

Ali M. Yazbeck<sup>1,2</sup> / Kifah R. Tout<sup>2</sup> / Peter F. Stadler<sup>1,3,4,5,6,7,8,9</sup> / Jana Hertel<sup>9</sup>

# Towards a Consistent, Quantitative Evaluation of MicroRNA Evolution

<sup>1</sup> Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany, E-mail: studla@bioinf.uni-leipzig.de

<sup>2</sup> Lebanese University, Doctoral School for Science and Technology, Rafic Hariri University Campus, Hadath, Lebanon

<sup>3</sup> German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, University Leipzig, Leipzig, Germany, E-mail: studla@bioinf.uni-leipzig.de

<sup>4</sup> Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany, E-mail: studla@bioinf.uni-leipzig.de

<sup>5</sup> Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany, E-mail: studla@bioinf.uni-leipzig.de

<sup>6</sup> Department of Theoretical Chemistry of the University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria, E-mail: studla@bioinf.uni-leipzig.de

<sup>7</sup> Center for RNA in Technology and Health, Univ. Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark, E-mail: studla@bioinf.uni-leipzig.de

<sup>8</sup> Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA, E-mail: studla@bioinf.uni-leipzig.de

<sup>9</sup> Helmholtz Centre for Environmental Research – UFZ, Young Investigators Group Bioinformatics and Transcriptomics, Permoserstraße 15, D-04318 Leipzig, Germany, E-mail: studla@bioinf.uni-leipzig.de

## Abstract:

The miRBase currently reports more than 25,000 microRNAs in several hundred genomes that belong to more than 1000 families of homologous sequences. Quantitative investigations of miRNA gene evolution requires the construction of data sets that are consistent in their coverage and include those genomes that are of interest in a given study. Given the size and structure of data, this can be achieved only with the help of a fully automatic pipeline that improves the available seed alignments, extends the set of available sequences by homology search, and reliably identifies true positive homology search results. Here we describe the current progress towards such a system, emphasizing the task of improving and completing the initial seed alignment.

**Keywords:** Alignments, Homology Search, miRBase, ascertainment biases

**DOI:** 10.1515/jib-2016-0013


**Received:** December 27, 2016; **Revised:** February 12, 2017; **Accepted:** February 16, 2017

## 1 Introduction

MicroRNAs (miRNAs) are an important class of abundant, endogenous small non-coding RNAs that are produced by almost all animals and plants as well as many unicellular eukaryotes. They function in post-transcriptional gene silencing making use of the evolutionarily ancient RNA interference pathways, through which double-stranded RNA can inactivate cognate sequences [1], [2]. The overwhelming majority of miRNAs is produced from mRNA-like primary precursor transcripts, the pri-miRNAs, through a common processing pathway. The hairpin-shaped pre-miRNAs are cut from the pri-miRNA in the nucleus. Pre-miRNAs are subject to stringent structural constraints that are different between major clades. The next processing steps differ substantially between animals and plants and lead to miRNA/miRNA\* duplex structures [3]. These are separated to produce mature miRNAs that are then incorporated in the Argonaute complex.

The innovation of miRNA-like endogenous RNAs clearly has occurred multiple times. Despite their functional analogy, there is no evidence for any homology between the plant miRNAs, animal miRNAs, and the miRNAs reported for several unicellular lineages [4], [5], [6]. Albeit, miRNAs and their hairpin precursors are most highly conserved genetic elements [7] within the single kingdoms. This makes it possible to accurately pinpoint the evolutionary origin of individual miRNAs [8]. Since miRNAs readily form paralogs by segmental or chromosomal duplications [9], [10] they often appear as members of families of homologous sequences, which

**Peter F. Stadler** is the corresponding author.

 ©2017, Peter F. Stadler, published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

form the basis of microRNA nomenclature used by miRBase [11]. The phylogenetic distribution of miRNA families appears to show that entire families are rarely lost. Hence they have been advertised as excellent, nearly homoplasy-free phylogenetic markers [12], [13], [14], [15], although there are some well described exceptions to this rule [16].

With the rapid increase of sequencing data covering an even denser distribution of animal taxa there has also been an increasing interest in quantifying innovation, turn-over, and loss of miRNAs and miRNA families [8], [15], [17], [18]. Despite best efforts, such quantitative questions are difficult to answer directly from the data provided e.g. by the miRBase. Recent critical evaluations of the miRBase inventory concluded that less than a third of the miRBase entries are robustly supported as miRNA genes [19], [20] and indicated that – without further curation – the database does not provide information that is sufficiently consistent for machine learning applications [21]. The reasons are manifold and, at least in part, a consequence of the non-trivial biology of small RNA genes. In fact, the term “microRNA” is far from being used in a well-defined and unambiguously manner in the literature. A broad range of biogenesis pathways for small RNAs have been discovered, reviewed e.g. in [22], [23]. These overlap with the canonical scheme of miRNA processing to various degrees. For instance, should *mir-451* [24], which is not processed by *Dicer*, or mirtrons, whose precursor hairpins are produced by the splicing machinery rather than with the help of *Drosha* [25], be counted as *bona fide* microRNAs? This issue is aggravated by small RNAs produced from structured RNAs such as tRNAs or snoRNAs. Here the consensus secondary structure often does not match the minimum free energy fold deposited in the database and/or deviates substantially from the criteria usually expected for a miRNA precursor [26], [27], [28].

