Non-protein-coding RNAs as a regulators of development in tunicates

Cristian Velandia

January 12, 2018

miRNA families origin and evolutionary perspective

miRNAs in clusters

Methods

The starting point to detect the miRNA clusters on those species was the generation of GFF3 files with the genome coordinates of the miRNA elements. At the same time, current genome annotations from each specie was retrieved in order to determine at the same time, the position of annotated elements on those genomes. Given those elements, for each one that are not part of miRNAs annotation was designed as $\mathcal{P}_j$ elements, and in another way ones that are annotated as miRNAs, was identified as $\mathcal{M}_i$. In this way, ordered elements $\mathcal{M}_i$ and $\mathcal{M}_{i+1}$ are part of a cluster $\mathcal{C}_k \neg \mathcal{M}_i \prec \mathcal{P}_j \prec \mathcal{M}_{i+1}$. in this way each $\mathcal{C}_k$ cluster was identified indepentendly and it was composed by sets of miRNA elements M_i with a $|P_k| \geq 2$. The M_i elements have a correspondent index T_i that maps into the hexadecimal alphabet. Those T_i elements could be aligned in the next steps by an implementation of a Needleman-Wunsch [**?**] algorithm, which have taken into account: $1:1$ matches and insertions $1:-$ or $-:1$, but not $1:1$ (mis)-matches, where the alignment program penalizes harder than a insertion. The alignment score was calculated by this score matrix:

$$D_{ij} = \max \begin{cases} D_{i-1,j-1} + 2 & \text{Matches} \\ D_{i-1,j-1} - (4) & \text{(Mis)-matches} \\ D_{i-1,j} - (1) & \text{Insertion}(a) \\ D_{i,j-1} - (1) & \text{Insertion}(b) \end{cases} \tag{1}$$

In this case, all the sequences that contains at least one M_i miRNA family represented by T_i were collected and next, aligned in a pairwise way with all possible combinations of sequences. All the resulting alignments for each T_i index and cleaned, taking only ones that reported alignments scores ($\mathcal{G}$) $\geq$ third quartile of the data (located in the 75% or greater percentage of the score distribution), creating a subset $\mathcal{B}$ with the best scored pairwise alignments.

With $\mathcal{B}$ an implementation of multiple alignments, implemented by Reztlaff (2017) was applied in order to detect the conserved blocks in all pairwise alignments. In this way the implementation allowed the detection of those conserved blocks of $M_i \subseteq \mathcal{B}$.

Results

Applying the last strategy to detect miRNA's clusters granted the option to study the conserved elements along chordate's genomes. As shown in Figure 1, directly with the location form those miRNA elements have been possible to identify the number and the length of those identified regions along all the studied genomes. In this case, the cluster that contains the greatest number of miRNAs elements (60) is located on *D. rerio* genome: `Chromosome` `4:28738556-28754891`, for tunicates on *C. intestinalis*: `Chromosome` `7:4153284-4156782` with 23 elements, and in *B. floridae*: `Bf_V2_118:` `216744-220351` only 5 miRNAs have been detected (for further details Table 1 describe the miRNAs families located inside the largest clusters for each specie). Additionally, is more frequent to found clusters that contain 2 miRNAs families, but is also important to know that the relation between the number of miRNAs inside a cluster and the cluster's length is higher on tunicates and

cephalochordates in comparison to vertebrates, it means that inside the identified clusters along vertebrates the miRNAs elements are more distant between them.

Clade	Specie	Chr	Start	End	Size(Mb)	No. miRNAs	Elements
C	*B. floridae*	Bf_V2_118	216744	220351	3607	5	bfl-mir-4869, bfl-mir-4857, bfl-mir-4862, bfl-mir-4856b, bfl-mir-4856a
T	*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
T	*B. schlosseri*	chrUn	40003	41320	1317	2	mir-233, mir-10
T	*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
T	*C. savignyi*	reftig_16	3924783	3925336	553	3	csa-mir-216b, csa-mir-216a, csa-mir-217
T	*C. savignyi*	reftig_1	1335375	1336487	1112	3	csa-mir-92b, csa-mir-92c, csa-mir-92a

