Contents

non-protein-coding RNAs as regulators of development in tunicates

Cristian A. Velandia-Huerto, Arjan Gittenberger, Federico D. Brown Almeida, Peter F. Stadler and Clara I. Bermúdez-Santana

1 Introduction

noncoding RNAs roles in tunicates development date earliest in the 90‘s from the works of Swalla & Jeffry in which RNAs localized in the yellow crescent or myoplasm, a cytoskeletal domain in oocytes of the ascidian *Styela clava* were discovered [37]. This yellow crescent or YC RNA identified to be present throughout embryonic development was the first example involved in envisioning the future of a growing family of ncRNAs that would play important roles in growth and development in tunicates[37].

This asymmetrically distributed ascidian RNAs were part of the set of many other RNAs known as maternally synthesized cytoplasmically localized RNAs, discovered first in oocytes of Xenopus [4]

2 miRNA families origin and evolutionary perspective

2.1 Origins and Evolution of MicroRNAs

MicroRNAs (miRNAs) have been described in almost all animals and plants as well as diverse unicellular eukaryotes. They are important post-transcriptional regulators of gene expression affecting a sizable fraction of all mRNAs [**?**]. Mechanistically, miRNAs depends on the presence of the evolutionarily even older RNA interference

Author Name
Name, Address of Institute, e-mail: writeemail@unal.edu.co

Author Name
Name, Address of Institute e-mail: writeemail@com

pathways [**?**, **?**] that leads to the suppression of double-stranded RNA molecules in a cell's cytoplasm.

Throughout animals, canonical miRNAs are the processed through a well-characterized pathway. The primary precursor transcript (pri-miRNA) is transcribed by pol-II. While in most cases the pri-miRNA is a long noncoding RNA, some miRNAs are processed from protein-coding transcripts, where they are mostly derived from introns [25]. In the next step, hairpin-shaped precursors, the pre-miRNAs, are extracted while the RNA is still residing in the nucleus. These are exported into the cytoplasm [26] and then processed further into miRNA/miRNA* duplexes. In the final step the single-stranded mature miR or its complement, the miR*, is incorporated in RISC complex. Sequence complementarity of miR and mRNA ensures the targeting specificity [3]. As as a consequence, miRNAs share a set of structural characteristics, most importantly the extremely stable secondary structure of the precursor hairpin and the 2-bp overhang of miR and miR* generated by Dicer processing. These features make it possible to reliably identify miRNAs from short RNA-seq data, see e.g. [21, 7, 22].

Most animal microRNAs are among the most highly conserved genetic elements. The most stringent selection pressure acts on the mature miR sequence. This is a consequence of the fact that a single miR typically targets a large number of mRNAs. Mutations in the mature sequence thus simultaneous affect many interactions, and thus are almost always selected against. In conjunction with the stringent requirements on the secondary structure, the entire precursor is under strong stabilizing selection [**?**], explaining the observed high levels of sequence conservation. As a consequence, even evolutionarily distant homologs of miRNAs can be readily detected despite the short sequence length. Most efficiently, `infernal` [30] is used for this purpose, since it makes use of both sequence and structure comparison. The evolution of miRNAs can thus be traced back in time with high accuracy [15].

Like other gene families, miRNAs form paralogs [**?**, **?**] and hence often appear as families as homologous genes. This forms the basis of the `miRBase` nomenclature [2]. A series of investigations into the phylogenetic distribution of miRNA families showed that miRNAs are infrequently lost at family level and thus serve as excellent phylogenetic markers [35, 13, 12, 44], although the massive restructuring of the miRNA complement of tunicates is an important exception to this rule [**?**].

The innovation of new miRNA families is an on-going process. Experimental surveys of the miRNA repertoir thus have reported a large number of very young and even species-specific miRNAs [**?**, **?**]. The process was studied quantitatively in fruit flies, where innovation rate of as many as 12 new miRNA genes per million years has been estimated [**?**]. This is consistent with the fact that stable hairpins are abundant structural elements in random RNAs, which makes is not only possible but actually quite likely that miRNA precursors appear by chance in transcribed genomic regions [**?**, **?**, **?**]. Of course, only a tiny fraction of these fortuitously processed hairpins have a function and hance are subject to selection, and an even smaller subset is conserved over long evolutionary time scales. Detailed studies showed that evolutionarily young miRNA have comparably low expression levels. Initially, they go through a phase of relatively fast sequence evolution [**?**, **?**], which

slows down as the selective pressures from a gradual increase in the number of target site increases. A large, diverse set of targets then protects against miRNA loss [**?**]. The rate of gain of miRNA families that retained essentially permanently amounts to only 1 per several million years. This number is consistent with divergence of the miRNA complements between animal phyla.

Many authors have observed that overall the miRNA repertoire has been expansing throughout animal evolution in a manner that at least roughly correlates with morphological complexity [15, 35, **?**, **?**, **?**, 13, **?**, **?**]. Several bursts of miRNA innovation have been observed [15, 13, **?**, 16], most notably at the root of the placental mammals, the ancestor of "free-living" nematodes, or the radiation of the drosophilids. Massive morphological simplification, on the other hand, is sometimes associated with a drastic loss of miRNA families. This has been observed most prominently for tunicates [**?**, **?**].

2.2 miRNA identification and validation

The first microRNA (miRNA) in tunicates was discovered in the year 2000 from the work of *Pasquinelli et al., 2000* when they were studying the temporal regulation of let-7 during development by using samples of small RNAs of a wide range of animal species, in which the ascidian *Ciona intestinalis* was included as well as other vertebrates, hemichordates, mollusc, annelids, arthropod and other bilateral and nonbilateria animals [34]. Later on the year 2003 the same team suggested that let-7 RNA may control the late temporal transitions during development across animal phylogeny [33] albeit it was not identified on basal metazoans such as cnidarians and poriferans.

Then after the era of genome sequencing became available, it was launched in 2005 the computational screening of whole-genomes of non-model organism as tunicates. Beginning with the Cionas *C. intestinalis* and *C. savigni* a profile-based strategy was implemented in the ERPIN program [24]. On that work were detected a set of new miRNAs candidates considered as *C. intestinalis* specific such as the members of the family miR-9 and miR-79 and as it was expected, other miRNA families were found homologous between both Cionas like the families miR-124;92;98;325;310-313 and let-7. Coincidentally, by the same year a whole-genomic comparative approach in the urochordate lineage was performed on the species *C. intestinalis, C. savignyi* and *O. dioica*. Using a computational screening of structured ncRNAs based upon homology between predicted precursor hairpin structures 41 miRNA candidates were detected including let-7 and other six known candidates in *C. intestinalis* [29]. After all, the same group in 2007 implemented a structure-based clustering approach in *C. intestinalis* predicted 58 miRNAs, of which only let-7, miR-7, miR-124, and miR-126 coincided with the previously annotated miRNAs [45].

