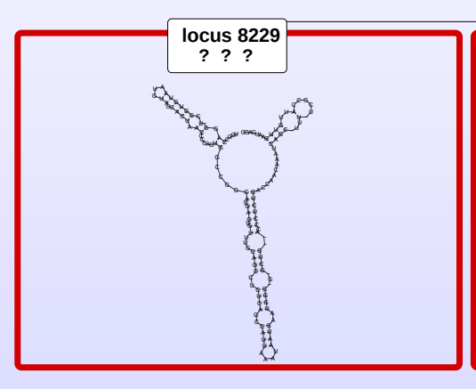
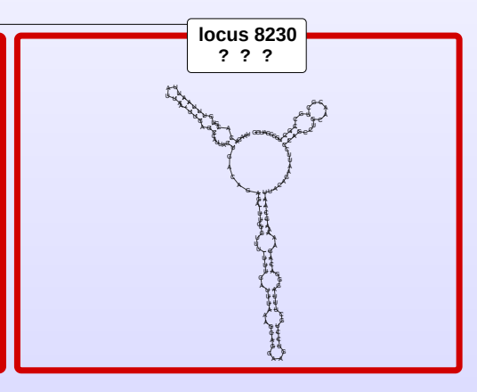
Duplicated RNA Genes In Teleost Fish Genomes

Dominic Rose*[a], Julian Jöris[a], Jörg Hackermüller[b], Kristin Reiche[a], Qiang LI[c], and Peter F. Stadler*[a,d,e]

[a] Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany
Email: {dominic, stadler}@bioinf.uni-leipzig.de
WWW: http://www.bioinf.uni-leipzig.de
[b] Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany
[c] T-Life Research Center, Fudan University, Shanghai, China
[d] Department of Theoretical Chemistry, University of Vienna, Austria
[e] The Santa Fe Institute, USA

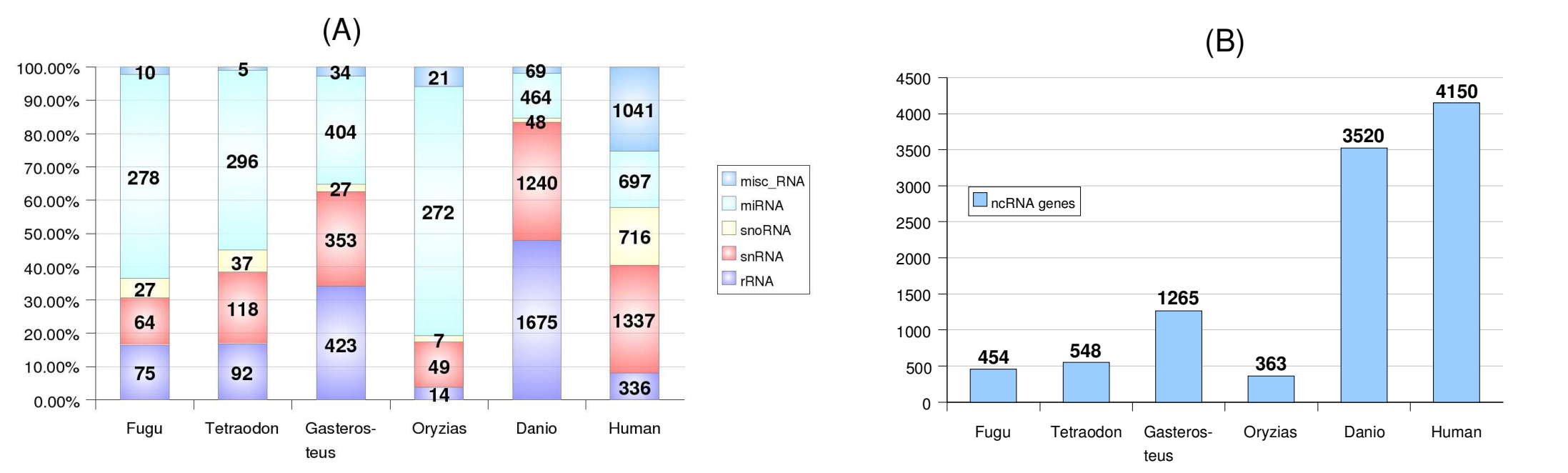
UNIVERSITÄT LEIPZIG





1. Introduction

- Recently, non-protein-coding RNAs (ncRNAs) have moved from a biochemical curiosity to a main research topic in molecular biology.
- High-throughput transcriptomics have established that ncRNAs in fact dominate the transcriptome and are implicated in a plethora of regulatory roles.
- Massive differences in coverage and biases in annotation between fairly closely related organisms.
- The list of non-protein coding RNAs seems still largely incomplete.