Additional issues plague a quantitative evolutionary analysis. First, there is a strong ascertainment bias favoring the miRNAs from heavily researched model organisms (human, mouse, fruit fly, *C. elegans*). Second, automatic homology searches are largely restricted to the genomes curated by *ensembl*. Third, some homologous groups of miRNAs have been assigned to different or no miRBase families as a consequence of sufficient sequence divergence in the mature sequences (e.g. *mir-100* and the “tunicate specific” *mir-1473* [10]). Pre-miRNAs, furthermore, are annotated with quite different lengths within the same family. This concerns both different paralogs from the same family and orthologs from different species. An unfortunate consequence of this inconsistency is that the distinction between orthologous and paralogous sequences is blurred. Families with uncertain conserved structures tend to result in gap-rich, unreliable alignments. These give rise to poor covariance models and thus less sensitive homology searches leading eventually to underestimating the family’s phylogenetic distribution. In response to all these difficulties, all large scale studies of miRNA evolution start with homology-based searches and/or the integration of alternative miRNA annotations, e.g. from *MIR-GENE*DB [20], to complete the data set, as well as extensive, usually manual, curation of the data; see e.g. [8], [29], [30], [31], [32], [18]. This approach, however, has reached its limits with more than a hundred genomes and thousands of groups of miRNAs.

In this contribution we give a preliminary report on progress towards a completely automated pipeline for the analysis of microRNA evolution. We focus here on the key step of processing the data from the miRBase into a consistent collection of miRNA alignments. These in turn are then extended by homology search across a large set of additional genomes, using the ideas of the initial data cleaning now as filter criteria for deciding on the inclusion of initial homology search candidates. Using the current set of metazoan miRNAs as an example we demonstrate that such an approach is feasible.

## 2 Methods

A quantitative analysis of miRNA evolution requires data sets that are comparable between miRNA families in two respects: (i) within each species, all paralogs should be included, and (ii) across species, homologs should be detected with comparable sensitivity and specificity to avoid strong correlations of coverage and completeness of the included sequence data e.g. with the evolutionary age of the miRNA families. The data included in any repository resource unavoidably is subject to ascertainment biases deriving from the research history of individual genes and gene families and necessarily lags behind the available genomic data that can be included in a particular analysis. Therefore, it is indispensable to prepare a consistent data set by means of homology search. We propose, based on previous experience [8], [18], to break down this task into well-separated steps that can be automatized independently and then combined to a single work-flow:

- The *completion and correction of (seed) alignments* requires the identification of incomplete and problematic sequences and their correction. We will focus on this step below.
- The *homology search* itself is conducted with covariance models using *infernal* and produced new candidates in all genomes, including those from which the seed alignments are taken. This has the potential to identify new paralogs that have not yet been included in the data source.

- The decision to include a homology search candidate into the data set is taken according to the same criteria that are applied to the initial data curation step.

As a consequence, the entire data processing pipeline can be run until no further candidates are detected and thus a self-consistent data set is obtained. To exploit all the available information, alignments, consensus structures, and homology searches are conducted at the level of miRNA precursors. The evaluation of the homology search candidates also explicitly uses the mature miRNA products.

Our first task is to construct consistent, high quality alignments of each miRBase family. Given a multiple sequence alignment we compute the average column-wise Shannon entropy

$$S(A) = -\frac{1}{m} \sum_{i=1}^n \left( \sum_{\alpha} P(\alpha; i) \ln P(\alpha; i) \right) \quad (1)$$

as a measure of alignment quality that is independent of the method and scoring system that has been used for the original construction of  $A$ . Here,  $m$  is the number of alignment columns,  $P(\alpha; i)$  is the frequency of the symbol  $\alpha$  (which may also be IUPAC nucleotide code or a gap character '-'), and  $n$  is the number of sequences in the alignment. A reduction of  $S()$  upon modification of  $A$  signifies an improvement of the alignment. At present we consider only column-wise entropies since these provide a good first order approximation to quantifying sequence conservation and thus also alignment quality. Given that dinucleotide distributions do matter also for miRNAs [33], one could conceivably extend the quality measure to incorporate mutual information measures of adjacent and/or paired columns.