Thus far, the primary focus to identify miRNAs into urochordate linage has been mainly toward the use of computational approaches but soon came up the use of

new hybrid strategies combining computational and experimental studies to validate candidate families previously detected. For instance the first bona fide record for *C. intestinalis* was registered in mirBase only in its Release 11. Those first miRNAs records were derived from the work published in 2007 by Norden-Krichmar et al., [32]. The authors searched for conservation with the seed region of the known mature miRNA sequences from miRBase release 2006 on the whole-genomic sequences of *C. intestinalis* and *C. savignyi*. Those miRNAs were aligned locally using the FASTA/ssearch34 program. Only matches of 90% identity or better were retained. In further steps these authors studied RNA sequences that folds like hairpin structures with the mature miRNA sequence in the stem region including other typical features exhibit in miRNA hairpins. By manual curation of the genomic sequences predicted by the software mfold which folded like hairpin structures, a set of 18 miRNA molecules were detected which appeared conserved in both Cionas. After all, using Northern blot analyses in the adult tissue of *C. intestinalis* the authors confirmed expression of let-7, miR-7, and miR-126, as well as 11 other conserved miRNA families.

Until 2008, most of the miRNAs annotations were concentrated in Cionas, but new annotation approaches for other species in tunicates were appearing slowly to increase then the repertory of new miRNAs families in urochordates. In this order of ideas, the first repertory of miRNAs based on non-Cionas species was published by Fu et al., in 2008 for the larvacean *O. dioica* [9]. At that time the authors were studying the temporal-spatial expression patterns of conserved miRNAs in different developmental stages of oocytes, 1-cell zygote, 2-8 cell embryons, blastulas, gastrulas, tadpoles (in different stages) and animals from 1 to 6 days from *O. dioica*. In this research, small RNAs were isolated, amplified by RT-PCR and rapid amplification of cDNA ends (RACE) of the developmental stages, cloned and sequenced. Blast searches using the sequences of cloned small RNA libraries were used to annotate small RNAs as miRNA candidates. In further steps the recovered genomic flanking sequences each side of those mapped candidates were used as input to predicted secondary structures by mfold v3.1. This step was used to detect candidates that folds like miRNA hairpins and aimed to decrease the set of false positive potential miRNAs in *O. dioica*. Finally, for this set of potential candidates a developmental miRNA array dot blot analyses were performed to detect miRNA expression. With this approach from 3066 sequenced small RNA clones only for 55 miRNAs was detected expression. As a conclusion the authors suggested that those candidates were expressed throughout the short life cycle of *O. dioica* showing that some of them were stocked as maternal determinants prior to rapid embryonic development. Besides the authors identified a set of sex-specific miRNAs that appeared as male/female gonad differentiation which became apparent and was maintained throughout spermatogenesis [9]. Unexpectedly, the majority of the miRNAs loci in *O. dioica* were located in antisense orientations into the hosted genes in opposite fashion observed in the majority of the known mammalian miRNAs at that time.

Between the years 2009 and 2015 the majority of the studies of miRNAs in tunicates were focused into the validation of expression of computational predicted miRNAs in Cionas specially focused in *C. intestinalis* as model organism of tuni-

cates or into the test of new computational approaches as miRTRAP, miRDeep2 and miRRim2 which used next-generation sequencing libraries of small RNAs derived from *C. intestinalis* to validate their algorithms. Then by the year 2016 the first comparative homology based search strategy let us to identify the repertory on miRNAs and other ncRNAs in the carpet sea squirt *Didemnum vexillum* with a preliminary comparative analysis of gain and losses of miRNA families on chordates which included the *Cionas*, *O. dioca* and the colonial tunicate *Botryllus schlosseri* [42]. By the same year, from the preliminary genome sequence assembled for the Southern Ocean salp, *Salpa thompsoni* (Urochordata, Thaliacea) a set of miRNAs families were detected [17] and in 2017 the prediction of miRNAs families were reported to the species *Halocynthia roretzi*. On the following two sections we will focus on those stages of the fascinated increased screening of the miRNA repertory in tunicates.

2.2.1 Validation and detection of miRNA families in Cionas in this decade

At the end of the last decade the application of next generation sequencing technologies to sequence small RNA libraries changed the common way used to detect expression of miRNAs in many organisms including the tunicates. This technology became in one of the most common approaches that supported methods like RT-PCR, microarrays or dot blotting which were previously used to validate miRNA expression in tunicates. In 2009 after preparing small RNA libraries from various developmental stages including unfertilized eggs, early embryos, late embryos and adults from *C. intestinalis* was performed high-throughput sequencing of cDNA with an Illumina 1G Genome Analyzer experiments. These sequencing led to document 80 miRNAs families for *C. intestinalis*. Unexpectedly, were detected a distinct species of small RNAs processed outside of the miRNA precursors which were termed as moRs or miRNA-offset RNAs [36]. Later on, after extracting non-coding conserved regions of whole genome alignments between *C. intestinalis* and *C. savigny* a set of 12 million sequences were computationally folded using RNAfold and mfold. Then after combining the following criteria: structure/sequence conservation, homology to known miRNAs, and phylogenetic footprinting the authors detected a set of 458 candidate sequences [18]. Then in order to validate those candidate, RT-PCR and PAGE were conducted to design a custom microarray. After screened them for miRNA expression were identifying that 244 of the 458 miRNA predictions were represented either in their microarray data or in the Illumina database constructed previously for small RNA derived from *C. intestinalis* by [36]. Although they failed to predict 39 previously characterized miRNAs, it was suggested in this work that *C. intestinalis* genome may encode about 300 miRNA genes. Then to increase the miRNAs collection in *C. intestinalis* a novel computational strategy for the systematic, whole-genome identification of microRNA from high throughput sequencing information was developed in 2010 by [14]. That method, known as miRTRAP, incorporated not only the mechanisms of microRNA biogenesis but also includes additional criteria regarding the prevalence and quality of small RNAs arising from

the antisense strand and the neighboring loci. With that approach, nearly 400 putative microRNAs loci were detected. In short words these strategy relies on the way how the the biochemical machinery processes pre-miRNA hairpins produces short RNA products. This approach is highly depended on the deep of the small RNAs mapped to a given locus and is highlighted by the authors that the approach requires an accurate assignment of small RNA sequences on their relative positions along the hairpin, that is, miR/miR*, moR/moR* and loop [14]. Again a new approach took advantage of importance to detect miRNAs from the high-throughput sequencing of small RNAs available from [36]. These approach known as miRDeep2 improved the algorithm of its first version miRDeep [8] and let to identify with an accuracy of 98.6% and 99.9% canonical and non-canonical miRNAs in different species. These approach reported 313 known and 127 novel ones miRNAs in *C. intestinalis*. In the same year the program miRRim2 [39] was applied to the *C. intestinalis* genome, in which some candidates identified from the work of [14] and the several promising candidates were detected. In 2013, [19] was investigated the expression patterns of the cluster miR-1 and miR-133 in *C. intestinalis* and in *C. savignyi*. RT-PCR amplification of miR-1/133 precursors were performed and PCR products were subcloned and sequenced. Whole-mount in situ hybridization to detect cin-miR-1/miR-133 primary transcript was performed and LNA Northern blotting was conducted on different developmental stages.

2.2.2 The new era to get deep insights into the repertory of miRNA in other urochordates

Since 2016 new approximations has increased our knowledge about new families in other tunicates thanks to the sequence of new urochordate genomes of the species *D. vexillum*, *S. thompsoni* and *H. roretzi* write here B. schlosseri because no-ncRNAs were reported, only on mtDNA and methodology to validate genes by RNA-seq from different tissues and it was reported on 2013...

