Evolution of the Vertebrate Y RNA Cluster

Axel Mosig[a,b], Meng Guofeng[a], Bärbel M. R. Stadler[b],
Peter F. Stadler[c,d,e,f,g,∗],

[a] *Department of Combinatorics and Geometry (DCG),*
MPG/CAS Partner Institute for Computational Biology (PICB),
Shanghai Institutes for Biological Sciences (SIBS) Campus, Shanghai, China

[b] *Max Planck Insitute for Mathematics in the Sciences,*
Inselstrasse 22, D-04103 Leipzig, Germany

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

[d] *Interdisciplinary Center for Bioinformatics, University of Leipzig,*
Härtelstraße 16-18, D-04107 Leipzig, Germany

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

[f] *RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie — IZI*
Deutscher Platz 5e, D-04103 Leipzig, Germany

[g] *Santa Fe Institute,*
1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Abstract

Relatively little is known about the evolutionary histories of most classes of non-protein coding RNAs. Here we consider Y RNAs, a relatively rarely studied group of related pol-III transcripts. A single cluster of functional genes is preserved throughout tetrapod evolution, which however exhibits clade-specific tandem duplications, gene-losses, and rearrangements.

Key words: Y RNA, non-coding RNAs, evolution, gene duplications

1 Introduction

Y RNAs were discovered as the RNA component of the Ro RNP particle a quarter of a century ago [12]. Y RNAs form a small family of short polymerase III transcripts [18, 14, 4, 13] that are located in close proximity in the genomes of human, mouse and xenopus. The molecules exhibit a characterstic secondary structure that has been extensively studied in the past both computationally [5] and by means of chemical probing [26]. The conserved structural features are required for binding the *Ro60* and *La* proteins [24, 8] and nuclear export [22]. A recent study [3] demonstrates a direct role of Y RNAs for DNA replication.

Although the conservation of Y RNAs among mammals has long been recognized [21, 6], there is to-date no systematic study of their molecular evolution. In addition to vertebrates, Y RNAs so far have been reported only in the nematode *C. elegans* [28] and the procaryote *Deinococcus radiodurans* [2]. In primates, Y RNAs are the founders of a family of about 1000 pseudogenes that constitute a class of L1-dependent non-autonomous retroelements [19]. In contrast, in almost all other species (with the notable exception of the guinea pig *Cavia procellus*) there are only a few Y-RNA derived pseudogenes.

In this short contribution we describe a comprehensive analysis of Y RNA evolution in vertebrates based on the currently available genomic data including un-assembled shot-gun traces.

2 Materials and Methods

The genomes in the `ENSEMBL` database (release 41) were searched with `blast` and `ssaha2` for orthologs of the human and xenopus Y RNAs compiled in the `Rfam` database. The resulting sequences were then used to search the vertebrate sequences contained in the NCBI trace archive for additional orthologs. The sequences were then aligned using both `clustalw` [27] and `dialign` [15], resulting in four unambiguously discernible paralog groups corresponding to the four classes of human Y-RNA genes (Y1, Y3, Y4, Y5) for all amniote species. Of the four Xenopus sequences, xY3 and xY4 can be unambiguously assigned to the Y3 and Y4 paralog groups, respectively. For fully or partially assembled genomes we furthermore recorded the genomic locations.

These alignments were then used to derive patterns for `fragrep` [16], a tool that searches genomic DNA for short highly conserved patterns with intervening sequences of variable length. We used here an improved version that matches position-specific weight matrices (PWMs) rather than exact sequence patterns against the genomic DNA. We used `fragrep` to search the genomes of teleost fishes as well as non-vertebrate deuterostomes for Y RNA homologs. Alignments of the members of the individual paralog groups are provided as electronic supplement. [1]

[1] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-004/`

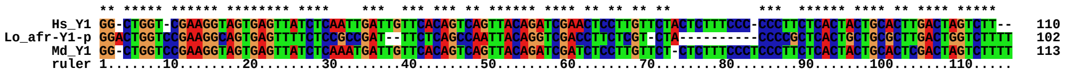

Fig. 1. In the elephant (*Loxodonta africana* genome, a divergent Y1 gene is located at positions 14960-15059 of scaffold 17199 (BROADE1) only 700nt downstream of the Y3 gene on the opposite strand. This is not unexpected, given that the distance between Y5 and Y4 is also reduced to less than 7kb. Compared to both the human and the Opossum Y1 genes, this locus is highly derived, exhibiting a substantial deletion in the loop regions.