The most frequently observed errors within the miRNA alignments are too short or too long sequences, and sequences that do not fit at all into the alignments. All these issues produce bad (high) entropy values, since there are more regions containing gaps or unmatched bases, and thus are readily identified by our pipeline. Our basic approach to improve the alignment is to correct for these errors. This is reasonable because the ends of the precursor sequences submitted to miRBase are in almost all cases not determined experimentally but are based on the individual contributor's perception of the highly conserved region, the region with conserved base pairs, or simply the extent of the *infernalis* hit included in an *ensembl* genome annotation. We therefore truncate database entries that are too long and complete sequences that are too short by retrieving the ostensibly missing flanking regions from the genomic DNA sequences. Finally, we remove sequences that are most likely false positives from a previous homology search step. These cases are recognizable as those that introduce a lot of heterogeneity into the alignment. This filtering step can lead to the complete removal of short input sequences that were extended in the previous step.

Upon modification of an input alignment by extension, cropping, or outright removal of one or more sequences, the sequences were realigned and the entropy  $S()$  was computed for  $A$ . We tested two re-alignment strategies: (i) sequence-based *clustalw* alignments (version 2.0.12) [34] and (ii) structure based *mlocarna* alignments [35], which in turn depend on secondary structure predictions using the *Vienna RNA* package [36]. Since the sequence-based approach usually resulted in larger entropy reductions in test sample of miRBase families, we opted for *clustalw*.

To determine plausible 5' and 3' ends for a precursor we use the mature miR and miR\* sequences as anchors. The latter are also provided by the *Rfam* database. To obtain a consistent query set we *define* our precursors to extent 10nt beyond the ends of the mature sequences and prune longer database entries. This step is made difficult by the fact that for many families mature sequences are annotated in some or all sequences on only one of the two arms of the precursor hairpin. For each miRNA family, the precursors that have both mature sequences are pruned to at most 10nt upstream of the start of the 5' mature sequence and 10nt downstream of the end of the 3' mature sequence, respectively.

In some cases precursor sequences in the initial data basis are reverse complements of each other. In a few well-described cases this is biologically correct. Famously, *iab-4* and *iab-8* are produced from almost exactly the same genomic location but opposite DNA strands in arthropods [37], [38], and a similar situation was reported for *mir-3120* and *miR-214* [39]. Such cases may also appear as artifacts, however, when a mature miR\* is erroneously identified with a miR produced from the opposite strand e.g. due to a chemical modification such as adenosine-inosine (A-I) editing [40], [41] in the mature product. To catch such cases we also test whether the replacement of a sequence with its reverse complement would lead to an improvement in the alignment quality.

At this stage, the improved alignment still contains some entries that are too short. It suffices to run *cm-search* with the covariance model for the improved alignment against the genome for which only a truncated precursor was available. The result of this search is then used to replace the original entry.

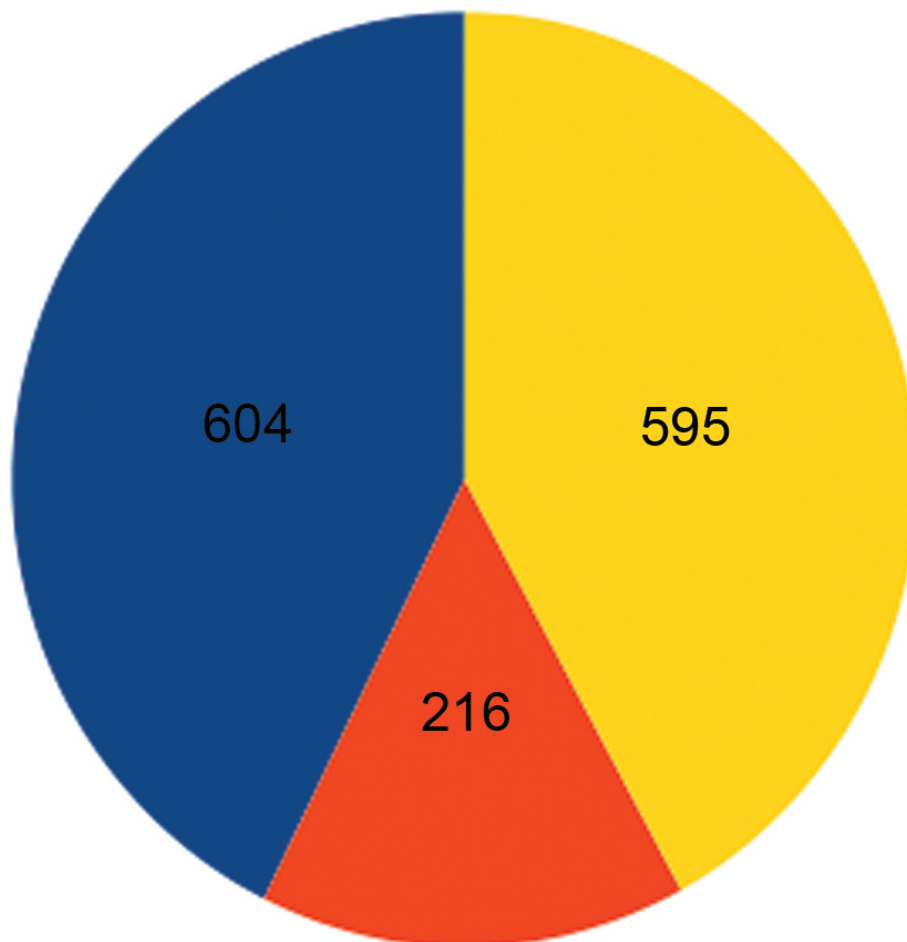
The sequences belonging to a family are finally re-aligned with *clustalw* and their consensus secondary structure is computed using *RNAalifold* [36]. The resulting structure-annotated alignment is then converted to *Stockholm* format and passed to *cm-build*, a component of the *infernalis* suite (release 1.1.1) [42], [43], to