(Please summary of our Dvexillum paper [42] including the first reported preliminary annotation for colonial tunicate *B. schlosseri* beside the one for Dvexillum.) For the draft genome sequence from *D. vexillum* an homology-based computational approach was applied [42]. Blast and HMMer searches were performed with annotated small ncRNAs sequences from metazoans and hidden markov models from RFAM[1] to obtain the sort of candidates at sequence level. Structural alignments of those sequences were performed by infernal (CITE), using metazoan-specific covariance models to annotate the small ncRNAs collection, which 57 families and 100 loci of miRNAs were found.

For the preliminary assembled of the genome sequence for the Southern Ocean salp *S. thompsoni* [17] were small RNA libraries constructed to be sequenced on an Illumina Hiseq 2000. After filtering data sets to 18-24 nt for miRNA and 28-32 nt for piRNA, the reads were aligned to *S. thompsoni* genome and miRNA gene

[1] http://rfam.xfam.org/

folding predictions were performed using RNAfold. In this initial survey of small RNAs, were revealed the presence of known, conserved miRNAs, as well as novel miRNA genes and mature miRNA signatures for varying developmental stages. Then in 2017, the prediction of 319 miRNAs candidates in *H. roretzi* were obtained through three complementary methods. The experimental validation suggested that more than half of these candidate miRNAs are expressed during embryogenesis. The expression of some of the predicted miRNAs were validated by RT-PCR using embryonic RNA. In this approach *C. robusta* small RNA-Seq reads derived from *C. robusta* [36] (previously known as *C. intestinalis* today reclassified) was used to identify conserved miRNAs in *H. roretzi* [43].

2.3 miRNA in clusters

One of the most interesting aspects about the patterns of genomic locations of miRNAs is to known whether those loci are randomly distributed throughout the genome as single copies or if they are arranged on consecutive locations or in tandem copies clustered to be expressed from polycistronic primary precursors or to be transcribed independently. Interestingly in *O. dioca* miRNAs are located in the antisense orientations of protein-coding gene and immediately downstream of its corresponding 3‘UTR region or even more in the sense strand of introns [9]. Nevertheless, after those conspicuous distributions some clusters have been also identified in *O. dioca*. For instance four miRNAs, miR-1490a, miR-1493, miR-1497d, and miR-1504, are reported by [9] to be presented as two copies, and miR-1497d-1 and miR-1497d-2 are included in the large miR-1497 cluster. See the current structure of this cluster in Table 1 although only one copy for the miR-1497 has been reported for *C. instestinalis* located in an intergenic region [9], [14] and one in *C. savigny* overlapped in an intron [9]. By testing real time PCR co-expression of some miRNAs, their host and adjacent genes in *O. dioca* by [9] it was discovered for the case of the cluster miR-1487/miR-1488 a not clear positive or negative correlation with the expression of its anti-sense hosting gene. In males this cluster expression was not associated with the expression of its adjacent ABCA3 gene by the same authors.

Table 1: Details of biggest miRNA cluster for chordate species.

Specie	**Chr**	**Start**	**End**	**Size (Mb)**	**No.**	**miRNAs detail**
B. floridae	Bf_V2_118	216744	220351	3607	5	bfl-mir-4869, bfl-mir-4857, bfl-mir-4862, bfl-mir-4856b, bfl-mir-4856a

O. dioica	scaffold_3	2222857	2223714	857	6	odi-mir-1497e, odi-mir-1497d-2, odi-mir-1497d-1, odi-mir-1497c, odi-mir-1497b, odi-mir-1497a
B. schlosseri	chrUn	40003	41320	1317	2	mir-233, mir-10
C. intestinalis	7	4153284	4156782	3498	23	cin-mir-4006d, cin-mir-4006c, cin-mir-4001b-2, cin-mir-4000i, cin-mir-4006g, cin-mir-4001e, cin-mir-4001d, cin-mir-4000g, cin-mir-4006f, cin-mir-4006b, cin-mir-4001b-1, cin-mir-4000c, cin-mir-4006e, cin-mir-4000b-2, cin-mir-4001a-1, cin-mir-4000b-1, cin-mir-4002, cin-mir-4000d, cin-mir-4001h, cin-mir-4000a-2, cin-mir-4006a-2, cin-mir-4006a-3, cin-mir-4006a-1
C. savignyi	reftig_16	3924783	3925336	553	3	csa-mir-216b, csa-mir-216a, csa-mir-217
C. savignyi	reftig_1	1335375	1336487	1112	3	csa-mir-92b, csa-mir-92c, csa-mir-92a

D. rerio	4	28738556	28754891	16335	60	dre-mir-430a-18, dre-mir-430c-18, dre-mir-430b-4, dre-mir-430a-15, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430a-15, dre-mir-430c-18, dre-mir-430b-3, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-8, dre-mir-430a-15, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430a-17, miR-430, dre-mir-430b-20, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430i-3, dre-mir-430c-18, dre-mir-430b-3, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-8, dre-mir-430a-11, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430i-3, dre-mir-430c-18, dre-mir-430b-19, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430a-17, miR-430, dre-mir-430b-20, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430i-3, dre-mir-430c-18, dre-mir-430b-19, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-5, dre-mir-430a-15, dre-mir-430c-18, dre-mir-430b-3, dre-mir-430a-10, dre-mir-430c-18, dre-mir-430b-8, dre-mir-430a-15, dre-mir-430c-18, dre-mir-430b-5

L. chalumnae	JH126646.1	1529355	1882777	353422	7	mir-233, mir-233, mir-233, mir-598, mir-672, MIR535, mir-233

In *C. intestinalis* some miRNAs are also located in introns and a small class of miRNAs are found to be deriving from mature mRNAs encoded within exons or UTR sequences [14] in contrast to the location of the loci in antisense orientations of protein-coding gene as seen in *O. dioca* but this antisense orientation is reported for some miRNAs loci which express antisense miRs derived from miRNA loci as antisense products and antisense moR products as the miR-2246. Only 44 loci appeared to be expressed as antisense products from the 300 miRNA loci predicted in 2010 by [14]. In Cionas have been also detected miRNAs organized in clusters, for example in *C. intestinalis* a putative cluster was detected by [18] using microarray analysis that shows a similar loci organization to the cluster let-7/miR-125/miR-100 observed in Drosophila. The miR-1473 was later classified as the orthologue of miR-100 in the analysis derived from the comparison of the evolution of this cluster conducted by [11]. The authors suggested that mir-100, mir125 and let7 are clustered in most of the bilaterian genomes including as 1473 as orthologue of mir-100.

Current analysis of this cluster shows that the distribution of miRNAs families on this let-7 cluster are distributed in all the studied chordate species. In vertebrate species like (*D. rerio* and *L. chalumnae*) exists more than one let-7 cluster, extending the loci definition which is not restricted only for one element but for a cluster of many locus with different length distributions. It is important to see that let-7 is organized sometimes with another let-7 locus or with another miRNA's loci families. The distribution of this cluster reported on amphioxus is composed by 2 let-7 and 3 mir-10 (1 bfl-mir-100, 1 bfl-mir-125a and 1 bfl-mir-125b), this cluster architecture almost conserved on vertebrates that apparently inverted the order and split the relation between let-7 and mir-10, creating two different cluster order groups: let-7 + mir-10 and let-7 + other families. In this way, tunicates reported the latter group, not including mir-10 on the cluster but including mir-233, mir-1473 or mir-125.