The final Y RNA dataset was then aligned using `dialign` with a small number of manual adjustments regarding mostly the incomplete trout sequence and the candidate from the stickleback. We opt here for a block-based, essentially local, alignment method since Y RNAs from different paralog groups differ by substantial in/dels. As a consequence, dynamic programming algorithms such as `clustalw` tend to preferentially cluster sequences of similar lenghts. In particular, `clustalw` fails to correctly align the stem region of the Y5 RNAs with the stem regions of the other groups. We used both the neighbor-joining and the neighbor-net algorithms as implemented in the `Splitstree` package [10] to assess the phylogenetic relationships.

In addition to the usual bootstrapping procedure we also use the following approach: Given a partition of the sequences into K credible monophyletic groups (in this case the tetrapod Y1, Y3, Y4, and Y5 genes, the teleost Y genes, and the single candidate sequence from branchiostoma), we construct a `dialign` alignment from each of the K groups and then compute the neighbor-joining (NJ) tree from this alignment. We use the strict consensus method as implemented in the `phylip` package [7] to compute the strict consensus tree and the frequencies of all splits that arise in the individual NJ trees.

3 Results

In Eutherian mammals we find the full complement of four Y RNAs in all major clades. The archetypic eutherian Y RNA cluster has the form Y5-Y4-Y3-$\overline{\text{Y1}}$, where Y1 is transcribed from the minus strand. Intergenic distances are rather well-conserved, randing from 17-32kb between Y5 and Y4, 6-22kb between Y4 and Y3, and only 3-6.5kb between Y3 and Y1.

Loss of members of the Y RNA family seems to be a fairly frequent phenomenon. In both rat and mouse, there is no trace of either Y5 or Y4, a fact that has been noticed already in previous studies [21]. In the closely related squirrel genome, on the other hand, Y4 is still unchanged, while only a divergent, probably pseudogenized, copy of Y5 has been found. In contrast, all four Y RNAs appear to be intact in *Dipodomys ordii*, the forth representative of the Sciurognathi. In all the three Cetartiodactyla (cow, pig, and bottlenose dolphin), no trace of a Y5 gene was found.

The afrotherian genomes (elephant, tenrec, and procavia) contain highly divergent Y1 sequences, which, judging from the increase rate of evolution, is likely to be a pseudogene, Fig. 1.

An interesting feature of the eutherian Y RNA cluster is the reverse orientation of the Y1 gene, Fig. 2. This character is not shared by both Platypus and Chicken. In these two species, as well as in the frog, all Y RNAs are colinear. In contrast, the opossum genome exhibits a different rearrangement. This suggests that platypus represents the ancestral state in amniota.

In xenopus, we find that xYa, xY4 and Y4 are linked with EZH-2 and CUL-1 as in amniotes. The Y3 gene is located on the same scaffold. It is however, separated by 270kb and 4 protein-coding genes.

In teleosts, only a single Y-RNA homolog is detectable [16]. The same loci have also been identified in ENSEMBL release 42 using `infernal` [17] in fugu, tetraodon, medaka, and zebrafish. In zebrafish, medaka, and stickleback, the Y RNA gene is located on the same chromosome as homologs of CUL-1, EZH2, and PDIA4, although not in close proximity to these genes.