construct revised covariance models. We used `cmscan` (another component of the `infernal` suite) to test the precursors of the same family with only one annotated mature sequence (if available) for membership.

Covariance models are used for homology search in more than 300 animal genomes in an ongoing study. The candidate sequences are added to query alignments using `cmalign` and then subjected again to data cleaning procedure. These steps can be repeated if necessary. The final data are converted to a presence/absence table that, for each species, reports the total number of paralogs of each miRNA family. The `ePoPe` tool [18] implements a Dollo parsimony approach that determines the most likely point of origin (as the lineage leading up to the last common ancestor of all observed paralogs) as well as the most parsimonious scenario of duplications and losses in each family.

### 3 Application and Results

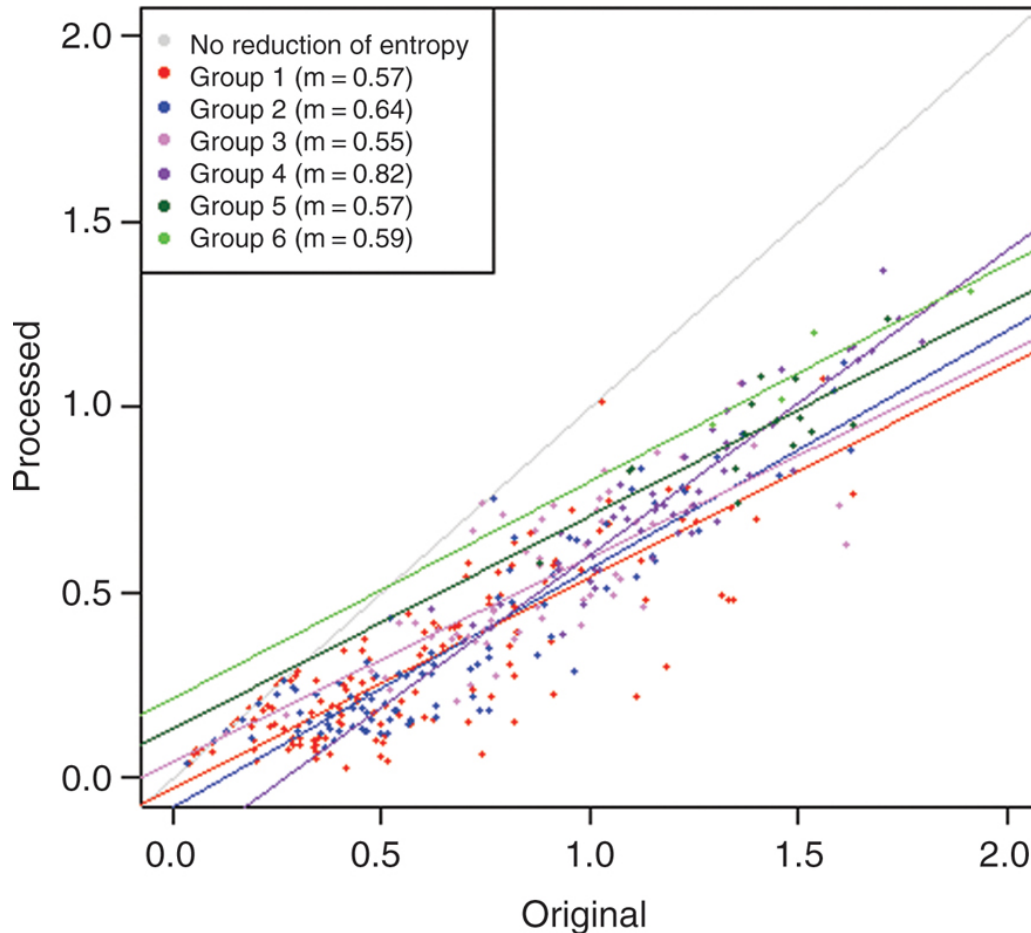
We focus here on the data cleaning steps for metazoan microRNAs as an illustrative example. The same pipeline, however, could just as well be applied to plant miRNA data. Our starting point is `miRBase` release 21. It comprises 21263 pre-miRNAs in animal species. A total of 14712 of these are assigned to a family and thus, come in at least two copies in this kingdom. The number of metazoan miRNA families is 1415. MiRNAs may have one or two annotated mature sequence, Figure 1. We selected the subgroup of miRNA families that have at least two pre-miRNAs with an annotated 5' and 3' mature sequence. These sum up to 367 families only. Albeit, these families comprise ~72 % of all metazoan pre-miRNAs. Of these we obtained improved alignments for 359 families. Only in 8 cases, the cleaning procedure did not lead to an improvement.