V	*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

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

Table 1: Details of biggest miRNA cluster for chordate species

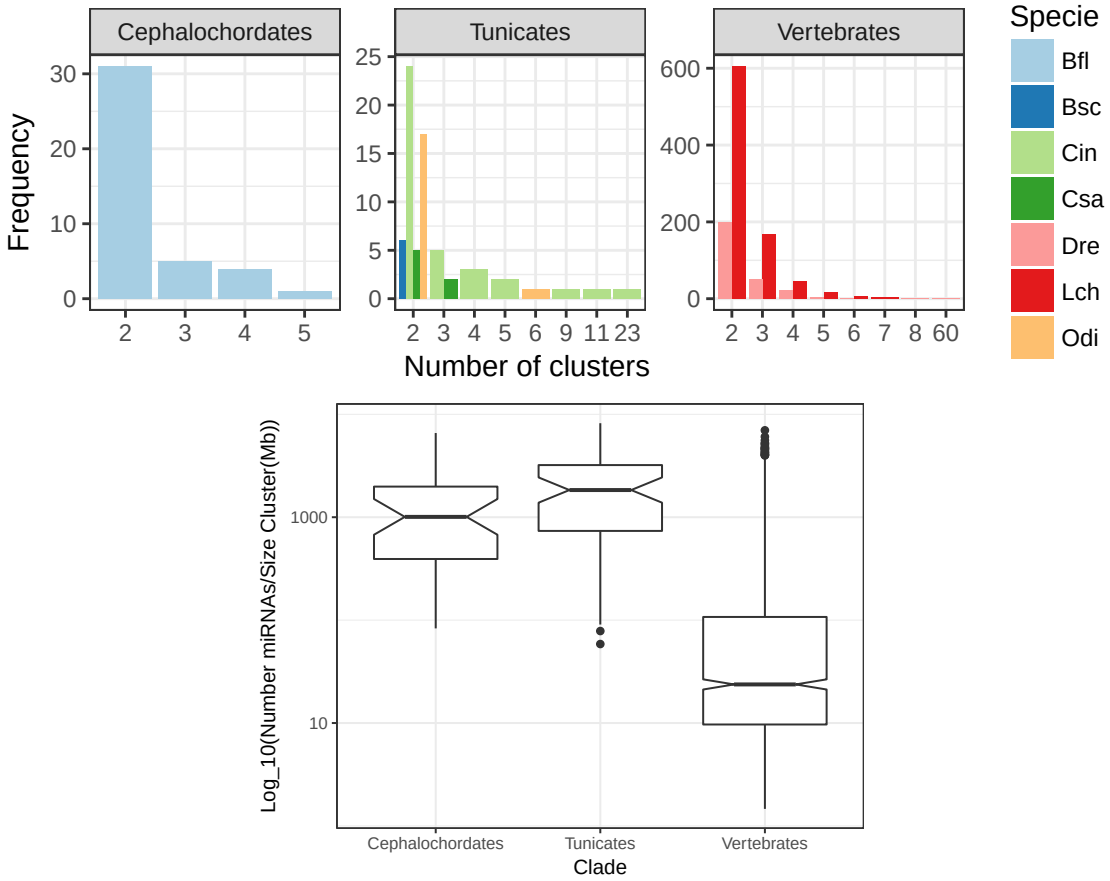

Figure 1: Analysis of the distribution, size and number's of cluster along chordate species.

Comparison between miRNA clusters have been calculated in order to access to the most conserved set of miRNA families inside chordate's clusters. As a result, the following miRNAs families were identified being part of specific clusters regions: let-7, mir-1, MIR1122, mir-130, mir-132, mir-133, mir-135, mir-146, mir-15, mir-17, mir-181, mir-183, MIR1846, mir-186, mir-19, mir-193, mir-216, mir-219, mir-23, mir-24, mir-242, mir-25, mir-27, mir-286, mir-29, mir-2985-2, mir-30, mir-34, mir-395, mir-454, mir-489, MIR535, mir-8, mir-9 (Figure **??**).

