DUPLICATED RNA GENES IN TELEOST FISH GENOMES – SUPPLEMENT –

Dominic Rose, Julian Jöris, Jörg Hackermüller[*], Kristin Reiche[*], Qiang LI[†], Peter F. Stadler[‡]

Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
{*dominic,julian,kristin,stadler*}*@bioinf.uni-leipzig.de, joerg.hackermueller@izi.fraunhofer.de, q.li@fudan.edu.cn*

Supplemental text

All annotatable ncRNA candidates are tagged with certain keywords describing the underlying annotation procedure. The following labels are used (e.g. in Fig. 3 of the main paper or the supplemental machine readable annotation files):

- **GIVEN**
  The candidate overlaps with a given, already known, ncRNA provided by `Ensembl 48`.
- **Rfam**
  A `blastn` search reveals that the candidate matches `Rfam` entries.
- **Noncode**
  A `blastn` search reveals that the candidate matches `NONCODE` entries.
- **ncRNAdb**
  A `blastn` search reveals that the candidate matches `ncRNAdb` entries.
- **miRBase**
  A `blastn` search reveals that the candidate matches entries of the `miRBase` that form hairpins.
- **mature**
  A `blastn` search reveals that the candidate matches the mature microRNA sequences provided by the `miRBase`.
- **miRNAmap**
  A `blastn` search reveals that the candidate matches sequences of the

[*]Primary Affiliation: Fraunhofer Institute for Cell Therapy and Immunology, Deutscher Platz 5e, 04103 Leipzig, Germany
[†]Primary Affiliation: T-Life Research Center, Fudan University, Shanghai 200433, China
[‡]Secondary Affiliations: Fraunhofer Institute for Cell Therapy and Immunology, Perlickstr. 1, 04103 Leipzig, Germany; 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

`miRNAmap`.

- **RNAmicro**
  The candidate is classified as microRNA by `RNAmicro`.
- **SnoReport**
  The candidate is classified as snoRNA by `SnoReport`.
- **tRNAscan**
  The candidate is classified as tRNA (or tRNA pseudogene) by `tRNAscan`.

Data sources

Known fugu ncRNAs (all except mitochondrial ncRNAs) are obtained from `Ensembl-48` (the term "GIVEN" labels these elements in annotation files). Furthermore, the following genomes are used in this study:

- *Takifugu rubripes*, FUGU 4.0, Jun 2005
- *Tetraodon nigroviridis*, TETRAODON 7, Apr 2003
- *Gasterosteus aculeatus*, BROAD S1, Feb 2006
- *Oryzias latipes*, HdrR, Oct 2005
- *Danio rerio*, Zv7, Apr 2007
- *Callorhinchus milii*, NCBI TraceDB, no version info available
- *Homo sapiens*, NCBI 36
- *Gallus gallus*, galGal3, May 2006
- *Canis familiaris*, CanFam 2.0
- *Mus musculus*, mm5, May 2004
- *Petromyzon marinus*, 3.0
- *Drosophila melanogaster*, BDGP Release 5
- *Caenorhabditis elegans*, WS183

Supplemental figures

Fig. 1. **Annotation statistics and coding potentials.**
(A) The figure illustrates the amount of annotated loci for each annotation approach. Overall, 637 loci can be annotated. Obviously, one locus can be annotated by several methods. This overlap between the diagrams pieces is not further illustrated here. As expected, because of the repeat-masked input, `tRNAscan` produces only two more tRNA predictions and identifies one pseudogene.
(B) The plot illustrates the distribution of coding potential scores obtained by the `Coding Potential Calculator CPC` [1]. The score actually represents the score of the programs underlying SVM (Support Vector Machine) which is the "distance" to the SVM classification hyper-plane in the feature space (the farther away the score is from zero, the more reliable the prediction is). Thereby, values smaller than -1 indicate non-coding and scores greater than $+1$ indicate coding elements. As expected, the majority is scored as non-coding.

