Molecular Evolution of a MicroRNA Cluster

Andrea Tanzer^{1,2} and Peter F. Stadler^{1,2*}

¹Lehrstuhl für Bioinformatik
am Institut für Informatik und
Interdisziplinäres Zentrum für
Bioinformatik, Universität
Leipzig, Kreuzstraße 7b
D-04103 Leipzig, Germany

²Institut für Theoretische
Chemie und Molekulare
Strukturbiologie, Universität
Wien, Währingerstraße 17
A-1090 Wien, Austria

Many of the known microRNAs are encoded in polycistronic transcripts. Here, we reconstruct the evolutionary history of the **mir17** microRNA clusters which consist of *miR-17*, *miR-18*, *miR-19a*, *miR-19b*, *miR-20*, *miR-25*, *miR-92*, *miR-93*, *miR-106a*, and *miR-106b*. The history of this cluster is governed by an initial phase of local (tandem) duplications, a series of duplications of entire clusters and subsequent loss of individual microRNAs from the resulting paralogous clusters. The complex history of the **mir17** microRNA family appears to be closely linked to the early evolution of the vertebrate lineage.

© 2004 Elsevier Ltd. All rights reserved.

*Corresponding author

Keywords: microRNA; mir-17; polycistronic; vertebrate evolution

Introduction

MicroRNAs (miRNAs) form a class of non-coding RNA genes whose products are small single-stranded RNAs with a length of about 22 nt. These are involved in the regulation of translation and degradation of mRNAs. We refer to the recent review¹ for a discussion of their functions and mechanisms as well as their history of discovery. Almost 800 microRNAs from different animal and plant species have been reported so far† and a dedicated database, the Rfam miRNA Registry² has been created to collect information about them. Despite recent efforts on microRNA target prediction^{3–7} little is known about the specific functionality of most miRNAs. The expression of miRNAs in animals involves at least two processing steps.⁸ A long primary transcript, called the primary microRNA (pri-miRNA), which may be polycistronic, is processed in the nucleus, yielding one or more hairpin precursor sequences (pre-miRNAs or stRNA, small temporal RNA). These are exported to the cytoplasm by means of the Exportin-5 pathway,⁹ where the mature miRNA is excised by the enzyme complex Dicer. The final destinations of the mature miRNAs are

different ribonucleoproteins (RNPs). RNA induced silencing complex (RISC) causes mRNA degradation or translational retention, whereas no specific function has been determined for miRNPs.¹⁰ Nelson *et al.* suggested the term RISC/miRNP endonuclease for the catalytic activity of RISC and miRNPs, since miRNAs with extensive sequence complementarity to their mRNA targets can trigger RNAi.¹ Furthermore, they propose the terms Mirgonaute and Sirgonaute for Argonaute proteins associated with protein complexes containing miRNAs and siRNAs, respectively. Further experimental verification of the components, functions, and subcellular localization of RISC and miRNPs will be necessary to clarify the mechanism of what might possibly be the same protein/RNA complex.

Many of the known microRNAs appear in clusters on a single polycistronic transcript.^{8,11–13} In this contribution we reconstruct the evolutionary history of the **mir17** cluster and its paralogs in detail.

The human **mir17**⁸ cluster contains six precursor miRNA within about 1 kb on chromosome 13 (Figure 1). In this Figure, *pre-mir-17* is the precursor of both *miR-17* on its 3'-arm¹⁴ and of *miR-91* (*miR-17-5p*) at its 5'-arm.¹¹ The other members of this cluster were already reported:^{11,14} *miR-18*, *miR-19*, *miR-19b*, *miR-20*, and *miR-92*. The **mir17** cluster is of particular interest because the human X-chromosome is known to contain paralogs of some of these micro RNAs: *miR-106* is a homolog of *miR-17*,¹¹ *miR-19b-2* is a homolog of *miR-19b-1*, and *miR-92-2* is a homolog of *miR-92-1*. In fact, these three miRNAs are located on a 1 kb interval

Abbreviations used: miRNAs, microRNAs; pri-miRNA, primary microRNA; RNPs, ribonucleoproteins; RISC, RNA induced silencing complex; DDC, duplication–degeneration–complementation.