Results

The distribution of miRNAs families on let-7 cluster are spanning all the studied chordate species, it is evident that in vertebrate species (*D. rerio* and *L. chalumnae*) exists more than one let-7 cluster, expanding the loci definition that is not restricted only for one element but for a cluster of elements that reports different length distributions. Is important to note that let-7 is clustered sometimes with another let-7 element or with another miRNA's families. The distribution reported on amphioxus is composed by 2 let-7 and 2 mir-10, 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: 1let-7 + 2 mir-10 and 2 let-7 + other families. In this way, tunicates reported the latter group, not including mir-10 on the cluster but including mir-233 or even mir-1473.

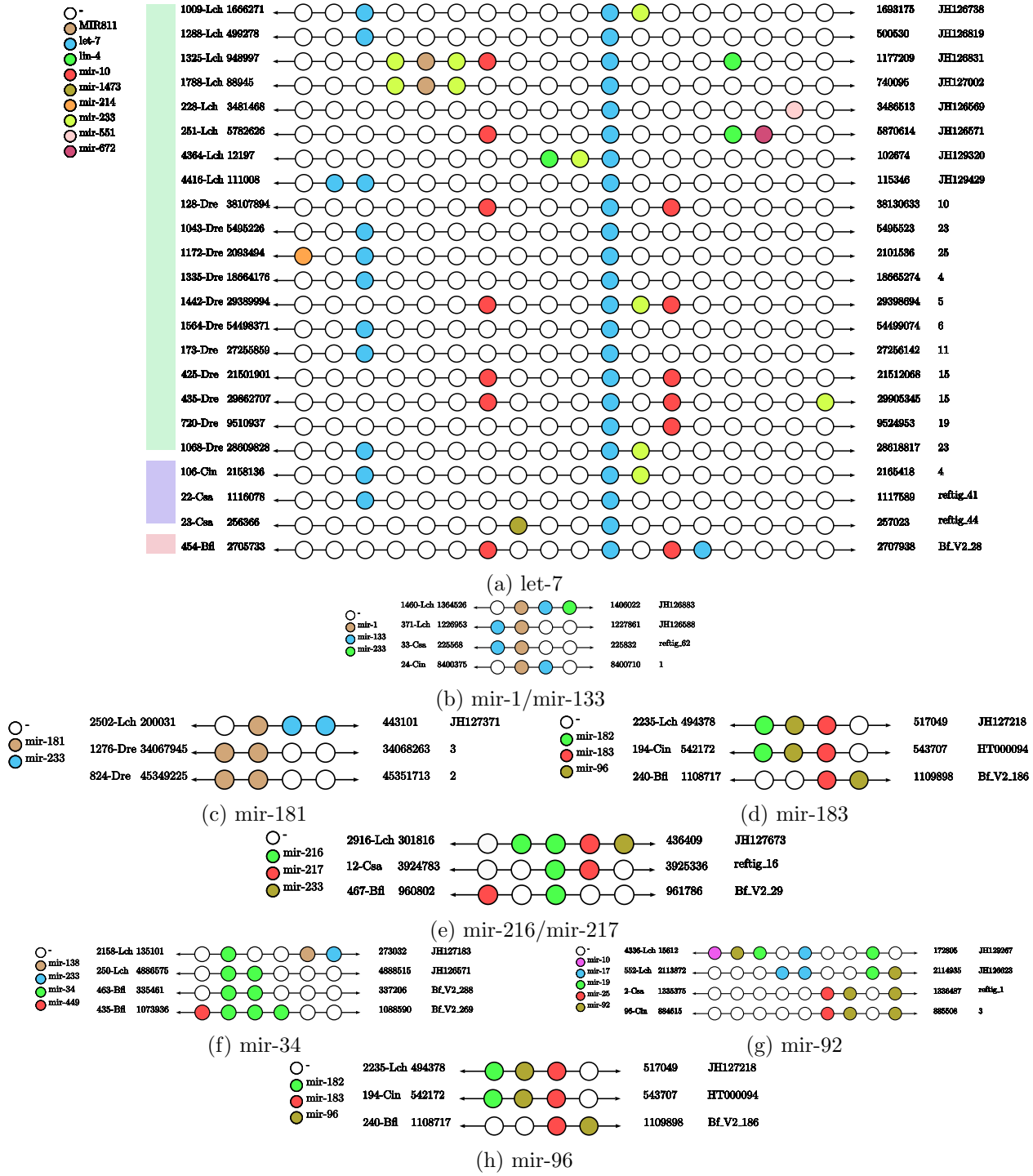

Figure 2: Multiple alignments of miRNA's clusters. **Prot**: Protostomata, **Brfl**: *B. floridae*, **Oidi**: *O. dioica*, **Dvex**: *D. vexillum*, **Ciin**: *C. intestinalis*, **Cisa**: *C. savignyi*, **Ciro**: *C. robusta*, **Sath**: *S. thompsoni*, **Mata**: *M. oculata*, **Mlta**: *M. occulta*, **Mlis**: *M. occidentalis*, **Bosc**: *B. schlosseri*, **Haro**: *H. roretzi*, **Pema**: *P. marinus*, **Dare**: *D. rerio*, **Lach**: *L. chalumnae*, **Xetr**: *X. tropicalis* and **Anca**: *A. carolinensis*.

To complete the tree of loss and gain of families

Methods

The initial eight chordates genomes from chordate species: *Branchiostoma floridae*, *Botryllus schlosseri*, *Ciona intestinalis*, *Ciona savignyi*, *Danio rerio*, *Didemnum vexillum*, *Latimeria chalumnae* and *Oikopleura dioica*, were analyzed through homology searches at sequence level and a posterior validation by secondary structure

alignments against pre-build metazoan-covariance models, as reported by [8]. The final set of miRNAs was generated on a GFF file, reporting all the genome coordinates from each miRNA element. Given these information, miRNA families names were obtained from miRBase, with the miFam.dat file, which contains the relationships between miRNA specific annotations for each specie and their correspondent miRNA family. These annotations were compared against the last report of miRNAs families reported by [4]. In case that the obtained miRNA family was not included neither miRBase or miRNAs matrix, a new label were designed to detect those specific elements. At the same time, from the reported matrix was considered the reported families from the following vertebrates: *Anolis carolinensis*, *Petromyzon marinus* and *Xenopus tropicalis*. At the same time, two new reports of miRNAs on tunicates were included on this matrix from: *Salpa thompsoni* [5] and *Halocynthia roretzi* [9]. At now, three new genomes from the *Molgula sp.* genus were reported [7] and the genome sequences have been obtained from ANISEED [1] [1].

