplicated loci, 964 of which have two to eight copies. Furthermore, we provide the number of `RNAz` hits conserved between fugu and human, which occur more often in fugu than in human (and the other way round). As an example, 135 ncRNA candidates, present as single copy in human, occur more than once in fugu; conversely, 81 candidates, not duplicated in fugu, appear multi-copied in human.

# copies	1	2	3	4	5	6	7	8	>8
$p > 0.5$	10,139	573	150	82	50	39	36	34	439
Annotated	379	85	45	32	25	8	8	4	51
Unknown	9,760	488	105	50	25	31	28	30	388
Teleost-specific	8,964	409	93	48	23	30	19	30	345
In tetrapods	1,175	164	57	34	27	9	17	4	94
In human	777	103	35	25	23	8	7	4	7
Fugu > human	135	34	18	1	0	0	0	0	0
Human > fugu	81	7	0	0	0	0	1	0	0
$p > 0.9$	3,689	253	67	75	44	25	25	18	211
Annotated	238	68	37	34	27	6	0	2	46
Unknown	3,451	185	30	41	17	19	25	16	165
Teleost-specific	3,180	162	26	38	12	18	14	16	171
In tetrapods	509	91	41	37	32	7	11	2	40
Fugu > human	94	28	17	1	0	0	0	0	0
Human > fugu	38	1	0	0	0	0	0	0	0

to the cluster (see supplement). This could be explained by a duplication/deletion mechanism in which cluster members are destroyed at random by mutation as the exact copy number is not under strong selection.

Comparing the number of ncRNAs that occur at most eight times in fugu with the copy number in human (cf. Table 5) reveals that, in 188 cases, teleosts contain more ncRNA copies than tetrapods, of which 74 are of unknown function. In turn, tetrapods contain a higher number of copies in only 89 cases. These data indicate that, with the most notable exception of microRNAs,[39] additional copies of non-coding RNAs are rarely retained in the aftermath of fish-specific genome duplication (FSGD). The loss appears to be more extensive than for protein-coding genes, where at least about 1,000 paralogs arising from FSGD have been reported in the genomes of fugu and tetraodon.[40] It is conceivable, however, that duplicated ncRNAs have diverged so far that they are not recognized as paralogs by `blast`-based methods.

To illustrate the fate of ncRNAs subsequent to duplication in more detail, we estimated the densities of the bivariate distribution of sequence versus structural similarity over duplicated and randomly chosen pairs of sequences (Fig. 7). Duplicated loci can be assigned to three distinct groups [Fig. 7(c)]: those where both sequence and structure are nearly identical (p1), a set with significant sequence divergence but negligible structural differences (p2), and a set which largely overlaps with the background distribution of randomly chosen pairs (p3). The ostensible discrepancy between the `E-value` of the initial `BLAST` search to identify duplicated pairs and the mean pairwise identity of the `ClustalW` alignment used here is a consequence of scoring differences between local and global alignments: subsequences of both loci are maintained with high similarity (which is what `blast` focuses on), whereas overall the loci may be quite highly divergent in terms of both sequence and structure. This might be interpreted as a process that maintains some local functional features, presumably those that are required to determine an RNA type (e.g. a protein-binding site for a guide RNA), whereas the functional role can diverge quickly upon duplication. A related behavior can be observed for duplicated miRNAs, which mainly make up for the p2 peak (cf. Suppl. Fig. 4). Upon duplication, miRNAs largely maintain their precursor structure but diverge on the sequence level (the affiliation to a particular functional class is maintained), whereas the functional role — the target specificity — changes. As expected, duplicated snoRNAs do not share such strict constraints; they are found at any hotspot position of the density plots. Similarly, the non-annotatable loci also appear widely spread throughout the landscape. No other annotated classes of ncRNAs are found in the set of duplicated pairs.

The small number of ncRNA candidates that have two paralogs produced by FSGD severely limits attempts to detect structured RNAs by comparing paralogous regions of the same genome. We have tested this in a preliminary study using paralogous sequences from the fugu genome as input for `RNAz`. With this ansatz, we recovered only 283 of the 454 fugu ncRNAs known at the time (`Ensembl-45`).

