Workflows for the Large-Scale Assessment of miRNA Evolution

Birth and Death of miRNA Genes in Tunicates

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt von

Master of Bioinformatics Cristian Arley Velandia Huerto
geboren am 21.09.1990 in Bogotá, Kolumbien

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Peter F. Stadler, Universität Leipzig, Leipzig, Germany
2. Professor Dr. Cedric Notredame, Center for Genomic Regulation
Barcelona, Spain

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 27.06.2022 mit dem Gesamtprädikat magna cum laude

Bibliographic Description

Title:	Workflows for the Large-Scale Assessment of miRNA Evolution
Subtitle:	Birth and Death of miRNA Genes in Tunicates
Type:	Dissertation
Author:	Cristian Arley Velandia Huerto
Year:	2022
Professional discipline:	Computer Science
Language:	English
Pages in the main part:	143
Chapter in the main part:	7
Number of Figures:	51
Number of Tables:	22
Number of Appendices:	4
Number of Citations:	327
Key Words:	microRNAs, Homology, Tunicates, Evolution, Synteny

This thesis is based on the following publications.

E. Parra-Rincón*, C. A. Velandia-Huerto*, A. Gittenberger, J. Fallmann, T. Gatter, F. D. Brown, P. F. Stadler und C. I. Bermúdez-Santana (Dez. 2021). „The Genome of the "Sea Vomit" Didemnum vexillum". In: *Life* 11.12, S. 1377. DOI: 10.3390/life11121377.

C. A. Velandia-Huerto, F. D. Brown, A. Gittenberger, P. F. Stadler und C. I. Bermúdez-Santana (Juli 2018). „Nonprotein-Coding RNAs as Regulators of Development in Tunicates". In: *Results and Problems in Cell Differentiation*. Hrsg. von M. Kloc und J. Kubiak. Bd. 65. Results Probl Cell Differ. Cham: Springer International Publishing, S. 197–225. DOI: 10.1007/978-3-319-92486-1_11.

C. A. Velandia-Huerto, J. Fallmann und P. F. Stadler (Feb. 2021). „miRNAture–Computational Detection of microRNA Candidates". In: *Genes* 12.3, S. 348. DOI: 10.3390/genes12030348.

C. A. Velandia-Huerto, J. Fallmann und P. F. Stadler (in prep.). „The bona fide miRNA complement on tunicates, based on an automatizated homology approach".

C. A. Velandia-Huerto, A. M. Yazbeck, J. Schor und P. F. Stadler (2022). „Evolution and Phylogeny of MicroRNAs — Protocols, Pitfalls, and Problems“. In: *miRNomics: MicroRNA Biology and Computational Analysis*. Hrsg. von J. Allmer und M. Yousef. New York, NY: Springer US, S. 211–233. ISBN: 978-1-0716-1170-8. DOI: 10.1007/978-1-0716-1170-8_11.

*The authors share first authorship.

Abstract

As described over 20 years ago with the discovery of the RNA interference (RNAi), double-stranded RNAs occupied key roles in regulation and as defense-line in animal cells. This thesis focuses on metazoan microRNAs (miRNAs). These small non-coding RNAs are distinguished from their small-interfering RNA (siRNA) relatives by their tightly controlled, efficient and flexible biogenesis, together with a broader flexibility to target multiple mRNAs by a seed imperfect base-pairing. As potent regulators, miRNAs are involved in mRNA stability and post-transcriptional regulation tasks, being a conserved mechanism used repetitively by the evolution, not only in metazoans, but plants and unicellular organisms.

Through a comprehensive revision of the current animal miRNA model, the canonical pathway dominates the extensive literature about miRNAs, and served as a scaffold to understand the scenes behind the regulation landscape performed by the cell. The characterization of a diverse set of non-canonical pathways has expanded this view, suggesting a diverse, rich and flexible regulation landscape to generate mature miRNAs. The production of miRNAs, derived from isolated or clustered transcripts, is an efficient and highly conserved mechanism traced back to animals with high fidelity at family level. In evolutionary terms, expansions of miRNA families have been associated with an increasing morphological and developmental complexity. In particular, the Chordata clade (the ancient cephalochordates, highly derived and secondary simplified tunicates, and the well-known vertebrates) represents an interesting scenario to study miRNA evolution. Despite clear conserved miRNAs along those clades, tunicates display massive restructuring events, including emergence of high derived miRNAs.

As shown in this thesis, model organisms or vertebrate-specific bias exist on current animal miRNA annotations, misrepresenting more diverse groups, such as marine invertebrates. Current miRNA databases, such as `miRBase` and `Rfam`, classified miRNAs under different definitions and possessed annotations that are not simple to be linked. As an alternative, this thesis proposes a method to curate and merge those annotations, making use of `miRBase` precursor/mature annotations and genomes together with `Rfam` predicted sequences. This approach generated structural models for shared miRNA families, based on the alignment of their correct-positioned mature sequences as anchors. In this process, the developed structural curation steps flagged 33 miRNA families from the `Rfam` as questionable.

Curated `Rfam` and `miRBase` anchored-structural alignments provided a rich resource for constructing predictive miRNA profiles, using correspondent hidden Markov (HMMs)

and covariance models (CMs). As a direct application, the use of those models is time-consuming, and the user has to deal with multiple iterations to achieve a genome-wide non-overlapping annotation. To resolve that, the proposed `miRNAture` pipeline provides an automatic and flexible solution to annotate miRNAs. It combines multiple homology approaches to generate the best candidates validated at sequence and structural levels. It increases the achievable sensitivity to annotate canonical miRNAs, and the evaluation against the human annotation shows that clear false positive calls are rare and additional counterparts are lying on retained-introns, transcribed lncRNAs or repeat families. Further development of `miRNAture` suggests an inclusion of multiple rules to distinguish non-canonical miRNA families.

This thesis describes multiple homology approaches to annotate the genomic information from a non-model chordate: the colonial tunicate *Didemnum vexillum*. Detected high levels of genetic variance and unexpected levels of DNA degradation were evidenced through a comprehensive analysis of genome-assembly methods and gene annotation. Despite those challenges, it was possible to find candidate homeobox and skeletogenesis-related genes. On its own, the ncRNA annotation included expected conserved families, and an extensive search of the Rhabdomyosarcoma 2-associated transcript (RMST) lncRNA family traced-back at the divergence of deuterostomes. In addition, a complete study of the annotation thresholds suggested variations to detect miRNAs, later implemented on the `miRNAture` tool. This chapter is a showcase of the usual workflow that should follow a comprehensive sequencing, assembly and annotation project, in the light of the increasing research approaching DNA sequencing.

In the last 10 years, the remarkable increment in tunicate sequencing projects boosted the access to an expanded miRNA annotation landscape. In this way, a comprehensive homology approach annotated the miRNA complement of 28 deuterostome genomes (including current 16 reported tunicates) using `miRNAture`. To get proper structural models as input, corrected `miRBase` structural alignments served as a scaffold for building correspondent CMs, based on a developed genetic algorithm. By this means, this automatic approach selected the set of sequences that composed the alignments, generating 2492 miRNA CMs. Despite the multiple sources and associated heterogeneity of the studied genomes, a clustering approach successfully gathered five groups of similar assemblies and highlighted low quality assemblies. The overall family and loci reduction on tunicates is notorious, showing on average 374 microRNA (miRNA) loci, in comparison to other clades: Cephalochordata (2119), Vertebrata (3638), Hemichordata (1092) and Echinodermata (2737). Detection of 533 miRNA families on the divergence of tunicates shows an expanded landscape regarding currently miRNA annotated families. Shared sets of ancestral, chordates, Olfactores, and specific clade-specific miRNAs were uncovered using a phylogenetic conservation criteria. Compared to current annotations, the family repertories were expanded in all cases. Finally, relying on the adjacent elements from annotated miRNAs, this thesis proposes an additional syntenic support to cluster miRNA loci. In this way, the structural alignment of miR-1497, originally annotated in three model tunicates, was expanded with a clear syntenic support on tunicates.

Zusammenfassung

Wie bereits vor über 20 Jahren mit der Entdeckung der RNA-Interferenz (RNAi) beschrieben wurde, haben doppelsträngige RNAs Schlüsselrollen als Vermittler der Regulation und als Verteidigungslinie in tierischen Zellen. In dieser Arbeit geht es um microRNAs (miRNAs) aus Metazoen. Diese kleinen nicht-kodierenden RNAs unterscheiden sich von ihren Verwandten, den small-interfering RNAs (siRNAs), durch ihre streng kontrollierte, effiziente und flexible Biogenese sowie eine größere Flexibilität bei der Ausrichtung auf mehrere mRNAs durch eine unvollkommene Basenpaarung. Als wirksame Regulatoren sind miRNAs an der mRNA-Stabilität und der posttranskriptionellen Regulierung beteiligt und sind ein konservierter Mechanismus, der von der Evolution nicht nur in Metazoen, sondern auch in Pflanzen und Einzellern immer wieder genutzt wurde.

Durch eine umfassende Überarbeitung des derzeitigen miRNA-Modells für Tiere hat sich gezeigt, dass der der kanonische Signalweg die umfangreiche Literatur über miRNAs dominiert. Dies diente als Gerüst, um die Hintergründe der von der Zelle durchgeführten Regulation zu verstehen. Die Charakterisierung einer Reihe von nicht-kanonischen Signalwegen hat diese Sichtweise erweitert und deutet auf eine vielfältige, reichhaltige und flexible Regulierungslandschaft hin, um reife miRNAs zu erzeugen. Die Produktion von miRNAs, die aus isolierten oder gebündelten Transkripten abgeleitet sind, ist ein effizienter und hochkonservierter Mechanismus, der bei Tieren zuverlässig bis auf die Familienebene zurück verfolgbar ist. Aus evolutionärer Sicht ist die Zunahme von miRNA-Familien assoziiert mit zunehmender Komplexität auf morphologischer und auf Entwicklungsebene. Insbesondere die Gruppe der Chordata (die alten Cephalochordaten, hochgradig abgeleitete und sekundär vereinfachte Manteltiere und die bekannten Wirbeltiere) stellt ein interessantes Szenario zur Untersuchung der miRNA-Evolution dar.

Trotz eindeutig konservierter miRNAs entlang dieser Kladen zeigen Manteltiere massive Umstrukturierungsereignisse, einschließlich der Entstehung von hoch abgeleiteten miRNAs. Wie in dieser Arbeit gezeigt wird, bestehen model- und vertebraten-spezifische Verzerrungen bei der aktuellen miRNA-Annotationen für Tiere, wodurch Annotationen für vielfältigere Gruppen, wie den wirbellosen Meerestieren, verfälscht werden. Aktuelle miRNA-Datenbanken, wie miRBase und `Rfam`, klassifizieren miRNAs nach unterschiedlichen Definitionen und verfügen über Annotationen, die sich nicht einfach verknüpfen lassen. Als Alternative wird in dieser Arbeit eine Methode vorgeschlagen, um Annotationen zu kuratieren und zusammenzuführen, wobei die miRBase-Vorläufer/Reife Annotationen und Genome zusammen mit den von `Rfam` vorhergesagten Sequenzen genutzt werden. Dieser Ansatz generiert strukturelle Modelle für gemeinsame miRNA-Familien, basierend auf

dem Alignment ihrer korrekt positionierten reifen Sequenzen als Anker. In diesem Prozess identifizierten die entwickelten strukturellen Kurationsschritte 33 miRNA-Familien aus dem `Rfam` als fragwürdig.

Kuratierte `Rfam`- und miRBase-Anker-Struktur-Alignments stellten eine reichhaltige Ressource für die Erstellung von prädiktiven miRNA-Profilen unter Verwendung entsprechender Hidden Markov (HMMs) und Kovarianzmodellen (CMs) dar. Als direkte Anwendung ist die Verwendung dieser Modelle zeitaufwändig. Der Benutzer muss mehrere Iterationen durchführen, um eine genomweite, nicht überlappende Annotation zu erhalten. Um dieses Problem zu lösen, bietet die vorgeschlagene `miRNAture`-Pipeline eine automatische und flexible Lösung für die Annotation von miRNAs. Sie kombiniert mehrere Homologieansätze, um die besten Kandidaten zu generieren, die auf Sequenz- und Strukturebene validiert sind. Sie erhöht die erreichbare Sensitivität bei der Annotation kanonischer miRNAs. Die Bewertung anhand der humanen Annotation zeigte, dass eindeutig falsch-positive Ergebnisse selten sind und das zusätzliche Gegenstücke auf Retained-Introns, transkribierten lncRNAs oder wiederholten Familien liegen. Für die Weiterentwicklung von `miRNAture` werden multiple Regeln eingeschlossen, um nicht-kanonische miRNA-Familien zu unterscheiden.

Diese Arbeit beschreibt mehrere Homologieansätze zur Annotation der genomischen Informationen eines Nicht-Modell-Chordaten: des kolonialen Manteltiers *Didemnum vexillum*. Eine hohe genetische Varianz und ein unerwartetes Ausmaß an DNA-Abbau wurden durch eine umfassende Analyse der Genom-Zusammensetzungsmethoden und der Genannotation deutlich. Trotz dieser Herausforderungen war es möglich, Homeobox- und Skeletogenese-bezogene Gene zu finden. Die ncRNA-Annotation selbst enthielt die erwarteten konservierten Familien. Eine umfassende Suche in der RMST lncRNA-Familie wurde bis zur Divergenz der Deuterostomier zurückverfolgt. Darüber hinaus legte eine vollständige Studie der Annotationsschwellenwerte Variationen bei der Erkennung von miRNAs nahe, die später in das `miRNAture`-Tool implementiert wurden. Angesichts der zunehmenden Forschung zur DNA Sequenzierung, ist dieses Kapitel ein Musterbeispiel des üblichen Arbeitsablaufs, welches auf ein umfassendes Sequenzierungs-, Assemblierungs- und Annotationsprojekt folgen sollte.

In den letzten 10 Jahren hat die bemerkenswerte Zunahme von ManteltierSequenzierungsprojekten den Zugang zu einer erweiterten miRNA-Annotation-Landschaft gefördert. Mit Hilfe von `miRNAture` annotierte auf diese Weise ein umfassender Homologieansatz das miRNA-Komplement von 28 deuterostomen Genomen (einschließlich der 16 derzeit berichteten Manteltiere). Um geeignete Strukturmodelle als Input zu erhalten, dienten korrigierte miRBase-Strukturalignments als Gerüst für den Aufbau entsprechender CMs, basierend auf einem entwickelten genetischen Algorithmus. Auf diese Weise wählte dieser automatische Ansatz die Sequenzen aus, die die Alignments bildeten und 2492 miRNA-CMs generierten. Trotz der vielfältigen Quellen und der damit verbundenen Heterogenität der untersuchten Genome gelang es mit einem Clustering-Ansatz erfolgreich fünf Gruppen mit ähnlichen Zusammenstellungen zu sammeln und Zusammenstellungen von geringer Qualität hervorzuheben. Die Reduzierung der Familien und Loci bei den Manteltieren ist bemerkenswert, da sie im Durchschnitt 374 miRNA-Loci im Vergleich zu den Cephalochordaten (2119), Wirbeltieren (3638), Stachelhäutern (1092) und Hemichordaten (2737) aufweisen. Die Erkennung von 533 miRNA-Familien bei der Divergenz der Manteltiere zeigt eine erweiterte Landschaft in Bezug auf die derzeitigen Familien. Gemeinsame Sets

von Vorfahren, Chordaten, Olfactoren und spezifischen kladenspezifischen miRNAs wurden anhand eines phylogenetischen Erhaltungskriteriums aufgedeckt. Im Vergleich zu den aktuellen Annotationen wurde das Repertoire in allen Fällen erweitert. Auf Grundlage der benachbarten Elemente von annotierten miRNAs unterstützt diese Arbeit zusätzlich einen syntänischen Ansatz, um miRNA-Loci zu clustern. Auf diese Weise konnte das strukturelle Alignment von miR-1497, das ursprünglich in drei Modell-Manteltieren annotiert wurde, um Syntänie in Bezug auf Manteltiere erweitert werden.

Acknowledgment

Through the past four years of this learning pathway, represented in the doctorate, I was supported, mentored and inspired by many people who I would like to express my gratitude to.

First to all, I wanted to thanks to my advisor Prof. Dr. Peter F. Stadler, for demonstrate by an open, humble, respectful and deep-though discussions the best path to do research. In the same way, for the guidance and supporting, specially in those moments where apparently there are no more ideas left. Demonstrating that doing science is, most of the time, being an *artist*.

Special thanks to my current and past colleagues of the Bierformatics group: Angel, Jörg, Steffi, Felix, Stephan, Sven, John, Iris, Carmen, Jenny, Carsten, Gabor, Zasha and all of them who take part of birthday/celebration' cakes, casual-chatting, annual seminars to share ideas, personal enthusiasm for doing science or simply helping me to improve my German/English/Spanish skills going through deep discussions about life.

I truly appreciate the heated scientific discussions with Dr. Jörg Fallman and his unconditional support through the development of `miRNAture` and the never ending bug correction for other projects. Thanks to Felix and his no-end interest in talk with me about science, Perl, VIM, and TEX. Thanks to Christiane Gärtner for proof-reading the draft and Fabian Gärtner for the TEX class of this thesis. The overseas support and excellent collaboration from Prof. Dra. Clara Bermúdez (Colombia), Prof. Dr. Federico Brown (Brazil) and Dr. Arjan Gittenberger (The Netherlands).

A deep grateful to Petra, Sven, Andrea and Jens for supporting our work from the administrative, technical and social side of the lab, as well as important as doing research.

Infinite thanks to my *Solecito* y mi *Praline/-cito* for being always by my side, loving, supporting, and smiling with me in this astonishing world. Thanks a lot for provide me much coffee/mate/tea litres to conclude this work. I would like to thank my family from the bottom of my heart. My lovely parents: Marlen and Gustavo, your life as inspiration to reach my highest goals. My sweetest sisters: Paula and Ale, your happiness motivates me to always be like a kid. ¡No saben cuanto los extrañé!

Finally, I would like to acknowledge funding and pleasant support from Deutscher Akademischer Austauschdienst (DAAD) which granted my PhD project at University of Leipzig, through the research grand: *Doctoral Programmes in Germany* 2017/2021 No. 57299294.

Contents

Part I

Background and context

— 1 —

Introduction

Contents

1.1 Uncovering the microRNA signals

The understanding of how genomic information is controlled, encoded and transmitted along biological entities is a resilient topic upon our days and an open question in molecular biology (J. M. Smith, 2000). Prove of that, the ontological concept of *information* did not reach a consensus and can be categorized by two historical paradigms (see (Barbieri, 2016) and references herein). One concept is endorsed by the *chemical paradigm*, which considers all the biological processes as a product of a linearly/sequentially defined chemical paths and physical quantities. Summarized in few words as '*life is chemistry*'. Another concept is based on the *information paradigm*, encompassing definitions that are grounded in terms distinct to physical/chemical entities, such as the order of genes in the double-helical structure of DNA and the linear organization of nucleotides, which grouped in codons, have the potential to generate aminoacids. Those processes are based on *information*, defining life as an *information-processing machine* (Barbieri, 2016; J. M. Smith, 2000). In this perspective, this paradigm can be summarized as: *life is chemistry plus information*. At the same time, Barbieri (2016) expanded those definitions in order to include an evident layer of *information* carried out by sequences and code rules, defining a *code paradigm*. That sustains the expanded idea that *life is chemistry plus information plus codes*.

As referred by Godfrey-Smith and Sterelny the cell by itself contains a complete machinery to express the contained *information*: response to signals, execution of programs, and interpretation of codes. Under these terms, life is an *artefact-making* by the evolution of the molecular machines, such as: *bondmakers*, *copymakers* and finally, *codemakers* (Barbieri, 2016). Inside the cellular environment, this involves the participation of i.e. RNA transcripts with a great range of structural conformations, sizes and tissue/temporal-expression patterns. As an example, genes by themselves carry *information* about their products and implicitly about the environments where those products are functional. Further *gene expression* of carried/inherited information depends on multiple surrounding factors, signals, or external messages (Godfrey-Smith and Sterelny, 2016).

Currently, the catalogue of molecules that play those functions have been largely recognized, described, and related by multiple interaction networks. The interplay between proteins and non coding RNAs (ncRNAs) has been recognized as a ubiquitous and a conserved mechanism intervening in transcription, translation and regulatory functions. The discovery of control functions performed by transcribed RNAs, in the form of ncRNAs, constituted a breakthrough in the understanding the initial insights about how the cell orchestrates the control of its transcripts. Since early 1990s, RNA interference (RNAi) mechanism led to uncover the potential outcome of selective interference processes in plants and animals, triggering an increased interest in the topic by their promising medical and technological uses. In the middle of this wave, miRNAs were characterized providing a better understanding of regulatory mechanisms, broad conservation patterns over multiple species, and evidence of additional actors such as Dicer or RNA-induced silencing complex (RISC).