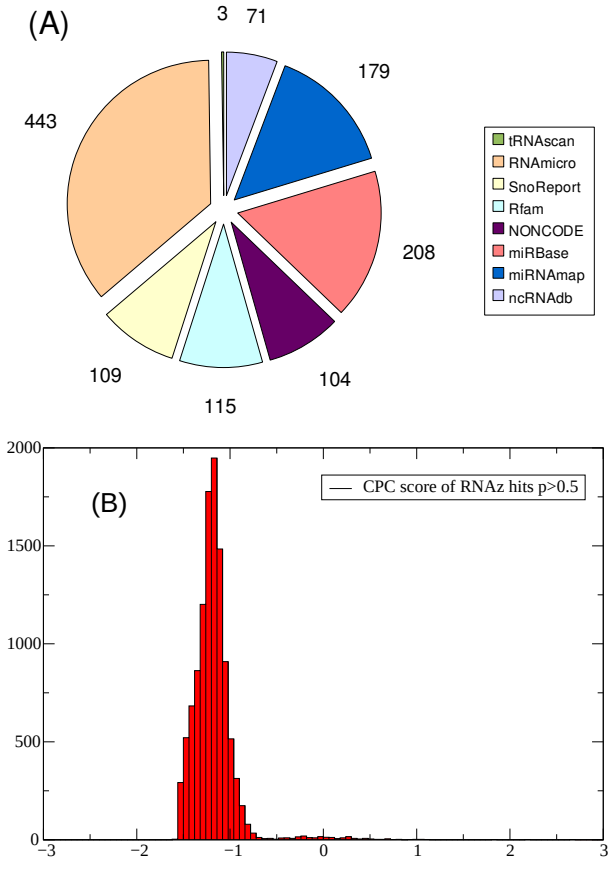
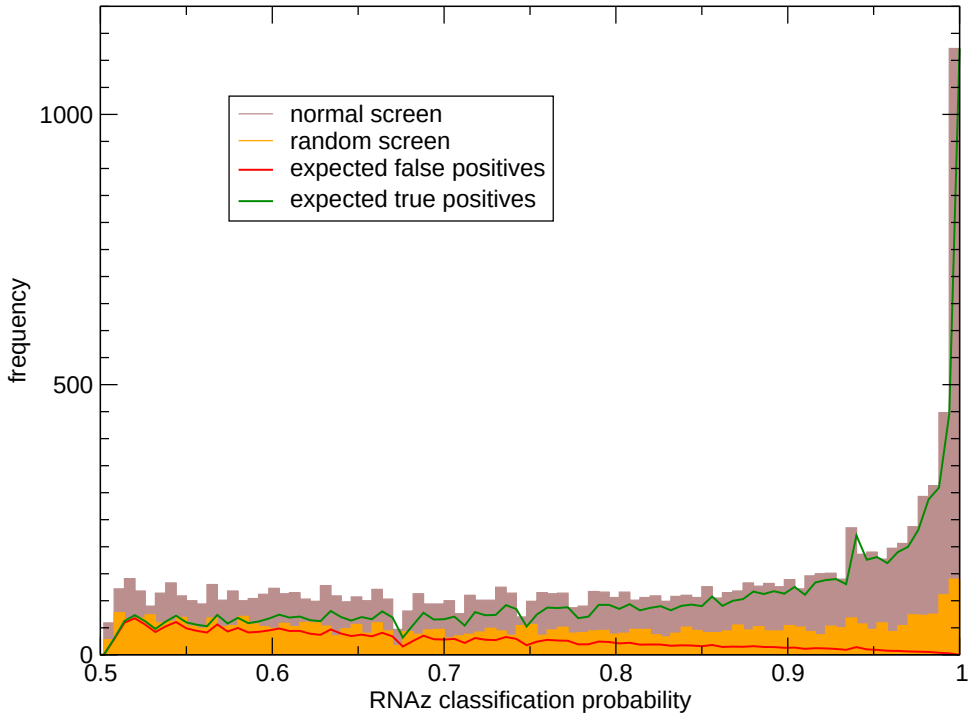


Fig. 2. **Histogram of the `RNAz` classification probability.**
The figure presents the distribution of p values obtained by `RNAz` for the normal and the random-control screen. 100% true positives are expected at p=1.0. Obviously, only a few loci display high p values in the control screen.