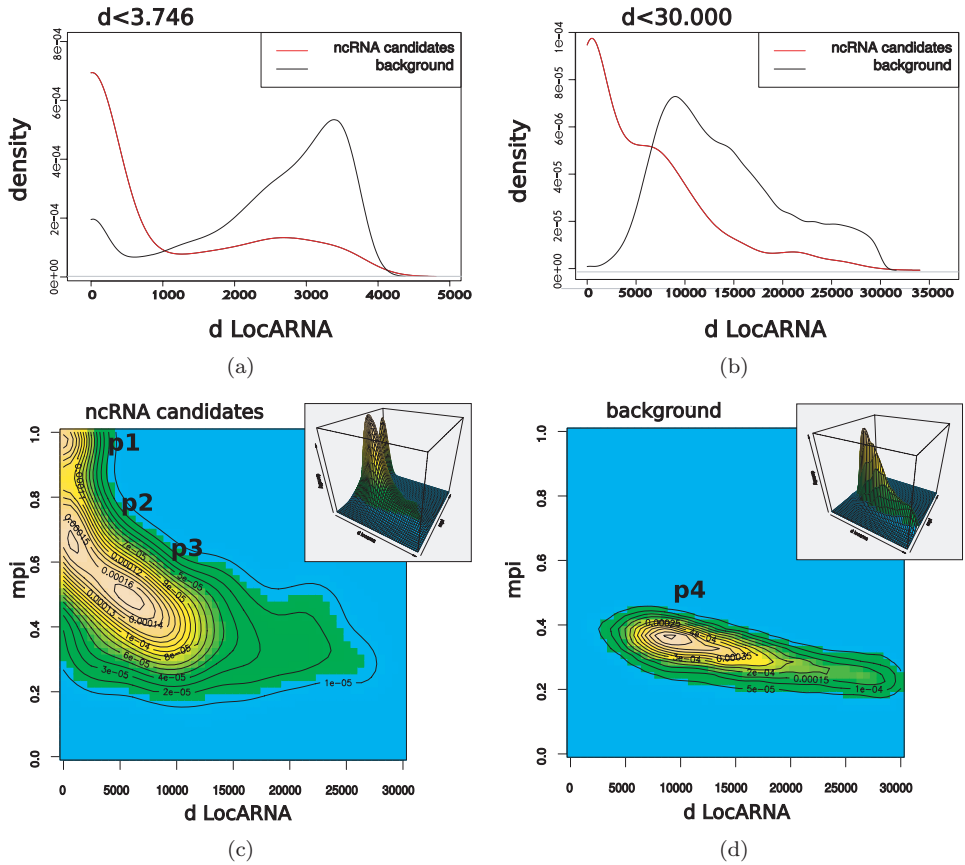


Fig. 7. Distribution of structure distances for duplicated and all ncRNA candidates. The figure illustrates the density plot of the distribution of all pairwise `LocARNA` distances of putatively duplicated pairs with recognizable sequence similarity (red curve) and the background distribution of randomly selected pairs (black curve) for all reasonable `LocARNA` alignments, i.e. (a) positive alignment scores and (b) the complete distribution excluding 25 outliers. Subsequent to duplication events, the majority of multi-copy genes preserve their secondary structure; however, a substantial fraction of genes displays highly diverged secondary structure, comparable to the distance of random pairs. (c) The bivariate density of mean pairwise identity versus `LocARNA` distance of pairs of duplicated ncRNA candidates displays three distinct peaks, corresponding to genes which are highly similar at both sequence and secondary structure levels (p1); genes with high structural similarity and diverged sequences (p2); and genes which show a degree of divergence at sequence and structure levels (p3), comparable to the background distribution of random pairs (p4) [see (d)]. In contrast to (a)–(c), where all possible pairs have been considered as background distribution, (d) includes only 150,000 randomly chosen pairs to reduce the computation time for `ClustalW` alignments.

Despite the inherent limitations of comparative genomics with a single genome due to massive loss of duplicated genes, we feel that further methodological improvements are worthwhile; these will focus in particular on increasing the sensitivity of detecting paralogous regions.

5. Discussion

We have reported here on an unbiased survey for evolutionarily conserved structured ncRNAs in the currently available genomes of teleosts. As in other metazoan animals, we found evidence of several thousand structured RNA motifs, of which only a small fraction can be annotated. Due to the large evolutionary distances among teleosts compared to mammals, our `RNAz` screen has a decreased sensitivity. The absolute value of 11,543 structured elements thus cannot be fairly compared with the much larger number of predictions for mammalian genomes.[8,9]

Furthermore, an overwhelming majority of the signals are specific to teleosts, i.e. almost no homologous sequences can be identified in invertebrates. Nevertheless, we have identified several hundred previously unannotated candidates that are shared between teleosts and tetrapods. Conversely, the majority of those `RNAz` hits that have homologous sequences in other vertebrates can still be recognized as structured RNAs at the expanded phylogenetic range. There is strong evidence for the existence of previously undescribed structurally defined ncRNA families from structure-based clustering (see Fig. 6).