Current definitions of miRNA paved the way to recent trends and research interest, positioning miRNAs as one of the most studied non-coding molecules. However, despite the growing interest most of the definitions, databases, annotations and genome-wide searches assumed a *canonical* biogenesis on the studied miRNAs. This affected profiling

methods, annotations and consequently evolutionary inferences. Through a revision of current state of those resources this thesis makes a comprehensive assessment of the current miRNA identification problem using computational approaches. Then, based on recognized pitfalls, multiple solutions are proposed covering multiple faces of these challenges. First, the databases' integration, using annotated information and extending them by the use of correct construction of structural miRNA profiles. In this way, the construction of an automatic homology approach to detect *canonical* miRNAs (`miRNAture`). Then, as a showcase to recurrent challenges on the way to represent the genome assembly of a non-model organism, the gene annotation and its biological meaning was studied from the invertebrate marine specie: the sea carpet squirt *Didemnum vexillum*. Finally, taking advantage of an improved genome representation the Tunicata clade, the homologous miRNA complement was accessed using the automatic annotation by `miRNAture`, expanding the set of shared miRNA families and providing a way to increase multiple miRNA structural alignments by the inclusion of syntenic regions.

1.1.1 Structure of this thesis

This work is divided in four parts: Part I covers the required biological background, basic computational terminology and general motivation imperative for the development of this thesis. In detail, Chapter 1 delineates a molecular framework of miRNAs with a brief summary of the structure of this thesis. Chapter 2 provides a comprehensive background about miRNAs, emphasizing on their definition, biogenesis and evolution in animal genomes. In particular, this chapter takes a closer look to the Tunicata clade due their well-documented morphological simplifications, genome particularities, and more importantly the limited miRNA annotation coverage detected in this clade. This description is developed disentangling the main biological and computational terms used widely through this work.

Part II states the challenge to annotate and classify miRNAs. In Chapter 3 detected pitfalls over main public databases, such as: `miRBase` and `Rfam`, concerning family definition and annotation of mature sequences are described. As a solution to those challenges, a curation of miRNA families is proposed based on a set of computational rules to further align structurally pre-miRNAs and their correspondent mature sequences. Those refined structural *mature-anchored* alignments served as a guide to build predictive Hidden Markov Models (HMMs) and Covariance Models (CMs), which together with the use of pairwise-alignments framed a rich homology source to detect miRNA candidates. The automatized detection of miRNAs using those multiple homology sources is approached using the program `miRNAture` as described in Chapter 4. This computational solution includes an outstanding annotation of mature sequences and an extensive set of filtering rules to detect the *bona fide* complement on animal genomic sequences.

In a broader context, current animal sequencing projects hold out a biased and partial coverage in terms of representations of assembled species. A tangible example is noted in the Tunicata clade, which could be benefited to a detailed dissection of the annotation of both, coding and non-coding elements despite inherent challenges detected in those marine organisms. In Chapter 5, based on the improvement obtained in the genome assembly for the colonial tunicate *Didemnum vexillum*, an annotation of their coding/non-coding elements was done, despite specie-specific challenges at DNA collection, subsequent genome

assembly, and computational analysis to assign biological meaning to annotated genes. Particularly, for the miRNA homology detection, the computational solutions to validate miRNA annotations based on their correct detection and positioning of their mature(s) elements are detailed.

In Part III, using `miRNAture` and the construction of miRNA family models from `miRBase`, the landscape of miRNA annotation on the currently available tunicate genomes is reported and analysed in relation to other deuterostomes, as seen on Chapter 6.

Lastly, in Part IV the overall landscape is drawn together with the future research prospects and open questions on Chapter 7.

1.1.2 Author contribution

Throughout the development and writing of this thesis, the pronoun *'we'* is used to refer to the joint work carried out throughout the publications, as a result of scientific collaborations. If a chapter written in this thesis is based on a publication to which I have contributed, my work and not from other contributors will be explicitly stated/referenced. This in accordance to the statement in the *Declaration of Independence*.

— 2 —

Background and related work

Contents

2.1 A *micro* introduction for an essential *RNA*

Undoubtedly, the plethora of heterogeneous and multiple tasks performed by the non coding RNAs (ncRNAs) led us to broaden our comprehension of the current molecular mechanisms carried out by the cell. Their discovered (or assigned) functions have been described in the last 70 years, detected ubiquitously along remarkable biological functions, such as: splicing, DNA replication, gene regulation, chromosome stability (see Campo-Paysaa et al. (2011) and references herein for more details). As pointed out by Cech and Steitz (2014), remarkable discoveries from RNA biology have shaped our understanding of molecular biology. In that way, this thesis underscore the importance of the posttranscriptional regulation and go deeper in a specific *momentum*, as defined by Cech and Steitz (2014), where the regulation mediated by RNAs and not by proteins was discovered. This turning point was reported in 1993 independently (Lee, Feinbaum, and Ambros, 1993; Wightman, I. Ha, and Ruvkun, 1993) associating a small RNA to the temporal control of postembryonic developmental genes in *Caenorhabditis elegans*. Later on in early 2000s, this molecular mechanism was fully described as microRNAs, after the characterization of the RNA interference (RNAi). They were found as a conserved molecular mechanism using an antisense translational repression and not constrained to the regulation of developmental genes (Lagos-Quintana, Rauhut, Lendeckel, et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). At the same time, previously described ways of regulation by short-interfering RNAs (siRNAs), contributed to refine the biogenesis model for miRNAs due their overlapping, in terms of maturation machinery, mediated by Argonaute (AGO). These key historical facts are further developed in Section 2.2.

In this way an overall context of the small silencing RNAs is exposed on Section 2.3.1. More specifically, as central topic of this thesis, current model of miRNAs reflecting multiple selection pressures that acted over its final sequence/structures are described thoughtfully in Section 2.3.2. Details about their biogenesis and the central role played by AGO-proteins are highlighted and extended in the light of the canonical/non-canonical pathways, in Section 2.3.3. The conservation of those pathways have been recognized as pivotal and the evolution of the miRNA complement has been traced back in animals (Sections 2.3.4 and 2.3.5). In general the landscape in animals shown a correlation between miRNAs diversification and morphological complexity patterns. However, this pattern has been disclosed in tunicates, due a dynamic reduction program reflected in their genome architecture, morphological simplification and interestingly, as revisited in Section 2.4, in their miRNA complement. At the same time, current knowledge of tunicate miRNA is reviewed in the same Section. This thesis explores and uncovers the potential miRNA complement over all recently reported tunicate genomes. To do so, key computational methods are briefly described in Section 2.5 and a possible miRNA annotation landscape using them is shown in Section 2.4.5 where the miRNA complement is accessed along with current data, reflecting an apparent miRNA reduction on tunicates.

2.2 miRNA revolution: uncovering the true boss

Through the study of the genes and proteins implicated in the post-embryonic developmental transitions in the nematode *Caenorhabditis elegans*, small ncRNAs transcripts

(about 21-61 nt) were detected modulating the LIN-14 protein expression. This relation based on the formation of RNA duplexes in the 3'untranslated region (3'UTR) from the lin-14 messenger RNA (mRNA) (Lee, Feinbaum, and Ambros, 1993; Wightman, I. Ha, and Ruvkun, 1993). Moreover, this small RNA gained particular attention due their recognition as a biological entity later in 2000, when Reinhart et al. (2000) recognized the same mechanism on the gene *lethal-7* (let-7) controlling additional *lin* and *daf* genes (Reinhart et al., 2000) (for a detailed review refer to Rougvie (2001)).

At that time, previous corroborated evidence was further strengthened with the discovered conservation of this regulatory mechanism not only in nematodes, but in another metazoans (Pasquinelli et al., 2000). As a result, a complete characterization and definition of miRNAs was done supported on the conserved mechanisms detected over other miRNA families: common sequence patterns recognized by AGO or RNA-induced silencing complex (RISC) (as determinants to further miRNA maturation): stable $\sim$ 60 nt stem-loop structure derived from transcribed precursors, high-conservation over multiple species (human, mouse and fruit fly) and the detection of diverse expression patterns from isolated locus or co-expressed clusters as polycistronic miRNA transcripts along developmental stages (Lagos-Quintana, Rauhut, Lendeckel, et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Later on 2003, Ambros et al. (2003) consolidated a set 5 expression and biogenesis criteria to classify *boda fide* miRNAs and distinguish them from other RNAs [1].

microRNAs were initially named as small temporal RNAs (stRNAs), due first described miRNAs, lin-4 and let-7, were involved in the regulation of developmental timing mediating a mRNA repression.

In parallel, additional examples categorized as RNAi, contributed to bring some missing miRNA-puzzle pieces. In one way, multiple efforts contributed to the identification of key actors involved in the biogenesis or processing mechanisms. In this sense, AGO was first described in plants in 1997, and the next year in fruit fly (Bohmert, 1998; Lin and Spradling, 1997; Moussian et al., 1998). Afterwards in 2000, through the double-stranded RNA (dsRNA) transfection of cultured fruit fly cells, Hammond et al. reported the discovery of an RNA enzyme that targets and processes mRNAs, defined as RISC, by its catalytic activity. Finally, in 2001 Bernstein et al. identified a conserved RNase III nuclease that shows specificity for dsRNAs, termed *Dicer* due its ability to produce homogeneous sequences of $\sim$ 22 nt from a dsRNAs.

As Lau et al. (2001) mentioned earlier, siRNAs similarly direct a mRNA cleavage during the RNAi process as the miRNAs maturation is generated via Dicer. In the same direction, Lagos-Quintana, Rauhut, Lendeckel, et al. (2001) described that protein members of AGO are evolutionary connected between RNAi and miRNA maturation. In parallel to those discoveries, the characterization of the RNAi mechanism in *C. elegans* by Fire et al. (1998), not only represented the Nobel Prize eight years later (Nobel Prize Outreach AB, 2021b), but boosted the basic research and therapeutical uses of the RNAi.

In a broader context, described historical facts can be compared to the word frequencies, found in printed sources from 1990 to 2019 (Michel et al., 2011) (Figure 1). The usage words that account for some RNA families, reflects a *first* characterized set of families, such as: transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs). Another set defined later, comprise: miRNAs, siRNAs, and long non-coding

[1] Detection of two 20-26 nucleotides (nt) by hybridization, identification of $\sim$ 22nt sequence in cDNAs and genomic locations, prediction of potential hairpin precursor containing the mature 22 nt sequence, with $>$ 16 nt complementarity, phylogenetic conservation of precursor and mature sequences, and evidence on precursor accumulation with reduced Dicer function (Ambros et al., 2003)

RNAs (lncRNAs). The increase of the usage is evident after a growing consensus and supporting evidence defined each of depicted RNA families. Observed peaks around 2006, coincide when Andrew Fire and Craig Mello were awarded with the Nobel Prize in Physiology or Medicine (Nobel Prize Outreach AB, 2021b). At that time, siRNAs were more frequent $\sim 1.6\times$ more than miRNAs. Next, around 2012 a general peak for all ncRNAs is explained by the ENCODE project results, which among others, reported that $\sim 80.4\%$ of human genome take part at least in one biochemical function related with RNA and/or chromatin (ENCODE Project Consortium, 2012). It turned out that after 2014 a growing interest started by lncRNAs and upon these days, miRNAs is not only the ncRNA with the highest frequency on text but after 10 years, overcome $2.3\times$ to siRNAs and $1.75\times$ old annotated families (rRNAs and tRNAs)[2].

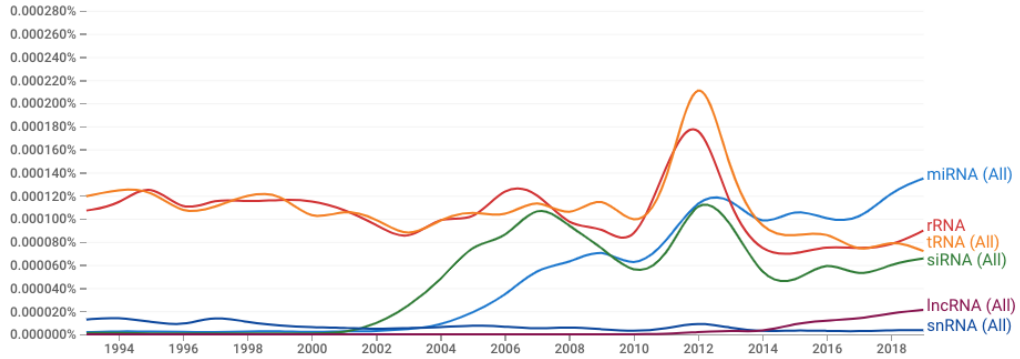


Figure 1: N-gram analysis (Michel et al., 2011) to get frequencies of common RNA families case-insensitive nouns (tRNA, rRNA, miRNA and lncRNA). Data retrieved from Google Books database from 1990 to 2019, smooth = 0, corpus = english_2019.

2.3 miRNAs: current model

Current definition of miRNAs is the result of an accumulated evidence that enhanced and refined the model, being the base to support all miRNA research. Herein, those details are explained, covering a wide definition of small RNA context until specific details related with structural features, biogenesis, and evolution.

2.3.1 small silencing RNAs

Animals make use of set of three small regulatory molecules to perform silencing and consequently, *confer a qualitative change in the way that cellular networks are managed*, as pointed by Wilson and Doudna (2013), namely: miRNAs, siRNAs and piwi-interacting RNAs (piRNAs) (see summary Table 1). Those molecules converged in the use of an effector, composed by an AGO protein together with a mRNA. This complex developed

[2] At the time of writing this thesis (01.11.2021), exists 101,095 publications with miRNA in their title or abstract reported in `PubMed`

specificity for target genes, regardless if their extracellular, cytoplasmic or nuclear origin (Wilson and Doudna, 2013).

Thanks to the action of two RNase III-type proteins: Drosha and Dicer, miRNAs are processed from endogenous precursor sequences. The mature result, is a short transcript of ~ 21 nt, subsequently bounded by AGO-proteins to complete their posttranscriptional regulations. As previously described, siRNAs are derived from dsRNAs and are dependent only on Dicer but not on Drosha. Finally, piRNAs encompass larger transcripts ($21-31$ nt in length) that are not dependent of Dicer, their action is related with transposon silencing through heterocromatin formation or RNA destabilization (Ghildiyal and Zamore, 2009; Lam et al., 2015; Ozata et al., 2018; Wilson and Doudna, 2013).

Based on the last points, AGO proteins are central of RNAi mechanisms, taking a pivotal part of overlapping functions. By the AGO proteins side, the ability of target and load foreign RNA was expanded by the incorporation of endogenous RNAs as an important part of the RNAi pathway, this step considered as a key evolutionary innovation (Dexheimer and Cochella, 2020). From the miRNA-pathway side, the use of ancestral silence machinery, allowed the processing of endogenous short hairpins for further processing into mature entities (Bartel, 2018). As a result, a shared use of the silencing machinery by both, siRNAs and miRNAs is currently evidenced: the former as a defence system, whereas the latter as a fine-tuner of gene expression (Creugny, Fender, and Pfeffer, 2018).

As expected, both molecules share similar physico-chemical properties as typical of Dicer products, such as: $20-25$ nt length, and 5'-phosphate, and 3'-hydroxyl (Lau et al., 2001). For their part, siRNAs diverge from miRNAs on the target specificity (fully complementary in siRNAs *versus* partial complementary in miRNAs), aspects of their biogenesis detailed in (Ambros et al., 2003; Lam et al., 2015), high transcript specificity in siRNAs, and the triggered gene regulation mechanisms (endonucleolyc cleavage in siRNAs and a way more in miRNAs as: translational repression, degradation, and endonucleolyc cleavage). For additional comparisons see Lam et al. (2015) and Mack (2007).

2.3.2 MiRNAs: current definition

The main-product of the miRNA maturation pathways (see Section 2.3.3) via Microprocessor (composed by one catalytic subunit DROSHA and two DGCR8 cofactors) is a short RNA (~ 22 nt) that mediate gene silencing by guiding AGO proteins, together referred as miRNA-induced silencing complexs (miRISCs), that targets mRNAs in most of the cases at 3'UTR (Gebert and MacRae, 2018; M. Ha and Kim, 2014; Kim, Han, and Siomi, 2009; Michlewski and Cáceres, 2019; Nguyen et al., 2015; Winter, Jung, et al., 2009).

As mentioned before, the role of Microprocessor is pivotal in the genesis of miRNAs from primary miRNAs (pri-miRNAs) and constitutes the decision point to distinguish a canonical or non-canonical processing (Bartel, 2018; Gregory et al., 2004). As depicted in Figure 2, the canonical miRNA processing model highlights specific sequence/structural patterns required for a Microprocessor recognition. In general, the hairpin-loop structure from pri-miRNAs can be split up into four regions: a basal single-stranded, lower and upper stem, and apical single stranded loop. This configuration has been detected optimal to be processed via Microprocessor (Bartel, 2018).

Inside those regions DROSHA and DGCR8 perform the recognition of specific sequence/structural patterns: In one side, the recognition of *basal junction* is performed

Table 1: Comparison to small silencing RNAs detected in animals: miRNAs, siRNAs and piRNAs based on selected features based on synthesised reports from Anzelon et al. (2021), Ghildiyal and Zamore (2009), Lam et al. (2015), and Ozata et al. (2018).

Feature	**miRNA**	**siRNA**	**piRNA**
Discovery	Lee, Feinbaum, and Ambros (1993) and Wightman, I. Ha, and Ruvkun (1993)	Hamilton and Baulcombe (1999)	Aravin et al. (2001)
Origin	dsRNA precursor	dsRNA precursor	long-single-stranded RNAs (ssRNAs) precursor
Length	19-25	21	24-30
Source	Pol II/III transcripts/Intron processing	Exogenous and Endogenous dsRNAs transcripts	piRNA precursor transcript
Prior Dicer processing	pre-miRNA with $70-100$ nt	dsRNA with $30-100$nt	Not processed
Function	Regulation of mRNA stability, translation mechanisms/Post-transcriptional regulation	Gene silencing line-defense based on mRNA targeting	Germline transposon regulation.
Guide	AGO proteins	AGO proteins	*P*-element *i*nduced *wi*mpy testis (PIWI) proteins
mRNA target	Multiple	One	Multiple, more specific than miRNAs.

by DROSHA, acting as a molecular rule of the distance between this region and the potential DROSHA cleavage site (11 nt upstream). On the other side, DGCR8 makes the recognition of the *apical junction* and enhance the RNA-binding affinity (Nguyen et al., 2015). Specifically, the contribution of multiple sequence patterns have been identified as checkpoints prone to be identified by DROSHA, as described in Auyeung et al. (2013), Fang and Bartel (2015), and Nguyen et al. (2015): a basal 'UG' and a mismatched 'GHG' (H corresponds to any nucleotide except G) close to the basal single-stranded (7-9) nt. Additionally, DGCR8 recognizes an apical 'UGU' motif. In addition, a 'CNNC' ('N'is any nucleotide) motif is recognized by the additional factors from Microprocessor: the splicing factor SRp20 or the DEAD-box helicase p72 (M. Ha and Kim, 2014). As early detected on Auyeung et al. (2013), not all miRNA families use those patterns and their use could be restricted to some species. Therefore, they should be treated as *enhancers* of processing (Bartel, 2018).

Together with sequence patterns, the analysis of the hairpin-loop structure led the discovery of additional patterns. For example, Roden et al. (2017) proposed that an optimum stem length is about $33-39$ nt (located along the lower and upper stems) analysing mammalian hairpins. Inside this stem region, two sub-regions are less tolerant to unpaired bases: $16-21$ nt and $28-32$ nt. Additionally, through directed mutations

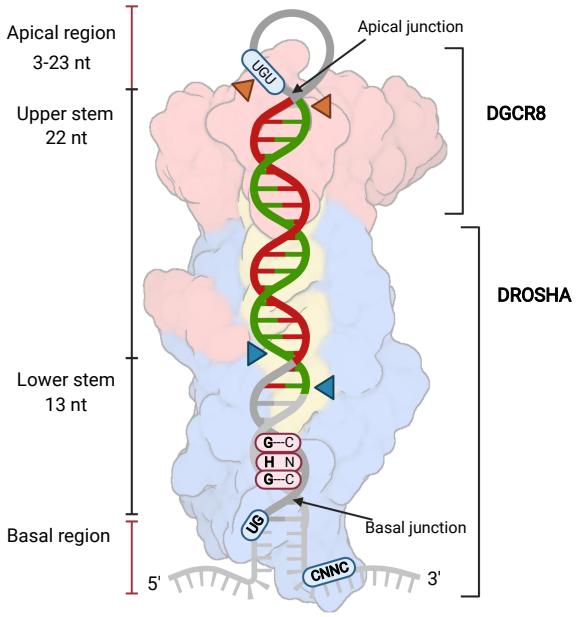


Figure 2: microRNA. Summary of current precursor miRNA (pre-miRNA) model and its recognition by Microprocessor (background model). Blue arrows define the Drosha cleavage sites. Orange arrows indicate the potential Dicer cleavage to remove the single stranded loop region. Additional sequence patterns are recognized by Drosha and Dicer are in red and blue boxes, respectively. At left margin, detail of canonical regions are indicated. Right margin delimit sub-regions in Microprocessor. Example of possible miR and miR* position are coloured by red and green, on the stem region. Figure created based on Bartel (2018) and Nguyen et al. (2015). PDB references: Drosha: 6V5B (10.2210/pdb6V5B/pdb). Created using BioRender.com.

over miR-21 and miR-30 precursors, Zeng and Cullen (2003) defined a range of $3 - 23$ nt at the apical loop region to optimal processing.