For the latter species, homology BLAST searches were performed. All hairpin sequences from miRBase (v. 21) [6] were used as queries against those genomes. After that, in order to obtain the best miRNA candidates, mature sequences were searched on previously detected hairpin candidates. This strategy also include the new reported genome from *Ciona robusta* reported on ANISEED.

In this way, the initial miRNA matrix from [4] was updated with the information retrieved from detection of miRNAs families and also, with the inclusion from candidates in new reported genomes (*H. roretzi*, *S. thompsoni*, *M. occidentalis*, *M. occulta* and *M. oculata*). Moreover, the phylogenetic distribution from Tunicate clade have been obtained from [3]. And the final tree has been completed with the inclusion of one cephalochordata (*B. floridae*) and five vertebrates (*A. carolinensis*, *D. rerio*, *L. chalumnae*, *P. marinus* and *X. tropicalis*).

Additionally, in the final matrix only families that have presence in at least two species were considered, except for the miRNAs families that belongs from Protostomata group (Prot). This updated matrix and the phylogenetic distribution of the species in Newick format were the input files for Count program [2] in order to reconstruct the corresponding miRNAs family history by the implemented Dollo parsymony.

Results

The updated matrix of miRNAs reported 691 families, including the miRBase annotations and the homology predictions validated by secondary structure alignments as shown in Figure **??**. For tunicates miRNAs, the majority have been detected/annotated in *C. intestinalis* (now referred as *C. robusta*) where exists miRBase annotations for specie-specific miRNAs families and also were annotated a set of miRNAs validated by secondary structure comparisons against metazoan-specific covariance models Cite Cristian Master Thesis. Along this distribution of miRNAs families, 2 families are conserved along all the species, even in Protostomata clade and vertebrates. Also, additional 8 families complement the 10 most conserved set of miRNAs that have been detected by this strategy, as follows: mir-124, mir-8, mir-153, mir-1, mir-216, mir-190, mir-133 and mir-31. From the most conserved miRNAs families *O. dioica* and *S. thompsoni* has been lost about 40% and 50% of those conserved families, respectively.