E-mail address of the corresponding author: studla@bioinf.uni-leipzig.de

† Rfam database Release 3.0: Jan. 2003, <http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml>

Figure 1. Human *mir17* microRNA clusters. Three paralog groups of microRNA precursors can be identified: *mir-17/18/18X/20/93/106a/106b/93* (yellow), *mir-19a/19b-1/19b-2* (blue) and *mir-92-1/92-2/25* (green). Names of the precursors are written below the boxes. Mammals contain at least three paralog clusters while teleost fishes have at least four copies. The position of the mature miRNA is indicated by a dark box with labels written in italic.

of the X-chromosome that also contains sequences homologous with *pre-mir-18* and *pre-mir-20* (Figure 1).

Homologous microRNAs have been reported also in the mouse: *mmu-miR-18* and *mmu-miR-20* are located on chromosome 14,^{15,16} while *mmu-miR-19b*, *mmu-miR-92-2* and *mmu-miR-106a*¹⁶ are located on the X-chromosome. A rat member of this family is *rno-miR-20*.¹⁷

A miRNA similar to *miR-106a* is reported¹⁶ both as *miR-94* and *miR-106b*. Inspection of the adjacent region of the mouse chromosome 5 shows that *miR-106b* is indeed part of a microRNA cluster that also contains *miR-93*.¹⁶ For both sequences human homologs mapping to chromosome 7 are known as well.¹¹ A third sequence, human *miR-25*,¹⁴ also belongs to this cluster. We find that *miR-25* is distantly related to *miR-92*. Parts of all three clusters are reported by Mourelatos *et al.*¹¹ (Figure 5).

Homologs of *mir-92* are also known in invertebrates. *Drosophila* contains two copies, *miR-92a* and *miR-92b*, which are located within some 5000 nt on chromosome 3R.^{13,18,19}

Results

From here on we use the following notation: here we denote throughout the miRNA precursors by *mir-#* and mature miRNAs by *miR-#* in accordance with the convention in Rfam. Note that in some cases the same mature miRNA is derived from different precursors, for instance, both *hsa-mir-19b-1* and *hsa-mir-19b-2* give rise to *hsa-miR-19b*. For microRNAs listed in Rfam the species is indicated by a three-letter abbreviation: *H. sapiens* (*hsa*), *M. musculus* (*mmu*), *R. norvegicus* (*rno*), *C. elegans* (*cel*) and *D. melanogaster* (*dme*). Since this notation does not indicate the location of paralog microRNAs on their cluster we use shorter designations in Tables and Figures consisting of a two-letter species name, the cluster name, and Rfam number. For example *Hs1-17* = *hsa-mir-17*, *HsX-106a* = *hsa-mir-106a*, and *Hs3-106b* = *hsa-mir-106b* (Figure 1). A complete list of correspondences is provided in the electronic supplement.

Figure 3 summarizes the results from blast

searches: mammals have three paralogs of the *mir17* cluster, teleost fishes contain four copies. Complete or nearly complete sequences of *mir17* clusters have been obtained for *Homo sapiens* (*Hs*), *Mus musculus* (*Mm*), *Rattus norvegicus* (*Rn*), *Pan troglodytes*, *Takifugu rubripes* (*Tr*), *Tetraodon nigroviridis* (*Tn*), and *Danio rerio* (*Dr*). In addition, homologs of partial clusters have been identified in the dog *Canis canis* (*Cc*) and frog *Xenopus tropicalis* (*Xt*) genomes. The 115 microRNA sequences and their locations are listed in the electronic Supplemental Material†.

Figure 4 provides an overview of the *mir17* sequences from different organisms. We can identify five paralog groups of precursor miRNAs whose order in the *mir17* clusters are preserved: 17/106, 18, 19a, 20/93, 19b, and 92/25. Only the 92/25 group has known invertebrate homologs, namely *mir-92a* and *mir-92b* in *Drosophila melanogaster* and *mir-235* in *Caenorhabditis elegans*. In Figure 3 and in the work by Lim *et al.*²⁰ the *C. elegans* microRNA *cel-mir-254* is suggested as relative of the 19a/19b group, the sequences are too distant to be certain about their homology, however.