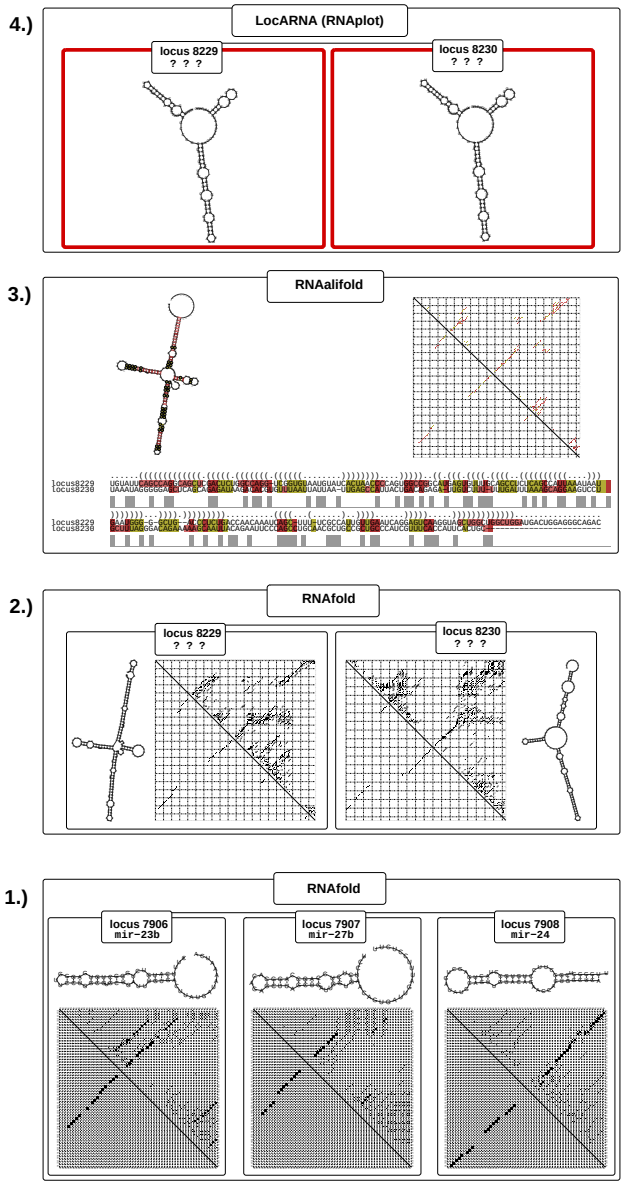


Fig. 3. **Benefits of local alignment and folding algorithms. 1.)** A well known miRNA cluster is smoothly recovered by application of the established `RNAz` and `RNAfold` algorithms. **2.)** Two exemplary `RNAz` hits that lack a valid functional assignment. `RNAfold` does not reveal significant common structural features. **3.)** The sequences poorly align using `ClustalW`, but at least `RNAalifold` suggests a stable consensus structure. **4.)** Intriguingly, folding the sequence/structure alignment obtained by `LocARNA` reveals structural identity of both loci. Locus IDs belong to a prior internal teleosten screen, they do not correspond with the screen of the main paper.