2.3.3 Biogenesis

A distinction between miRNAs is done based on their biogenesis pathways. In one way, a *canonical* pathway makes use of Microprocessor and Dicer cleavage to generate mature miRNA products. In another way, some miRNAs do not require either Microprocessor or Dicer processing, taking part of the *non-canonical* pathways. To further reference Figure 4 summarizes both pathways.

Canonical miRNAs

In animals, mature miRNAs are initially produced from primary precursor transcripts, the pri-miRNAs, which are transcribed by pol-II. As a pol-II products, the pri-miRNA is 5' capped and sometimes are polyadenylated at 3'-end. Subsequent processing steps mediated by Microprocessor are performed on the nucleus, which recognizes using the DGCR8 dimer and cleavages upper the basal junction ~ 11 bp at 3'-end, and about 22 nt away from the apical junction, leaving a 2 bp offset, transcribed hairpin-loops via the endonuclease Drosha. The result is a ~ 60 pre-miRNA, preserving the stem-loop secondary structure that is subsequently exported and protected to nucleolytic attack via exportin-5 (EXP5) and RAN-GTP to the cytoplasm (Bartel, 2018; M. Ha and Kim, 2014).

Translocation is done through nuclear pore complex and the pre-miRNA release is the result of the disassembly of the complex mediated by the GTP hydrolysis (GDP) (Bartel, 2018; M. Ha and Kim, 2014). In the cytoplasm, pre-miRNA is cleaved close to the apical junction by Dicer, and its RNase III domains and additional co-factors. The recognition is done by the helicase domains of Dicer, which interact with the terminal loop. At the same time PIWI-AGO-ZWILLE (PAZ) domain recognize the basal region. The cleavage sites are determined by a fixed cut of $21-25$ nt length in both, 5' and 3': one based on the binding at 5' phosphorylated end, meanwhile the other from the dsRNA. The 2-nt overhang is recognized by PAZ, which contains two pockets where both ends fit (M. Ha and Kim, 2014). Multiple cofactors have been identified in the process, in flies Loquacious (Loqs) which is the homolog in mammals for TAR RNA-binding protein (TRBP), improving the efficiency at processing and by intervening in the measure of mature miRNAs. The product is a small dsRNA, containing the miRNA (which guide the silencing complex) and their corresponding passenger miRNA* strand (which will be discarded and degraded) and at both ends a $\sim 2-3$ nt overhang.

This small dsRNA is loaded to AGO protein by a conformational opening promoted by the chaperone proteins HSC70/HSP90. Here it is crucial the binding orientation from the dsRNA to AGO, that could fit on the pocket that binds the 5'-nucleoside monophosphate with preference for pU or pA binding. The identity assignment between strands relies on the stability of the 5'-ends, the less stable appear to be assigned as guide strand (Rüegger and Großhans, 2012). The pre-RISC complex releases the passenger strand, resulting in the mature miRISC (Bartel, 2018; M. Ha and Kim, 2014).

In terms of kinetics of tightly controlled miRNA biogenesis, Reichholf et al. (2019) found that this is a fast process: in average the top ten highly expressed miRNAs accounted $\sim 58 \pm 14$ molecules per minute. That rate is higher in comparison to reported messenger RNA (mRNA) rate, that accounts 8 molecules per minute. As a consequence, exists an accumulation of miRNA-duplexes prior AGO loading, essentially acting as a limiting step to miRISC formation. In addition, this loading mechanism do not load all miRNA-duplexes, given its average decay about $\sim 40\%$ of miRNAs do not load into AGO (Reichholf et al., 2019).

A successfully loaded guide strand inside the miRISC, will target mRNAs by a partial Watson-Crick complementarity, in animals, usually at 3'UTR. This targeting is translated into mRNA posttranscriptional repression, which in mammalians, it is done in $\sim 66-90\%$ of cases by mRNA destabilization and in a lower quantity, via translational repression and even AGO-catalyzed cleavage (Bartel, 2009, 2018; M. Ha and Kim, 2014; Rüegger and Großhans, 2012). The importance of the targeting relation between miRNA:mRNA is discussed broadly by Bartel (2009, 2018), who highlights the role of the complementarity extension in the *seed* region. This region is located in the 5' side of the miRNA, at nucleotides 2-7 (see Figure 3). The identification of animal miRNA targets is challenging due the recognition of either imperfect matches (contiguous 6 nt) or matches that display an offset, which have shown some degree of repression (Bartel, 2018).

This fact is depicted in Figure 3, showing a classification of types of miRNA target sites (Bartel, 2009, 2018; Ellwanger et al., 2011; Friedman et al., 2008). The match degree is classified by the contiguous length of the seed nucleotides to the target mRNA. Ellwanger et al. (2011) quantified the frequency of functional target sites, finding that a 67% resembled a 6mer-$\alpha/\beta/\gamma$ site, 23% 7mer-α/β and complete 8mer-α accounted for

9%. The 6mer-α was additionally included by Ellwanger et al. (2011) and have not been previously considered in Bartel (2009). In addition, compensatory roles have been detected on the 3' side (nucleotides 13-16), serving as additional support of the 5' seed matching but detected with lower proportion and efficacy (see Bartel (2009) for an extended discussion). The overall gain doing this target detection is an additional classification layer for miRNAs, as a direct link to understand their repression functions (Friedman et al., 2008).

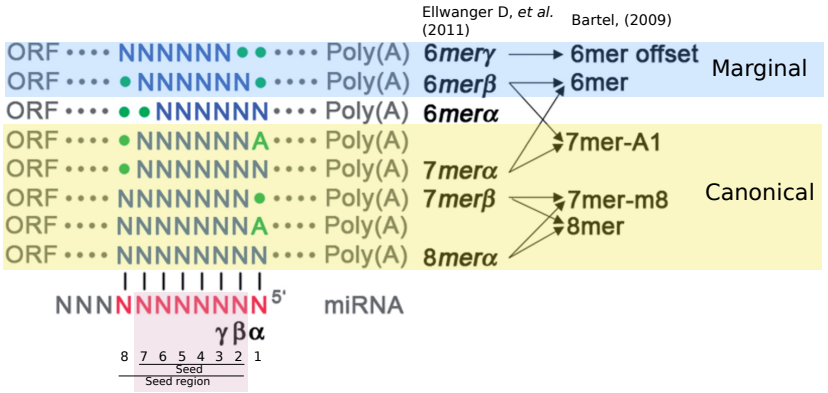


Figure 3: Canonical and marginal target miRNA sites. Reported sites by Ellwanger et al. (2011) are compared to those reported in Bartel (2009). Canonical and marginal target sites are coloured by yellow and blue background boxes, respectively. MiRNA seed is highlighted in a red box (nucleotides 2-7). Ellwanger et al. defined position types, based on their start: $\alpha = 1$, $\beta = 2$, and $\gamma = 3$. Watson-Crick complementary is represented by vertical bars between the miRNA and the target mRNA. Modified figure from Ellwanger et al. (2011), including details from Bartel (2009, 2018). ORF: Open reading frame.

At the same time, once produced, miRNA are susceptible to be modified via *single nucleotide polymorphisms*, *regulation of tailing*, *RNA editing* and *regulation of stability level*. This kind of modifications have a direct impact on proposed miRNA family classifications, as described by Hertel, Langenberger, and P. F. Stadler (2013). Details about these modifications are described in (M. Ha and Kim, 2014).

Non-canonical miRNA

In other way, non-canonical miRNAs comprises a larger diverse of miRNA biogenesis modes, due this classification encompasses miRNAs apart from the canonical biogenesis model due the bypass of Microprocessor or Dicer. Microprocessor-independent miRNAs where described in flies and roundworms as *mirtrons*, because they are derived from intronic regions as splicing lariats products, subsequently maturated with help of the enzyme *debranching* that opens the 2'-5' linkage of the lariat (Ruby, Jan, and Bartel, 2007; T. Treiber, N. Treiber, and Meister, 2018). Some *mirtrons* require a trimming step before export using nuclear exosomes, due they possess additional nucleotides at 3' and/or 5': defined as *tailed mirtrons*. They are exported by EXP5 in a folding that resembles pre-miRNAs (T. Treiber, N. Treiber, and Meister, 2018). This mechanism was proposed

as ancient, that emerged before the miRNA processing machinery (Ruby, Jan, and Bartel, 2007) (see an extensive characterization of mammalian *mirtrons* in Wen et al. (2015))

In the same way, chimeric hairpins derived from other non-coding RNA (ncRNA) families or Pol III transcripts, as small nucleolar RNAs (snoRNAs) or transfer RNAs (tRNAs), that could be subsequently be processed by Dicer. In case of snoRNAs, they are processed into short-stable miRNA-like fragments, called small nucleolar RNA-derived RNAs (sdRNAs). For example, Ender et al. (2008) described the snoRNA ACA45, which is processed by Dicer and subsequently associated with AGO proteins, generating a regulation similar to miRNA on the CDC2L6 gene. Endogenous short-hairpin RNAs (shRNAs) are transcribed by Pol II and are 7-methylguanosine (m^7G)-capped, which promotes cytoplasmic export by exportin-1 (EXP1) and biased the RNA-induced silencing complex (RISC) loading to the 3'-arm (T. Treiber, N. Treiber, and Meister, 2018; M. Xie et al., 2013).

Dicer-independent pathway was discovered by the detection of miR-451, which plays key roles in erythrocyte maturation in zebrafish, levels were resistant to the Dicer-knockdown but reduced when mutated Ago2 (Cheloufi et al., 2010; Cifuentes et al., 2010). This miRNA differs to the canonical model shown in Figure 2 due its conserved hairpin with a 42 nt and $\sim$ 17 nt stem, extended 3' over the loop region with length about $20-30$ nt, and after nucleotide 30 longer transcripts are uridinated. The binding and subsequent cleavage is promoted by Ago2, with a slicer catalytic activity, independent of Dicer (Cifuentes et al., 2010). The removal of additional uridines is made by a nuclease. However, this is not restricted to miRNA, Langenberger et al. (2012) identified loci strongly dependent of Dicer, such as: vault RNAs, snaR ncRNAs, tRNAs and H/ACA snoRNAs.

2.3.4 MicroRNAs genomic architecture and evolutionary implications

Increasing the local *scale* to a larger genomic architecture, miRNAs take part of a dynamic, rich and conserved genomic context. The combination of both, the current availability of a diverse and large number of genome sequences and the current miRNA biology knowledge, are a proper scaffold to understand the miRNA evolution (Berezikov, 2011). In this respect, current understanding of miRNA genes genomic locations are depicted in Figure 5. As an isolated transcriptional unit with own promoters, *isolated* miRNAs are transcribed by pol-II (see **a**). In case that promoters are shared with a host gene (**b-f**), miRNAs could be arranged as a *cluster* or be isolated, based on the proximity respect to other miRNAs without the interruption of other genomic elements. This promoter-sharing miRNAs can be located in intergenic (**b**), intronic (**c-e**) or exonic regions (**f**) (Berezikov, 2011; M. Ha and Kim, 2014; Monteys et al., 2010).

In detail, clusters could resemble polycistronic transcripts composed by multiple miRNA families, mostly $2-3$ but in few cases, with a higher number of elements as evidenced in the imprinted miRNA human 14q32 locus/mouse distal 12 domain cluster, which holds about 46 miRNAs in $\sim$ 1 Mb (Seitz, Royo, et al., 2004; Seitz, Youngson, et al., 2003). Additionally, members from the same cluster could target common mRNAs (Y. Wang et al., 2016). In general, the evolution and generation of miRNA clusters has been reported via tandem or non-local duplication or even through, indels to intronic or

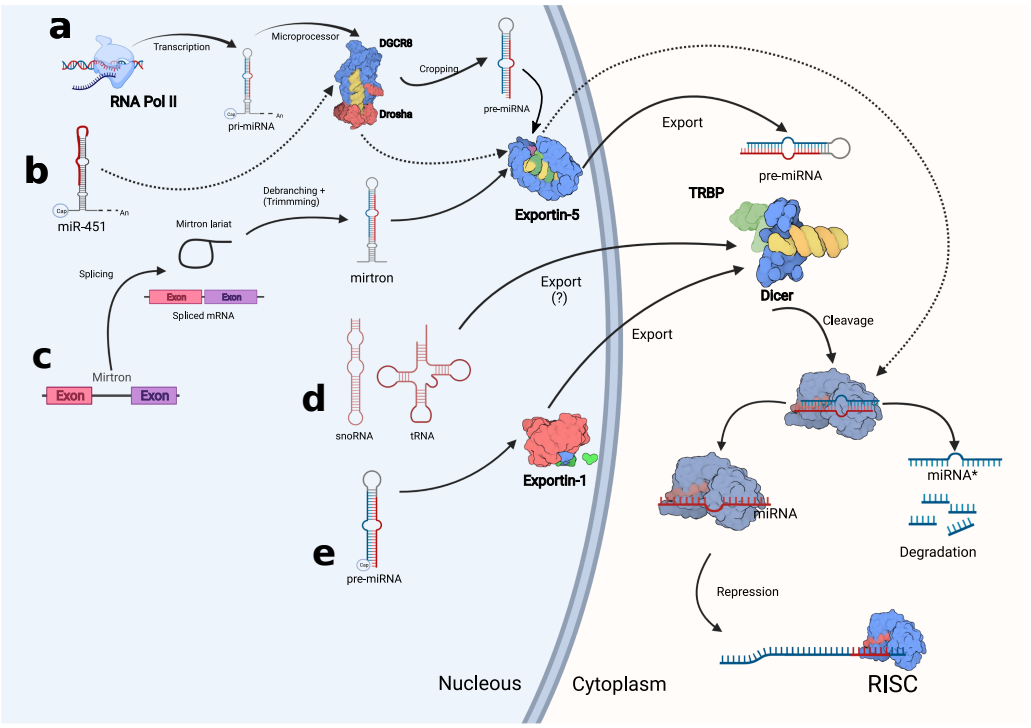


Figure 4: Canonical (**a**) and Non-canonical (**b-e**) miRNA biogenesis. **a)** The canonical pathway makes use of Microprocessor (generating a pre-miRNA) and Dicer (loop cleavage). **b)** miR-451 is processed at the same time by Microprocessor and exported by Exportin-5, being incorporated directly to RISC. **c)** Takes advantage of intron processing, that generates a hairpin structure exported by Exportin-5 bypassing Microprocessor and being cleaved by Dicer. **d)** chimeric hairpins are derived from other ncRNAs are Microprocessor-independent and their export to cytosol is uncertain. **e)** Endogenous shRNAs skip Microprocessor but are exported by Exportin-1. **d,e** make use of Dicer to be incorporated into RISC. Figure re-drawn based on (Bartel, 2018; Cifuentes et al., 2010; M. Ha and Kim, 2014; T. Treiber, N. Treiber, and Meister, 2018; M. Xie et al., 2013). PDB models: Drosha (6V5B `10.2210/pdb6V5B/pdb`), Exportin-5 (3A6P, `10.2210/pdb3A6P/pdb`), Exportin-1 (5JLJ, `10.2210/pdb5JLJ/pdb`), Dicer (6BU9, `10.2210/pdb6BU9/pdb`), and RISC (4W5N, `10.2210/pdb4W5N/pdb`). Created using `BioRender.com`.

intergenic regions (Berezikov, 2011; Campo-Paysaa et al., 2011; Marco et al., 2013; Seitz, Royo, et al., 2004; Seitz, Youngson, et al., 2003).

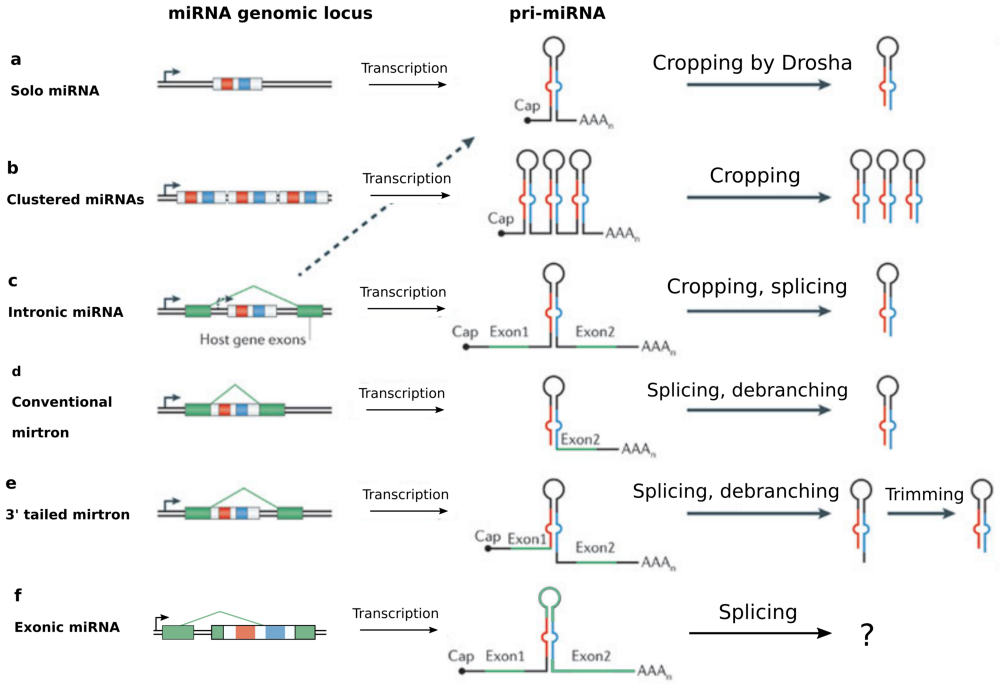


Figure 5: Genomic locations of miRNAs. Figure obtained from Berezikov (2011) and modified including exonic derived miRNAs.

The intrinsic relation between miRNAs and their host genes (in **c-e** examples in Figure 5) is not casual. As noted by França, Vibranovski, and Galante (2016), when studying the age of miRNA-host genes, the genomic location of mammalian miRNA influenced their expression divergence. As seen in primates, there is a bias on the miRNA emergence over old-age host genes. Comparing features from old and young host genes, old ones are broadly expressed, have higher intron density, evolve slower and have been subjected to strong purifying selection (Albà and Castresana, 2004; Park et al., 2011; Wolf et al., 2009). Intronic miRNAs located in old-/middle-age genes tend to be more similarly expressed than intergenic ones (França, Hinske, et al., 2017; França, Vibranovski, and Galante, 2016).

A **conservative phase** involves a greater integration into transcriptional networks and slowing evolutionary rates

In addition, the generation of new miRNA families is a continuous and dynamic process. This emergence could start from a random local stem-loop structure with a low tissue-/specie-specific expression, with fast evolution rates at sequence level (Berezikov et al., 2006; Marco et al., 2013; Tanzer and P. F. Stadler, 2004). The small subset of selected candidates going through a conservative phase yield a broadly tissue-expression and higher transcription rates (França, Vibranovski, and Galante, 2016; Hertel and P. Stadler, 2015; H. Liang and W.-H. Li, 2009).

2.3.5 Animal evolution and relation with miRNA

With the advent of *deep-sequencing* and associated computational methods, the ways to discovery and annotate small RNAs has changed and improved (Berezikov, 2011). The miRNA evolution can be traced back in time with high accuracy thanks to the broad detection efforts in many species. As a result, it has been possible to define high conserved conservation patterns, as demonstrated in computational analysis performed by Hertel, Lindemeyer, et al. (2006) and Hertel and P. Stadler (2015). In addition to detect them, a particularity is that miRNAs are likely to form paralogs, that in studied cases as let-7 (Hertel, Bartschat, et al., 2012) or miR-17 (Tanzer and P. F. Stadler, 2004) the conservation allowed to propose an evolutionary scenario of the families at loci level. However, the correct orthology assignment for all reported families is the major difficulty to reconstruct the complete evolutionary history for all miRNA families.