At the same time, the state of those conserved elements in Protostomata, Cephalochordata and Vertebrata show evident lost about $\sim 10\%$ of the families. The general trend in Craniata shows an increment of conserved families, sometimes with the possibility of trace the presence from *P. marinus* and identifying 16 conserved candidates along all species in the clade that are not reported in another clades (mir-15, -181, -23, -199, -24, -204, -128, -221, -205, -192, -132, -138, -145, -143, -451 and -456). Specifically, exists 44 miRNAs that have been identified at least on 2 species of tunicates but not in Cephalochordata or Vertebrata [2], but from those candidates only 7 (mir-1497, -281, -1473, -200, -92, -1502 and -4079) are previously reported as tunicate specific, the other ones have been reported also in vertebrates (mir-1277, -297, -3533, -466, -467, -568, -374, -450, -876, -8915, -3149, -355, -340, -553) and in insects (mir-3 and mir-11). In complement Figure 5, compares the shared families of miRNAs in all the studied chordate species. Is important to annotate that the candidates from *P. marinus*, *D. rerio*, *L. chalumnae* and *A. carolinensis* have been grouped with the Vertebrates label. In this

[1] `https://www.aniseed.cnrs.fr/`

[2] mir-368, mir-450, mir-3, mir-340, mir-335, mir-297,mir-466, mir-287, mir-11, mir-374, mir-664, mir-467, mir-568, mir-876, mir-1497, mir-281, mir-553, mir-200, mir-1473, mir-1469, mir-92_2, mir-1502, mir-4079, mir-3149, mir-1277, mir-8915, mir-3533, cin-mir-4034, cin-mir-4047, cin-mir-4049, cin-mir-4052, cin-mir-4053, cin-mir-4054, cin-mir-4065, cin-mir-4072, cin-mir-4086, cin-mir-4093, cin-mir-4101, cin-mir-4123, cin-mir-4171, cin-mir-4220, cin-mir-4000c, cin-mir-4010, cin-mir-4029

case, in overall is possible to found 47 miRNAs families that are at least in one specie in Vertebrata clade and is not shared with the other chordates. The homology searches allowed scan the annotated miRNAs families on new reported genomes assemblies from tunicates, as *H. roretzi* and *S. thompsoni*, including additional possible candidates to the reported set, that in case of *H. roretzi* has been reported as experimental candidates from *C. intestinalis* reported families. At the same time, *D. vexillum* share a high number of elements with *B. schlosseri* and species included on Vertabrata (9) and specifically with vertebrates (13). At the same time, remain a set of candidates old candidates that are shared between *Protostomata*, *B. floridae* and *Vertebrata*: mir-193, mir-375 and mir-182. Different to the set of candidates those are basal in this study: *Protostomata* and *B. floridae*, whose share 2 families: mir-71 and mir-242_2.