Both the mosquito *Anopheles gambiae* and the nematode *Caenorhabditis briggsae* have distant relatives of the *mir-92* group, see electronic supplement for details. The mosquito genome also contains two paralogs that are separated by about 20 kb.

We were not able to find unambiguous homologs of members of the *mir17* family in the *Ciona intestinalis* and *Strongylocentrotus purpuratus* genomes. In the light of the extreme conservation of many of the members of the *mir17* family among vertebrates this suggests that their regulatory role(s) are predominantly vertebrate specific.

The *mir17* sequences from mammalian species belong to three clusters. One of them is homologous to the cluster containing the *hsa-mir-17* (*Hs1-17*) sequence, the second one, which contains *hsa-mir-106a* (*HsX-106a*) is located at the X-chromosome in man, mouse, and rat. We label

† URL: <http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/04-002/>

Figure 2. Predicted consensus secondary structure of the **mir17** cluster I-X located at the mammalian X-chromosomes. The computation was performed with RNAfold and alidot from a CLUSTALW alignment as described.³² (a) The five precursor miRNAs appear as well-conserved hairpins in the consensus structure shown in color. (b) The color-coded mountain plot⁴⁰ shows each base-pair between sequence positions i and j as a slab from i to j with a thickness proportional to the probability that the base-pair is formed. Thus, the x -axis is the sequence position in the RNA, the height of a base-pair is the (weighted) number of base-pairs by which it is enclosed, i.e. its distance from the base of structural elements. Sharp peaks thus correspond to stem-loop structures while multi-peaked mountains indicate branchings (so-called multi-branch loops) in the secondary structure. The saturation of the color indicates whether there are sequences in the alignment that cannot form a base-pair so that base-pairs with conflicts in one or two sequences appear pale. The color indicates the consistent and compensatory mutations: a conserved base-pair is shown in red, two types of base-pairs are shown in yellow, three in green. We see some structural ambiguities (pale base-pairs) usually next to interior loops or bulges or at the base of the stem-loop structures. *mir-106a* and *mir-18X* are almost perfectly conserved in sequence, while the structures of *mir-20X*, *mir-19b-2* and *mir-92-2* receive additional support from a number of compensatory mutations. Peaks in the mountain plot (b) correspond to colored precursor miRNAs in (a).

these two clusters as I-1 and I-X, respectively, and collectively refer to them as type-I clusters. They are closely related and their component microRNAs can be easily aligned. The third mammalian cluster, II-3, is a distant relative containing *hsa-mir-93* (Hs3-93). It cannot be aligned well with the type-I clusters; henceforth we will refer to this cluster as type-II. The frog genome also contains orthologs of both the I-1 and II-3 cluster.

A search from homologous microRNA sequences in the three teleost fish genomes revealed four microRNA clusters of which three, I-A, I-B, and I-C are homologous to the mammalian type-I clusters, and a single type-II cluster II-D that is clearly orthologous to the mammalian cluster II-3. To the extent that data are available, paralog clusters are located on different chromosomes. Type-I clusters identified in both, *T. rubripes* and *T. nigroviridis*, are identical in genomic organization and exhibit high similarity in precursor miRNA sequences. Interestingly, the third type-I cluster seems to consist of only two miRNAs, *mir-18* (TrC-18, TnC-18) and *mir-19b* (TrC-19b, TnC-19b). Compared to the human type-II cluster, the type-II cluster in *T. rubripes* has retained *mir-19b* (TrD-19b). However, no type-II cluster has been found in *T. nigroviridis*. We believe that this is

most likely because the genomic sequence is still not complete. The third teleost species investigated, *D. rerio*, contains four clusters (three type-I and one type-II), all incomplete compared to *H. sapiens*, *T. rubripes* and *T. nigroviridis*. The *D. rerio* cluster DrA, located on contig NA10065 in Zv3, for instance, contains a stretch of 607 non-identified bases (N) in the DNA sequence. Adjacent to this region a sequence homologous to the 5' end of *mir-19b* was identified. With ongoing genome sequencing we thus expect that a few additional members of **mir17** clusters will be found.