It is assumed that miRNA families are grouped based on an independently evolutionary descent. Authors as Tarver, Sperling, et al. (2013) suggested that miRNAs are suited candidates to be included as markers in phylogenetic studies. Their arguments are listed as follows: 1) the canonical biogenesis is conserved and let to define structural criteria for detection, 2) the miRNA annotation is an ongoing and accrual process, 3) after a family emergence exists a low level of secondary loss along metazoans (Berezikov, 2011; Tarver, Taylor, et al., 2018), 4) the mature sequence have a low substitution rate, 5) and convergent evolution is an event with a low probability and certainly generated by a false-positive annotation (Tarver, Sperling, et al., 2013). In contraposition to those arguments, Thomson et al. (2014) questioned the use of miRNAs in phylogenetics based on apparent pervasive loss, but later Tarver, Taylor, et al. (2018) argued that this apparent issue is an artefact related to high homoplastic loss and sampling errors.

In general, the repertory of miRNAs has been expanding through animal evolution about at the same rate of increase of morphological complexity (Berezikov, 2011; Sempere et al., 2006). The accumulated evidence points out that this expansion is correlated with a broad interaction with control networks and, as a result, with an increased morphological complexity. In detail, Hertel and P. Stadler (2015) described multiple hotspots of miRNA innovation associated with the origin of vertebrates, at the root of the placental mammals, the ancestor of "free-living" nematodes, or the radiation of the drosophilids. Additional high peaks of innovation have detected on Amniota, Eutheria, Boreotheria, Muridae and Catarrhini.

However, an exception to this rule has been reported in Tunicates (Fu, Adamski, and E. M. Thompson, 2008). In this way, the massive simplification observed at genomic and morphological level in tunicates, is correlated with massive miRNA loss/restructuring (Dai et al., 2009; Fu, Adamski, and E. M. Thompson, 2008; Hertel, Bartschat, et al., 2012).

2.4 Absence of evidence is not evidence of absence: Tunicate miRNA annotation

The interesting pattern in tunicates regarding their massive loss or high divergence of miRNA families (Section 2.3.5), is explained by their genomic features as: genome compactness, gene loss, fast rate evolution, genome rearrangements, and poor synteny

conservation (Berná and Alvarez-Valin, 2014; Dai et al., 2009; Denoeud et al., 2010; Fu, Adamski, and E. M. Thompson, 2008). In spite their extensive divergence, the basic developmental kit resembles vertebrate ones (Lemaire and Piette, 2015; Tsagkogeorga, Turon, et al., 2010). In the following sections, tunicates are contextualized as key clade to understand chordate evolution and further connecting their current state at miRNAs annotation.

2.4.1 What should be considered a tunicate?

As a marine filter feeders, Tunicata (or Urochordata) designate a diverse monophyletic clade of invertebrate organisms (Swalla, C. B. Cameron, et al., 2000). Grouped by the homoplastic character related to the possession of an extracellular *tunic*, composed by cellulose, as a means of protection of adult forms (zooids) (L. Z. Holland, 2016; Lemaire and Piette, 2015; Nakashima et al., 2004). As chordates, they have a notochord, dorsal neural tube (in larva and adult stages), and gill slits (as adults) (Satoh and Levine, 2005). Back in 2006, they have been recognized as sister group of vertebrates, resembling the *Olfactores* clade, implying that tunicates were subject to secondary simplification process given the cephalochordate ancestor (i.e loss of metameric segmentation) (Delsuc, Brinkmann, et al., 2006). At the same time, there is about 13 recognized synapomorphies in the Olfatores clade, for example: *Brachyury* expression in the notochord and *Pax1/9* expression in the pharynx, the resembled sensory vesicle to the vertebrate forebrain, for complete descriptions refer to Ruppert (2005). They are classified in three groups: ascidians, appendicularians (larvaceans) and thaliaceans (Lemaire and Piette, 2015; Satoh and Levine, 2005).

The **cellulose synthase** gene was acquired by a horizontal gene transfer from a bacterium, designed by Nakashima et al. (2004) as *Ci-CesA* gene in the tunicate *Ciona robusta*.

In general terms, their fast evolution has been identified at protein, nucleic and mitochondrial level and consequently their ecology reflect this plasticity. Relaxed constraints in the evolution of genomes and developmental trajectories in the tunicates may have been responsible for the plethora of reproductive strategies, morphologies, and life histories observed in the group (L. Z. Holland, 2014; Velandia-Huerto, Brown, et al., 2018). They are dispersed into multiple marine habitats, including shallow waters, shore to open ocean and the deep sea (L. Z. Holland, 2016). At the same time some species have been categorized as invasive species in Europe, the Americas, and New Zealand (G. Lambert, 2009). It negatively affects established benthic species and damages ship hulls as well as the infrastructure in marinas, ports, and shellfish farms (Parra-Rincón et al., 2021).

2.4.2 ncRNA annotation on tunicates

As pointed out by Velandia-Huerto, Brown, et al. (2018), the annotation landscape of non coding RNAs (ncRNAs) in tunicates needs to be completed. Current studies are biased to experimental models such as *O. dioica*, *Ciona robusta*[3], and *Ciona savignyi*. More recently, species as *Didemnum vexillum* (Parra-Rincón et al., 2021; Velandia-Huerto, Gittenberger, et al., 2016), *Halocynthia roretzi* (K. Wang et al., 2017), and *Salpa thompsoni* (Jue et al., 2016) have benefited to genome annotation of ncRNA elements. Despite the morphological,

[3] In accordance with use in the ascidian community in this thesis the term *Ciona robusta* reflecting that "Morphological evidence that the molecularly determined *C. intestinalis* type A and type B are different species: *Ciona robusta* and *C. intestinalis*" (Brunetti et al., 2015).

reproductive modes and heterogeneous life histories recognized in tunicates (L. Z. Holland, 2016), the consensus point out an evident lost of conserved miRNAs and a recognizable number of specie-specific gains over all species, as previously hypothesized in Fu, Adamski, and E. M. Thompson (2008).

2.4.3 miRNA identification and validation on tunicates

Along with the recognition of miRNAs as control mechanism conserved over multiple species done by Pasquinelli et al. (2000), let-7 was first identified on the tunicates *C. robusta* and *Herdmania curvata*, showing expression signals. Then, with the publication of the first complete genome of the solitary specie *C. robusta* (Dehal et al., 2002), computational methods were used to detect miRNAs as a complement strategy to experimental methods.

As summarized in Figure 6, since 2005 computational methods such as `ERPIN` (A. Lambert et al., 2004) and `BLAST` (Altschul et al., 1990) have used previous data deposited on the *microRNA Registry* (`miRBase`) (Griffiths-Jones, 2004) to annotate tunicate miRNAs (Legendre, A. Lambert, and Gautheret, 2004). At the same time, Missal, Rose, and P. F. Stadler (2005) used `RNAz` to search conserved ncRNAs structures, including miRNAs, on *C. robusta*, *C. savignyi* and *O. dioica*. As a result, conserved miRNA families, such as: miR-124, miR-92, miR-98, miR-325, the miR310-313, and let-7 were identified. Additionally, ciona-specific miRNAs, such as: miR-9, miR-78 and many others were annotated (see the extended list by Velandia-Huerto, Brown, et al. (2018)).

Two years after, Norden-Krichmar et al. (2007) employed a homology approach, using the program `FASTA/ssearch34`, and considered the `miRBase` mature and precursor sequences on the *C. robusta* genome to annotate Ciona spp. miRNAs. Strict conservation parameters (conservation seed $\geq 90\%$ + conservation $\geq 90\%$ between Cionas + matches with human and *C. elegans*) yielded a number of 257 matching miRNAs in both solitary species. Subsequent structural evaluation using `mfold`(Zuker, 2003) gave a conserved list of 14 miRNA families in both tunicates [4]. Eight subsequently validated by Northern blot analyses.

The study of miRNAs on the appendicularian *O. dioica* was approached by Fu, Adamski, and E. M. Thompson (2008), using a hybrid approach experimental an experimental approach that started with isolation of small RNA, amplification of cDNA ends by RT-PCR (RACE), followed by cloning and sequencing. This protocol was made to study temporal-spatial expression patterns of conserved miRNAs in multiple developmental stages in *O. dioica*. As complement, computational tools were used to map cloned small RNA libraries to annotate candidate miRNAs on the reference genome, considering sequences with ≥ 15 nt or seed matches, which were extended and subject to structural folding using `Mfold` (Zuker, 2003). Expression was checked for 55 miRNAs using array dot blot analysis. Sex-specific expression was reported for miR-1487 and miR-1488. Additionally, opposed to vertebrates (which at that time the intron location of miRNAs was estimated $\geq 80\%$) *O. dioica* reported about $22 - 27\%$. Most of them found in antisense strand of protein-coding genes, enabling the idea to a profound reorganization of the miRNA repertoire (Fu, Adamski, and E. M. Thompson, 2008).

[4] let-7/miR-98, miR-72/miR-31, miR-25, miR-153, miR-47, miR-34, miR-126, miR-141, miR-200, miR-7, miR-33, miR-302a, miR-452*, and miR-520d.

Between the years 2009 and 2015 the majority of the studies of miRNAs in tunicates focused on the validation of expression of computational predicted miRNAs in Ciona spp. with the special focus on *C. robusta* as model organism of tunicates or testing new computational approaches as `miRTRAP`, `miRDeep2` and `miRRim2`, which used next generation sequencing (NGS) libraries of small RNAs derived from *C. robusta* to validate their algorithms (for an extensive description refer to Velandia-Huerto, Brown, et al. (2018)).

Since 2016 new approximations has increased our knowledge about new families in other tunicates thanks to the sequence of new tunicate genomes of the species *D. vexillum*, *S. thompsoni* and *H. roretzi*. A detailed homology-based computational survey of ncRNAs was performed on the preliminary draft genome of *D. vexillum* (Velandia-Huerto, Gittenberger, et al., 2016). `Blast` and `HHMer` searches were performed with annotated small ncRNAs sequences from metazoans and Hidden Markov Models (HMMs) from `Rfam`[5] to obtain the sort of candidates at sequence level. Structural alignments of those sequences were performed by `infernal` (Nawrocki and S. R. Eddy, 2013), using metazoan-specific Covariance Models (CMs) to annotate the small ncRNAs collection, that accounted 57 families and 100 loci of miRNAs.

Small RNA libraries for the Southern Ocean salp *S. thompsoni* were sequenced with an Illumina Hiseq 2000 (Jue et al., 2016). After filtering data sets to $18-24$ nt for miRNA and $28-32$ nt for piwi-interacting RNA (piRNA), the reads were aligned to *S. thompsoni* genome and miRNA gene folding predictions were performed using `RNAfold` (Lorenz et al., 2011). In this initial survey of small RNAs, were revealed the presence of known, conserved miRNAs, as well as novel miRNAs 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 searching methods. The experimental validation suggested that more than half of these miRNAs candidates are expressed during embryogenesis. The expression of some predicted miRNAs were validated by RT-PCR using embryonic RNA. In this approach *C. robusta* small RNA-Seq reads (Shi et al., 2009) were used to identify conserved miRNAs in *H. roretzi* (K. Wang et al., 2017).

2.4.4 miRNA genomic organization in tunicates

As previously indicated in Section 2.3.4, miRNAs organization spans from isolated genes to clustered sets (Figure 5). Through the annotation of putative miRNAs in *C. robusta* Hendrix, Levine, and Shi (2010) found that nearly one-third of its miRNAs reside on introns, intergenic regions, or in a really few proportions belonged from exonic positions. Additionally, the bidirectional transcription described for the *D. melanogaster* miR-iab-4 (Stark et al., 2008) was found recurrently in Ciona spp. and in the immediately flanking regions of mature miR/miR* the prevalence of miRNA-offset RNAs (moRs) (Hendrix, Levine, and Shi, 2010). In *O. dioica* most of the miRNAs are single-copy and are located in the antisense orientation ($\sim 50\%$), often located on introns or downstream the 3'untranslated region (3'UTR) (Fu, Adamski, and E. M. Thompson, 2008).

In this regard, multiple *clusters* have been identified in tunicate species *C. robusta*, *C. savigny* and *O. dioica*, as reviewed in Velandia-Huerto, Brown, et al. (2018). A conserved

[5] `https://rfam.xfam.org/`

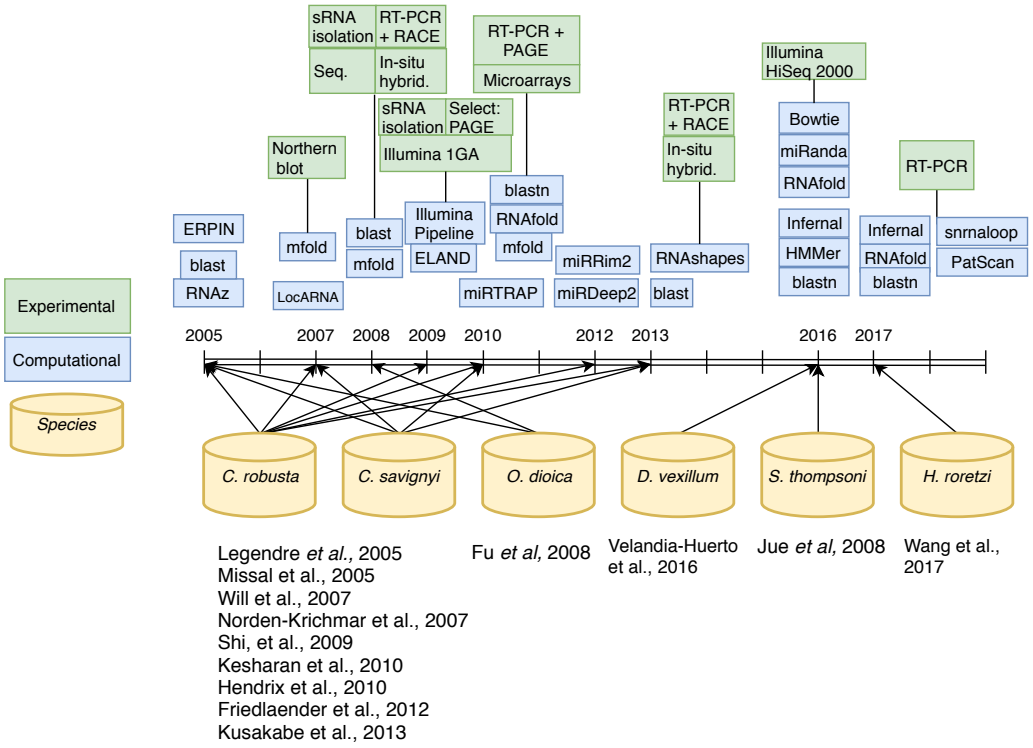


Figure 6: Current miRNA annotations over tunicate genomes associated with their computational and experimental methods according summarized from Velandia-Huerto, Brown, et al. (2018).

cluster, previously reported in fruit fly (reviewed in Roush and Slack (2008)) have been characterized in Ciona spp. : the let-7/miR-125/miR-100 cluster described by Griffiths-Jones, Hui, et al. (2011) and studied in detail on metazoans by Hertel, Bartschat, et al. (2012). Along reported elements in Ciona spp. the miR-1473 was suggested as an ortholog of miR-100. On the appendicularian *O. dioica*, the same cluster was identified missing miR-125 (Hertel, Bartschat, et al., 2012).

As depicted in Figure 7, the cluster miR-96/miR-182/miR-183 has been characterized as sensor-specific miRNA polycistronic cluster in mouse and in a conserved synteny with human (Lagos-Quintana, Rauhut, Meyer, et al., 2003; Xu et al., 2007). In addition, this cluster was detected on lancelet, sea squirt, and coelacanth by Velandia-Huerto, Brown, et al. (2018). In this update, `miRBase` annotations complemented with homology searches where used to get contained miRNAs in the cluster. Specifically, the coelacanth did not have `miRBase` annotations, and by using homology searches the cluster was reconstructed as: miR-182/miR-96/miR-183. In case of the sea squirt, all miRNAs were annotated by `miRBase`, as: cin-mir-182/cin-mir-96/cin-mir-183. Finally, on lancelet the miR-96 is missing and reported as an inverted miRNA organization: bfl-183/bfl-96. In the pacific sea squirt (*C. savignyi*) those elements are dispersed, since their genome assembly has much

less contiguity, the cluster relation can not be inferred. In addition to those findings, the conservation at genomic level has been tracked back to bilaterian divergence, correlated at the same time with its important expression patterns characterized in zebrafish, mice, fruit fly and nematodes (see Dambal et al. (2015) and references herein).

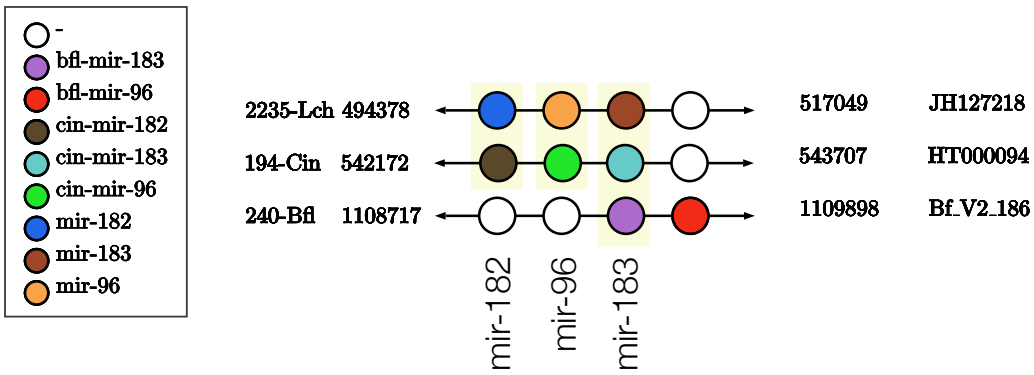


Figure 7: Multiple alignment of miR-182/miR-183 clusters. Specific names from annotations and homology predictions are described in the legend. Names from `miRBase` families are reported at the bottom of the aligned elements. White circles represent a non-detected/deleted miRNA represented as a gap: '-'.

Cases of specie-specific families were reported extensively in Ciona spp. and *O. dioica*. The genomic organization of those clusters span less than 3 miRNAs, which are composed by families that differ on the seed region on 1 nt, consequently reporting high homology. In *C. robusta* two large clusters: one containing 25 miRNAs from 3 families (Ci-mir-2200, Ci-mir-2201, and Ci-mir-2203) and other with 11 miRNAs, derived from 4 paralogous families (Ci-mir-2200, Ci-mir-2201, Ci-mir-2204, and Ci-mir-2217) were reported by Hendrix, Levine, and Shi (2010). For *O. dioica* five compact clusters (with 15 miRNAs) have been also identified, most of them recently duplicated carrying homologous miRNAs. For instance four miRNAs, miR-1490a, miR-1493, miR-1497d, and miR-1504, are reported by to be present as duplicated, and miR-1497d-1 and miR-1497d-2 are included in the large miR-1497 cluster, which in the largest region contains 6 miRNAs (Fu, Adamski, and E. M. Thompson, 2008), see more details in Appendix A: Table 16.

2.4.5 Current miRNA annotation status: preliminary study using available homology approaches

Using computational methods, the prediction of miRNAs over tunicates extended current miRNA annotation landscape. This repertoire is depicted as a Dollo parsimony (Farris, 1977), based on the final miRNAs family matrix retrieved from Hertel and P. Stadler (2015) and complemented by homology methods developed on Velandia-Huerto, Gittenberger, et al. (2016). Specifically for *S. thompsoni* and *H. roretzi* `Blast+` (Camacho et al., 2009) searches with structural alignments with `INFERNAL` (Nawrocki and S. R. Eddy, 2013) where applied on the reported candidates in Jue et al. (2016) and K. Wang et al. (2017) (Figure 8).