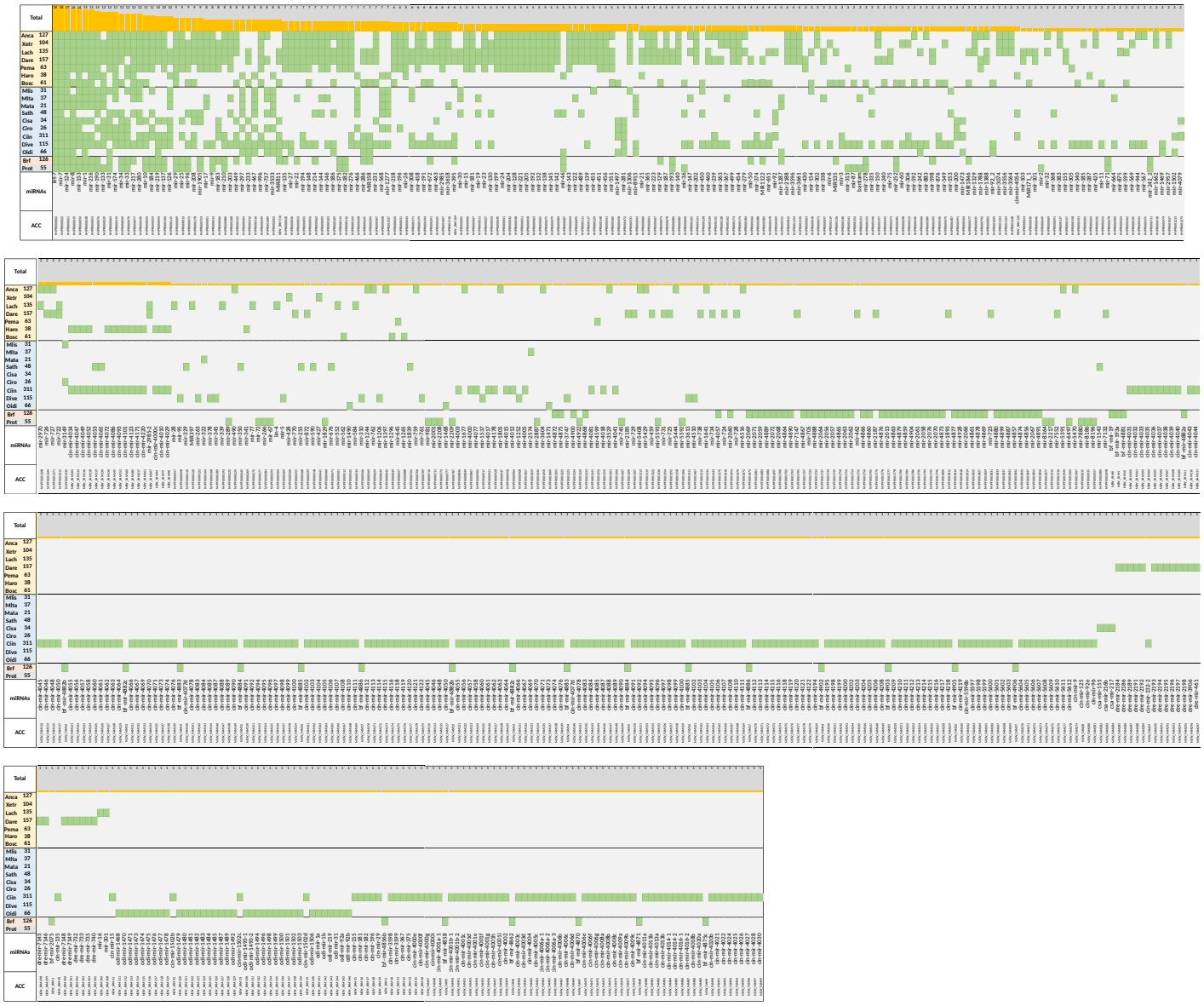

Figure 3: Absence/Presence Matrix of miRNAs families along Bilaterian species. **Prot**: Protostomata, **Brfl**: *B. floridae*, **Oidi**: *O. dioica*, **Dvex**: *D. vexillum*, **Ciin**: *C. intestinalis*, **Cisa**: *C. savignyi*, **Ciro**: *C. robusta*, **Sath**: *S. thompsoni*, **Mata**: *M. oculata*, **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*.