A single Y RNA candidate was identified in the genome of the lancelet *Branchiostoma floridae*. Fig. 3 shows that the sequence has plausible homology with the human Y-RNAs, in particular at the ends, i.e., in the characteristic stem region [5, 26]. Using `RNAfold`, one finds that the sequence can form a stem-loop structure as a suboptimal structure just about 2kcal above a V-shaped groundstate structure. A search across the genomes of the urochordates *Ciona intestinalis* and *Ciona savignyi* and of the echinoderm *Strongylocentrotus purpuratus* was not successful.

The phylogenetic analysis of the Y RNA sequences is complicated by the short sequence length (between 74 and 114bp) and by the divergence of the paralog groups outside the very well-conserved stem-region. The pattern of in/dels strongly suggests that the tetrapod Y1, Y3, Y4, and Y5 groups are monophyletic. The xY5 and xYa sequences are clearly recent paralogs and are both consistenly placed within the Y5 group by both visual inspection of alignments and the phylogenetic analysis in Fig. 4. The bootstrap support for monophyletic Y1, Y3, Y4, and Y5 groups increases if the lancelet candidate is excluded from the analysis (data not shown). The support for a monophyletic clade of teleost Y RNA increases to more than 30% when the partial trout sequence is removed from the dataset.

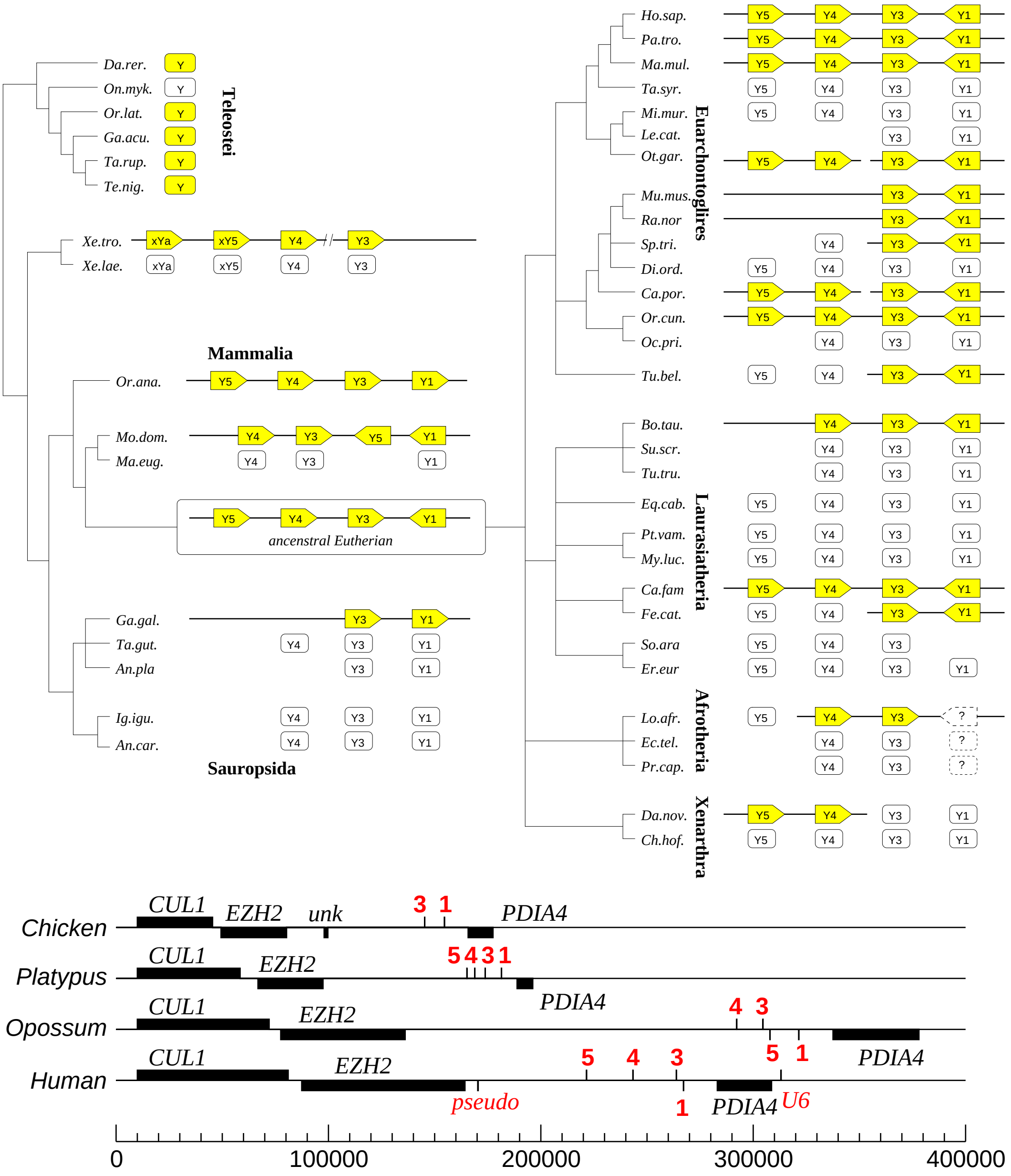

Fig. 2. Evolution of the vertebrate Y locus.
(Top) schematic overview. With the exception of Xenopus, the functional Y RNA genes are located in a single cluster in all sufficiently assembled genomes (symbols with arrows on a line marking an uninterrupted piece of genomic DNA). For most species, only short scaffolds or shotgun traces are available (white symbols without direction).