ENSEMBL 45 annotation of human and teleost fish genomes.

- Why teleosts? They have undergone a complete genome duplication (*Fish Specific Genome Duplication*), just before the radiation of the crown-group teleosts.
- Purpose of this contribution:
  (1) Revealing undescribed, clade-specific, non-coding RNAs.
  (2) Analysing the fate of ncRNAs in the wake of genome duplications.

2. Prediction of Structured RNAs

- Applied genomes:
  Takifugu rubripes, fugu, Tr
  Tetraodon nigroviridis, pufferfish, Tn
  Gasterosteus aculeatus, stickleback, Ga
  Oryzias latipes, medaka, Ol
  Danio rerio, zebrafish, Dr
- NcDNAlign [1] is used to build genome-wide alignments of non-coding regions of the five teleosts.
- RNAz [2] is applied to predict structured ncRNAs in these alignments.

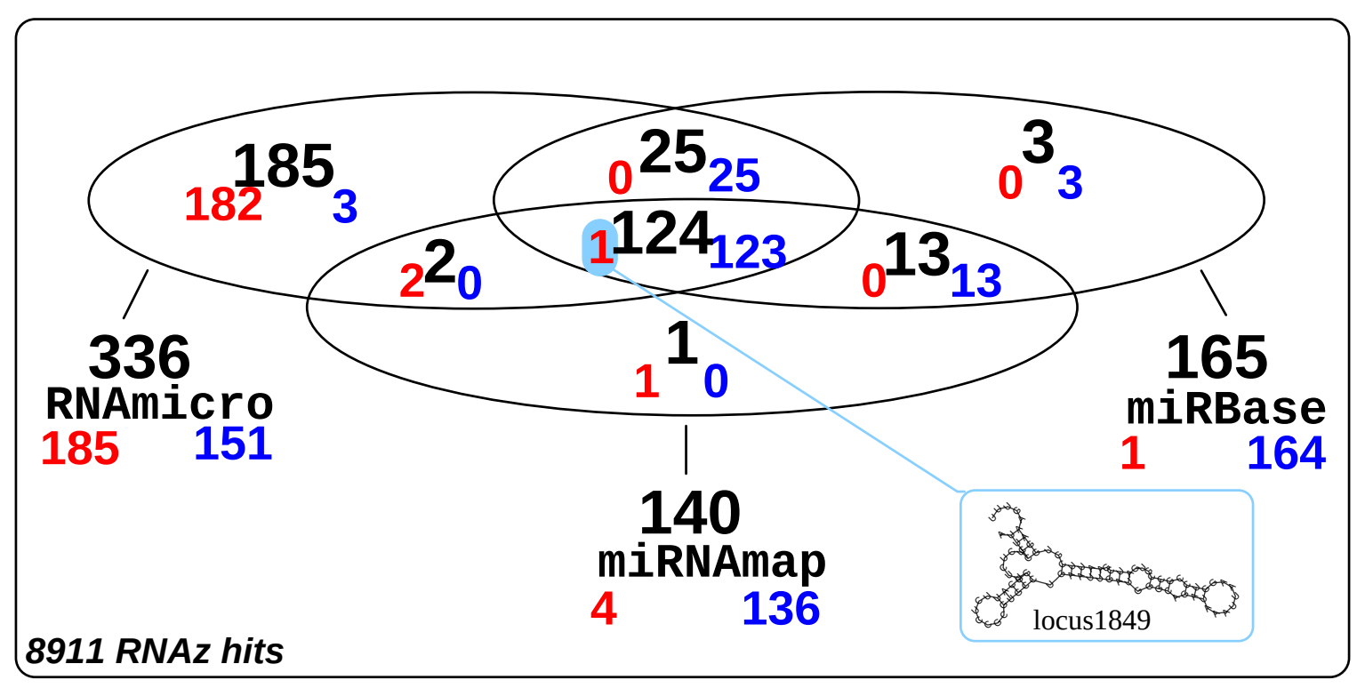
Species	Tr	Tn	Ga	Ol	Dr
genome size [Mb]	393	342	447	700	1 763
without CDS [Mb]	244	263	199	596	969
non-coding alignments	54 115	41 856	37 713	36 602	7 089
aligned DNA [Mb]	8.45	6.55	5.63	5.44	0.96
scored by RNAz [Mb]	8.17	5.84	5.41	4.99	0.79
RNAz p>0.5	8 911	6 948	6 210	6 321	954
[Kb]	1 117	876	772	790	116
RNAz p>0.9	3 551	2 766	2 388	2 558	468
[Kb]	471	328	277	299	54
FDR p>0.5, sequence	35.9	37.4	37.5	37.4	40.7
FDR p>0.9, sequence	23.0	26.3	27.4	25.0	23.5
FDR p>0.5, windows	26.4	26.8	26.9	25.8	24.6
FDR p>0.9, windows	25.7	17.1	17.7	15.6	11.9