A second miRNA cluster consisting of the miR-182 and miR-183 was also detected in *C. intestinalis* in 2010 by [18] which is in the current predictions is reported another member locus the miR-96 organized in the middle of those loci as is shown in the plot 7. Here the authots also found five additional paralogs of let-7 within a 1-kb stretch, but it is important to know that those elements had been identified on chromosome 4q on Ensembl release 54 version, at the current version only two of those elements have been identified by homology approaches (Figure 1).

The cluster miR-1/miR-133, expressed specifically in Cionas muscle tissues was also reported by [19]. The authors reported that one copy of this cluster is presented in both Cionas. As is shown in the plot 3 a copy is also presented in *L. chalumnae*. In 2012 a new cluster was proposed in *C. intestinalis* by [39] located on the cromosome 10q and composed by the mir-4054 locus and the mir-4091. In the current

distribution of this cluster a new annotated family the mir-4008 with three paralogous is located on the middle of those loci. This current distribution in shown of Table 2 whose loci were validated by [32], [9], [14], and [39]. As was mentioned by [14] is not very common to find related miRNAs organized in clusters composed by closely related families that differ in just a single nucleotide in the seed sequence as was found on the cluster composed by nine Ci-mir-2200, seven Ci-mir-2201 and nine Ci-mir2203 which were previously reported under that putative names. They also found a second large cluster composed of 11 miRNAs that gather into 4 paralogous families three of Ci-mir-2200, three Ci-mir-2201, four Ci-mir-2204 and two Ci-2217. Current distribution of miRNAs families in *C. intestinalis* and curated annotations indicate than in other regions of the chromosome 7q are also organized miRNAs in tandem copies of families. For instance the big cluster presumably re-named from [14] today is known to be built by the families miR4000, miR-4001, miR4002, and miR4006 located on chromosome 7q (Table 2). Another cluster is also located on the same chromosome composed by the families miR-4003, miR4005 and miR4077 in Table 2. Some other cluster are also found on the cromosome 1a, 10q and 3p. See this structure on Table 2, most of them validated by [14].

Some other clusters shared between both Cionas are the cluster 92, 124 and 200 validated by [32], [9], [14] which the structure is seen on Table 2.

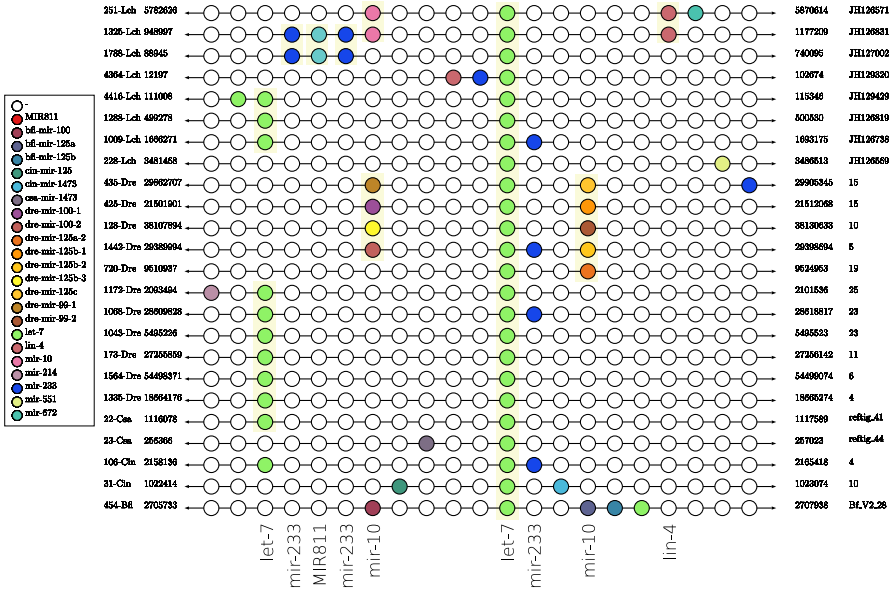

Fig. 1 Multiple alignment of let-7 clusters. Specific names from annotations and homology predictions are described at the legend. Names from miRBase families are reported at the bottom of the aligned elements.

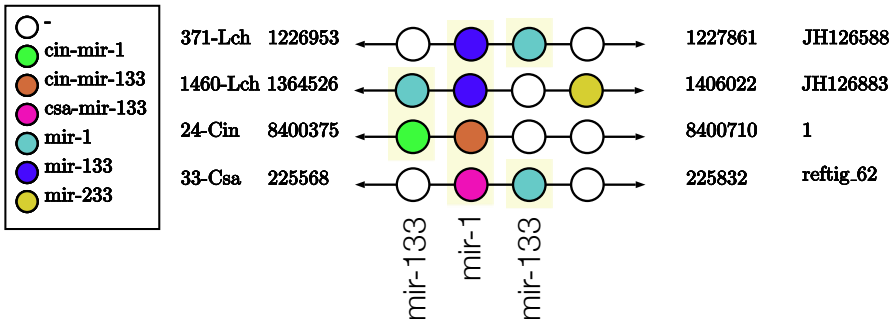

Fig. 2 mir-1/mir-133

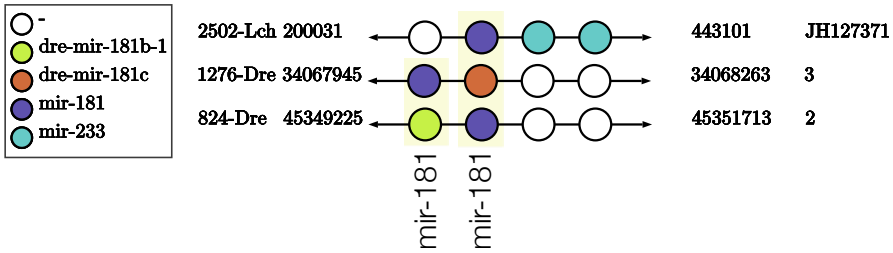

Fig. 3 mir-181

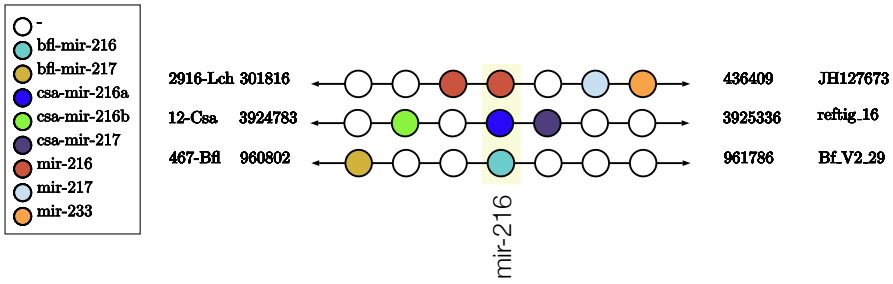

Fig. 4 mir-216/mir-217

Table 2: Reported clusters on literature. Bold text represent those miRNAs elements that are currently annotated and validated, but could not possible to detect by homology strategies.