(Below) Comparison of Amniote Y RNA loci drawn to scale together with their flanking genes.

In addition to a classical NJ tree, Fig. 4a, we also computed a phylogenetic network to highlight the relatively large noise level in our data. Despite the rather high noise level, Fig. 4b also shows well-separated clades Y1, Y3, Y4, Y5, and teleost-Y. In order to assess the branching order of the major groups

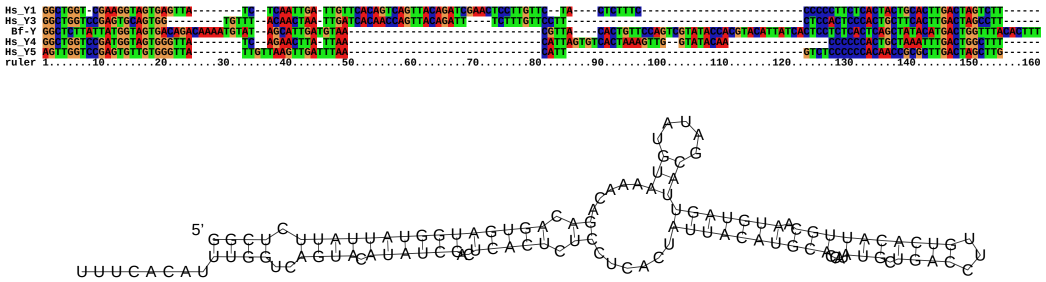

Fig. 3. A single candidate Y RNA sequence was found in the genome of *Branchiostoma floridae* using `fragrep`. The top panel shows an unedited `dialign` alignment of the lancet Y RNA with the four human Y RNA sequences. The panel below shows the predicted RNA secondary structure which was obtained using the most exterior base pair as a constraint. This suboptimal structure is about 2kcal/mol above the ground state.

we constructed individual `dialign` alignments containing one member of each clade as detailed in the Methods section. Fig. 4c displays the 7 most frequently supported splits. The three best supported ones define the strict consensus tree that places the teleost Y RNAs as sister group to a (Y1,Y3) subgroup of tetrapods. The grouping of teleost-Y with tetrapod-Y3 is in the NJ is very poorly supported; in the light of the split analysis, it is most likely just noise.

4 Discussion

Our analysis can be summarized by the following scenario, which is consistent with and at least weakly supported by all data. Starting with a single ancestral Y RNA a first tandem duplication produced pre-(Y4,Y5) and pre-(Y1,Y3) in this order on the genomic DNA. Presumably this event happened before the split of actinopterygians and sarcopterygians. The pre-(Y4,Y5) appears to have been lost in extant teleosts. Note that this first step ist the one with the least direct support.