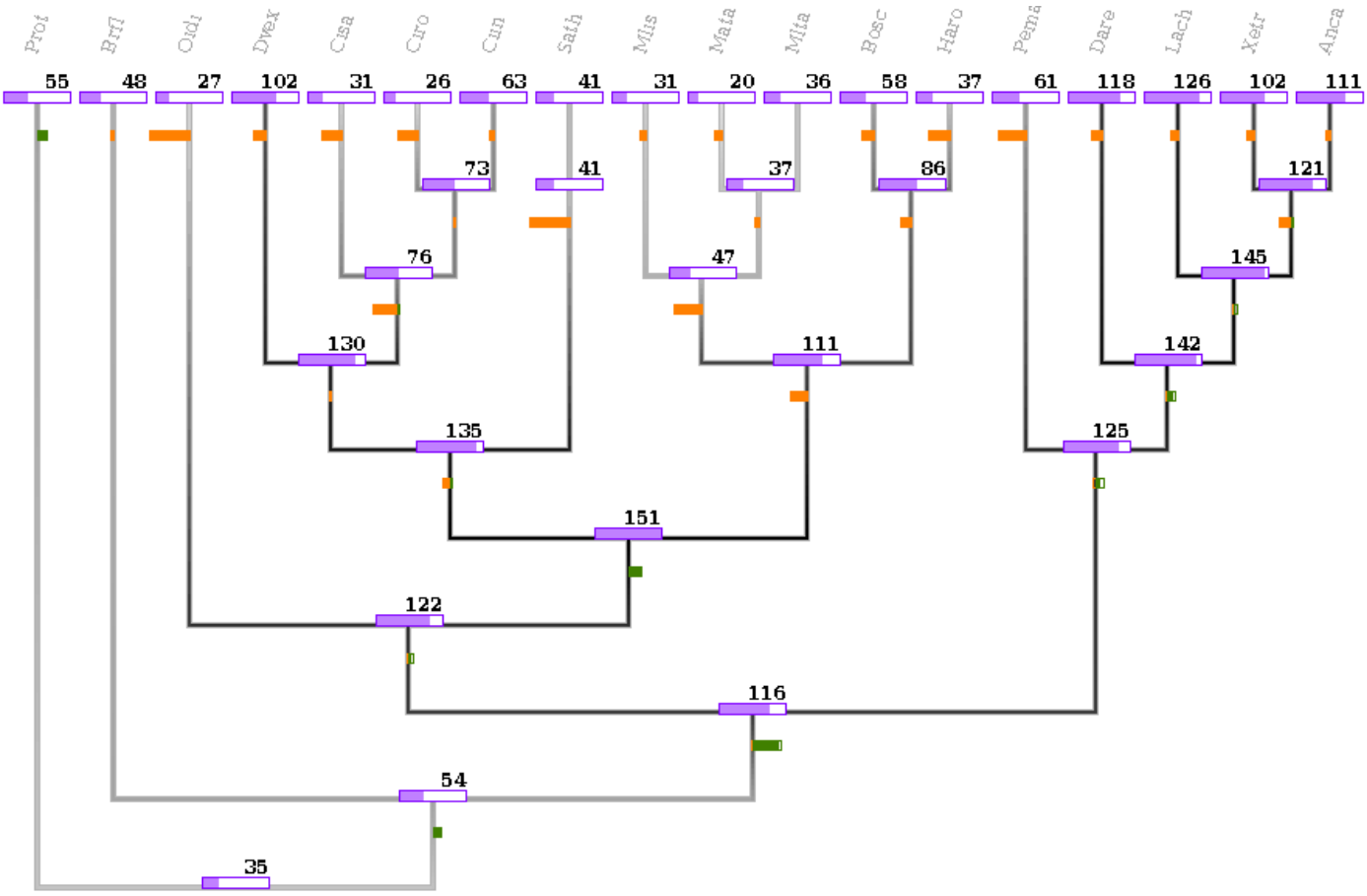

Figure 4: Dollo parsymony of miRNAs families distribution in some chordates genomes

.

References

[1] Matija Brozovic, Christelle Dantec, Justine Dardaillon, Delphine Dauga, Emmanuel Faure, Mathieu Gineste, Alexandra Louis, Magali Naville, Kazuhiro R. Nitta, Jacques Piette, Wendy Reeves, Céline Scornavacca, Paul Simion, Renaud Vincentelli, Maelle Bellec, Sameh Ben Aicha, Marie Fagotto, Marion Guéroult-Bellone, Maximilian Haeussler, Edwin Jacox, Elijah K. Lowe, Mickael Mendez, Alexis Roberge, Alberto Stolfi, Rui Yokomori, C. Titus Brown, Christian Cambillau, Lionel Christiaen, Frédéric Delsuc, Emmanuel Douzery, Rémi Dumollard, Takehiro Kusakabe, Kenta Nakai, Hiroki Nishida, Yutaka Satou, Billie Swalla, Michael Veeman, Jean-Nicolas Volff, and Patrick Lemaire. Aniseed 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. *Nucleic Acids Research*, page gkx1108, 2017.