Summary of the RNAz screen.

3. Sensitivity

class	RNAz	input	Ensembl	sensitivity (%)
rRNAs	55	75	75	73
snRNAs	42	64	64	66
snoRNAs	9	26	27	35
miRNAs	252	255	278	99
other	8	9	10	89
all	366	429	454	85

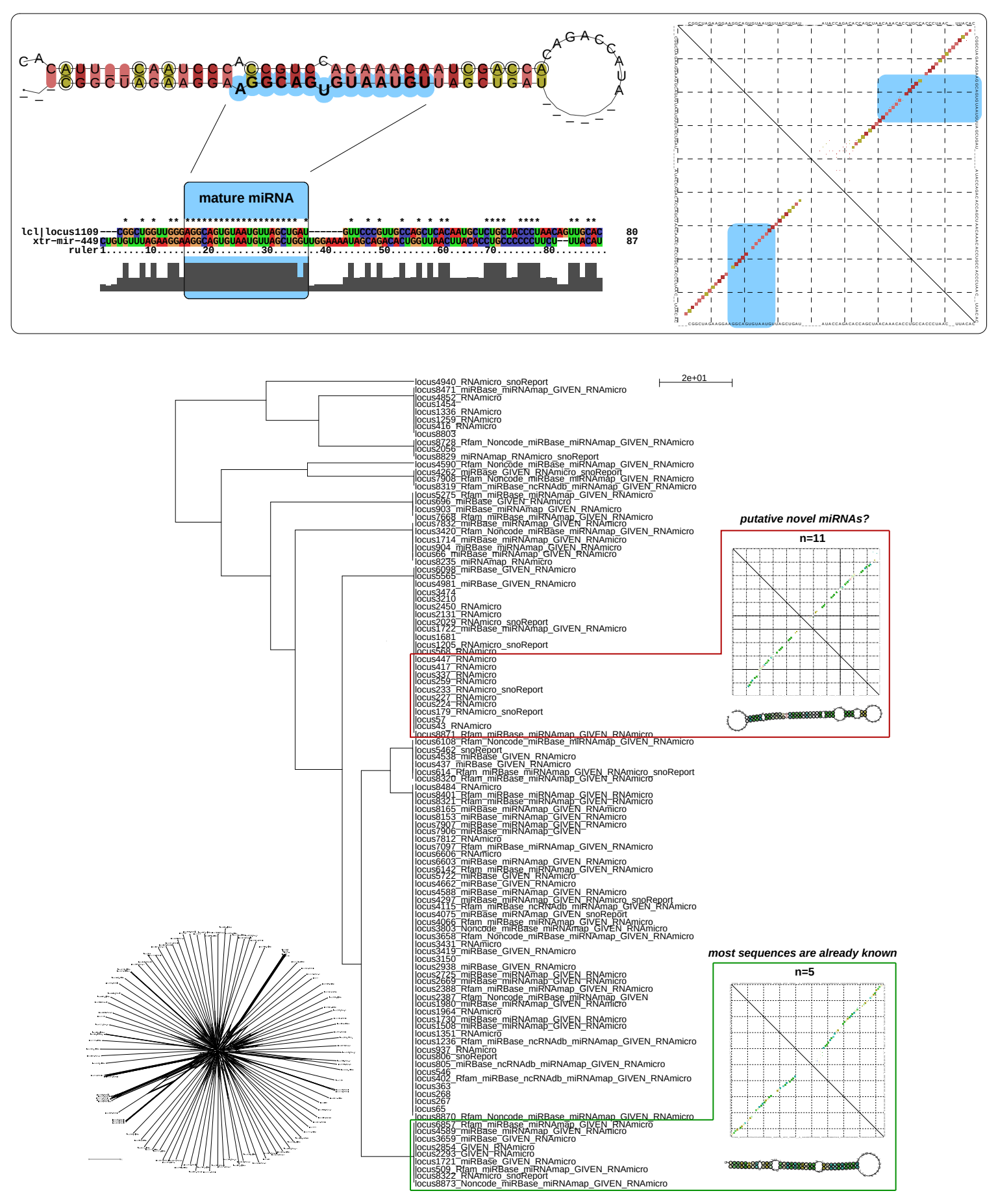
Sensitivity of RNAz on the annotated fugu ncRNAs. Denoted percentages refer to the number of recovered ncRNAs using RNAz over the number of ncRNAs present in the input alignments.