In the tetrapod ancestor, pre-(Y4,Y5) and pre-(Y1,Y3) then gave rise to the ancestral arrangement Y5→Y4→Y3→Y1. An additional duplication of Y5 occured in the xenopus lineage. Loss of Y RNAs is an abundant phaenomenon: Y1 was lost in xenopus, Y4 and Y5 was repeatedly lost in various amniote lineages (see Fig. 2). In mammals we furthermore observe different rearrangements of the Y RNA locus in both metatheria and in the eutherian ancestor.

As is the case with other non-coding RNAs, in particular pol-III transcripts, Y RNAs give rise to extensive families of (retro)pseudogenes [19]. In contrast to other classes, however, the locus of the functional originals has remained surprisingly stable. Despite frequent retrotransposition events and the fact

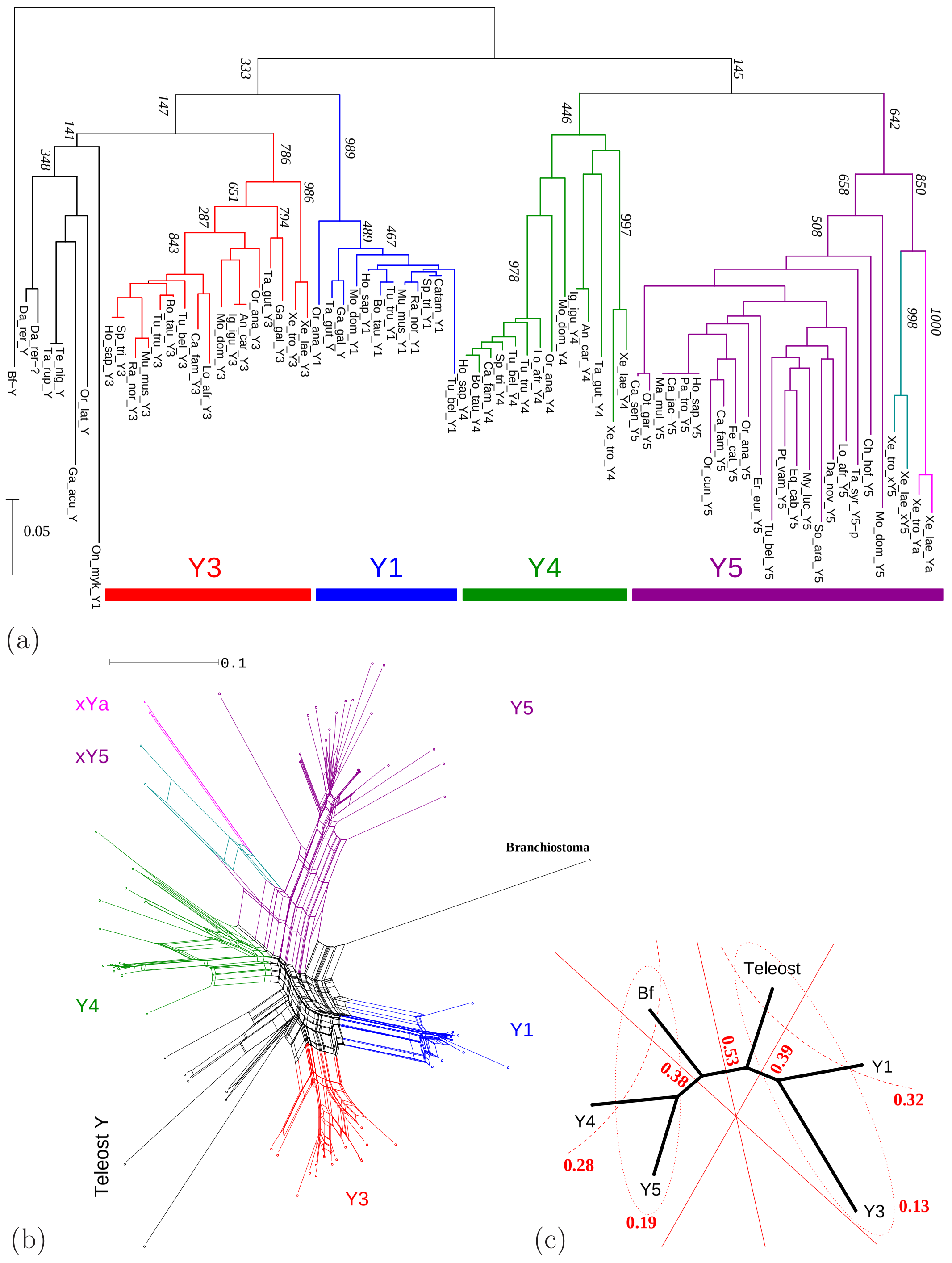

Fig. 4. Phylogenetic Analysis of Y RNA sequences.
(a) Neighbor-joining tree constructed from a `dialign` alignment
(b) Neighbor-net based on the same`dialign` alignment.
(c) Most prominent splits derived from individual `dialign` alignments containing one sequence from each of the six groups.
The data are consistent with the existence of four distinctive paralog groups (Y1, Y3, Y4, and Y5) in tetrapoda (with an additional tandem duplication of Y5 group in xenopus leading to the xY5 and xYa subgroups).

that the Y-RNA genes are independent pol-III transcripts, the functional Y RNA genes have remained within single uninterrupted clusters (with the exception of the situation in *Xenopus tropicalis*) throughout the evolution of tetrapods. In this respect, Y RNAs appear to behave differently compared other non-coding RNAs whose evolution has been studied in some details, including microRNAs [25, 9, 23, 20] and snoRNAs [1, 29, 30]. In microRNAs, preservation of genomic clustering is easily explained by the fact that the clustered miRNAs are processed from a single polycistronic primary transcript, see e.g. [11]. Other ncRNAs, in particular tRNAs, snoRNAs, and snRNAs behave more like mobile genetic elements [1, 29].