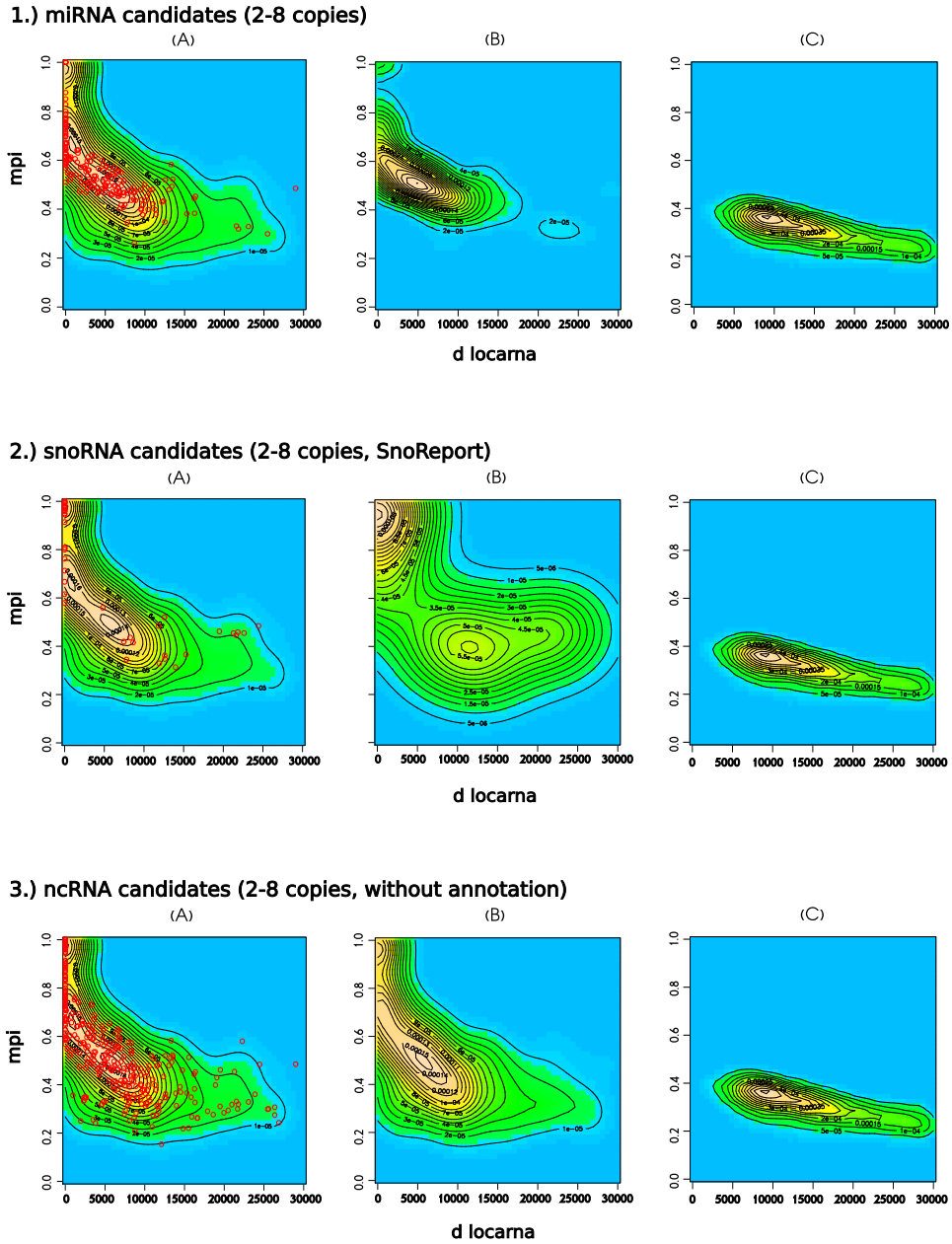


Fig. 4. **Density plot of duplicated ncRNA candidates considering pairwise `LocARNA` distances and mean pairwise identity (mpi).**
(1) Contour plot of microRNA-annotated `RNAz` hits.
(2) Contour plot of snoRNA-annotated `RNAz` hits.
(3) Contour plot of `RNAz` hits that are not annotatable.
(A) Density of all pairs with red circles marking a specific ncRNA class of (1), (2), or (3).
(B) Density of pairs restricted to the specific ncRNA class.
(C) Background distribution.

Supplemental tables

Copy numbers of housekeeping RNAs in teleosts and tetrapods.

The copy number of the housekeeping RNAs in fugu does not provide information on the fate of these ncRNAs after the teleostean duplication because of the high variation between relatively closely related species. Ribosomal RNAs were searched with `blastn`, tRNAs we retrieved using `tRNAscan-SE`, snRNA data are extracted from an unpublished manuscript by M.Marz, T.Kirsten, P.F. Stadler. The numbers are based on sequence similarity and analysis of promoter structures.

Table 1. rRNAs

	5S_rRNA	5.8S_rRNA	28S_rRNA	18S
fugu	87	0	2	1
danio	4154	7	11	10
human	574	6	19	4
mus	202	2	13	4
gallus	10	2	12	1
xeno	254	6	15	9

Table 2. tRNAs

Genomes	fugu	human	mouse	chicken	xenopus	platypus
tRNAs decoding Standard 20 AA:	584	505	2856	195	2688	708
Selenocysteine tRNAs (TCA):	1	3	5	1	3	678
Possible suppressor tRNAs (CTA,TTA):	0	3	2	0	4	2
tRNAs with undetermined/unknown isotypes:	10	3	425	2	12	447
Predicted pseudogenes:	121	109	22976	5	187	83688
Total tRNAs:	716	623	26264	203	2894	85523

Table 3. snRNAs, taken from [2].

	total	U1	U2	U4	U5	U6	U11	U12	U4atac	U6atac
fugu	27/62	5/5-10	5/9-15	3/3-8	6/0-18	4/5-6	1/1	1/1	1/1	1/2
human	31/?	8/10-35	3/2-17	2/2-16	5/2-7	7/24-?	1/1-8	1/2-3	3/1-16	1/2-8
mouse	33/112	7/5-29	6/4-19	1/2-5	6/2-11	7/33	1/1-5	2/1-3	1/0-1	2/0-6
chicken	13/46	1/3-5	1/0-9	1/2-4	2/6-13	4/4-6	1/1-2	1/1	1/1	1/3-5
xenopus	21/248	5/11-66	1/5-39	3/2-41	2/31-63	5/12-22	1/1	1/4-9	1/1-5	2/2
platypus	23/168	5/3-59	2/4-10	2/3	4/0-6	6/6-49	1/1-2	1/2	1/1	1/3-36

Table 4. Structural distances of duplicated ncRNA candidates grow with higher copy-number.

candidate copy-number	2	3	4	5	6	7	8
avg `LocARNA`-distance	5 409	11 635	15 653	15 533	17 619	19 524	18 321

References

1. L. Kong, Y. Zhang, Z. Ye, X. Liu, S. Zhao, L. Wei, G. Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine., Nucleic Acids Res 35 (2007) W345–W349.
2. M. Marz, T. Kirsten, P. F. Stadler, Molecular Evolution of Spliceosomal snRNA Genes in Metazoan Animals, In Preparation.