A neighbor-joining tree, Figure 5(a) summarizes the relationships of the type-I clusters. It follows that the various paralog type-I clusters share a common ancestor that pre-dates the divergence of the sarcopterygian and the actinopterygian lineages. The type-I clusters arose from this ancestor by means of duplications of the entire clusters. The type-II clusters simply reproduce the known species tree (see inset in Figure 5(a)), indicating that they also originated very early in vertebrate evolution.

Clustering all 115 microRNA sequences of the **mir17** family by means of the z -score defined in equation (1) shows that the microRNAs *mir-17*,

Figure 3. *mir17* clusters of several species. Precursor miRNAs are drawn in black, mature miRNAs in blue. Mammals have three paralogs of the *mir17* cluster, whereas teleosts contain four copies. Since the dog and frog genome sequencing projects are still in progress, only fragments of miRNA clusters are available, missing sequence information is indicated by dotted lines. In the *D. rerio* DrA cluster only the 3' end of *mir-19b* was found (gray box).

mir-106a, *mir-106b*, *mir-93*, *mir-20*, and *mir-18* are ancient paralogs. The same holds for *mir-19a*, *mir-19b*, and *mir-119b* as well as *mir-25* and *mir-92* (Figure 5(b)). It follows that the ancestral type-I and type-II clusters themselves arose through a complex sequence of tandem duplications, cluster duplications, as well as loss of microRNAs. A plausible scenario that is consistent with the history of the individual microRNAs in Figures 4 and 5(a) and with the consensus tree of the clusters (Figure 5(b)) is shown in Figure 6.

Discussion

The microRNAs of the *mir17* clusters arose through a complex history of duplication and loss of individual members as well as duplications of entire clusters. It consists of three large groups of seemingly non-homologous microRNAs. Since we did not find homologs of *mir-17* and *mir-19* in invertebrates we suggest that these microRNAs

are vertebrate innovations. In contrast, *mir-92* is homologous to microRNAs from both *D. melanogaster* and *C. elegans*.

We can only speculate how the ancestors of *mir-17* and *mir-19* arose. The stem-loop structures of microRNA precursors are among the most common local structures. In a recent simulation study²¹ all locally stable RNA structures were computed using the RNALfold program²² for the genome of *C. elegans*. More than 65,000 subsequences were found in a single chromosome that are at least as similar to the *cel-mir-1* structure as the other known microRNA precursors in terms of their tree-edit distance²³ to the *cel-mir-1* precursor structure. It seems plausible that a part of the transcript containing the ancestral *mir-92* may have accidentally formed a second hairpin structure that was processed and exported from the nucleus. If such a *de novo* precursor happens to be functional it could have been retained by stabilizing selection, otherwise it would have soon disappeared again as random drift destroys the

Figure 4. WPGMA clustering of all available **mir17** sequences using the z-score defined in equation (1) as similarity score. The main groups are highlighted. Short negative-length branches arise, since the z-scores are computed by means of sampling.

pre-mir-like structural features. This accidental formation of hairpins might be a rather general mechanism for the origin of polycistronic microRNA transcripts. It could explain the large fraction of microRNAs that appear clustered.^{11–13}

The ancestral **mir17** cluster probably contained the ancestors of the *mir-17* group, the *mir-19* group, as well as the ancient *mir-92*. The first detectable duplication event was the split of *mir-18* from the *mir-17* group, preceding the duplication of the entire cluster that created the ancestors of the 17/18/19a/20/19b/92 family (type-I cluster) and the 106b/93/25 family (type-II cluster). WPGMA clustering (Figure 5(b)) shows that the first duplication of the ancestral **mir17** cluster must have been preceded by a complex series of tandem duplications: the cluster dupli-

cation could not have happened before the split of *mir-19a* and *mir-19b* since the *mir-19* paralog in the type-II cluster clearly belongs to the *mir-19b* group. Similarly, *mir-106b* is orthologous to the ancestor of *mir-17* and *mir-20*, or possibly to *mir-20*. The origin of the *mir-93* group thus also predated the cluster duplication. The large evolutionary distance between *mir-92* and *mir-25* is either the result of an enhanced rate of evolution in the *mir-92* group after the divergence of type-I and type-II clusters, or *mir-92* and *mir-25* might have been ancient tandem copies in the ancestral cluster. In the latter case a different paralog was lost in each of the two paralog clusters after the duplication of the ancestral cluster.