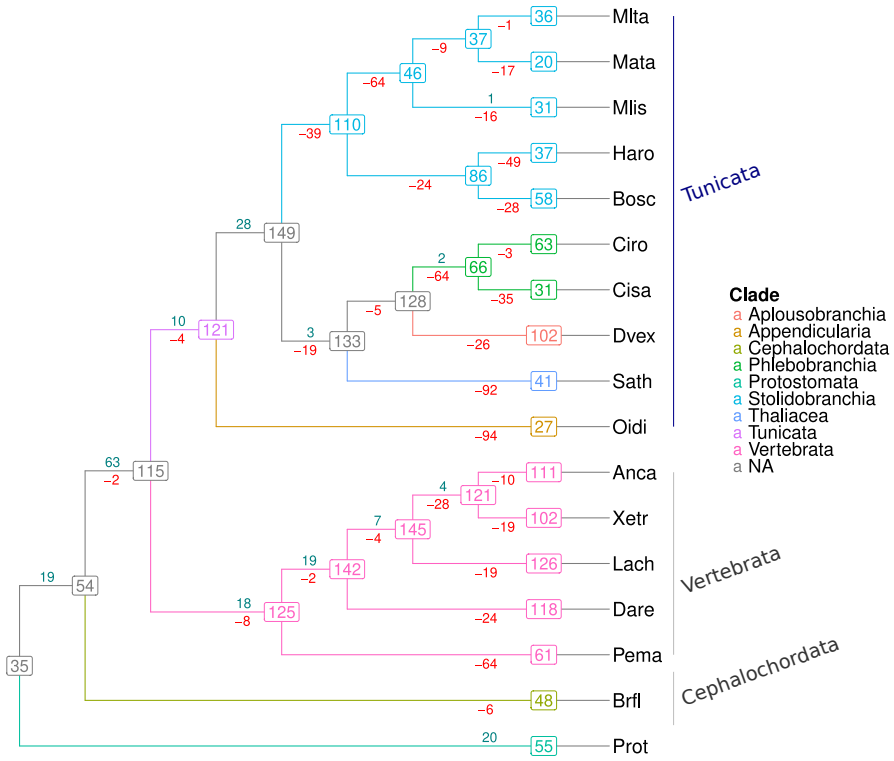


Figure 8: Dollo parsimony of miRNAs families distribution in chordate genomes. **Prot**: Protostomata, **Brfl**: *B. floridae*, **Oidi**: *O. dioica*, **Dvex**: *D. vexillum*, **Ciro**: *C. robusta*, **Cisa**: *C. savignyi*, **Sath**: *S. thompsoni*, **Mata**: *M. oculata*, **Mlta**: *M. occulta*, **Mlis**: *M. occidentalis*, **Bosc**: *B. schlosseri*, **Haro**: *H. roretzi*, **Pema**: *P. marinus*, **Dare**: *D. rerio*, **Lach**: *L. chalumnae*, **Xetr**: *X. tropicalis* and **Anca**: *A. carolinensis*. The phylogenetic distribution of this species was obtained from (Delsuc, Philippe, et al., 2018; Kocot et al., 2018). Dollo parsimony calculated using `Count` program (Csurös, 2010).

2.5 Computational approaches to discover homology relations

In order to understand the associated complexity of the genomic information, huge efforts have been done to infer and assign a biological meaning. Recently, with the development of next generation sequencing (NGS) techniques that generate large proportion of data in regard their low associated cost, the use of potent computational resources and fast/efficient algorithms is essential. Currently, according to the `NCBI`[6], since 1982 the number of bases on the `GenBank` has doubled every 18 months. The number of reported sequences deposited on the `GenBank` release 247 (December 2021) accounted 2.345×10^8 sequences representing 1053.3 Gbp. In this sense, as summarized by Gauthier et al. (2018) derived from the analysis of protein sequences, the development of computer programs was promoted together with the creation of a new branch of research that combine computational methods to resolve biological/medical questions. Thus, problems as protein assembly, determination of protein primary structure (Dayhoff and Ledley, 1962), and even comparisons of strings were addressed between 1950-1970. Further development of the *orthology* concept by Fitch (1970), with the interest to develop sequence comparison methods, promoted the creation of algorithms to align sequences and account their differences, later accounted as *indels*. In this section, a brief revision of computational methods to infer homology relations are described.

2.5.1 Homology and its link to evolutionary relations

Homology is the term designated to describe the relation between two characters that have descended from an ancestral character, derived from their cenancestor (Fitch, 2000). Translated to a practical statistical point of view, under the ancestor assumption characters that share *significant* similarity can be considered as *homologous*, without detailing their evolutionary scenario (Koonin, 2005; Pearson, 2013) and without assert that *similarity* means directly homology (Fitch, 2000). The inference of homology by computational means using primary, secondary or tertiary structure is based on the detection a similarity signal higher than expected by chance. This is related to the dependency of the sequences, that is assumed to be derived from a common ancestor (Pearson, 2013).

2.5.2 Pairwise alignment as first approach to link homology

Computationally speaking, comparisons between two strings are encompassed by the string edit problem, which search the minimum set of operations (edit distance) required to transform one string to another (Sung, 2009). Given the growing dataset of biologically meaningful sequences, fast and efficient algorithms have been developed to discover the homology or orthology relations. In those cases, similarities and not differences, are the particular interest when analyse biological sequences. The algorithms are classified as heuristic and optimal, depending on the increase of sensitivity or speed, respectively. The objective is to find regions of *similarity* between a *query* and *target* pair of sequences. The nature of this alignment could be classified as local, global or glocal. For optimal solutions and local alignments the most popular implementation was done by T. Smith

[6] `https://www.ncbi.nlm.nih.gov/genbank/statistics/`, accessed on 10.12.2021

and Waterman (1981), the global one by Needleman and Wunsch (1970), and *glocal* development was made by Brudno et al. (2003). In other way, heuristic methods find sub-optimal solutions but increasing the search speed. The best know algorithm proposed by Altschul et al. (1990) is `BLAST` together with the first one: `FastA` (Pearson and Lipman, 1988).

Needleman-Wunch algorithm

To avoid finding the optimal solution over all possible alternatives when aligning a pair of sequences by a divide-and-conquer strategy Needleman and Wunsch (1970) broken the problem making a progressive alignment of two amino acids at time. Using a dynamic programming approach, it generates: 1) every possible alignment accounting nucleotide/amino acid comparisons for various combinations of matched, mismatched or insert/delete pairs and 2) a score matrix to score the generated alignment (Mount, 2004). In this problem the objective function is to maximize the similarity score between sub-sequences. The dynamic programming approach computes optimal sub-solutions in a matrix D, where an entry $D_{i,j}$ represents the best score for aligning the residues $a_{1..i}$ with $b_{1..j}$. The according recursions are shown in Equation 2.1.

$$D_{i,j} = \max \begin{cases} D_{i-1,j-1} + s(a_i, b_j) \\ D_{i-1,j} + s(a_i, -) \\ D_{i,j-1} + s(-, b_j) \end{cases} \tag{2.1}$$

To recover the optimal alignment, a backtracing step must be calculated. A diagonal, horizontal or vertical arrow is drawn for each $D_{i,j}$ if $D_{i,j}$ equals $D_{i-1,j-1} + s(a_i, b_j)$, $D_{i-1,j} + s(a_i, -)$, or $D_{i,j-1} + s(-, b_j)$, respectively. The backtracing start from $D_{i,j}$ to $D_{0,0}$. The time complexity of this algorithm was calculated as $O(nm)$ time and $O(nm)$ space, with nm entries in the matrix $D_{i,j}$ (Sung, 2009).

Basic Local Alignment Search Tool (BLAST)

The significance of the alignment was a recurrent problem when aligning sequences (Mount, 2004). The problem was addressed by Altschul et al. (1990) with the `BLAST` program, designed to be faster than `FastA` at cost of lowering sensitivity (Sung, 2009). Based on the identification of local short *high scoring hits*, `BLAST` identifies and extends recognized matches (seeds) between pairs of homologous sequences. As described in Pertsemlidis and Fondon (2001), `BLAST` discriminates between random background and significant homologous sequence with a collection of raw scores, bit scores and *E-values*. In detail, *Raw scores* are the sum up of all contributions in the maximal-scoring segment pair (MSP). As a comparable variable, Bit score (S') accounts the log base of a scoring matrix (λ) and the scale of search space (K), related as:

$$S' = \frac{\lambda S - \ln K}{2 \ln} \tag{2.2}$$

Additionally, the *E-value* (E) is related with the Bit score (S'), as:

$$E = mn2^{-S'} \tag{2.3}$$

Where n is the sequence query length and m the size of the sequence database.

2.5.3 Hidden Markov Models and Covariance Models: modelling based on states

The Hidden Markov Models (HMMs) were applied the first time in 1970, focused on studies related to speech recognition (see Rabiner (1989) and references herein). The main idea was to characterize observed output as signals, through a model. This model, once constructed, can describe theoretically a system and could be used to modify system variables to generate desired outcomes, additionally through the model inherent features from the source can be inferred. As a type of statistical stochastic models, Markov and Hidden Markov processes, allow that signals can be characterized as parametric random process (Rabiner, 1989). In brief, to understand the nature of HMMs, the definition of a Markov chain are defined in terms of *states* and the *probabilities* of taking on values. Given the state variables $\mathrm{Q_s} = q_1, q_2, \ldots, q_i$, it makes use of the *Markov assumption*, as described by Equation 2.4. It requires that the probabilistic description of the current state (q_i) depends on exclusively to its predecessor q_{i-1}.

Markov chain

Markov assumption

$$P(q_i = a|q_1 \ldots q_{i-1}) = P(q_i = a|q_{i-1}) \tag{2.4}$$

In addition, the Markov chain have a transition probability matrix $\mathrm{A_{ij}}$ between the $\mathrm{Q_s}$ (Equation 2.5) satisfying $\sum_{j=1}^{n} aij = 1, \forall i$. At the same time, an initial probability distribution defined as $\pi = [\pi_1, \pi_2, \ldots, \pi_N]$, with $\sum_{i=1}^{N} \pi_i = 1$.

$$A_{i,j} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & a_{i,2} & \cdots & a_{i,j} \end{pmatrix} \tag{2.5}$$

In the same way, in cases when the events are not observable (hidden), it is useful to addition to the previous Markov chain parameters, a set of observations $\mathcal{O} = O_1, O_2, \ldots, O_T$ and their associated likelihood or emission probabilities $\mathcal{B} = b_i(O_T)$, to define a Hidden Markov model as $\lambda = (A_{ij}, \mathcal{B}, \pi)$. An additional assumption is considered with HMMs, concerned with the exclusive dependence of the output observations (O_i) with its corresponding producing state (q_i) as shown in Equation 2.6 (Jurafsky and Martin, 2020; Rabiner, 1989).

Hidden Markov Model

$$P(O_i|q_1 \ldots q_{i-1}) = P(O_i|q_i) \tag{2.6}$$

The correspondence between a multiple alignment and a HMM, can be modelled by a *profile* HMMs. Once build, it is possible to search the model for other sequence candidates in large databases (Krogh, 1998). In this regard, the *states* ($\mathrm{Q_s}$) are: *match* (**M**), *insert* (**I**), and *delete* (**D**), and additional start and end states (see Figure 9). Subsequent comparison between the HMM to sequences, that are not included in the alignment, is done asking about the best sequence of hidden states $\mathrm{Q_s}$ given a query sequence. An efficient solution is performed by the `Viterbi` dynamic programming algorithm, yielding

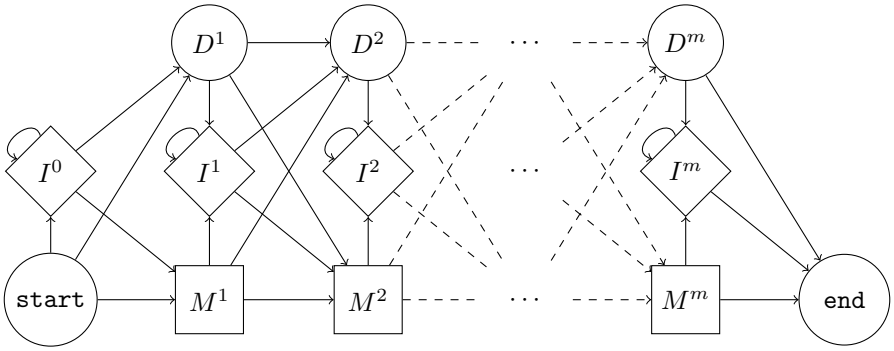


Figure 9: Profile HMM illustration without showing the transition and emission probabilities. The start is modelled as column 0. Other consensus columns are modelled with their corresponding states: match (M), Insertion (I) or deletion (D). For M and I, exists 20 or 4 emission probabilities in case of proteins or RNA/DNA sequences, respectively. Meanwhile, for D there is no emission probabilities. Transition probabilities are represented as arrows. End state is included. Figure based on (S. R. Eddy, 1998).

the probability of the sequence in that model, that can be translated into a log-odds score (Krogh, 1998).

By its own, Covariance Models (CMs) adds the conservation of well-nested (not including pseudoknotted basepairs) secondary structure in case of RNAs (S. R. Eddy and Durbin, 1994). They are a type of profile stochastic context-free grammar (SCFG), represented by an ordered tree that captures both, sequence and structure information. Minor structural variations are captured based on additional states: begin, bifurcation, insert-left, insert-right, match-pairwise, match-left, match-right, and delete. At the same time, *transition* and *emission* probabilities are considered as explained before for HMMs (S. R. Eddy and Durbin, 1994). In comparison to HMMs, the additional states in CMs are: bifurcations and pairwise states.

2.6 Genome sequencing: translate DNA information into raw data

For more than 30 years ago the *chain termination* or Sanger method has been used widely to determine the nuclear sequence of genes and complete genomes, being recognized with the Nobel Prize in Chemistry in 1980 (Nobel Prize Outreach AB, 2021a; Sanger, Nicklen, and Coulson, 1977). Up today, this method can generate read-lengths ~ 1000 bp, and a per-base accuracy $> 99.99\%$ (Shendure and Ji, 2008). However, nowadays it is not longer the most used sequencing method. Initial enthusiasms was dampened as technological challenges arose, after the completion of the human genome sequencing (International Human Genome Sequencing Consortium, 2004) such as the high associated cost, and the increasing interest to get more sequenced genomes, which pushed up to improve the overall sequencing protocol.

In 1996, the development of the sequencing-by-synthesis (know as pyrosequencing) by Ronaghi et al. and subsequently in 2005 its commercialization by 454 Life Sciences (Margulies et al., 2005), constituted a milestone to infer nucleotide identity making use of parallel reactions and being recognized later as the beginning of *second-generation sequencing* technologies. As a response of previous challenges recognized in Sanger sequencing, second-generation sequencing methods reduced their associated cost (due a reduction of reaction volumes and parallelization), and the read size, but increased the number of sequencing reactions (Schuster, 2007). This constant optimization has been reflected on the subsequent development of additional sequencing platforms and methodologies. A classification can be produced by means of the method strategy, as pointed by Goodwin, McPherson, and McCombie (2016): Sequencing by ligation (SBL) and Sequencing by synthesis (SBS).

In SBL, a step of hybridization and ligation is performed between a labelled probe and anchor sequences to a free DNA strand. Ligation is initiated by an adapter sequence as preliminary step for ligation and posterior identidication of the base(s) in the probe. Example platforms are SOLiD platform (Applied Biosystems) and Complete Genomics (BGI) (Goodwin, McPherson, and McCombie, 2016). By its way, SBS comprised a broad category of techniques dependent of DNA-polymerase, such as cyclic reversible termination (CRT) and single-nucleotide addition (SNA). For CRT is included the Illumina platform and for SNA the described 454 and Ion Torrent. It turned out that the data management for those sequencing platforms started to be challenging, due the increased quantity of data.

Towards a longer, high-throughput generation of data the third-generation sequencing methods, started with the development of PacBio, enabling sequencing fragments up to 30 kb - 50 kb or longer, at single-molecule level and in real-time. This is achieved by the Single Molecule Real Time (SMRT) sequencing technology, which ligate adapters to double-stranded DNA to generate a circular library (SMRTbell®), called: circular consensus sequencing (CCS) and continuous long read sequencing (CLR). Then, this library is immobilized into a special nano-photonic chamber, referred as Zero Mode Waveguide (ZMW), where the real-time nucleotide addition is performed in a proper aluminium background to detect the emitting fluorescent light, excited by a laser light. This is generated when the polymerase grows the new synthesised chain, by incorporating distinguishable fluorescently labeled deoxyribonucleosides triphosphates (dNTPs) (Eid et al., 2009). The duration of this light impulse is in milliseconds, and the time between emissions are termed inter-pulse duration (IPD). This parameter allows the detection of base modifications, such as methylated bases (Eid et al., 2009; Rhoads and Au, 2015; Slatko, A. F. Gardner, and Ausubel, 2018). This sequencing method can be complemented and enhanced using hybrid strategies, for example to reduce the high median error rate reported by PacBio ($\sim 11\%$ associated with deletions or insertions) (Korlach, 2015), it can be combined with a short high-accuracy reads, as obtained by Illumina. And using the long-reads as scaffold (Rhoads and Au, 2015). Finally, four-generation sequencing methods have been developed as biological membranes that directly sequence the DNA/RNA, promoted by Oxford Nanopore Technologies with the portable MinION®. This technology measures the electric current in the flow cells containing a hole (nanopore), which the current disruption caused by a molecule (squiggle) is decoded to define the DNA/RNA sequence, for detailed explanation refer to Oxford Nanopore Technologies (2022).

At August 2021, the associated cost per megabase was US$0.006 and per genome about US$562, according to the NHGRI Genome Sequencing Program (Dunham, 2005). As described by Shendure and Ji in 2008 this processes 'democratizing the sequencing field, putting the sequencing capacity of a major genome center in the hands of individual investigators'.

Part II

Ways to challenge and improve current miRNA annotation

— 3 —

miRNA annotation and current pitfalls

Contents

3.1 Sources and state of miRNA annotation

Despite the vast number of references related to microRNAs (miRNAs) and the overall description of multiple species, still an efficient and consistent miRNA computational detection method is missing. This is challenging when a new sequencing project faced an initial approach to annotate protein-coding or non-coding elements in an interest species.

In this chapter, a thoughtful analysis of current miRNA annotations was used to evidence up-to-date pitfalls identified on public databases and in miRNA annotation *per se*, as published on Velandia-Huerto, Yazbeck, et al. (2022).

3.1.1 Current miRNA annotations and databases

Historically, the repository of animal miRNA annotations has been the *microRNA Registry* (`miRBase`) (Griffiths-Jones, 2004), which in its release v.22.1, contained information of 38,589 miRNAs belonging from 285 species, classified as: Metazoa (158), Viridiplantae (86), Viruses (34), Chromalveolata (5), Alveolata (1), and Mycetozoa (1). In Metazoa, almost all species are bilaterians, except for 3 Cnidaria and 3 Porifera. The Bilateria clade is represented by species contained in Deuterostomia (84), Ecdysozoa (53), and Lophotrochozoa (15). The overrepresented clades in Ecdysozoa are Hexapoda and Nematoda, and in Lophotrochozoa: the Platyhelminthes. In Deuterostomia an overrepresentation of chordates, specifically on vertebrates (74), is evident when compared to other chordates: 2 species of cephalochordates and 3 from tunicates. To annotate a miRNA, the source of annotations depends on submitted sequences and published works, most of the time focused on model species. Once a miRNA is submitted, a set of rules to annotate it have been stated by Ambros et al. (2003) and later formalized by Griffiths-Jones, Saini, et al. (2008). This classification included, for some loci, a family annotation [1] and a name assignment.

In this respect, `miRBase` v.22.1 accounted for 1983 families. However, about 50% (19,326) of annotated precursors have not been assigned to a miRNA family. Additionally, a proper delimitation of *canonical* or *non-canonical* miRNAs is missing. As a consequence, multiple reports pointed out an outstanding number of 'false positives', when compared them as *canonical miRNAs* (see Velandia-Huerto, Yazbeck, et al. (2022) for detailed examples). Last but not least, `miRBase` have not been updated since 2019 (Kozomara, Birgaoanu, and Griffiths-Jones, 2019), despite the large increase of miRNA annotations.

In parallel, the annotation of non coding RNAs (ncRNAs) (including miRNAs) has been approached by the `Rfam` database (Griffiths-Jones, 2003). It has taken advantage of the construction of multiple structural alignments of ncRNA families as a direct link of a conserved functionality. From the multiple structural alignments, structure profiles as Covariance Models (CMs) has been built to collect curated RNA families. In the release 1.0 it contained *seed* alignments [2] from ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), a dataset of bacterial RNAses, and miscellaneous RNAs, see details in (Griffiths-Jones, 2003). At this release, the construction

[1] Since `miRBase` release 8.1, the relation between loci and their family assignment is reported on the `miFam.dat` file.

[2] Here the *seed* refers to the representative hand-curated set of sequences that compose a family structural alignment.

of a general miRNA model was perceived as a challenge. By the use of the *seed* alignments, `Rfam` search over all the sequences deposited on the `RFAMSEQ` database and perform an iterative process to assembly a representative set of sequences of a ncRNA family. Then, a multiple structural alignment is build with its corresponding CM (Kalvari, Argasinska, et al., 2018). By this method, in the last release 14.6 the `Rfam` reported 4070 families[3], which 1506 are classified as miRNAs. However, in a short-term, miRNA annotations would be constantly updated in `Rfam`, which started to integrate `miRBase` families. The first phase of the project added 973 new and updated 152 miRNA families[4].