**Figure 1:** Distribution of mature microRNA sequence entries for `miRBase` (v. 21) families. The majority of families reports a miR and miR\* for all (red) or at least some (blue) of the family members. Families in which only a single mature product is reported for every member are shown in yellow.

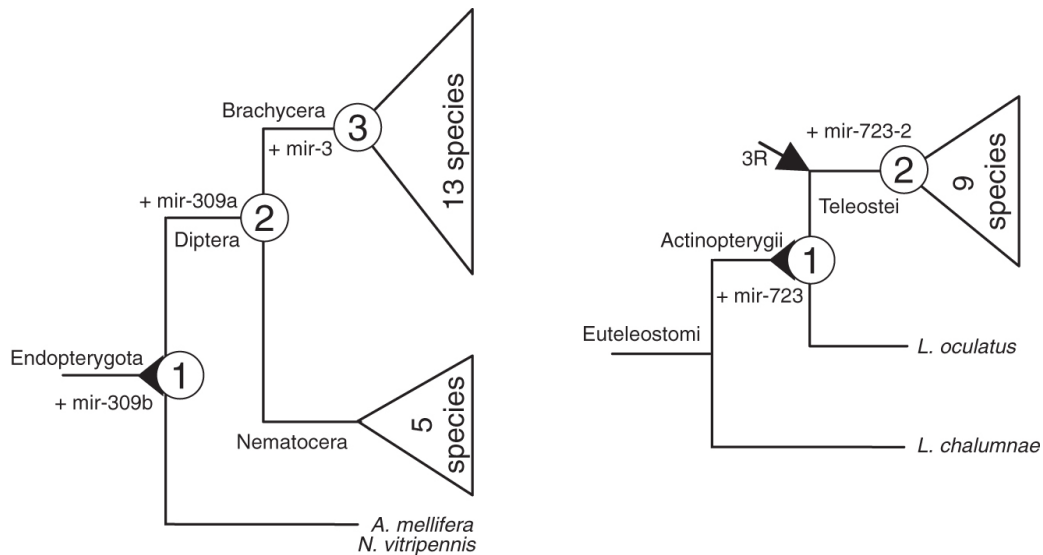
Figure 2 compares the entropies  $S()$  and  $S()$  before and after processing of the alignments. We find that the entropy systematically increases roughly proportional to the per column entropy in the input. Not surprisingly, noisy, high-entropy alignments are thus improved more than alignments that already show very little variation

upon input. The average improvement is about 33 % and does not vary dramatically with number of sequences per alignment. This constant rate of improvement suggests that the input alignments with higher entropy not only contain more diverse sequences but also substantial levels of alignment problems. Only the large alignments (group 4 in Figure 2), which often correspond to well-studied miRNA families, have been improved by only about 20 % on average. It is likely that this outlier can be explained by a more extensive manual curation of the corresponding miRBase alignments.



**Figure 2:** Comparison of  $S_0$  for original and processed alignments. Each data point represents one miRNA family. Data are stratified into six groups of miRBase families depending on the number of members in the initial alignment (1: 2-10 pre-miRNAs, 139 families; 2: 11-20 pre-miRNAs, 92 families; 3: 21-40 pre-miRNAs, 57 families; 4: 41-100 pre-miRNAs, 52 families; 5: 101-200 pre-miRNAs, 15 families; 6: >200 pre-miRNAs, 4 families), indicating that the improvements in alignment quality depends much more strongly on the entropy of the input alignment than on the number of sequences in the miRNA family.

To demonstrate the value of consistent data for quantitative analyses we select the miRNA family mir-3 (MIPF0000140) from the miRBase database. It comprises at total of 30 entries of the two paralogous groups traditionally named mir-3 and mir-309. While mir-3 can only be observed in *Drosophila* species, mir-309 is also found in Nematocera. A naïve analysis of the miRBase data thus would result in mir-3 being identified as the result of a drosophilid-specific duplication of an ancestral mir-309 gene. Processing the MIPF0000140 alignment into a revised query eventually results in homologs in most of the 35 sequenced genomes of Hexapoda, 20 of which were not present in the miRBase. In particular, the pipeline uncovered unambiguous mir-309 homologs in *Apis mellifera* and *Nasonia vitripennis*, pushing the origin of this family back from the Diptera to the stem of the Endopterygota. In the bees and wasps, however, miRs are observed only on the 3' side of the hairpin. As a second example we briefly discuss the mir-723 family, which in miRbase is annotated in the three teleost species, namely zebrafish, atlantic halibut and atlantic salmon. Homology search with the automatically improved alignment uncovered homologs not only in additional teleosts but also in the gar *Lepisosteus oculatus*. This pushed the origin of the miRNA family back to the stem of the Actinopterygii. Furthermore, the 3R genome duplication, which pre-dates the radiation of the teleosts, is clearly visible in the data, see Figure 3.