Family	Specie	Chr	Start	End	miRNAs	Comments	Source DB	Ref.
let-7	Ciin	4q	2082260	2083286	**cin-let-7a-1, cin-let-7f, cin-let-7b, cin-let-7c, cin-let-7a-2**	Reported on miRBAse and annotated on Ensembl	miRBase	[14], [9]

	Cisa	reftig_41	1114139	1117597	**csa-let-7c-1, csa-let-7b, csa-let-7a, csa-let-7c-2**	Reported on miRBase and does not detected by homology strategies.	miRBase	[9]
mir-4091	Ciin	10q	3226200	3228884	**cin-mir-34**, cin-mir-4091, cin-mir-4008a, cin-mir-4008c, cin-mir-4008b, cin-mir-4054	NA	miRBase and Homology	[32], [9], [14], and [39]
7qother	Ciin	7q	4828431	4835967	cin-mir-4077b, cin-mir-4003b, cin-mir-4005b, cin-mir-4077d, cin-mir-4003a-1, cin-mir-4003c, cin-mir-4077a, cin-mir-4003a-4, cin-mir-4003d	NA	miRBase and Homology	[14]
3p	Ciin	3q	567478	571031	cin-mir-4001f, cin-mir-4000e, cin-mir-4001c, cin-mir-1502d, cin-mir-4018a, cin-mir-4019, cin-mir-1502b, cin-mir-1502a, cin-mir-4007, cin-mir-4000f, cin-mir-4001i, cin-mir-4018b, cin-mir-1502c	Inclusion of cin-mir-4019 and cin-mir-4007	miRBase and Homology	[14]
1q	Ciin	HT000037.1	4884	5250	**cin-mir-367**, **cin-mir-4009c**, **cin-mir-4009b**, cin-mir-367, cin-mir-4009c, cin-mir-4009a, cin-mir-4009b	Non-highlighted names could be found in the current genome at HT000037.1 scaffold	miRBase and Homology	[14]

10q	Ciin	10q	3226200	3228884	**cin-mir-34**, cin-mir-4091, cin-mir-4008a, cin-mir-4008c, cin-mir-4008b, cin-mir-4054	NA	miRBase and Homology	[14]
mir-92	Ciin	3q	884615	885508	cin-mir-92a, cin-mir-92d, cin-mir-92c	NA	miRBase and Homology	[14]
	Cisa	reftig_1	1335375	1336487	csa-mir-92b, csa-mir-92c, csa-mir-92a	NA	miRBase and Homology	[9]
	Oidi	scaffold_1	3086369	3086586	odi-mir-92b, odi-mir-92a	NA	Homology	[9]
mir-124	Ciin	7q	4969691	4969912	cin-mir-124-1, cin-mir-124-2	NA	miRBase and Homology	[14], [9]
	Cisa	reftig_262	49392	49620	csa-mir-124-1, csa-mir-124-2	NA	miRBase and Homology	[9]
mir200	Ciin	HT000325.1	8331	8778	cin-mir-200, **cin-mir-3575**, cin-mir-141, **cin-mir-5611**	NA	miRBase and Homology	[32], [9], [14], [8]
	Cisa	reftig_613	31353	31949	csa-mir-200, csa-mir-141	NA	miRBase and Homology	[9]

7qf	Ciin	7q	4153284	4156782	cin-mir-4006d, cin-mir-4006c, cin-mir-4001b-2, cin-mir-4000i, cin-mir-4006g, cin-mir-4001e, cin-mir-4001d, cin-mir-4000g, cin-mir-4006f, cin-mir-4000h*, cin-mir-4006b, cin-mir-4001b-1, cin-mir-4006e, cin-mir-4001a-1, **cin-mir-4001a-2**, cin-mir-4002, cin-mir-4001h, cin-mir-4000a-2, cin-mir-4006a-2, cin-mir-4006a-3, cin-mir-4006a-1, **cin-mir-4006e**, **cin-mir-4001a-1**, **cin-mir-4006b**, cin-mir-4000c, **cin-mir-4006e**, cin-mir-4000b-2*, **cin-mir-4001a-1**, cin-mir-4000b-1*, **cin-mir-4001a-2**, **cin-mir-4002**, cin-mir-4000d*, **cin-mir-4001h**, **cin-mir-4006a-3**, **cin-mir-4006a-1**, **cin-mir-4000a-1**	Elements marked with * are identified by homology strategies at the same cluster, but in another order reported by miRBase.	miRBase and Homology	[14]

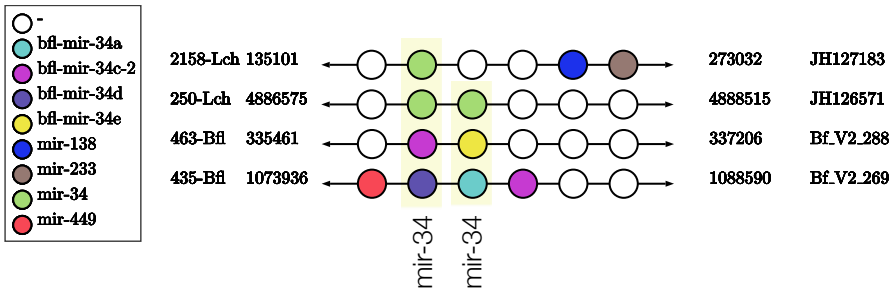

Fig. 5 mir-34

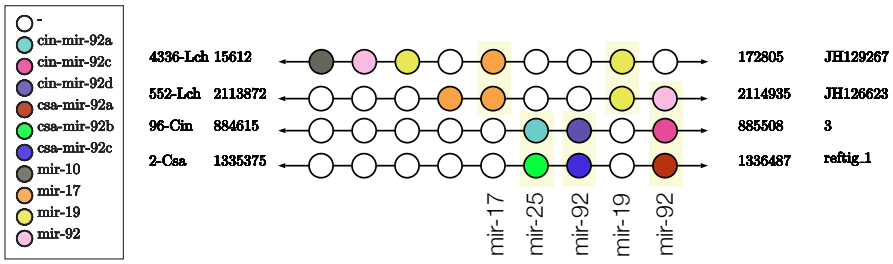

Fig. 6 mir-92

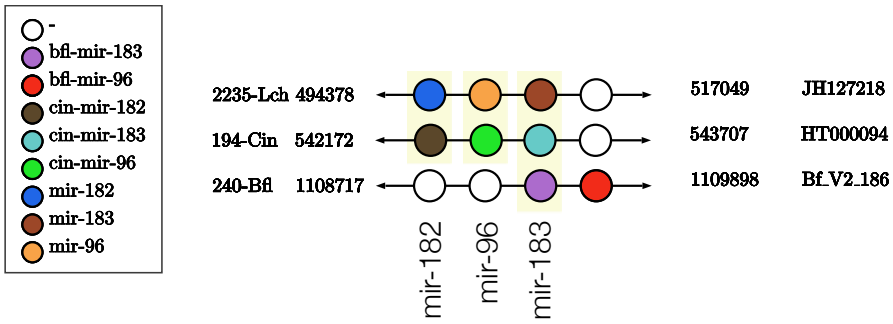

Fig. 7 mir-182/mir-96/mir-183

2.4 To complete the tree of loss and gain of families

Our miRNA families updated with the new two annotated miRNAs Salpa and Halocyntia