Acknowledgments

Discussions with Michael M. Müller is gratefully acknowledged. This work has been funded, in part, by the German DFG Bioinformatics Initiative BIZ-6/1-2.

References

[1] A. F. Bompfünewerer, C. Flamm, C. Fried, G. Fritzsch, I. L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Müller, S. J. Prohaska, B. M. R. Stadler, P. F. Stadler, A. Tanzer, S. Washietl, and C. Witwer. Evolutionary patterns of non-coding rnas. *Th. Biosci.*, 123:301–369, 2005.

[2] X. Chen, A. M. Quinn, and S. L. Wolin. Ro ribonucleoproteins contribute to the resistance of *Deinococcus radiodurans* to ultraviolet resistance. *Genes Dev.*, 14:777–782, 2000.

[3] C. P. Christov, T. J. Gardiner, D. Szüts, and T. Krude. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol. Cell. Biol.*, 26:6993–7004, 2006.

[4] A. D. Farris, J. K. Gross, J. S. Hanas, and H. J. B. Genes for murine Y1 and Y3 Ro RNAs have class 3 RNA polymerase III promoter structures and are unlinked on mouse chromosome 6. *Gene*, 174:35–42, 1996.

[5] A. D. Farris, G. Koelsch, G. J. Pruijn, W. J. van Venrooij, and J. B. Harley. Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis. *Nucl. Ac. Res.*, 27:1070–8, 1999.

[6] A. D. Farris, C. A. O'Brien, and J. B. Harley. Y3 is the most conserved small RNA component of Ro ribonucleoprotein complexes in vertebrate species. *Gene*, 154:193–198, 1995.

[7] J. Felsenstein. Phylip – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[8] C. D. Green, K. S. Long, H. Shi, and S. L. Wolin. Binding of the 60-kDa

Ro autoantigen to Y RNAs: evidence for recognition in the major groove of a conserved helix. *RNA*, 4:750–765, 1998.

[9] J. Hertel, M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I. L. Hofacker, P. F. Stadler, and Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. *BMC Genomics*, 7:25, 2006.

[10] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23:254–267, 2006.