**Figure 3:** Distribution of miRNA homologs of miRNA families mir-3 (left) and mir-723 (right). Numbers at the nodes denote potential observations of the number of paralogs. The black triangle assigns the LCA to the phylogenetic tree. '3R' denotes the third round of whole genome duplication events that is assumed to have happened during the evolution of vertebrates.

## 4 Discussion

The current version 21 of miRBase contains 28,645 entries in 223 species [44]. Nevertheless the data set is far from providing a complete catalog of miRNAs for any given species. A quantitative analysis of miRNA evolution therefore requires substantial computational efforts to complete the data by homology search. This in turn requires extensive efforts to curate the miRBase data, facilitate homology search and to make the data comparable between families. In the present report we have focused on the strategies to obtain virtually complete sets of sequences and plausible alignments for each of the microRNA families. With more than 1000 families, a number that is rapidly increasing further, it is clear that a fully automatic pipeline is indispensable – and it is indeed under development. In this preliminary report we concentrated on the basic strategy and demonstrated in a pilot application to animal microRNAs that a fully automatized work-flow is feasible. The detailed evaluation of several example families showed, furthermore, that the last common ancestor of a miRBase family is frequently older than what would be inferred from the raw miRBase data themselves. A quantitative comparison will be described elsewhere.

The data cleaning procedure developed so far does not solve all problems and certainly will be subject to further improvements. Solutions to several issues remain under construction at this point. We therefore advocate to abandon the goal of computationally completing the data base itself. We suspect that discrepancies between manual and computational curation cannot be entirely avoided. The aim for data consistency in quantitative analyses and the aim of representing the published knowledge hence will remain at odds, at least on a – as one would hope – small subset of the data. We therefore aim instead at the construction of a pipeline that can simply be rerun when updates of the miRBase or the relevant sequenced genomes become available. This also allows us to seek to improve pipeline components gradually that address specific biases and technical limitations.

Very small initial seed alignments, and in particular miRNAs that are not incorporated into a miRBase family, for instance, remain a problem. The reason is that covariance models cannot be trained from a single sequence or a few very similar sequences, and the prediction of secondary structures is much less reliable from a single sequence than for an alignment of sequences with moderate mutual differences. A simple `blastn`-based homology search in closely related species, however, usually provides an easy remedy.

There is no guarantee, however, that the results of pre-processing the data by extending small families retrieves a *bona fide* miRNA family. In fact miRBase also occasionally includes fragments from misclassified ncRNAs other than miRNAs. For instance, mir-1940, which is a fragment of the snoRNA SCARNA4 [44] and a survey of many more such cases can be found in [27]. Upon identification, these erroneous miRBase families are usually quickly retired [44]. A check for miRNA properties with a tool such as `RNAmicro` [45] is nevertheless strongly advised. Such a filter, on the other hand, will tend to reject also special cases such as mir-451. The mature mir-451 sequence is not produced from the stems in the canonical way but originates from the loop [24],

hence it is (correctly) rejected by tools that are designed to recognize canonical miRNAs. The “fuzzy” limits of the notion of microRNAs not only affects machine learning efforts [21] but also has a potential impact for any quantification of microRNA evolution since there is no reason to assume that related classes of small RNAs should show the same patterns in evolution. Since miRBase will likely continue to hold a diverse collection of “miRNA-like” genes, we suggest that subclassifications that e.g. identify mirtrons, semi-mirtrons, etc., will be required. Whether annotation on the side of the database will provide this information or whether it will also need to be retrieved and/or computed by post-processing pipelines is still an open question.

We have focused here on properly treating miRNA precursors for which both miR and miR\* sequences are known. In cases where a mature product is reported for only one of the two arms of the precursor, a different strategy is required. The most natural approach is to start from the predicted secondary structure for the precursor, obtained e.g. with RNAalifold. The predicted structure and the known position of the miR implies the position of the hypothetical miR\* by taking into account that processing produces 2nt overhangs. After this extra processing step the alignment improvement can proceed as outlined in the previous section.

Conceivably, further improvements to the miRNA alignments could be achieved by explicitly using secondary structure information already in the alignment process. The small size of the precursors certainly would allow the use of one of the many variants of simultaneous alignment of folding algorithms [46]. However, the very definition of families as groups of miRNAs that are unambiguous homologs implies a sufficiently high level of sequence similarity to obtain decent sequence-based alignments. In addition, the positioning of the mature miRs yields a powerful anchor for the alignment that is consistent with the structure due the mechanism of *Dicer* processing. We suspect, therefore, that there is not much to be gained for the data curation pipeline itself. The presence of an appropriate consensus secondary structure relative to the positioning of mature products and the proper matching of patterns of sequence conservation extracted from the sequence conservation with the consensus structure becomes an important issue for the discrimination of *bona fide* microRNAs from their more distant small RNA relatives.