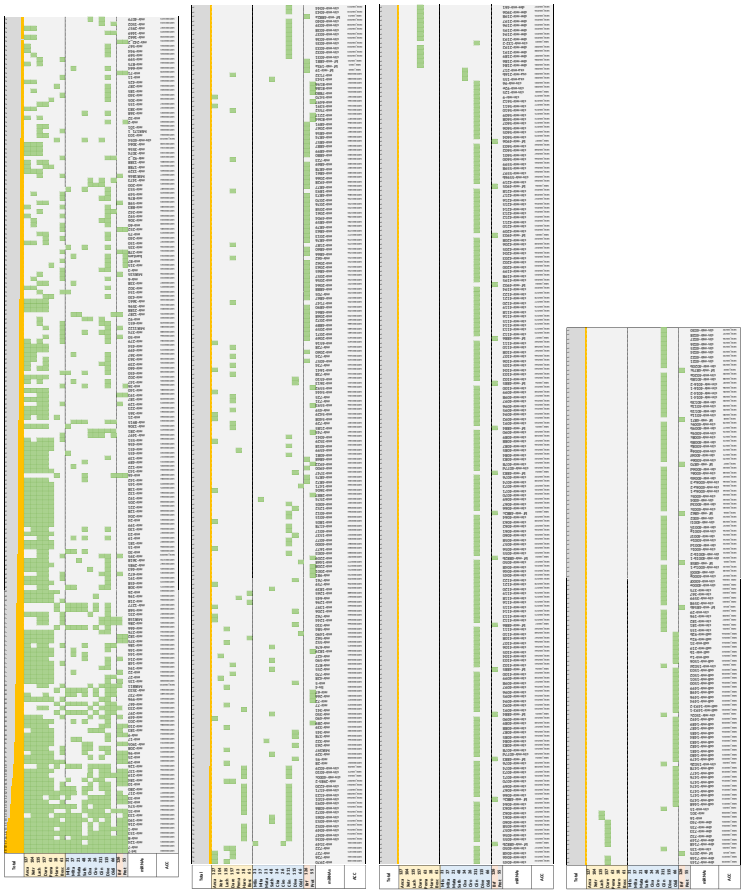

Fig. 8 Absence/Presence Matrix of miRNAs families along Bilaterian species. **Prot**: Protostomata, **Brfl**: *B. floridae*, **Oidi**: *O. dioica*, **Dvex**: *D. vexillum*, **Ciin**: *C. intestinalis*, **Cisa**: *C. savignyi*, **Ciro**: *C. robusta*, **Sath**: *S. thompsoni*, **Mata**: *M. oculata*, **Mlta**: *M. occulta*, **Mlis**: *M. occidentalis*, **Bosc**: *B. schlosseri*, **Haro**: *H. roretzi*, **Pema**: *P. marinus*, **Dare**: *D. rerio*, **Lach**: *L. chalumnae*, **Xetr**: *X. tropicalis* and **Anca**: *A. carolinensis*.

3 miRNAs and its rol in development

3.1 miRNAs discovery and development

4 Other ncRNAs associated to development

4.1 Yellow Crescent RNA

Yellow crescent RNA, i.e. YC RNA, concerns an about 1.2 kb long polyadenylated RNA, which can be present throughout the embryonic development of ascidians

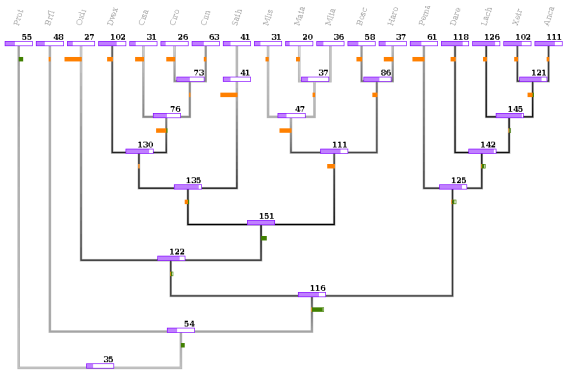

Fig. 9 Dollo parsymony of miRNAs families distribution in some chordates genomes

.

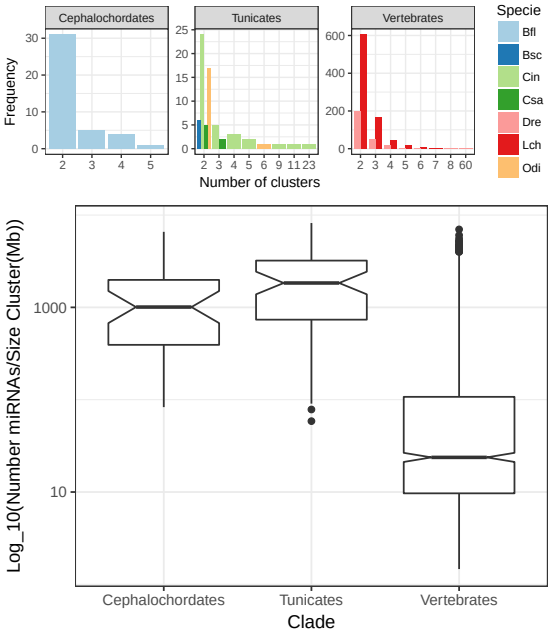

Fig. 10 Analysis of the distribution, size and number's of cluster along chordate species.

[37]. Its name refers to the fact that in situ hybridization confirmed that YC RNA is localized in the yellow crescent region of one-cell zygotes. The YC transcripts are actually already found in the cortex of unfertilized eggs, segregating with the myoplasm to the yellow crescent after fertilization [37]. Subsequently most YC transcripts enter the primary muscle cell lineage after cleavage and are also present in the secondary muscle cell lineage [37]. YC RNA was first discovered in the club tunicate Styela clava [37]. As the presence of the 1.2-kb RNA in oocytes and early cleaving embryos indicates that it is a maternal transcript, YC RNA is considered to

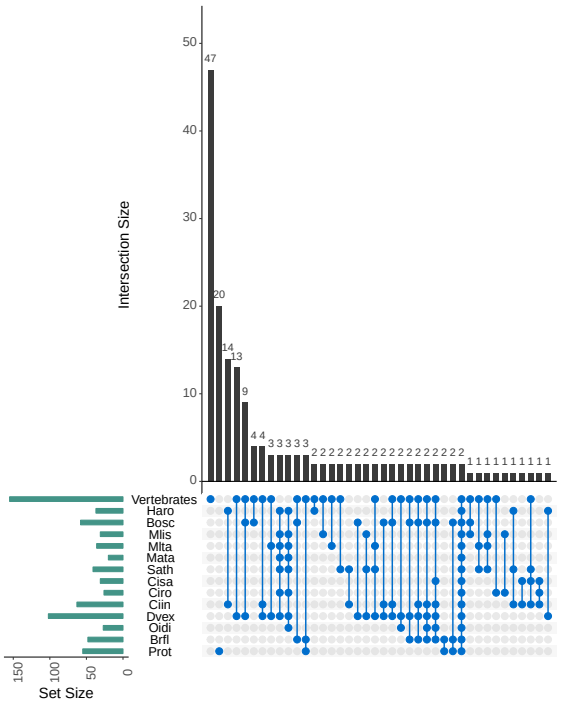

Fig. 11 Comparsion between miRNAs families along Bilaterian species. Same labels from Figure 8 were used. In this case *vertebrates* group the following species: **Pema**: *P. marinus*, **Dare**: *D. rerio*, **Lach**: *L. chalumnae*, **Xetr**: *X. tropicalis* and **Anca**: *A. carolinensis*

be a maternal RNA [37]. It is associated with the cytoskeleton and segregates to the muscle cells during ascidian embryogenesis. Although the YC ORF encodes for a putative polypeptide of 49 amino acids, this protein is relatively small and does not show any significant homology to any known proteins. As the YC RNA shows various features indicating that it actually functions as an RNA rather than as a protein coding molecule, it is considered to be a noncoding RNA that may play an important role in growth and development [37].