Comparison: RNAmicro vs. miRBase vs. miRNAmap; As an example, a *mir-124* homolog (Locus 1849) is indicated; legend: ${}_{y}x_{z}$ with $y + z = x$

x: number of microRNA precursors in the intersection
y: number of miRNAs *not* annotated in Ensembl
z: number of corresponding miRNAs annotated in Ensembl

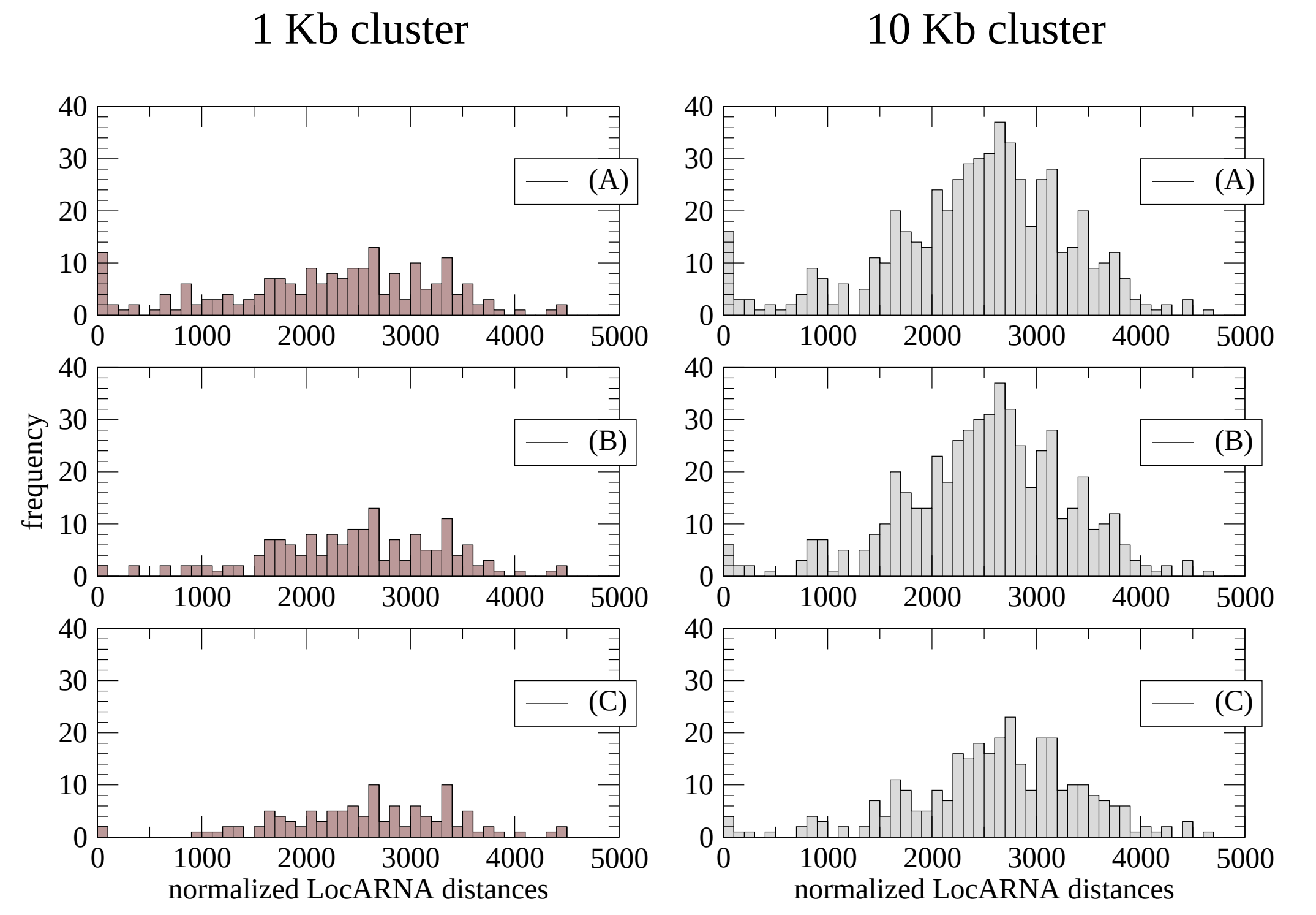
4. Novel microRNAs



Top: *fru-miRNA-449*, homolog of a known miRNA not annotated in Ensembl, contains mature miRNA *xtr-miR-449*.