It is important to keep in mind that the exact size of the pre-miRNA stem-loop has not been determined experimentally for most miRNAs. In the absence of direct information, we have opted for a simple heuristic to delimit the alignments, giving precedence to consistency. There are some exceptional cases, however, where the 3' and 5' ends of the pre-miRNA are known precisely. This is in particular the case for mirtrons. These are introns that constitute pre-miRNAs [25], [47]. Our pragmatic “10nt from the miR and miR\*” rule will in general not respect this information. Ideally, the relative entropy criterion should be able to determine the boundaries of the precursor. This is bound to fail, however, in cases such as mirtrons, where the surrounding coding exons may exhibit an even larger level of sequence conservation than much of the pre-miRNA sequence. Further research into the patterns of sequence conservation and turnover around the boundaries of pre-miRNAs will be required to arrive at improved criteria. The pipeline outline here can, however, be easily adapted to such a task since it is a trivial matter to retrieve and align the precursor sequence with more extensive flanking regions. It is not difficult, however, to use pattern-based tools such as MaxEntScan [48] to identify conserved splice junctions close to the estimated ends of the precursor hairpin.

Albeit a fully automatized pipeline to process miRBase alignments into a complete and consistent set of miRNA sequences has proved to be feasible, much work still lies ahead. In addition to the incorporation and improvement of components that address more difficult special cases and exceptions, it also remains a non-trivial task to devise informative benchmarking strategies to test and optimize methods for the quantitative analysis of microRNA evolution.

## Acknowledgement

A.Y. was funded by a doctoral stipend of the National Council for Scientific Research of Lebanon (CNRS-L).

**Conflict of interest statement:** Authors state no conflict of interest. All authors have read the journal’s Publication ethics and publication malpractice statement available at the journal’s website and hereby confirm that they comply with all its parts applicable to the present scientific work.

## References

- [1] Cerutti H, Casas-Mollano JA. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet.* 2006;50:81–99.
- [2] Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol.* 2008;23:578–87.