4.2 MicroRNA-offset RNAs

MicroRNA-offset RNAs, i.e. moRNAs, concern about 20 nucleotides long RNAs that lie adjacent to pre-miRNAs. They can originate from both ends of these pre-miRNAs, although prevalently they are derived from the 5' arm [6]. During a study focused on identifying miRNAs in the simple chordate *C. intestinalis* moRNAs were first discovered [36]. Unexpectedly, half of the *C. intestinalis* miRNA loci that were detected in this study turned out to encode for previously uncharacterized small RNAs, in addition to conventional miRNA and miRNA* products. This new class of RNAs was hereafter referred to as 'moRNAs', for miRNA-offset RNAs. It became clear that these moRNAs are probably produced by RNAse II-like processing

and are observed, like miRNAs, at specific developmental stages [36]. These results and subsequent studies gave rise to the hypothesis that moRNAs concern a new class of functional regulators whose qualitative alteration and/or expression dysregulation might even impact human diseases [6]. Evidence supporting this hypothesis is still fragmentary however. After the discovery in *Ciona*, moRNAs were also found in human cells by deep sequencing analysis. Hereby it was reported that moRNAs from 78 genomic loci were weakly expressed in the prefrontal cortex [20]. Additional indications that moRNA have a distinct function include the fact that some moRNAs are as conserved as miRNAs and are in fact conserved across species to an extent that correlated with expression level [36]. The expression level of certain moRNAs can even be greater than for their corresponding miRNA [40]. Finally, it can be argued [6] that it is likely that moRNAs might represent a functional class of miRNA-related agents as moRNAs are prevalently produced by the 5' arm of the precursor, independent of which arm produces the most expressed mature miRNA [20, 40]. What functions moRNAs may have, varies. For example, moRNA expression was recorded in solid tumours, together with other small RNAs [28]. In addition the fact that an 18-fold enrichment of moRNAs was observed in the nucleus [38] indicates that at least some moRNAs may have functions related to nuclear processes [6]. Although these studies do provide good indications, the potential functional roles that moRNAs can play, remain still largely unknown.

4.3 Long Noncoding RNA RMST

Long noncoding RNAs, i.e. lncRNAs, are abundantly found within mammalian transcriptomes. One of the known groups of lncRNAs, includes the rhabdomyosarcoma 2-associated transcript (RMST), which is indispensable for neurogenesis [31]. Human RMST was shown as being responsible for the modulation of neurogenesis as its expression is regulated by the transcriptional repressor REST while it increases during neuronal differentiation [31]. Hereby it was found that RMST is actually necessary for the binding of SOX2 to promoter regions of neurogenic transcription factors. SOX2, a transcription factor known to regulate neural fate, in combination with RMST were actually found to coregulate a large pool of downstream genes implicated in neurogenesis, i.e. more than 1 000 genes were differentially expressed upon RMST knockdown [31]. These results illustrated the role of RMST as a transcriptional coregulator of SOX2 and a key player in the regulation of neural stem cell fate [31]. A further confirmation of the importance of RMST came with the discovery of a homologue of this lncRNA in the simple chordate *D. vexillum*, i.e. the carpet sea-squirt [42]. While homologues of "human" lncRNAs are rarely found across all chordates due to their low levels of sequence conservation, a plausible homolog of RMST 9, the conserved region 9 of the Rhabdomyosarcoma 2 associated transcript known for its interaction with SOX2, was found in *D. vexillum*. Subsequently putative homologs were also found in the genomes of the ascidians *C. intestinalis*, *C. savignyi* and *B. schlosseri* and the Florida lancelet *B. floridae*,

illustrating that RMST lncRNA are thus conserved across chordates, making them one of the best conserved lncRNAs known to date [42].

4.4 Splices-leader RNA

mRNA 5'leader trans-splicing is a mode of gene expression in which the 5' end of a pre-mRNA is discarded and replaced by the 5' segment of a spliced leader (SL) RNA [41]. Spliced-Leader RNAs, i.e. SL RNAs, hereby consist of a 5' exon and a 3' intron with a conserved consensus 5' splice donor site at the exon-intron boundary [10]. SL RNA trans splicing has not only been described for euglenoids, kinetoplastids, cnidarians, nematodes, and Platyhelminthes [10], but also for deuterostomes like the simple chordate *C. intestinalis* [41] and the appendicularian *O. dioica* [10]. Hereby *O. dioica* was shown to not only trans-splice SL RNAs to mRNAs, as does *C. intestinalis*, but also to use trans splicing in resolving polycistronic transcripts [10]. During trans splicing, the capped SL RNA exon moiety is covalently linked to the 5' ends of mRNAs, forming a leader sequence ranging from 16 nt in *C. intestinalis* to 41 nt in trypanosomatids [10]. The role of SL trans-splicing is still unknown in many cases. SL trans-splicing may potentially having functions varying from the mediation of mRNA stability or translatability [27] and the resolution of polycistronic pre-mRNAs [1, 5], to the production of functional mRNAs from RNA polymerase I transcripts [23].

References

1. Nina Agabian. Trans splicing of nuclear pre-mrnas. *Cell*, 61(7):1157 – 1160, 1990.
2. V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl. A uniform system for microrna annotation. *RNA*, 9:277–279, 2003.
3. David P. Bartel. MicroRNA target recognition and regulatory functions. *Cell*, 136:215–233, 2009.
4. Arash Bashirullah, Ramona L Cooperstock, and Howard D Lipshitz. RNA localization. *Annu. Rev. Biochem*, 67:335–394, 1998.
5. Thomas Blumenthal. Trans-splicing and polycistronic transcription in caenohabditis elegans. *Trends in Genetics*, 11(4):132 – 136, 1995.
6. Stefania Bortoluzzi, Marta Biasiolo, and Andrea Bisognin. Microrna-offset rnas (mornas): by-product spectators or functional players? *Trends in Molecular Medicine*, 17(9):473–474, 2011.
7. Marc R. Friedländer, Sebastian D. Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microrna genes in seven animal clades. *Nucleic Acids Res.*, 40:37–52, 2012.
8. Marc R. Friedländer, Sebastian D. MacKowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52, 2012.