Critical references pointed a higher proportion of false negatives in `miRBase` are reported in Fromm, Billipp, et al. (2015). In particular, a careful evaluation of the sequence and structural features from miRNAs annotations in `miRBase` shows that only about 16% of the metazoan annotations are robustly supported (Fromm, Billipp, et al., 2015). The authors defined a set of rules that relies on an evaluation of the expression data, including: expression of both arms, 2-nt offset, homogeneity at 5' starts, and evolutionary conservation (Fromm, Billipp, et al., 2015). As a result, the database `MirGeneDB`[5], re-evaluated `miRBase` annotations, and proposed new annotations based on small RNA-seq analysis. In their last release, `MirGeneDB` v.2.1 reported >1500 miRNA families from only 75 metazoan species (Fromm, Høye, et al., 2021).

As an effort to centralize and integrate multiple ncRNA annotations in one resource, `RNAcentral`[6] acts as a front-end of multiple databases that have annotations on same species. One of the advantages is the assignment of a stable unique identifier (URS) assigned to the identified molecules together with the visualization in a unique genomic browser, that allows to explore reference genomes annotations for 560 species, derived from 44 RNA resources and > 13 million sequences, in its release 19 (Consortium et al., 2020).

3.1.2 Homology search as a link to detect miRNAs

Despite their short sequence length, it is usually not too difficult to identify homologs of known miRNAs. Their mature sequences are nearly perfectly conserved and thus are convenient anchors for sequence-based search methods, such as `blastn`. The sequences of the precursor hairpins are usually also quite well conserved, evolving at rates comparable to coding sequences. The selective constraints on the mature sequences are stronger by up to an order of magnitude (Nozawa, Miura, and Nei, 2010). The structural constraints on the precursor hairpin, finally, imply an approximate complementarity between the miR and miR* sequences[7], even if the miR* sequence is not functional in its own right. As a consequence, simple, sequence-based methods are usually successful at least at phylum level. A `blastn` search with the human miR-10a precursor, for instance, readily yields top hits with E-values $< 10^{-40}$ in mammals, $< 10^{-20}$ in sauropsids, and $< 10^{-8}$ in the duplicated genomes of teleost fishes. The methods quickly loses power outside the vertebrates, however, significant hits $E < 10^{-5} \ldots 10^{-3}$ are found in *some* echinoderms and

[3] Accessed at November 11^{th}, 2021
[4] `https://rfam.org/microrna`
[5] `https://mirgenedb.org/`
[6] `https://rnacentral.org/`
[7] See explanation of miR and miR* in Chapter 2 Figure 2

protostomes. Sensitivity and specificity can be increased by optimizing `blast` parameters and by comparing "hit lists" obtained with different parameter settings, as approached by Velandia-Huerto, Gittenberger, et al. (2016) to the annotation of ncRNAs.

MicroRNA precursors also feature well-conserved secondary structure. The phylogenetic scope of homology searches can therefore be expanded by employing CMs (S. R. Eddy and Durbin, 1994). CMs are a generalization of Hidden Markov Models (HMMs) that incorporates the co-variation of paired bases. Thus, the specificity of CM-basic homology search with `infernal` (Nawrocki and S. R. Eddy, 2013; Nawrocki and S. Eddy, 2005) is considerably increased compared to sequence-only methods such as `blast`, full dynamic programming alignments (Hertel, Jong, et al., 2009), and HMMs. CMs are trained from sequence alignments annotated by a consensus structure for the aligned sequences. The `Rfam` database (Kalvari, Argasinska, et al., 2018) provides such miRNAs alignments. It is not comprehensive in its coverage of miRNAs families, and it does not report mature miRNAs. On the other hand, `mirBase` (Kozomara, Birgaoanu, and Griffiths-Jones, 2019; Kozomara and Griffiths-Jones, 2013) provides a much more complete coverage of the miRNA precursor and mature sequences. In addition, miRNA family alignments in `miRBase` have to be manually curated or extended by additional members.

In this way, the general workflow for a *de novo* construction of a miRNA family alignment is as follows:

(1) Obtain a set of *seed* sequences covering as evenly as possible the phylogenetic range of family members that are known already.

(2) Construct a multiple sequence alignments (MSAs) of the *seed* sequences, using one of the many tools such as `Muscle` (Edgar, 2004), `MAFFT` (Katoh, 2002), `ClustalW` (J. D. Thompson, Higgins, and Gibson, 1994), `t-coffee` (Notredame, Higgins, and Heringa, 2000), or `dialign` (Morgenstern, 2004),

(3) Compute the consensus structure e.g. with `RNAalifold` (S. H. Bernhart et al., 2008).

(4) In general, a curation of the 3' and 5' ends is necessary. Ideally, this step includes the evaluation annotated ends of the precursor relative to the location of the mature sequence and in relation to secondary structure. If sequences are extended or trimmed, steps (2) and (3) should be repeated.

MSAs are sufficient to construct HMM, which provide an alternative to searching homologs of the seed sequences individually. A convenient and efficient implementation is `nhmmer` (Wheeler and S. R. Eddy, 2013). The `infernal` package provides the necessary tools to convert a MSA, that is annotated by a consensus structure, into a CM to identify score thresholds for significant matches, and to search a nucleotide sequence for approximate matches.

Instead of starting from sequence alignments, it is also possible to first predict secondary structure for each sequence with `RNAfold` (Lorenz et al., 2011) and to use as a second step a structure-based alignment tool such as `LocARNA` (Reiche and P. F. Stadler, 2007; Will et al., 2012) or `MARNA` (Siebert and Backofen, 2005) to obtain an alignment together with a consensus structure. The resulting MSA will in general require some level of user intervention to identify any obvious alignment errors, incorrect or corrupted sequences, and other errors.

In many cases, only a single query sequence is available at the outset. This is in particular often the case when miRNAs are identified from small RNA-seq data. The natural first step is then to approximate the precursor hairpin. This can be achieved by extracting about ±100 nt of genomic flanking sequence on both sides, which is sufficient to ensure that the precursor hairpin is completely contained in the extracted sequence. The precursor hairpin, if it exists, can then be identified by computing the secondary structures. An elegant method for this purpose is `RNAplfold` (S. Bernhart, Hofacker, and P. F. Stadler, 2006), which allows the extraction of locally stable structures by restricting the base-pair span to 80 or 100 nt. Since miRNAs fold into much more stable structure compared to random RNA structures with the same sequence composition, see e.g. some examples described in Clote et al. (2005), Freyhult, P. P. Gardner, and Moulton (2005), and B. H. Zhang et al. (2006), the most stable *local* stem-loop structure identifies the precursor miRNA (pre-miRNA) hairpin. In Yazbeck, P. F. Stadler, et al. (2019), instead a simple rule is used to estimate the precursor sequence depending on whether the miR is assumed to be located on the 5'- or 3'-side of the hairpin; both are then folded and the more stable hairpin is retained. Once a plausible precursor has been found, closely related genomes can be conveniently searched with `blastn` for an initial set of homologous examples, from these a structure-annotated alignment can then be computed as above.

Figure 10 summarizes the outline of the workflows, starting from a single candidate *query*. Targets for the search are typically genomic sequences (at various stages of assembly) retrieved from `Ensembl` or `NCBI`. Typically, the iterative searches use relaxed parameters optimized for sensitivity and necessarily produce large numbers of false positive candidates. These require extensive curation steps, which in part are performed by automatic filtering procedures (considering quantitative measures such as E-values, sequence identity, the presence of highly conserved miR sequences, etc.), and in part rely on user intervention and manual inspection. Most studies use the scheme in Figure 10 as a guideline rather than as a fully integrated pipeline. A notable exception is the recent survey of tunicate ncRNAs Velandia-Huerto, Gittenberger, et al., 2016, which combined `blastn`-based searches with several combinations of parameters with sequence-based searches using HMMs to generate candidate sets. These are then filtered using CMs for the known miRNA families as a first automatic curation step.

3.1.3 Computational identification of miRNAs

Homology searches on large phylogenetic groups are usually performed in an iterative way, e.g., using new candidates as a means of expanding and refining the search. Purely `blast`-based approaches usually use the new candidates as additional queries. In HMM or CM based approaches, the alignments are expanded with the new candidates. This can either be done by re-aligning the entire set of homologs, or by aligning the new sequence(s) to the HMM or CM using specialized tools. The augmented alignments are then used to re-train the model. Typically, this process is iterated until no further candidate homologs can be found.

Iterative homology search requires the evaluation of the candidate hits, and typically involve an update to the MSA and its consensus structure. For the latter task, there are at least two distinct strategies:

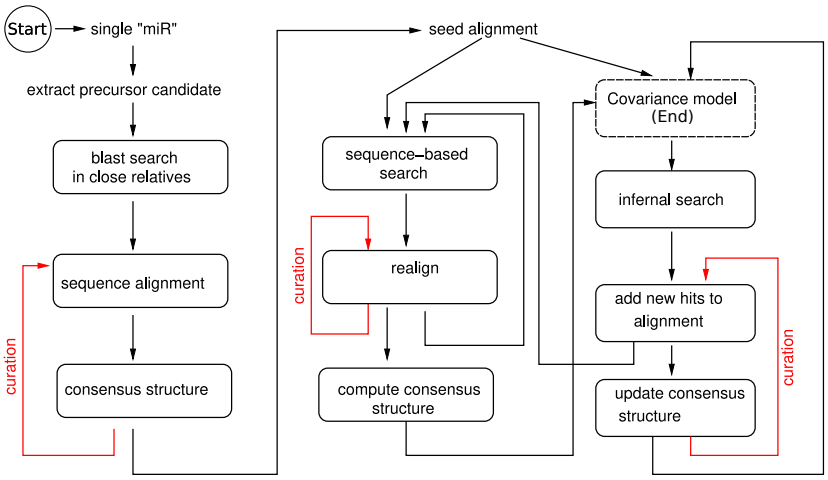


Figure 10: General workflow for homology search for miRNAs and other structured RNAs. In the initial phase, the goal is to obtain a seed set of trusted homologs starting from a single small RNA (obtained e.g. by sequencing) or a predicted precursor structure. This seed set is then expanded iteratively. Often a sequence-based search can efficiently expand the phylogenetic scope considerably, leading to a collection of homologs with sufficient diversity to allow validation of the consensus structure by patterns of sequence co-variation. Sequence-based searches may be performed e.g. using `blast`, full dynamic programming alignment tools such as `gotohscan` (Hertel, Jong, et al., 2009), or Hidden Markov Models (inferred from the sequence alignment). Alignments annotated with a consensus structure allow the construction of CMs. Often these are more sensitive than purely sequence-based models. Importantly, putative homologs need to be curated either manually or with the help of automatic means to avoid the inclusion of false positives into the next iteration of the search. A possible 'end' can be at the construction of the CM (see dashed box). Modified plot from Velandia-Huerto, Yazbeck, et al. (2022).

(a) If the search started from a trusted seed alignment, new hits are often added individually. Many sequence alignment tools offer an option of align individual sequences to a given MSA. Similarly, individual sequences can be aligned to HMMs and CMs. Since there is a correspondence between the positions in the model and the columns of the seed alignment, this also determines how the candidate fits to the MSA.

(b) If no trusted seed is available it may be preferable to completely realign the union of query and candidate sequences. As for the seed alignment, this can be done either purely sequence-based alignment methods or with the help of a structure-based alignment method.

The update of the alignment in general also prompts an update of the consensus structure.

Recently, pipelines implementing this workflow have become available for general use. `RNAlien` (Eggenhofer, Hofacker, and Höner zu Siederdissen, 2016) and `GLASSgo` (Lott et al., 2018) are primarily intended for the small RNAs of procaryotes. It has not been

applied to animal miRNAs so far, although conceptually it should be suitable for this task as well. A partial solution for miRNA families in which at least one representative has a documented miR and miR* sequence has become available in (Yazbeck, Tout, et al., 2017).

The particular structure of miRNA precursors can also be utilized to devise a miRNA-specific approach. In the first step, near exact matches of the mature miR are retrieved. In the second step, the flanking sequences of the initial candidates are retrieved and investigated for the presence of a significantly stable hairpin structure. The pre-miRNA candidates that satisfy the filtering criteria can then be treated as above. A quite general workflow, `MIRfix`, combining these structure-based criteria with annotation of miR and miR* sequence is described in (Yazbeck, P. F. Stadler, et al., 2019). This method has been used for the annotation of miRNAs in tunicates (Parra-Rincón et al., 2021).

3.2 How to combine current annotation resources?

3.2.1 Curation of Rfam miRNA families

Since pre-miRNA alignments and their CMs in `Rfam` did not specify the position of *mature* sequences, the annotation of those regions on annotated pre-miRNA were performed using the available mature sequences from `miRBase`. Extending the methodology reported on Parra-Rincón et al., 2021, a combined strategy using the information retrieved from `RNAcentral`, `Rfam` and `miRBase`, allowed the assignment to the best available mature sequence to `Rfam` pre-miRNAs, using the source of matures and their correspondent hairpin sequences obtained from `miRBase`, as shown in Figure 11. Input data was automatically pre-processed with a `Perl` script to subsequently use them as `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019) input[8]. This program facilitated key steps towards the correct positioning and prediction of the mature sequences, by the following process:

- Based on the family mature database, define the best *mature* that fit into the pre-miRNA sequence.
- Defined this sequence, reported the best position of the *mature* sequence.
- Given the positions of all *mature* sequences, calculate a new anchored-multiple alignment, given the positions determined by annotated mature regions.
- If detected some sequences that could be inverted or bad positioned, those sequences were corrected to be included on the final family alignment.
- Reported at the end the set of coordinates of the candidate mature sequences with their corresponding sequence and precursor sequence, too.

By one way, using the reported sequences from each `Rfam` miRNA family and their associated `RNACentral` unique accession numbers, via `Posgresql` public service, served as a bridge to associate `miRBase` data (precursors + mature sequences). By one way,

[8] All the annotations and corrections of the mature sequences were performed using `MIRfix` v2.0.1, modified in this thesis project.

those Rfam families that reported at least one sequence with mature annotation were analysed with MIRfix. As explained before, this pipeline was used to correct the position of their annotated mature sequence(s) and correct their pre-miRNA sequence. Next, taking together the corrected mature and precursor sequences, the remaining sequences without previous mature annotation were subject to detection of best mature region in behalf of available mature sequences using MIRfix.

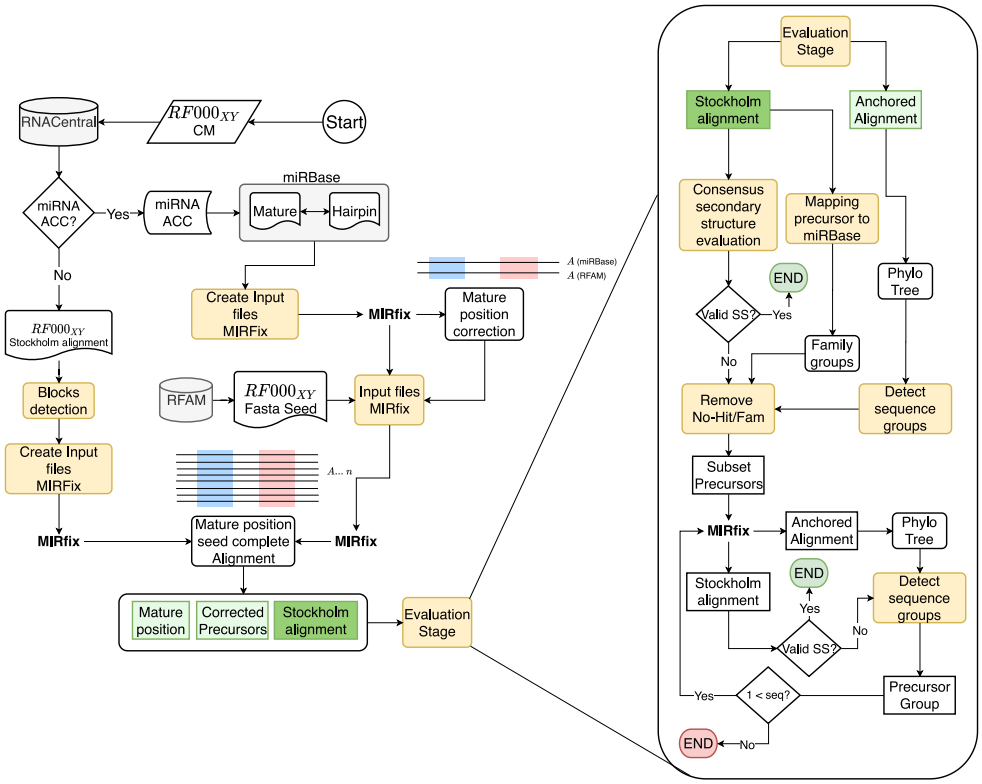


Figure 11: Complete workflow of annotation of mature miRNA Rfam sequences. Iteration over a Rfam miRNA family starts locating their annotation in RNACentral and miRBase. In one way, when identified mature sequences in miRBase led to the creation of an anchored alignment (blue and red regions) including the sequences from the Rfam family. On the other way, the missing mature reference is inferred from the structural conservation. As a result MIRfix corrects the sequences, gives mature positions and generate a corrected structural alignment. This input data is suitable to perform the evaluation stage, as described in text.

By the other way, in case that any of the sequences reported a mature annotation, an inference of the mature regions was performed based on the analysis of the consensus structural sequence, provided by the structural Rfam alignment, which considered conserved 5' and 3' stem regions. The curation process performed on the Rfam families allowed the refinement of the family pre-miRNA sequences, its alignment, and the identification of the position of candidate mature sequences. In the same way, it allowed the detection

of certain sequences that did not fit into the re-defined alignment, which were discarded or even corrected. Those sequences that missed annotating mature sequences reported a minimum free energy (MFE)> -10 or were excluded of the anchored family alignment.

3.2.2 Evaluation of structure consistency

At the end of the mature annotation stage the mature positions, corrected precursors, and final anchored structural alignments were obtained for each evaluated family. Using this calculated data, a further *Evaluation Stage* was done focused on obtained consensus secondary structure from the anchored structural alignments. In one hand, the consensus secondary structure was evaluated in terms of number of matching mir and mir* columns (> 20 nt) and an evaluation of the secondary structure to be similar to a hairpin-like structure, with both mir and mir* in the stem region with a central loop.

3.2.3 Rescue of non-structured miRNA models

Basic criteria have been defined to curate multiple structural alignments, which led to the improvement of alignments and rescue of some miRNA alignments, founded on the evaluation and dynamic modification of their structural alignments.

Filtering sequences as path to rescue alignments

For those families that did not have a valid secondary structure, a classification in comparison to `miRBase` v.22.1 precursors was accessed by `blastn`. When a family assignment was found, the `miRBase` family annotation were kept to use as label of the `Rfam` miRNA sequence. In parallel, based on the generated anchored alignment by `MIRfix`, a phylogenetic tree was generated using `UPGMA` clustering by `ClustalW` (J. D. Thompson, Higgins, and Gibson, 1994) (see Listing 3.1). Next, families were evaluated in terms of their assigned families. Those sequences that did not report a label or matched to a sequence without classification, were removed from the set of sequences. Then, the subset of sequences was submitted for another iteration of mature annotation with `MIRfix` and phylogenetic tree reconstruction based on the resulted alignment. Additional round of consensus secondary structure determined if the current set of sequences were representative for the studied miRNA family. If not, based on the phylogenetic tree, the sequences were divided into labeled groups or by the obtained clades (composed by $2 >$ sequences), depending on whether were assigned more than 2 labels in the family or not. Next, each detected group of sequences was processed again with `MIRfix` and the subsequent analysis until reached one of this end points: a valid secondary structure or the number of sequences were exhausted.

```
clustalw -in <align_file> -tree -outputtree=phylip -clustering=UPGMA
```

Listing 3.1: Creation of UPGMA clustering trees using `clustalw`

3.3 Anchored-structured alignments to curate miRNA families

3.3.1 Rfam as source of miRNA families

To get an idea how the miRNA CMs are represented in the Rfam v.14.2, were identified a set of 529 miRNA CMs, built based on *seed* alignments. In one hand, 6168 hairpin sequences composed the miRNA *seed* alignments. In the other hand a larger set of 209,080 sequences composed the *full* alignments, inferred from an iterative search along the RFAMSEQ database, with the pre-built *seed* CMs (Kalvari, Argasinska, et al., 2018).