Below: Structure-based clustering reveals a collection of 108 hairpin structures (61 miRBase entries, 32 RNAmicro predictions).

5. Genomic Clusters

Altuvia *et al.*[3]: miRNA precursors occur in close vicinity (<3 000 nt)



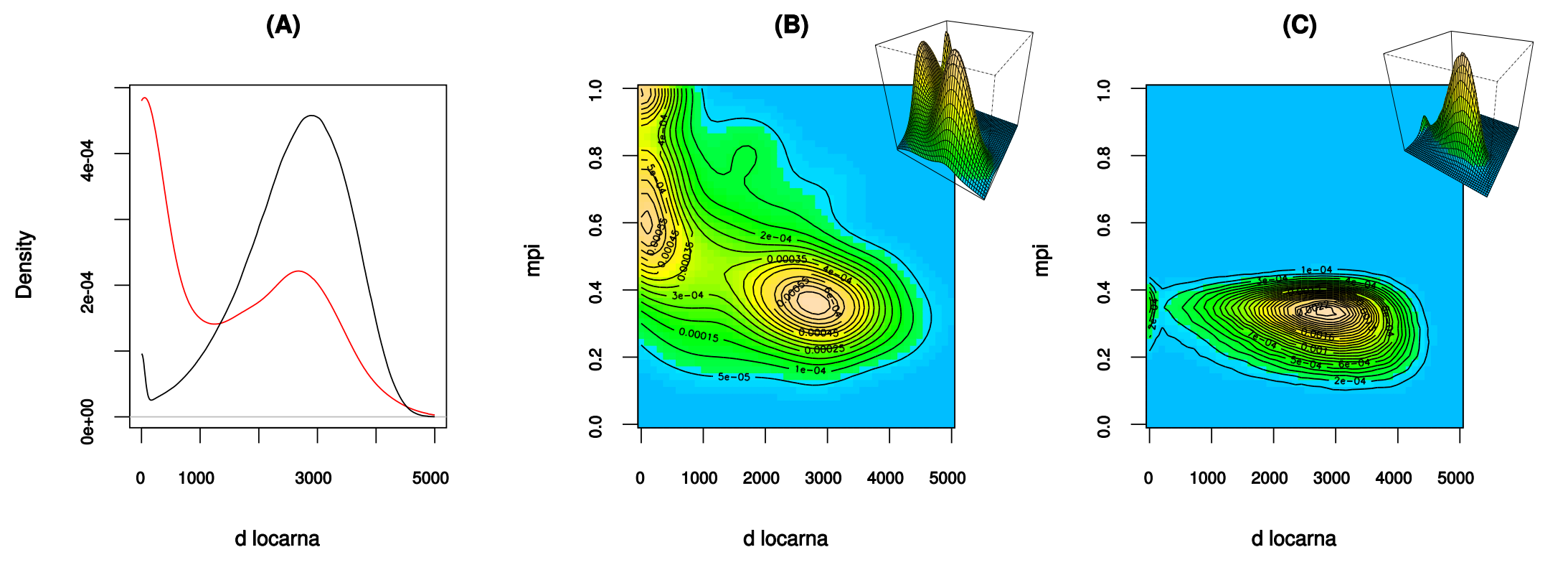
Distribution of structure distances between pairs of adjacent high confidence ($p_{\text{RNAz}} > 0.9$) RNAz hits with a maximum distance of 1 kb and 10 kb, resp.
(A): all genomic clusters of RNAz predictions
(B): restricted to the clusters containing at least one unannotated signal
(C): completely unannotated clusters only.