9. Xianghui Fu, Marcin Adamski, and Eric M. Thompson. Altered miRNA repertoire in the simplified chordate, Oikopleura dioica. *Molecular Biology and Evolution*, 25(6):1067–1080, 2008.
10. Philippe Ganot, Torben Kallesøe, Richard Reinhardt, Daniel Chourrout, and Eric M. Thompson. Spliced-leader rna trans splicing in a chordate, oikopleura dioica, with a compact genome. *Molecular and Cellular Biology*, 24(17):7795–7805, 2004.
11. Sam Griffiths-Jones, Jerome H L Hui, Antonio Marco, and Matthew Ronshaugen. MicroRNA evolution by arm switching. *EMBO reports*, 12(2):172–177, 2011.
12. A M Heimberg, R Cowper-Sal·lari, M Sémon, P C Donoghue, and Kevin J Peterson. MicroRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc. Natl. Acad. Sci. USA*, 107:19379–19383, 2010.
13. A. M. Heimberg, L. F. Sempere, V. N. Moy, P. C. J. Donoghue, and K.J. Peterson. MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl. Acad. Sci. USA*, 105:2946–2950, 2007.
14. D Hendrix, M Levine, and W Shi. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol*, 11(4):R39, 2010.
15. Jana Hertel, Manuela Lindemeyer, Kristin Missal, Claudia Fried, Andrea Tanzer, Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and The Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. *BMC Genomics*, 7:15 [epub], 2006.
16. Jana Hertel and Peter F. Stadler. The expansion of animal microRNA families revisited. *Life*, 5:905–920, 2015.
17. Nathaniel K. Jue, Paola G. Batta-Lona, Sarah Trusiak, Craig Obergfell, Ann Bucklin, Michael J. O'neill, and Rachel J. O'neill. Rapid evolutionary rates and unique genomic signatures discovered in the first reference genome for the southern ocean salp, salpa thompsoni (Urochordata, Thaliacea). *Genome Biology and Evolution*, 8(10):3171–3186, 2016.
18. Raja Keshavan, Michael Virata, Anisha Keshavan, and Robert W. Zeller. Computational Identification of Ciona intestinalis MicroRNAs. *Zoological Science*, 27(2):162–170, 2010.
19. Rie Kusakabe, Saori Tani, Koki Nishitsuji, Miyuki Shindo, Kohji Okamura, Yuki Miyamoto, Kenta Nakai, Yutaka Suzuki, Takehiro G. Kusakabe, and Kunio Inoue. Characterization of the compact bicistronic microRNA precursor, miR-1/miR-133, expressed specifically in Ciona muscle tissues. *Gene Expression Patterns*, 13(1-2):43–50, 2013.
20. David Langenberger, Clara Bermudez-Santana, Jana Hertel, Steve Hoffmann, Philipp Khaitovich, and Peter F. Stadler. Evidence for human microrna-offset rnas in small rna sequencing data. *Bioinformatics*, 25(18):2298–2301, 2009.
21. David Langenberger, Clara Bermudez-Santana, Peter F. Stadler, and Steve Hoffmann. Identification and classification of small RNAs in transcriptome sequence data. *Pac. Symp. Biocomput.*, 15:80–87, 2010.
22. David Langenberger, M. Volkan Çakir, Steve Hoffmann, and Peter F. Stadler. Dicer-processed small RNAs: Rules and exceptions. *J. Exp. Zool: Mol. Dev. Evol.*, 320:35–46, 2012.
23. Mary Gwo-Shu Lee and Lex H. T. Van der Ploeg. Transcription of protein-coding genes in trypanosomes by rna polymerase i. *Annual Review of Microbiology*, 51(1):463–489, 1997. PMID: 9343357.
24. Matthieu Legendre, André Lambert, and Daniel Gautheret. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 21(7):841–845, 2005.
25. Shi-Lung Lin, Joseph D. Miller, and Shao-Yao Ying Ying. Intronic MicroRNA (mirna). *J Biomed Biotechnol.*, 2006:26818, 2006.
26. E Lund, S Güttinger, A Calado, J E Dahlberg, and U Kutay. Nuclear export of microRNA precursors. *Science*, 303:95–98, 2004.
27. P A Maroney, J A Denker, E Darzynkiewicz, R Laneve, and T W Nilsen. Most mrnas in the nematode ascaris lumbricoides are trans-spliced: a role for spliced leader addition in translational efficiency. *RNA*, 1(7):714–723, 1995.

28. Eti Meiri, Asaf Levy, Hila Benjamin, Miriam Ben-David, Lahav Cohen, Avital Dov, Nir Dromi, Eran Elyakim, Noga Yerushalmi, Orit Zion, Gila Lithwick-Yanai, and Einat Sitbon. Discovery of microrna and other small rnas in solid tumors. *Nucleic Acids Research*, 38(18):6234–6246, 2010.
29. Kristin Missal, Dominic Rose, and Peter F. Stadler. Non-coding RNAs in Ciona intestinalis. *Bioinformatics*, 21(SUPPL. 2):77–78, 2005.
30. Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29:2933–2935, 2013.
31. Shi-Yan Ng, Gireesh K. Bogu, Boon Seng Soh, and Lawrence W. Stanton. The long noncoding rna rmst interacts with sox2 to regulate neurogenesis. *Molecular Cell*, 51(3):349–359, 2013.
32. Trina M Norden-Krichmar, Janette Holtz, Amy E Pasquinelli, and Terry Gaasterland. Computational prediction and experimental validation of Ciona intestinalis microRNA genes. *BMC genomics*, 8(1):445, 2007.
33. A. E. Pasquinelli, A. McCoy, E. Jimenez, E. Salo, G. Ruvkun, M. Q. Martindale, and J. Baguna. Expression of the 22 nucleotide let-7 heterochronic RNA throughout the Metazoa: a role in life history evolution? *Evol. Dev.*, 5(4):372–378, 2003.
34. a E Pasquinelli, B J Reinhart, F Slack, M Q Martindale, M I Kuroda, B Maller, D C Hayward, E E Ball, B Degnan, P Müller, J Spring, a Srinivasan, M Fishman, J Finnerty, J Corbo, M Levine, P Leahy, E Davidson, and G Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, 2000.
35. L F Sempere, C N Cole, M A McPeek, and K J Peterson. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zoolog B Mol Dev Evol.*, 306B:575–588, 2006.
36. Weiyang Shi, David Hendrix, Mike Levine, and Benjamin Haley. A distinct class of small RNAs arises from pre-miRNA–proximal regions in a simple chordate. *Nature Structural & Molecular Biology*, 16(2):183–189, 2009.
37. B J Swalla and W R Jeffery. A maternal RNA localized in the yellow crescent is segregated to the larval muscle cells during ascidian development., 1995.
38. Ryan J Taft, Ken C Pang, Timothy R Mercer, Marcel Dinger, and John S Mattick. Non-coding rnas: regulators of disease. *The Journal of Pathology*, 220(2):126–139, 2010.
39. Goro Terai, Hiroaki Okida, Kiyoshi Asai, and Toutai Mituyama. Prediction of Conserved Precursors of miRNAs and Their Mature Forms by Integrating Position-Specific Structural Features. *PLoS ONE*, 7(9):1–11, 2012.
40. Jennifer L. Umbach and Bryan R. Cullen. In-depth analysis of kaposi's sarcoma-associated herpesvirus microrna expression provides insights into the mammalian microrna-processing machinery. *Journal of Virology*, 84(2):695–703, 2010.
41. Amanda E. Vandenberghe, Thomas H. Meedel, and Kenneth E.M. Hastings. mrna 5-leader trans-splicing in the chordates. *Genes & Development*, 15(3):294–303, 2001.
42. Cristian A. Velandia-Huerto, Adriaan A. Gittenberger, Federico D. Brown, Peter F. Stadler, and Clara I. Bermúdez-Santana. Automated detection of ncrnas in the draft genome sequence of a colonial tunicate: the carpet sea squirt didemnum vexillum. *BMC Genomics*, 17(1):691, Aug 2016.
43. Kai Wang, Christelle Dantec, Patrick Lemaire, Takeshi A. Onuma, and Hiroki Nishida. Genome-wide survey of miRNAs and their evolutionary history in the ascidian, Halocynthia roretzi. *BMC Genomics*, 18(1):314, 2017.
44. B M Wheeler, A M Heimberg, V N Moy, E A Sperling, T W Holstein, S Heber, and K J Peterson. The deep evolution of metazoan microRNAs. *Evol. Dev.*, 11:50–68, 2009.
45. Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, 3(4):680–691, 2007.