Following FSGD, we observe that very few ncRNAs retain recognizable duplicates. Indeed, only a small minority of structured ncRNAs appear in a copy number between 2 and 8. The overwhelming majority are single-copy loci, while at the same time about 12% of the loci are accounted for by multi-copy gene families (note that the latter number is an underestimate, since it excludes, for instance, tRNA genes). It appears, therefore, that with the exception of a few RNA classes, most notably microRNAs, large-scale duplication events do not lead to a corresponding increase in the ncRNA repertoire — at least as far as RNAs which depend on a well-defined structure are concerned. One immediate implication is that comparative approaches within the same genome, i.e. comparisons between paralogous regions, will have very limited sensitivity for ncRNA discovery.

6. Supplementary Material

Machine-readable sequence and annotation files of the *Takifugu rubripes* `RNAz` predictions as well as supplemental text, figures, and tables are available at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-024/.

Acknowledgments

This work was funded, in part, by the German DFG under the auspices of the Bioinformatics Initiative BIZ-6/1-2 and the SPP1174 "Deep Metazoan Phylogeny", and by the 6th Framework Programme of the European Union (SYNLET). Q. Li's visit to Leipzig in the fall of 2006 was supported in part by the "Vereinigung von Förderern und Freunden der Universität Leipzig e.V." and by grants to Prof. Bailin Hao (Fudan University, Shanghai, China). We are grateful to Tanja Gesell and Stefan Washietl for their offer to use `SISSIz` before it was published.

References

1. The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature* **447**:799–816, 2007.
2. Carninci P, Constructing the landscape of the mammalian transcriptome, *J Exp Biol* **210**:1497–1506, 2007.
3. Pheasant M, Mattick JS, Raising the estimate of functional human sequences, *Genome Res* **17**:1245–1253, 2007.
4. Kapranov P, Cheng J, Dike S *et al.*, RNA maps reveal new RNA classes and a possible function for pervasive transcription, *Science* **316**:1484–1488, 2007.
5. Lau NC, Lai EC, Diverse roles for RNA in gene regulation, *Genome Biol* **6**:315, 2005.
6. Mattick JS, Makunin IV, Non-coding RNAs, *Hum Mol Genet* **15**:R17–R29, 2006.
7. Bompfünewerer AF, Flamm C, Fried C *et al.*, Evolutionary patterns of non-coding RNAs, *Theory Biosci* **123**:301–369, 2005.
8. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF, Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome, *Nat Biotechnol* **23**:1383–1390, 2005.
9. Washietl S, Pedersen JS, Korbel JO *et al.*, Structured RNAs in the ENCODE selected regions of the human genome, *Genome Res* **17**:852–864, 2007.
10. Torarinsson E, Sawera M, Havgaard J, Fredholm M, Gorodkin J, Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure, *Genome Res* **16**:885–889, 2006.
11. Meyer A, Van de Peer Y, From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD), *Bioessays* **27**:937–945, 2005.
12. Crow KD, Stadler PF, Lynch VJ, Amemiya CT, Wagner GP, The fish specific Hox cluster duplication is coincident with the origin of teleosts, *Mol Biol Evol* **23**:121–136, 2006.
13. Missal K, Rose D, Stadler PF, Non-coding RNAs in *Ciona intestinalis, Bioinformatics* **21**(Suppl 2):ii77–ii78, 2005.
14. Missal K, Zhu X, Rose D, Deng W, Skogerbo G, Chen R, Stadler PF, Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*, *J Exp Zoolog B Mol Dev Evol* **306**:379–392, 2006.
15. Rose D, Hackermüller J, Washietl S, Findeiß S, Reiche K, Hertel J, Stadler PF, Prohaska SJ, Computational RNomics of drosophilids, *BMC Genomics* **8**:406, 2007.
16. Steigele S, Huber W, Stocsits C, Stadler PF, Nieselt K, Comparative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions, *BMC Biol* **5**:25, 2007.
17. Rose D, Hertel J, Reiche K, Stadler PF, Hackermüller J, `NcDNAlign`: Plausible multiple alignments of non-protein-coding genomic sequences, *Genomics* **92**:65–74, 2008.
18. Washietl S, Hofacker IL, Stadler PF, Fast and reliable prediction of noncoding RNAs, *Proc Natl Acad Sci USA* **102**:2454–2459, 2005.
19. Washietl S, Hofacker IL, 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. Babak T, Blencowe BJ, Hughes TR, Considerations in the identification of functional RNA structural elements in genomic alignments, *BMC Bioinformatics* **8**:33, 2007.
21. Gesell T, Washietl S, Dinucleotide controlled null models for comparative RNA gene prediction, *BMC Bioinformatics* **9**:248, 2008.