Since the genomic data of *X. tropicalis* are still incomplete we cannot decide whether the two

Figure 5. (a) Neighbor-joining tree of the combined *mir17* type-I clusters and type-II clusters (inset). The fragment of the dog CcX cluster has been inserted based on a separate tree reconstruction leaving out the Dr14/pufferfish-C cluster, since the absence of homologous sequences between the dog cluster fragment and some teleost clusters causes missing data in the distance matrix that cannot be dealt with properly by the Phylip package. (b) WPGMA tree of the evolution of the paralog groups in the *mir17* family.

clusters, one type-I and one type-II cluster, represent the complete inventory or whether the frog genome also includes an ortholog of the mammalian X-chromosome cluster. The teleost sequences, unfortunately, are diverged so far that

the duplication history of their three type-I clusters could not be resolved unambiguously. It is very likely, however, that at least the teleost I-A and I-B clusters arose during the teleost-specific genome duplication.²⁴ The homology of Dr14

Figure 6. A plausible scenario for the evolution of the *mir17* family. The inset shows the duplication histories of the three members of the primordial cluster. For details see the text.

with TnC and TrC is based on the cluster organization, the sequences are too short to provide an unambiguous signal. The tree in Figure 5(a) suggests that the teleost I-A and I-C clusters are first order paralogs that subsequently lost some of their microRNAs, thereby resolving the redundancy that arose in the duplication event as predicted by the duplication–degeneration–complementation (DDC) model.²⁵ In brief, the DDC model predicts that after a gene duplication either one of the two paralogs is lost or, if both are retained, then the two paralogs will evolve to perform complementary sub-functions. Thus, we expect loss of some microRNAs subsequent to duplication. The model further predicts that duplicate clusters that still contain redundant microRNAs should differ in their spatio-temporal expression patterns.

It is tempting to speculate that the expansion of the **mir17** family is linked to the origin of vertebrates and, more specifically, is associated with the two genome-wide duplications^{26,27} that also caused the quadruplicate *Hox*-clusters in vertebrates.²⁸ A resolution of this question, however, will have to await data on microRNAs in lower vertebrates and amphioxus.

Methods

The publicly available genome databases (see Appendix A for URLs) were searched using Blastn²⁹ against all precursor miRNAs of the **mir17** family mentioned in Introduction. Parameter settings are tabulated in the electronic supplement. Conversely, the entire Rfam MicroRNA Registry, version 3.0, was compared against the genomic sequences near the putative family members. Exact locations of homologs of known miRNAs were identified using CLUSTALW³⁰ alignments and subsequent prediction of the secondary structure using Vienna RNA Package,³¹ in particular the programs RNAfold, RNAalifold,³² RNALfold,²² and alidot,³³ in order to verify the hairpin structure of the precursor (Figure 2).

Phylogenetic trees were reconstructed both with Maximum Parsimony and Neighbor-joining using the Phylip package³⁴ with standard parameters. The phylogeny of the entire clusters was computed using a concatenation of the alignments of the individual paralogous microRNAs according to their order in the cluster, and treating microRNAs that are not present in a particular cluster as missing data. This ensures that distances are measured based on nucleic acid substitution frequencies, not based on changes of cluster organization.

In principle, correlations between paired nucleotides ought to be taken into account when deriving phylogenies from RNA with highly conserved secondary structures,^{35,36} in particular when using maximum likelihood approaches. Nevertheless, this is rarely done in practice, probably because reasonable trees are routinely obtained, e.g. from ribosomal RNAs without considering the RNA structure at all. Multiple alignments of each of the groups of homologous pre-miRNAs show a bias in sequence conservation between the 3' and 5' arm of the pre-miRNAs (see electronic supplement). In general, the

sequence similarity is larger on the arm that contains the functional microRNA. The program MIRscan³⁷ uses this difference between the two arms to validate microRNA candidates. It appears that this difference can be attributed mostly to the almost perfect conservation of the mature miRNA itself. The mature miRNA thus carries very little phylogenetic information and at least a large part of the stem-loop structure is free of correlated mutations because only the arm opposite to the mature miRNA is mutated. Standard tree-reconstruction methods are therefore applicable as long as the miRNA precursors are not too distantly related.