In general terms, most of the *seed* alignments are composed by a set of heterogeneous species. However, this is widely dominated by sequences from Chordata (3961), Magnoliopsida (895) and Arthropoda (781). Those clades together sum up 91.4% of the miRNA *seed* sequences. In this thesis, the analysis of *bona fide* miRNAs is restricted to metazoan sequences, which covered 83.4% (5144) of the miRNA *seed* sequences distributed along 449 target metazoan CMs. Additionally, a subset of 462 CMs contained at least one miRNA sequence from metazoans.

Given this annotation context, the space of metazoan sequences from Rfam was compared to other miRNA databases, in terms of shared and/or non-shared annotated accession numbers, such as: miRBase (Kozomara and Griffiths-Jones, 2010) and MirGeneDB (Fromm, Billipp, et al., 2015). As shown on Figure 12A, the comparison between those databases, included the RNACentral database (RNAcentral Consortium, 2019), as an integrative resource of almost all reported non coding RNAs (ncRNAs). It shows that Rfam contained a high number of predicted miRNAs that were not shared on the other evaluated databases, but only visible on RNACentral (71,573). Regarding common sequences, were identified 328 miRNAs shared along Rfam, miRBase and MirGeneDB with their correspondent unique identifier (URS) in RNACentral. In contrast, 524 sequences were found as *orphans*, since were not related on RNAcentral and are Rfam-specific.

Through this approach were identified sequences that share the same URS, with other sequences in RNAcentral. Those are designated as *rnacentralMissing* in Figure 12. As an example, the sequence recognized on RNAcentral as URS0001BC2ADC_9606[9] annotated for human as mir-1296, mapped for three Rfam sequences: AADB02013109.1/161463-161352, AADC01091081.1/15063-14952 and AC022022.10/21212-21101 from the *seed* family RF01921, mir-1296 from Rfam.

Constraining the analysis to *seed* sequences (Figure 12B), the most abundant references are annotated in miRBase (29,604), exceeding 4.6× the number of references on other databases. Most of the *seed* Rfam sequences were identified with an URS (4189) but not reported in another database, which means that 81.14% from the *seed* sequences are currently Rfam-specific sequences. Moreover, 5.4% (277) have been reported along all databases with an URS. As seen before, a reduced set of 285 pure Rfam-specific sequences were found without URS. At the same time, 136 *seed* sequences were labelled with the same URS.

3.3.2 Curation of Rfam miRNA families through multiple database integration

Since Rfam reported a precursor that contains the mature miRNA sequence, available CMs did not contain information about the position of anchored mature sequences: miR

[9] https://rnacentral.org/rna/URS0001BC2ADC/9606

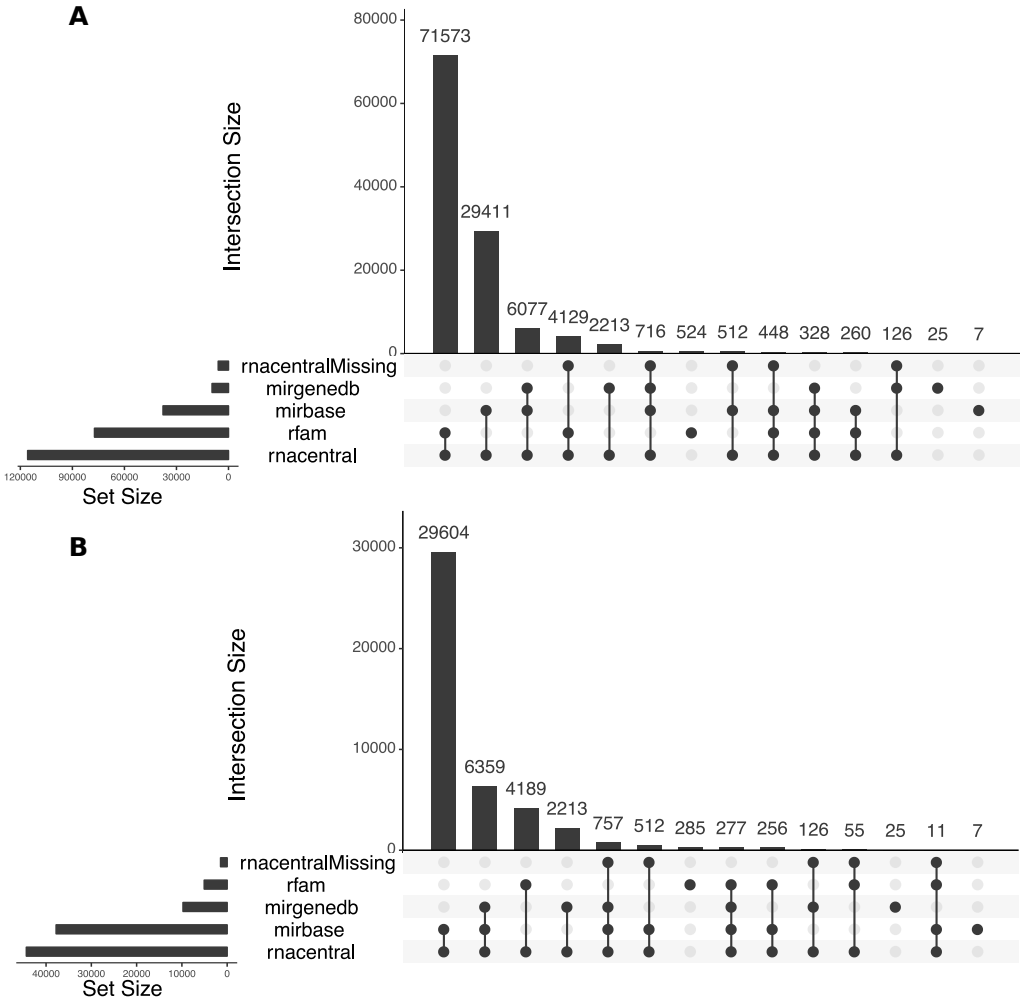


Figure 12: Distribution of miRNA annotations along three ncRNA databases: `miRBase` (miRNAs), `MirGeneDB` (metazoan miRNAs), and `RFAM` (ncRNAs). Additional comparisons were also performed including the database `RNAcentral`, which centralized on the most of the cases the reported data. Left panel with *Set size*, distribution accounted for the total number of annotated miRNAs, for `RNAcentral` accounted for the total number of common registers with all other databases. At the same time, were detected a set of candidates that shared the same URS from `RNAcentral` with other candidates (*rnacentralMissing*). **A.** Comparisons for all the sequences reported from `Rfam`: *full* alignments. **B.** Only taking into account the *seed* sequences.

and miR*. This information was inferred using `miRBase` annotated mature sequences and multiple anchored structural alignments were constructed, as described in detail in Section 3.2.2 and Figure 11.

To do so, from the previous selected 449 metazoan miRNA models, the annotation of mature sequences detected on the first round of annotation was successful for 437. The remaining set of 12 miRNA families were in most of the cases discarded due the lacking of reliable mature annotations to be annotated along all the miRNA family *seed* sequences: mir-31 (RF00661), mir-198 (RF00681), mir-458 (RF00750), mir-257 (RF00788), mir-42 (RF00974), mir-1302 (RF00951), mir-604 (RF01041), and mir-1803 (RF02094). At the same time, some of them shown misleading structural alignments: mir-1419 (RF001919), mir-2518 (RF001944), and mir-56 (RF02214) or even contained only one valid sequence annotated in human as mir-BART3 (RF00866), from which was not possible to build a posterior multiple structural alignment (for details see Appendix B:Table 17).

From the remaining set of 437 models, an additional workflow to evaluate the resulting multiple anchored alignments is depicted in Figure 11. A round of structure evaluation detected 79.9% valid candidates. For the remaining families that failed the first structural evaluation (88), a subset of sequences from each analysed alignment were selected in terms of their annotated miRNA family gene, found using `miRBase` precursor annotation as a reference, and the reconstructed phylogenetic tree from the anchored alignment produced at the mature annotation step. In general, mapped sequences into a non-family sequences or did not generate a hit were removed. At the same time, those that were clustered as outliers into the phylogenetic tree (see examples for validated and corrected miRNA families in Figures 14, 15, and 16).

The rescued 55 miRNA families using the described methodology reported a valid secondary structures and annotated mature regions. However, through the validation the remaining set of 33 miRNA families failed the validation step by detection of an invalid long-hairpin structures, a number of sequences < 2 with a correct `miRBase` family assignment, reported multiple family assignments that were classified in the same family, or were discarded on the curation process (see discarded examples at Table 2).

In more detail, Figure 13 shows the overall results from curated set of `Rfam` miRNA families. A distinction between sequences that were part of a `Rfam` families with mature annotation in `miRBase` (*Ann_mature*: *Annotated mature*) and those which this region was inferred (*Inf_mature*: *Inferred mature*). Three different categories were considered to describe sequence and model features. In a first place, *Input* panel shown the proportion of sequences that belonged from the *full* or *seed* alignments. As shown earlier, the reduced proportion of *seed* in comparison to *full* sequences confirmed the nature of the first group as a representative set of the large available space of sequences on `Rfam` database. At the same time, almost equal proportion of sequences were found when compared the *seed* set, that belongs to family models that contained sequences with (2393) and without (2102) mature annotations in `miRBase`.

In the next panel (*MatureMatch*) accounted the number of *seed* models that reported a mature annotation on `miRBase`. Were found 509 sequences that reported mature annotation on `miRBase`, reported in the *Annotated model* set. On the other hand, 4926 miRNAs completely missed mature annotations (*no_mature*), but from which 2828 had at least one sequence with annotations as reference to perform the mature annotation, the other 2098 sequences lacked of this reference. In this case, the mature annotation was

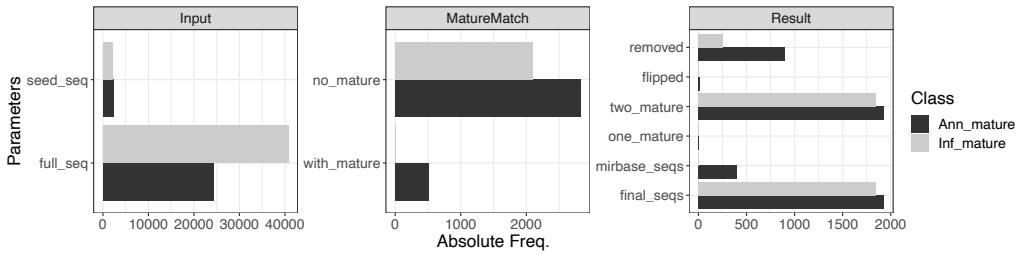


Figure 13: Absolute frequency of processing stages on the evaluated metazoan sequences from `Rfam`. Three processing stages (*Input*, *MatureMatch*, and *Result*) distribute the obtained frequencies for all `Rfam` families. *Class* denoted the distinction between those families that reported annotated mature sequences on `miRBase` (*Ann_mature*) or not (*Inf_mature*). The *Input* category described the number of sequences annotated as seed: *seed_seq* or full: *full_seq* sequences on Rfam. Consequently, in the next panel *MatureMatch* described the number of sequences that used an annotated mature sequence to infer the position of their own mature region (*with_mature*), meanwhile *no_mature* inferred those mature regions based on the secondary structural alignment. Final results were described on the *Result* panel, which each category accounted the frequency over all sequences resulted from mature annotation analysis, including those sequences that imported their mature annotation from `miRBase` (*mirbase_seqs*).

inferred from the reported structural alignment, as described in Section 3.2.1.

Final results are reported in panel *Result*, where 3777 sequences were validated, mapping multiple features (see y axis). From them, 400 reported mature annotations on `miRBase`. At the same time, was possible to annotate both candidate mir and mir* regions on all the corrected precursors. A number of 10 `Rfam` families contained 15 flipped sequences: let-7 (RF00027, 4), mir-192 (RF00130, 2), mir-124 (RF00239, 1), mir-33 (RF00667, 2), mir-143 (RF00683, 1), mir-433 (RF00748, 1), mir-490 (RF00792, 1), mir-280 (RF00801, 1), mir-488 (RF00861, 1), and mir-668 (RF00890, 1). At the final of this strategy, those corrected sequences were included into the family structural alignment. As result of the curation, were removed 1149 sequences from the final sequence set, due lacking of correct hairpin-structure folding or the missing of a correct annotation of mature regions on the provided precursor sequence.

Curation examples on `Rfam` miRNA families

As a successful example of the mature annotation workflow, Figure 14 represents the final alignment for the model *lin-4* (RF00052), which generated a valid final consensus secondary structure. Additionally, the labelling process using `miRBase` resulted on an additional layer of information which allowed to detect that the sequences included in this alignment belonged from the families: *lin-4* and *miR-10*. Those sequences from *lin-4* family came from Caenorhabditis sp. and the remaining classified as *miR-10*, from: *Drosophila melanogaster*, *Gorilla gorilla*, *Homo sapiens* and *Mus musculus*. Two human specific sequences (AB232081.1/345-415 and AP001359.4/67794-67724) with 100% of identity, mapped into sequences without miRNA classification, but restricting the `miRBase` sequence space to the *high confidence* set, resulted on a mapping to the *miR-10* family.

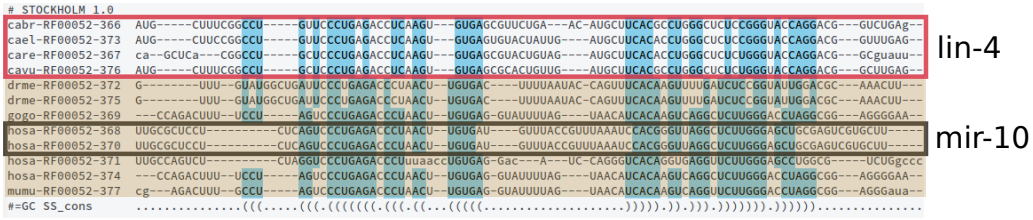


Figure 14: Disentangling of the lin-4 (RF00052) model. Blue columns show the conservation at secondary structure level, see SS_cons line. For the lin-4 model 2 `miRBase` families were detected by the explained mapping methods: lin-4, highlighted by a round red box, and miR-10 shaded by orange. Black line box, indicated the human specific sequences (AB232081.1/345-415 and AP001359.4/67794-67724), which reported 100% identity.

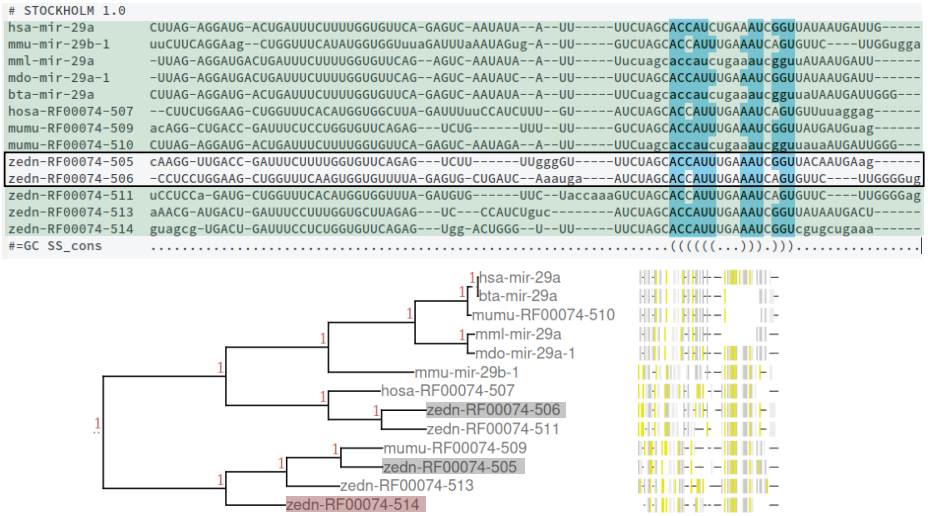
On the other hand, an example for the iterating process to validate a miRNA family is depicted on Figure 15, where the model for miR-29 (RF00074) generated a non-hairpin folding secondary structure after the first iteration of mature annotation. In this case, two zebrafish sequences (`BX001041.6/158851-158788` and `BX001041.6/159008-158937`) did not report a hit and have mapped onto a sequence in `miRBase` without a family classification (dre-mir-29b3, `MI0039521`). Additionally, the sequence from zebrafish zedn-RF00074-514 (`CR749762.6/80332-80264`), mapped exclusively to the `miRBase` family ipu-mir-29a (MI0024673), without any candidate from the high confidence set of sequences. In this case, to improve the folding of the model, those sequences were removed (Figure 15-2). In this iteration, the consensus structure shown a non-hairpin folding, for that reason a multiple set of iterations over the clades were performed to clean the alignment, as shown on Figure 15-3 the final alignment shows a correct folding composed by sequences that reported a uniform family assignment.

Finally, some families were discarded based on the same iterative process that at the end reported a miRNA family that should be revisited. The family mir-638 (RF00978) reported a wrong alignment on the anchored structural alignment, as seen in Figure 16. Next, two sequences: `ABRN01352993.1/11133-11034` and `ABRQ01087111.1/310-211` from *Tursiops truncatus* and *Procavia capensis* were identified and discarded due lack of `miRBase` family labels. At the same time, the reconstructed phylogenetic trees reported branches with only one sequence, which means that at the end only existed 1 branch with 2 sequences. The final secondary structure (Figure 16-3) with those sequences has shown a long hairpin-like structure with 40 paired nucleotides. Looking into the `miRBase` mapping results, were found perfect matches for both sequences to: MI0003653 from *H. sapiens* and MI0007882 from *M. mulatta*, correspondingly. Both sequences in the `miRBase` database were classified as non-high confidence sequences and reported a few numbers of reads on deep sequencing experiments (see Figure 16-4).

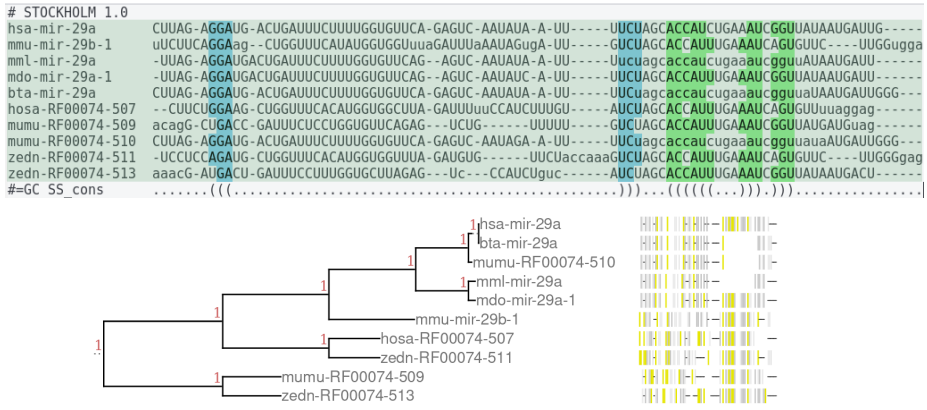
Families that failed structural filters: rescued or filtered

From the complete 88 miRNA families that failed the structural evaluation (see Section 3.3.2), where improved 54 families by the described curation strategy. The other discarded set of 33 families were collected over all the curation process, as summarized on Table 2.

1



2



3

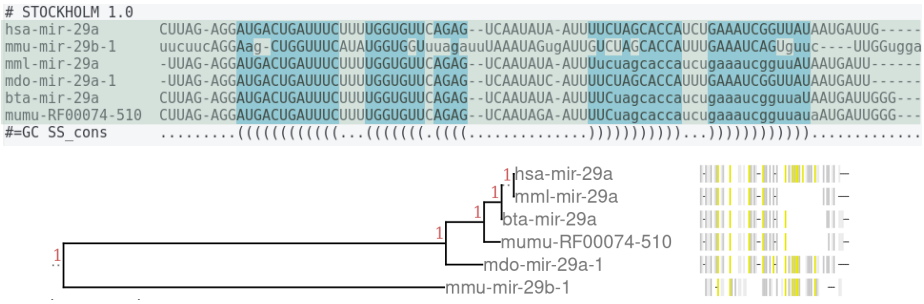


Figure 15: Mir-29 (RF00074) model: mir-29 family reported a non-valid miRNA hairpin folding, all the sequences except two paralog zebrafish sequences (BX001041.6/158851-158788 and BX001041.6/159008-158937, highlighted by the black box in the alignment and grey shade in the tree), were detected as part of the mir-29 `miRBase` defined family (shade green on the alignments). Red shade on the tree pointed out the sequence CR749762.6/80332-80264, which generated an exclusive mapping with non-high confidence sequences. Numbers on the up left side, indicate the rounds of processing the curated alignment (1) into the valid one (3).