[11] Y. Lee, K. Jeon, J. T. Lee, S. Kim, and V. N. Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.*, 21:4663–4670, 2002.

[12] M. R. Lerner, J. A. Boyle, J. A. Hardin, and J. A. Steitz. Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. *Science*, 211:400–402, 1981.

[13] R. Maraia, A. L. Sakulich, E. Brinkmann, and E. D. Green. Gene encoding human Ro-associated autoantigen Y5 RNA. *Nucl. Acids Res.*, 24:3552–3559, 1996.

[14] R. J. Maraia, N. Sasaki-Tozawa, C. T. Driscoll, E. D. Green, and G. J. Darlington. The human Y4 small cytoplasmic RNA gene is controlled by upstream elements and resides on chromosome 7 with all other hY scRNA genes. *Nucl. Acids Res.*, 22:3045–3052, 1994.

[15] B. Morgenstern. `DIALIGN 2`: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.

[16] A. Mosig, K. Sameith, and P. F. Stadler. `fragrep`: Efficient search for fragmented patterns in genomic sequences. *Geno. Prot. Bioinfo.*, 4:56–60, 2005.

[17] E. P. Nawrocki and S. R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comp. Biol.*, 2007. in press.

[18] C. A. O'Brien, K. Margelot, and S. L. Wolin. *Xenopus* Ro ribonucleoproteins: Members of an evolutionarily conserved class of cytoplasmic ribonucleoproteins. *Proc. Natl. Acad. Sci. USA*, 90:7250–7254, 1993.

[19] J. Perreault, J.-F. Noël, F. Brière, B. Cousineau, J.-F. Lucier, J.-P. Perreault, and G. Boire. Retropeudogenes derived from human Ro/SS-A autoantigen-associated hY RNAs. *Nucl. Acids Res.*, 33:2032–2041, 2005.

[20] S. E. Prochnik, D. S. Rokhsar, and A. A. Aboobaker. Evidence for a microRNA expansion in the bilaterian ancestor. *Dev. Genes Evol.*, 217:73–77, 2007.

[21] G. J. M. Prujin, P. A. E. T. M. Wingens, S. L. M. Peters, J. P. H. Thijsen, and W. J. van Venrooij. Ro RNP associated Y RNAs are highly conserved among mammals. *Biochim. Biophys. Acta*, 1216:395–401, 1993.

[22] S. A. Rutjes, E. Lund, A. van der Heijden, C. Grimm, W. J. van Venrooij, and G. J. M. Pruijn. Identification of a novel *cis*-acting RNA element involved in nuclear export of hY rnas. *RNA*, 7:741–752, 2001.

[23] L. F. Sempere, C. N. Cole, M. A. McPeek, and K. J. Peterson. The phy-

logenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zoolog B Mol Dev Evol.*, 306B:575–588, 2006.

[24] F. H. Simons, S. A. Rutjes, W. J. van Venrooij, and G. J. Pruijn. The interactions with Ro60 and La differentially affect nuclear export of hY1 RNA. *RNA*, 2:264–273, 1996.

[25] A. Tanzer and P. F. Stadler. Molecular evolution of a microRNA cluster. *J. Mol. Biol.*, 339:327–335, 2004.

[26] S. W. M. Teunissen, M. J. M. Kruithof, A. D. Farris, J. B. Harley, W. J. van Venrooij, and G. J. M. Pruijn. Conserved features of Y RNAs: a comparison of experimentally derived secondary structures. *Nucl. Acids Res.*, 28:610–619, 2000.

[27] J. D. Thompson, D. G. Higgs, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.

[28] D. J. van Horn, D. Eisenberg, C. A. O'Brien, and S. L. Wolin. *Caenorhabditis elegans* embryos contain only one major species of Ro RNP. *RNA*, 1:293–303, 1995.

[29] M. J. Weber. Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet*, 2(12):e205, 2006.

[30] A. Zemann, A. op de Bekke, M. Kiefmann, J. Brosius, and J. Schmitz. Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Research*, 34:2676–2685, 2006.