[2] Miklós Csűös. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, 2010.

[3] Frederic Delsuc, Herve Philippe, Georgia Tsagkogeorga, Paul Simion, Marie-Ka Tilak, Xavier Turon, Susanna Lopez-Legentil, Jacques Piette, Patrick Lemaire, and Emmanuel J. P. Douzery. A phylogenomic framework and timescale for comparative genomics and evolutionary developmental biology of tunicates. *bioRxiv*, 2017.

[4] Jana Hertel and Peter Stadler. The Expansion of Animal MicroRNA Families Revisited. *Life*, 5(1):905–920, 2015.

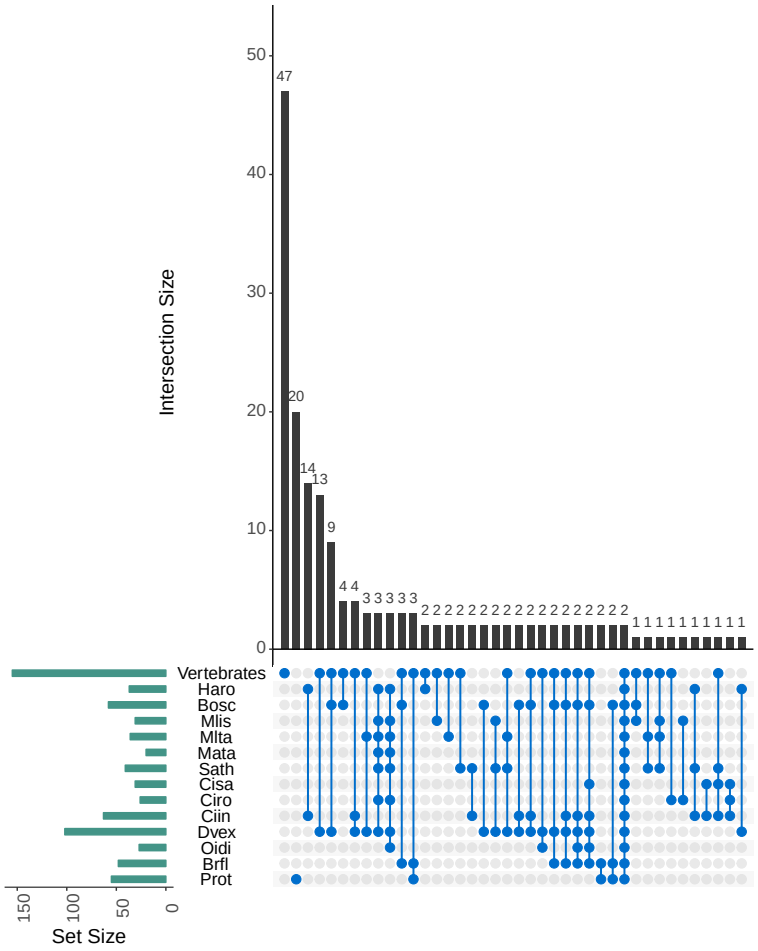

Figure 5: Comparsion between miRNAs families along Bilaterian species. Same labels from Figure 3 were used. In this case *vertebrates* group the following species: **Pema**: *P. marinus*, **Dare**: *D. rerio*, **Lach**: *L. chalumnae*, **Xetr**: *X. tropicalis* and **Anca**: *A. carolinensis*

.

[5] 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.

[6] Ana Kozomara and Sam Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73, 2014.

[7] Alberto Stolfi, Elijah K Lowe, Claudia Racioppi, Filomena Ristoratore, C Titus Brown, Billie J Swalla, and Lionel Christiaen. Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *eLife*, 3:e03728, sep 2014.

[8] C. A. Velandia-Huerto, A. A. Gittenberger, F. D. Brown, P. F. Stadler, and C. I. Bermudez-Santana. Automated detection of ncRNAs in the draft genome sequence of a colonial tunicate: the carpet sea squirt Didemnum vexillum. *BMC Genomics*, 17:691, Aug 2016.

[9] 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.