6. Paralogs

Distribution of paralogs of RNAz hits in the fugu genome: Paralogous groups are determined by means of BLAST **(A)** or by re-aligning and evaluating the sequences with the NcDNAlign/RNAz pipeline **(B)**.

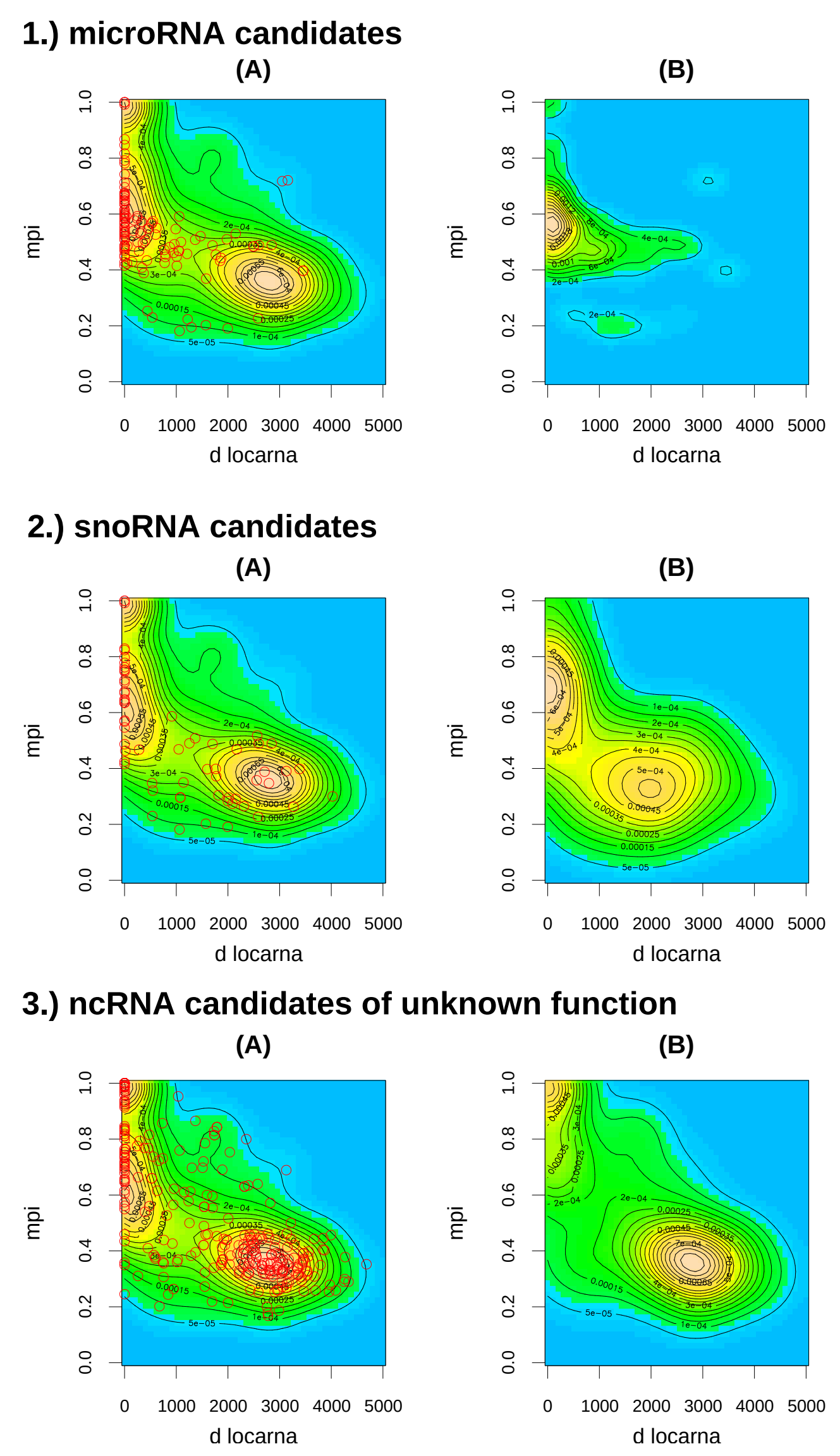
	# copies	1	2	3	4	5	6	7	8	>= 8
	p>0.9	2 955	203	68	33	30	21	15	18	208
	annotated	515	85	31	12	11	1	0	0	53
	unknown	2 440	118	37	21	19	20	15	18	155
(A)	teleost-specific	2 656	130	47	26	19	21	15	18	208
	in tetrapods	299	73	21	7	11	0	0	0	0
(B)	teleost-specific	2 789	130	47	25	21	21	14	18	207
	in tetrapods	166	73	21	8	9	0	1	0	1