22. Kong L, Zhang Y, Ye Z, Liu X, Zhao S, Wei L, Gao G, CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine, *Nucleic Acids Res* **35**:W345–W349, 2007.
23. The Athanasius F. Bompfünewerer Consortium, RNAs everywhere: Genome-wide annotation of structured RNAs, *J Exp Zoolog B Mol Dev Evol* **308**:1–25, 2007.
24. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A, Rfam: Annotating non-coding RNAs in complete genomes, *Nucleic Acids Res* **33**:D121–D124, 2005.
25. Liu C, Bai B, Skogerbo G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R, NONCODE: An integrated knowledge database of non-coding RNAs, *Nucleic Acids Res* **33**:D112–D115, 2005.
26. Griffiths-Jones S, miRBase: The microRNA sequence database, *Methods Mol Biol* **342**:129–138, 2006.
27. Hsu PWC, Huang H, Hsu S, Lin L, Tsou A, Tseng C, Stadler PF, Washietl S, Hofacker IL, miRNAMap: Genomic maps of microRNA genes and their target genes in mammalian genomes, *Nucleic Acids Res* **34**:D135–D139, 2006.
28. Szymanski M, Erdmann VA, Barciszewski J, Noncoding RNAs database (ncRNAdb), *Nucleic Acids Res* **35**:D162–D164, 2007.
29. Lowe TM, Eddy SR, `tRNAscan-SE`: A program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res* **25**:955–964, 1997.
30. Hertel J, Hofacker IL, Stadler PF, `SnoReport`: Computational identification of snoRNAs with unknown targets, *Bioinformatics* **24**:158–164, 2008.
31. Hertel J, Stadler PF, Hairpins in a haystack: Recognizing microRNA precursors in comparative genomics data, *Bioinformatics* **22**:e197–e202, 2006.
32. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R, Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering, *PLoS Comput Biol* **3**:e65, 2007.
33. Ritchie W, Legendre M, Gautheret D, RNA stem-loops: To be or not to be cleaved by RNAse III, *RNA* **13**:457–462, 2007.
34. Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H, Clustering and conservation patterns of human microRNAs, *Nucleic Acids Res* **33**:2697–2706, 2005.
35. Quezada C, Navarro C, San Martin R, Alvarez M, Molina A, Vera MI, Genomic organization of nucleolin gene in carp fish: Evidence for several genes, *Biol Res* **39**:353–365, 2006.
36. Mosig A, Guofeng M, Stadler BMR, Stadler PF, Evolution of the vertebrate Y RNA cluster, *Theory Biosci* **126**:9–14, 2007.
37. Perreault J, Perreault JP, Boire G, Ro-associated Y RNAs in metazoans: Evolution and diversification, *Mol Biol Evol* **24**:1678–1689, 2007.
38. Woolfe A, Goodson M, Goode DK *et al.*, Highly conserved non-coding sequences are associated with vertebrate development, *PLoS Biol* **3**:e7, 2005.
39. Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, Students of Bioinformatics Computer Labs 2004 and 2005, The expansion of the metazoan microRNA repertoire, *BMC Genomics* **7**:25, 2006.
40. Jaillon O, Aury J, Brunet F *et al.*, Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype, *Nature* **431**:946–957, 2004.

Dominic Rose received his Diploma in Computer Science from the University of Leipzig, Germany, in 2006. He is now working on his Ph.D. dissertation in Bioinformatics at the same university, focusing on ncRNA detection and annotation.

Julian Jöris is a Master's student of Computer Science, focusing on Bioinformatics, at the University of Leipzig, Germany.

Jörg Hackermüller received his Ph.D. in Chemistry from the University of Vienna, Austria, in 2004. After postdoctoral positions at the Novartis Institutes for BioMedical Research, Vienna, and at the Fraunhofer Institute for Cell Therapy and Immunology (IZI), Leipzig, Germany, he is currently leading the RNomics group at Fraunhofer IZI.

Kristin Reiche received her Ph.D. in Computer Science from the University of Leipzig, Germany, in 2007, focusing on ncRNA detection and annotation. She is now a staff scientist with the RNomics group at the Fraunhofer Institute for Cell Therapy and Immunology (IZI), Leipzig, Germany.

Qiang Li is a Ph.D. student focusing on plant ncRNA detection working with Prof. Bailin Hao, at the T-Life Research Center, Fudan University, Shanghai, China.

Peter F. Stadler received his Ph.D. in Chemistry from the University of Vienna, Austria, in 1990, and then worked as Assistant Professor and later Associate Professor for Theoretical Chemistry at the same school. In 2002, he moved to the University of Leipzig, Germany, as Full Professor for Bioinformatics. Since 1994, he has been an External Professor at the Santa Fe Institute, USA. He also helped to found the RNomics group at the Fraunhofer Institute for Cell Therapy and Immunology (IZI), Germany, in 2005.