In order to identify distant sequence similarities between precursor miRNAs from different paralog groups we compute a similarity score based on the significance of the alignment score: the identity score $s(I, J)$ for the pairwise alignment of two precursor miRNAs I and J is computed using the implementation of the fast approximate Wilbur–Lipman algorithm³⁸ from the CLUSTALW program. Then the mean score m and the variance v are estimated from a sample of 1000 alignments of sequences I_π and J_π that are obtained by randomly permuting the positions of I and J independently of each other. The z-score:

$$z(I, J) = \frac{s(I, J) - m}{\sqrt{v}} \quad (1)$$

is then used as similarity measure of I and J for WPGMA clustering.³⁹ Since $z(I, J)$ is computed by means of sampling it is possible that for closely related sequences I and J the estimate for $z(I, I)$ is slightly smaller than that for $z(I, J)$. This can lead to short negative-length branches in the WPGMA tree displayed in Figure 4. Nevertheless, this method produces robust similarity scores in regimes where reliable global alignments cannot be obtained.

The duplication history of the **mir17** family was reconstructed “by hand” based on the following assumptions: genomic rearrangements are (a) duplications of individual microRNAs within a linked cluster, (b) the deletion of a microRNA, and (c) the duplication of an entire cluster. In other words, we explicitly exclude the possibility of recombination between paralog clusters within an organism and copying of individual microRNAs from one cluster to another. The available data do not contain any evidence that such processes might play a role.

Acknowledgements

Useful comments by Claudia Fried, Sonja J. Prohaska, and two anonymous referees, are gratefully acknowledged. This work was supported in part by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, Project No. P15893, and by the German DFG Bioinformatics Initiative BIZ-6/1-2.

References

1. Nelson, P., Kiriakidou, M., Sharma, A., Maniatakis, E. & Mourelatos, Z. (2003). The microRNA world: small is mighty. *Trends Biochem. Sci.* **28**, 534–540.
2. Griffiths-Jones, S. (2004). The microRNA registry. *Nucl. Acids Res.* **32**, D109–D111. Database issue..

3. Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D. S. (2003). MicroRNA targets in *Drosophila*. *Genome Biol.* **5**, R1.
4. Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
5. Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. (2003). Identification of *Drosophila* MicroRNA targets. *PLoS Biol.* **1**, 397–409.
6. Doench, J. & Sharp, P. (2004). Specificity of microRNA target selection in translational repression. *Genes Dev.* **18**, 504–511.
7. Rajewsky, N. & Socci, N. (2004). Computational identification of microRNA targets. *Dev. Biol.* **267**, 529–535.
8. Lee, Y., Jeon, K., Lee, J. T., Kim, S. & Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* **21**, 4663–4670.
9. Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E. & Kutay, U. (2004). Nuclear export of microRNA precursors. *Science*, **303**, 95–98.
10. Schwarz, D. S. & Zamore, P. D. (2002). Why do miRNAs live in the miRNP. *Genes Dev.* **16**, 1025–1031.
11. Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L. *et al.* (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* **16**, 720–728.
12. Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A. & Tuschl, T. (2003). New microRNAs from mouse and human. *RNA*, **9**, 175–179.
13. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. (2003). Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4**, R42.
14. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
15. Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W. & Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. *Curr. Biol.* **12**, 735–739.
16. Houbaviy, H. B., Murray, M. F. & Sharp, P. A. (2003). Embryonic stem cell-specific microRNAs. *Dev. Cell*, **5**, 351–358.
17. Kim, J., Krichevsky, A., Grad, Y., Hayes, G. D., Kosik, K. S., Church, G. M. & Ruvkun, G. (2004). Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. *Proc. Natl Acad. Sci. USA*, **101**, 360–365.
18. Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B. T. G. *et al.* (2003). The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell*, **5**, 337–350.
19. Sempere, L. F., Sokol, N. S., Dubrovsky, E. B., Berger, E. M. & Ambros, V. (2003). Temporal regulation of microRNA expression in *Drosophila melanogaster* mediated by hormonal signals and Broad-Complex gene activity. *Dev. Biol.* **259**, 9–18.
20. Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W. *et al.* (2003). The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991–1008.
21. Obernosterer, G., Tanzer, A., Hofacker, I. L., Stadler, P. F. & Höchsmann, M. (2003). Searching genomes for small temporal RNAs. In *The 18th Dubrovnik International Course & Conference on the Interfaces among Mathematics, Chemistry and Computer Science (MATH/CHEM/COMP/2003)* (Graovac, A., Pokrivic, B. & Snreficki, V., eds), 65. Rudjer Bofiskovici Institute, Zagreb, Croatia. Poster.
22. Hofacker, I. L., Priwitzer, B. & Stadler, P. F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 191–198.
23. Höchsmann, M., Töller, T., Giegerich, R. & Kurtz, S. (2003). Local similarity in RNA secondary structures. In *Proc of the Computational Systems Bioinformatics Conference, Stanford, CA, August 2003 (CSB 2003)* pp. 159–168.
24. Taylor, J., Braasch, I., Frickey, T., Meyer, A. & Van De Peer, Y. (2003). Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* **13**, 382–390.
25. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-I. & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531–1545.
26. Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A. & Sidow, A. (1994). Gene duplication and the origins of vertebrate development. *Development, Suppl.*, **125**–133.
27. Panopoulou, G., Hennig, S., Groth, D., Krause, A., Poustka, A. J., Herwig, R. *et al.* (2003). New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.* **13**, 1056–1066.
28. Bailey, W. J., Kim, J., Wagner, G. & Ruddle, F. H. (1997). Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Mol. Biol. Evol.* **14**, 843–853.
29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
30. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). 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.
31. Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucl. Acids Res.* **31**, 3429–3431.
32. Hofacker, I. L. & Stadler, P. F. (1999). Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. Chem.* **23**, 401–414.
33. Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066.
34. Felsenstein, J. (1989). Phylip—phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.
35. Schöninger, M. & von Haeseler, A. (1999). Towards assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. *J. Mol. Evol.* **49**, 691–698.
36. Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M. & Higgs, P. G. (2003). RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol. Phylogenet. Evol.* **28**, 241–252.
37. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. (2003). Vertebrate microRNA genes. *Science*, **299**, 1540–1540.
38. Wilbur, W. J. & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl Acad. Sci. USA*, **80**, 726–730.
39. Sokal, R. R. & Michner, C. D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **38**, 1409–1438.
40. Hogeweg, P. & Hesper, B. (1984). Energy directed folding of RNA sequences. *Nucl. Acids Res.* **12**, 67–74.

Appendix A: Web references

Rfam: <http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml>

Genomes

Man (Hs): http://www.ensembl.org/Homo_sapiens/
Chimpanzee (Pt): http://pre.ensembl.org/Pan_troglodytes/
Mouse, (Mm): http://pre.ensembl.org/Mus_musculus/
Rat (Rn): http://www.ensembl.org/Rattus_norvegicus/
Dog (Cc): <http://www.ncbi.nlm.nih.gov/genome/seq/CfaBlast.html>
Chicken (Gg): <http://genome.wustl.edu/blast/client.pl>
Frog (Xt): <http://genome.jgi-psf.org/xenopus0/xenopus0.home.html>
Tetraodon (Tn): <http://www.genoscope.cns.fr/externe/tetraodon/>
Fugu (Tr): http://www.ensembl.org/Fugu_rubripes/
Zebrafish (Dr): http://pre.ensembl.org/Danio_rerio/
Tunicate (Ci): <http://genome.jgi-psf.org/ciona4/ciona4.home.html>
Sea Urchin (Sp): <http://www.hgsc.bcm.tmc.edu/projects/seaurchin/>
Drosophila (Dm): http://www.ensembl.org/Drosophila_melanogaster/
Mosquito (Ag): http://www.ensembl.org/Anopheles_gambiae/
Nematode (Ce): http://www.ensembl.org/Caenorhabditis_elegans/
Nematode (Cb): http://www.ensembl.org/Caenorhabditis_briggsae/

Edited by J. Doudna

(Received 26 January 2004; received in revised form 23 March 2004; accepted 26 March 2004)