Density plot of the distribution of all pairwise LocARNA distances of putatively duplicated pairs with recognizable sequence similarity:



A *(red curve)*, B: duplicated pairs; A *(black curve)*, C: background.

7. Distribution of structure distances



8. Conclusions

- Unbiased survey for evolutionary conserved structured ncRNAs in teleost fish genomes
- Evidence for several thousand structured RNA motifs
- Only a small fraction can be annotated
- Strong evidence for the existence of previously undescribed structurally defined ncRNA families from structure-based clustering
- Very few ncRNAs retain recognizable duplicates
- Large scale duplication events do not lead to a corresponding increase in the ncRNA repertoire (at least as far as RNAs are concerned that depend on a well-defined structure.)
- One immediate implication is that comparative approaches within the same genome, i.e., comparisons between paralogous regions, will have very limited sensitivity at least for ncRNA discovery.

Acknowledgements

. . . for contributing:

- Julian Jöris, Jörg Hackermüller, Kristin Reiche, Qiang LI, Peter F. Stadler

· · · for funding:

- German DFG Bioinformatics Initiative BIZ-6/1-2
- 6th Framework Programme of the European Union (SYNLET)
- “Vereinigung von Förderern und Freunden der Universität Leipzig e.V.”
- Prof. Bailin Hao, Fudan University

References

[1] D. Rose, J. Hertel, K. Reiche, P. F. Stadler, J. Hackermüller, NcDNAlign: Plausible Multiple Alignments of Non-Protein-Coding Genomic Sequences, Submitted.

[2] S. Washietl, I. L. Hofacker, P. F. Stadler, Fast and reliable prediction of noncoding RNAs, Proc. Natl. Acad. Sci. USA 102 (2005) 2454–2459.

[3] Y. Altuvia, P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M. J. Brownstein, T. Tuschl, H. Margalit, Clustering and conservation patterns of human microRNAs., Nucleic Acids Res 33 (2005) 2697–2706.

Gedruckt im Universitätsrechenzentrum Leipzig