- [3] Bologna NG, Schapire AL, Palatnik JF. Processing of plant microRNA precursors. *Brief Funct Genomics*. 2013;12:37–45.
- [4] Militello KT, Refour P, Comeaux CA, Duraisingh MT. Antisense RNA and RNAi in protozoan parasites: working hard or hardly working?. *Mol Biochem Parasitol*. 2008;157:117–26.
- [5] Braun L, Cannella D, Ortet P, Barakat M, Sautel CF, Kieffer S, et al. A complex small RNA repertoire is generated by a plant/fungal-like machinery and effected by a metazoan-like Argonaute in the single-cell human parasite *Toxoplasma gondii*. *PLoS Pathog*. 2010;6:e1000920.
- [6] Avesson L, Reimegård J, Wagner EC, Söderbom F. MicroRNAs in Amoebozoa: deep sequencing of the small RNA population in the social amoeba *Dictyostelium discoideum* reveals developmentally regulated microRNAs. *RNA*. 2012;18:1771–82.
- [7] Price N, Cartwright RA, Sabath N, Graur D, Azevedo RB. Neutral evolution of robustness in drosophila microRNA precursors. *Mol Biol Evol*. 2011;28:2115–23.
- [8] Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, et al. The expansion of the metazoan microRNA repertoire. *BMC Genomics*. 2006;7:15.
- [9] Tanzer A, Stadler PF. Molecular evolution of a microRNA cluster. *J Mol Biol*. 2004;339:327–35.
- [10] Hertel J, Bartschat S, Wintsche A, Otto C. The students of the Bioinformatics Computer Lab 2011. Evolution of the let-7 microRNA family. *RNA Biol*. 2012;9:231–41.
- [11] Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. *RNA*. 2003;9:277–9.
- [12] Sempere LF, Cole CN, McPeck MA, Peterson KJ. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol*. 2006;306B:575–88.
- [13] Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson K. MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci USA*. 2007;105:2946–50.
- [14] Heimberg AM, Cowper-Sal-lari R, Sémon M, Donoghue PC, Peterson KJ. MicroRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci USA*. 2010;107:19379–83.
- [15] Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, et al. The deep evolution of metazoan microRNAs. *Evol Dev*. 2009;11:50–68.
- [16] Fu X, Adamski M, Thompson EM. Altered miRNA repertoire in the simplified chordate, *Oikopleura dioica*. *Mol Biol Evol*. 2008;25:1067–80.
- [17] Tarver JE, Sperling EA, Nailor A, Heimberg AM, Robinson JM, King BL, et al. miRNAs: small genes with big potential in metazoan phylogenetics. *Mol Biol Evol*. 2013;30:2369–82.
- [18] Hertel J, Stadler PF. The expansion of animal microRNA families revisited. *Life*. 2015;5:905–920.
- [19] Castellano L, Stebbing J. Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. *Nucleic Acids Res*. 2013;41:3339–51.
- [20] Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, et al. A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu Rev Genet*. 2015;49:213–42.
- [21] Saçar MD, Hamzeiy H, Allmer J. Can MiRBase provide positive data for machine learning for the detection of mirna hairpins?. *J Integr Bioinf*. 2013;10:215.
- [22] Miyoshi K, Miyoshi T, Siomi H. Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. *Mol Genet Genomics*. 2010;284:95–103.
- [23] Okamura K. Diversity of animal small RNA pathways and their biological utility. *Wiley Interdiscip Rev RNA*. 2012;3:351–68.
- [24] Cifuentes D, Xue H, Taylor DW, Patnode H, Mishima Y, Cheloufi S, et al. A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science*. 2010;328:1694–8.
- [25] Curtis HJ, Sibley CR, Wood MJ. Mirtrons, an emerging class of atypical miRNA. *Wiley Interdiscip Rev RNA*. 2012;3:617–32.
- [26] Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev*. 2010;24:992–1009.
- [27] Langenberger D, Bartschat S, Hertel J, Hoffmann S, Tafer H, Stadler PF. MicroRNA or not MicroRNA?. In: de Souza ON, Telles P, Palakal MJ, editors. *Advances in bioinformatics and computational biology, 6th Brazilian Symposium on Bioinformatics, BSB 2011, volume 6832 of Lecture notes in computer science*. Berlin, Heidelberg: Springer, 2011:1–9.
- [28] Meng Y, Shao C, Wang H, Chen M. Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants. *RNA Biol*. 2012;9:249–53.
- [29] Copeland CS, Marz M, Rose D, Hertel J, Brindley PJ, Santana CB, et al. Homology-based annotation of non-coding RNAs in the genomes of *Schistosoma mansoni* and *Schistosoma japonicum*. *BMC Genomics*. 2009;8(10):464.
- [30] Tanzer A, Riester M, Hertel J, Bermudez-Santana CI, Gorodkin J, Hofacker IL, et al. Evolutionary genomics of microRNAs and their relatives. In: Caetano-Anolles G, editors. *Evolutionary genomics and systems biology*. Hoboken, NJ: Wiley-Blackwell, 2010:295–327.
- [31] Niehuis O, Hartig GH, Grath S, Pohl H, Lehmann J, Tafer H, et al. Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. *Curr Biol*. 2012;22:1309–13.
- [32] Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-to-teleost comparisons. *Nat Gen*. 2016;48:427–37 Corrigendum: *Nat Gen* 2016;48:700. doi:10.1038/ng0616-700c.
- [33] Fang Z, Du R, Edwards A, Flemington EK, Zhang K. The sequence structures of human microRNA molecules and their implications. *PLoS One*. 2013;8:e54215.
- [34] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 1994;22(22):4673–80.
- [35] Will S, Missal K, Hofacker IL, Stadler PF, Backofen R. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*. 2007;3:e65.
- [36] Lorenz R, Bernhart SH, Höner zu Siederdisen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol*. 2011;6:26.
- [37] Tyler DM, Okamura K, Chung W, Hagen JW, Berezikov E, Hannon G, et al. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev*. 2008;22:26–36.



- [38] Hui JH, Marco AM, Hunt S, Melling J, Griffiths-Jones S, Ronshaugen M. Structure, evolution and function of the bi-directionally transcribed *iab-4/iab-8* microRNA locus in arthropods. *Nucleic Acids Res.* 2013;41:3352–61.
- [39] Scott H, Howarth J, Lee YB, Wong LF, Bantounas I, Phylactou L, et al. MiR-3120 is a mirror microRNA that targets heat shock cognate protein 70 and auxilin messenger RNAs and regulates clathrin vesicle uncoating. *J Biol Chem.* 2012;287:14726–33.
- [40] Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, et al. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol.* 2006;13:13–21.
- [41] Heale BSE, Keegan LP, McCurk L, Michlewski G, Brindle J, Stanton CM, et al. Editing independent effects of ADARs on the miRNA/siRNA pathways. *EMBO J.* 2009;28:3145–56.
- [42] Eddy SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinf.* 2002;2(3):18.
- [43] Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* 2015;43:D130–7.
- [44] Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 2014;42:D68–73.
- [45] Hertel J, Stadler PF. Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics.* 2006;22:e197–202.
- [46] Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math.* 1985;45:810–25.
- [47] Wen J, Ladewig E, Shenker S, Mohammed J, Lai EC. Analysis of nearly one thousand mammalian mirtrons reveals novel features of dicer substrates. *PLoS Comput Biol.* 2015;11:e1004441.
- [48] Yeo C, Burge C. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol.* 2004;11:377–94.