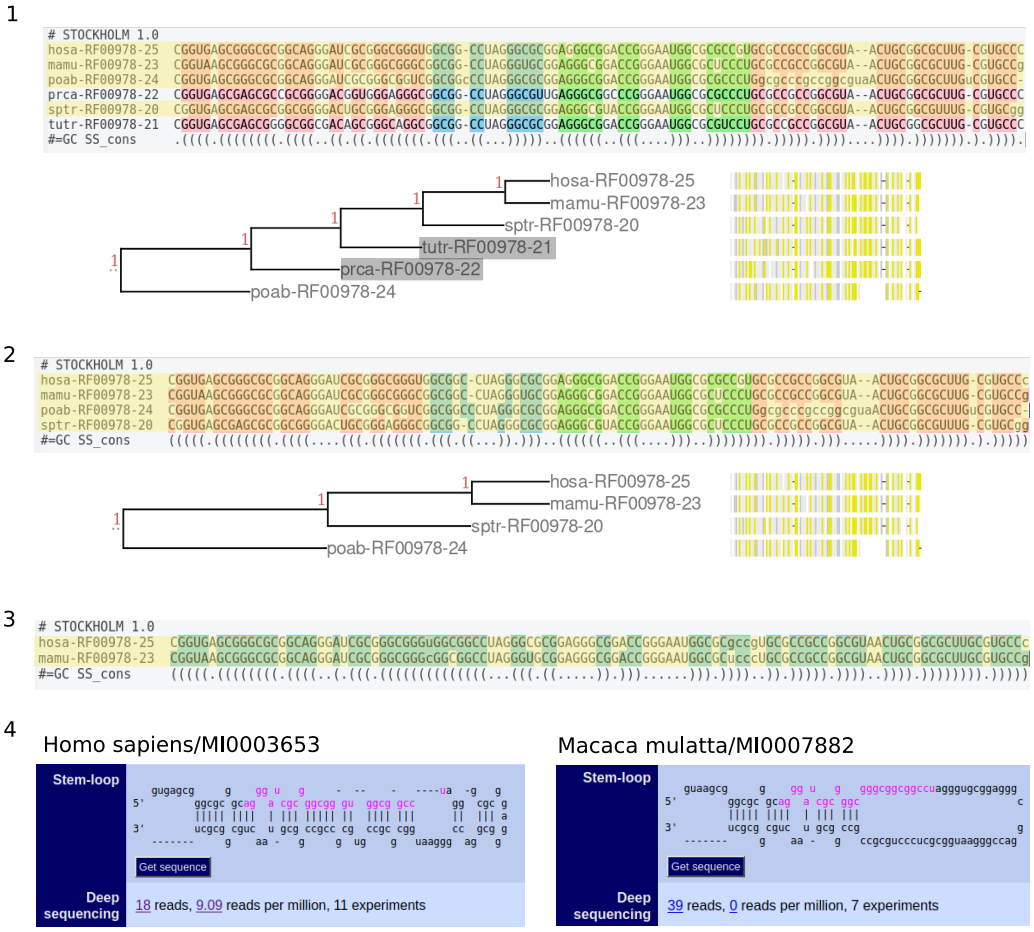


Figure 16: Processing of mir-638. The mir-638 `miRBase` family was highlighted in yellow, those candidates without any `miRBase` labels were shaded in gray in the phylogenetic tree. After two rounds of iteration was conserved the sequences `AADB02020051.1/340879-340978` and `AANU01122668.1/894-795`, from *H. sapiens* and *Macaca mulatta*.

In detail, those miRNA families were not able to be accepted due the report of unusual hairpin long secondary structures (13), insufficient number of sequences after a family assignment step (7), ended detected as discarded along the iteration process (12). A particular case was found in the family miR-154 (RF00641), which reported 9 sets of `miRBase` families, including some candidates without hits and no family assignments, as seen on Appendix B:Figure 45. Additionally, due multiple branches on the calculated phylogenetic tree, it was not trivial to define the set of representative sequence clade that composed this family, for that reason it was not included into the curated dataset.

3.4 Discussion

Quantitative studies are not only hampered by the limits of homology search, but also suffer from other problems with the available data. First, there is a massive ascertainment bias of empirical studies, which concentrate on a few model species whose microRNA (miRNA) complements are very well studied, in particular human, mouse, fruit fly and *C. elegans*. On the other hand, there is a substantial ambiguity what exactly constitutes a miRNA as opposed to another family of small non-coding RNAs. Using a stringent view requiring canonical processing by Drosha and Dicer as criteria, a significant fraction of the entries in recent releases of the `miRBase` are presumably "false positives" (Tarver, Taylor, et al., 2018). This is difficult to decide, however, since `miRBase` is not very explicit about where exactly it draws the boundary of "`miRBase`-miRNAs". In contrast, a quite specific set of rules to identify canonical miRNAs has been adopted by the `MirGeneDB` (Fromm, Domanska, et al., 2019), based on the analysis small-RNA-seq data from a selection of from few metazoan species.

A related source of errors is unrecognized homologies, such as tunicate mir-1473, which is homologous to the ancient mir-100 family (Hertel, Bartschat, et al., 2012). Such cases lead to the erroneous prediction of both an innovation and possibly many loss events. Distant homologies are not trivial to recognize due to limited size and thus information content of miRNA precursors. So far, two computational approaches have been explored. A comparison of similarity scores of pairwise alignments with the randomly shuffled sequences was proposed in Tanzer and P. F. Stadler (2004) as a means to identify mir-25/mir-92 and mir-17/mir-18/mir-20/mir-93/mir-106, respectively, as ancient homologs. A more sophisticated approach is `CMCompare`, which directly compares the Covariance Models (CMs) representing two miRNA families (Siederdissen and Hofacker, 2010). The method, which is also available as a web service Eggenhofer, Hofacker, and Höner zu Siederdissen, 2013, identified e.g. the plant families MIR806, MIR811, MIR812, MIR821, MIR1023, and MIR1151 as likely homologs. A systematic analysis of distant homologies among miRNA families has not been conducted in recent years.

A clear outlier in `miRBase` is mir-451, whose mature product is processed by Argonaute (AGO) from the loop region of the precursor (Cifuentes et al., 2010). It is usually treated as miRNA, even though it is not recognized by many automatic annotation tools. Recent studies show that there are many alternative biogenesis pathways that overlap more or less with the canonical one eventually producing small RNAs that incorporate into the Argonaut complex and perform miRNA-like functions in gene silencing Kim, Han, and Siomi, 2009; L. Li and Y. Liu, 2011; Okamura, 2011.

Table 2: Discarded non-structure `Rfam` miRNA families. **Acc.**:`Rfam` reference.

Fail reason	**Acc.**	**miRNA Family**
Invalid structures with long hairpins (13)	RF00786	miR-289
	RF00787	miR-288
	RF00824	miR-50
	RF00838	miR-252
	RF00858	miR-306
	RF00900	miR-255
	RF00940	miR-327
	RF00985	miR-640
	RF01040	miR-573
	RF01314	miR-1227
	RF01901	miR-284
	RF02014	miR-1178
	RF02015	miR-1287
Insufficient number of sequences (7), remove no-label sequences	RF00131	miR-30
	RF00805	miR-351
	RF00875	miR-692
	RF01035	miR-887
	RF01045	miR-544
	RF01942	miR-1937
	RF02002	miR-720
Structural issues after iteration (12)	RF00783	miR-484
	RF00790	miR-358
	RF00799	miR-354
	RF00834	miR-268
	RF00968	miR-626
	RF00977	miR-600
	RF00998	miR-562
	RF00999	miR-924
	RF01031	miR-639
	RF01036	miR-567
	RF01922	miR-654
	RF02244	miR-785
Multiple miRNA families (1)	RF00641	miR-154

Mir-451 is the prototypical representative of a larger class of Argonaut2-processed *loop-miRNAs* whose mature product, the "miR-loop", is excised from the loop rather than the precursor stem. MiR-loop RNAs are also produced from the precursors of some canonical miRNAs including mir-33a, mir-34a, mir-192, mir-219-2 (Okamura et al., 2013;

Winter, Link, et al., 2013). It does not seem too difficult to adapt e.g. the workflow of (Yazbeck, P. F. Stadler, et al., 2019) to loop-miRNAs; so far, however, no such tool seems to be available.

First, there is still no universal consensus of what exactly distinguishes a *bona fide* miRNA from one of the many types of other RNAs of comparable size. This is not merely a question of nomenclature or the decision defining the scope of a data repository such as `miRBase`. The distinction between (canonical) miRNAs and other small RNAs, as well as a classification of miRNAs in canonical miRNAs and several types of non-canonical ones is of practical importance for the construction of computational methods. Machine learning approaches, in particular, need clearly defined test and training data. This issue is particularly important in the context of lineage-specific miRNAs and miRNAs associated with or derived from repetitive elements.

The wealth of miRNA data available in `miRBase` and `Rfam`, in particular when augmented by (semi)automatic pipelines to curate and complete the data by homology search set the stage for investigating a wide array of questions. While it seem to have a decent understanding of the evolution of miRNAs at the family level, much less is known about the histories of the individual paralogs. It seems natural to ask, therefore, whether it is possible to devise automatic methods to distinguish orthologs reliably and to phylogenetically map duplication and loss events within miRNA families. So far, only changes in the number of family members have been investigated systematically (see Hertel and P. Stadler (2015)).

— 4 —

Automatic miRNA detection based on homology signals: `miRNAture`

Contents

4.1 miRNA profiling and detection: current challenge

Successful miRNA profiling and quantification has opened the door to an expanded RNA expression landscape. As recognized by Baker (2010) and Dong et al. (2013) in case of miRNAs this could be exploited as a source of meaningful comparisons or miRNA-based biomarkers, e.g. between healthy and diseased cells populations, in the diagnosis of certain types of diseases, or even as molecular diagnostic tool. Despite recent advances of high-throughput sequencing and previous development of cloning and microarray methodologies, still the most reliable way for a high sensitive and selective identification of the miRNA transcription is missing. As reviewed in more detail by Aldridge and Hadfield (2012), Dong et al. (2013), and Pritchard, Cheng, and Tewari (2012), unique miRNA features complicate current analysis, such as: their short length, low abundance, discrimination between miRNA biogenesis entities, variable GC content, associated to variance in melting temperature (T_m) of primers and probes, RNA enzymes favouring certain sequences over others, the high degree of homology of miRNA families, and the high discovery rate of new miRNAs.

Early small RNA cloning methods detected abundant high confidence miRNAs. Subsequently, those samples were validated using northern blotting assays, microarrays or even quantitative RT-PCR (qRT-PCR) (Backes et al., 2015; Dong et al., 2013; Lagos-Quintana, Rauhut, Lendeckel, et al., 2001). The selection of the protocol relies on the quality of starting samples, associated cost and desired precision. Additionally, given the wide range of miRNA expression patterns that are tissue/cell dependent, those traditional techniques have detected only a portion of the miRNA landscape (Backes et al., 2015).

Recently, with the fast and low-cost generation of high-throughput sequencing data using i.e. next-generation sequencing methods, an expanded profiling landscape of known and novel miRNAs was uncovered, and together with the parallel development of efficient *ab initio* computational methods and repository databases (Y. Li et al., 2012). Once (small—mi)RNA-seq data are available, those tools deal with a mapping problem, guiding the short reads back to the genome and subsequently, a posterior characterization by structural folding over those mapped loci is required to define a candidate miRNA (Hu, Lan, and Miller, 2017; Y. Li et al., 2012).

The most recent release of `miRBAse` (v.22.1) (Kozomara, Birgaoanu, and Griffiths-Jones, 2019) lists 1984 human miRNA precursors, and a recent extrapolation estimates about 2300 mature microRNAs for human (Alles et al., 2019). These numbers are much larger than those reported for other mammals, suggesting that our knowledge of the miRNA repertoire of animal genomes is still far from complete. On the other hand, Fromm, Billipp, et al. (2015) accounted only 519 "confidently identified canonical miRNA genes", see also Bartel (2018). The discrepancy derives both from the level of experimental evidence required to confidently identify a non-coding RNA (ncRNA) gene and from the *definition* of what constitutes a canonical miRNA, as opposed to a member of a wider class of small RNAs associated with the RNA-interference pathways (see for example Fromm, Domanska, et al. (2019), Okamura (2011), and Velandia-Huerto, Yazbeck, et al. (2022)).

In this regard, as a toy example, the annotation of human miR-100 (MI0000102) in `miRBase` is depicted on Figure 17. Current definition of mature sequence is based on the distribution of mapped reads, originated by multiple RNA-seq libraries, that forms

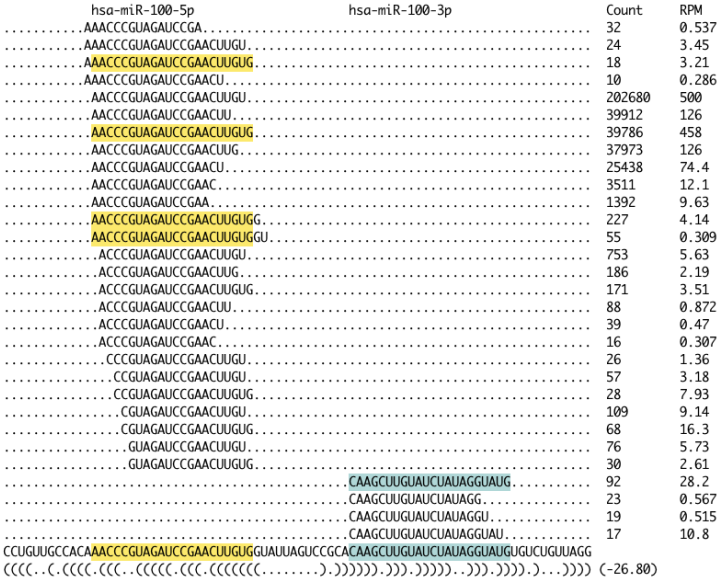


Figure 17: Mir-100 sequence heterogeneity from human (MI0000102). Data extracted from `miRBase`. Annotated mature 5p and 3p miRNAs are coloured in yellow and blue, respectively. Reads with Count < 10 were removed. Predicted secondary structure with folding energy are calculated below in the last line.

a multimodal distribution that corresponds to the canonical expression of both, 5p and 3p ends. Once defined the mature regions, the harbouring precursor can be inferred with their characteristic hairpin-loop folding. A discrete single length for miRNAs can not be defined since the processing by-products could have variable sizes, and the 22 nt is suggested as average length in the distribution. Post-transcriptional modifications, nucleotide deletion and/or addition, at 5' or 3' are the most plausible explanation for the length variability observed on mapped reads (Pritchard, Cheng, and Tewari, 2012). At this point, all the miRNA classification of miRNAs relies on a correct mature definition. This pivotal definition involves a correct mature delimitation and consequently, their enveloped seed sequence, which could be challenging by the existence of isomiRs and annotation methods that only relies on RNA-seq data to detect conserved or new miRNAs. An optimal approach should combine computational predictions and validation by RNA-seq experiments.

Moreover, the non-availability of RNA-seq experiments in the majority of current and new sequencing projects is the rule. On those cases, the prediction of miRNAs can be benefited by the use of computational methods to bring a preliminary number of miRNAs candidates. In this way, based on an automatic homology strategy to detect miRNAs: `miRNAture`, this idea became reality. The findings of this chapter are based on Velandia-Huerto, Fallmann, and P. F. Stadler (2021).

4.1.1 miRNA profiling by experimental means

Despite the challenging features recognized on miRNAs, well-established experimental approaches can be summarized as: quantitative reverse transcription PCR (qRT-PCR), hybridization-based methods (microarrays, *In situ* hybridization) and high-throughput sequencing (RNA-seq) (Hammond, 2006; Pritchard, Cheng, and Tewari, 2012; Yaylak and

Akgül, 2022). Those techniques are briefly described in following sections in addition to the outline for the tunicate miRNA annotation context depicted in Chapter 1:Figure 6.

The historical Northern Blotting to discover miRNAs

The first described *lin-4* miRNA, by Lee and Ambros (2001), was characterized using a modified northern blot protocol. In brief, this is based on the migration of secondary structure disrupted RNA samples, denatured (i.e. with formaldehyde), through a polyacrylamide gel to separate them by size, using electrophoresis. Next, a nylon membrane transference (blotting) allow the targeting of specific molecules by an antisense hybridization probe and posterior visualization (Yaylak and Akgül, 2022). This method is particularly useful to detect specific miRNA entities, mature and precursor sequences. Disadvantages came with the parallel profiling, due specific miRNA differences such as GC content, directly related with the T_m, single nucleotide differences between miRNA families (that restrict their distinction by the normal protocol) and the presence of *isomiRs* that modify the length of mature sequences (Pritchard, Cheng, and Tewari, 2012). Enhancements of the original technique allowed the distinction of those entities from short-interfering RNAs (siRNAs), a detailed characterization of miRNA biogenesis, and used to get detailed information about 5' and 3' ends combined with other techniques as rapid amplification of cDNA ends (RACE) (Koscianska et al., 2011).

Hybridization-based methods: localization and quantification

To achieve an increasing high-throughput, simultaneous sample processing, and fast miRNA profiling, an extension of hybridization techniques was required. A solution was developed by C.-G. Liu, Calin, Meloon, et al. (2004) for miRNA profiling by oligonucleotide miRNA microarrays. They modified the microarray protocol in terms of their oligo probes, the attachment of the probes, the sample labeling and signal-detection methods (C.-G. Liu, Calin, Volinia, et al., 2008). The principle is based on the design of oligo probes that are derived from annotated miRNAs. Next, in a *printing* step, that takes the oligo-probes to be fixed in a physical activated slides, served as substrate for further hybridization. The attached probes consist on a linker sequences, poly(dT) or poly(dA) with an amine-modified terminus, attached to the glass or beads (W. Li and K. Ruan, 2009), and capture antisense-sequences that will hybridize the biotin-labeled miRNA or labeled products of reverse transcription of miRNA targets (to more details about multiple labeling options see W. Li and K. Ruan (2009)). To detect the signals, a laser conjugates and detects specific Streptavidin stains (C.-G. Liu, Calin, Meloon, et al., 2004; C.-G. Liu, Calin, Volinia, et al., 2008). This method is limited to detect known miRNAs. In the same way to study localization and miRNA abundance, the *in-situ* hybridization (ISH) could be used directly on tissues or histological samples, using an improved affinity and consequent specificity using locked nucleic acid (LNA) modified DNA/RNA probes, that open the door to create short probes that support high temperatures ($\geq 70°C$), playing a fundamental role in the established ISH for miRNAs (Song, Ro, and Yan, 2010). Posterior visualization is done by the use of DIG-labeled LNA, which can be used to quantify the miRNA amount. Direct fluorescence can be obtained using LNA with an DIG as 5' end.

This probe is antisense for an anti-DIG antibody generating a signal that can be visualized by fluorescent microscopy (Yaylak and Akgül, 2022).

qRT-PCR

A more efficient method to monitor miRNA levels with high sensitivity and specificity is by using Quantitative Real-time PCR (qRT-PCR). This is composed by a combination of reverse transcription coupled with a real-time PCR and measures the accumulation of PCR products by the quantification of a fluorogenic probe (Heid et al., 1996). Due to their low requirements of initial RNA, this method is balanced in terms of cost and the quality of validation for candidates from microarrays or RNA-seq experiments (Yaylak and Akgül, 2022). As a first step a miRNA size extension is required, due short length of mature products, to subsequent perform a reverse transcription and PCR amplification. For that means, mature miRNAs are polyadenylated at 3' end and using a universal poly(dT) primer, it enables the cDNA synthesis, overcoming variability at mature ends by the presence of isoforms. This primer can be a specific-designed as a stem-loop, too. The cDNA is quantified using qPCR, using sequence-specific fluorescent reporters, such as TagMan®. This reporter function is activated when the complementary sequence is hybridized and its fluorescence signal is proportional to the length of template nucleic acid (Yaylak and Akgül, 2022). This method is suited to amplify known miRNAs.

RNA-seq

Current offer of next generation sequencing (NGS) platforms are continuously evolving and short miRNA have tractable patterns prone to be recognized, despite the high-throughout of those methodologies. The steps particularly include the library preparation, sequencing, and data analysis (Pritchard, Cheng, and Tewari, 2012). In detail, a cDNA library preparation is build from small RNAs, which previously were purified from a polyacrylamide gel, as a previous enrichment step. To complete the library construction, single adapter (Real-Seq®[1]) or two-adapters (5' and 3') are ligated for subsequent sequencing, by Illumina/Solexa platforms.

The quantification of read accumulation follows a digital approach, detecting the overlapping reads as a link to estimate miRNA abundance. This recognition includes the entire miRNA population on a specific transcriptional time-frame, including not only the *boda fide* miRNAs, but the putative ones (Pritchard, Cheng, and Tewari, 2012; Svoboda, 2015). The analysis of the data comprises: obtaining of raw data from specific sequencing platform. Specifically the output RNA can be modified considering the depth of the sequencing, for example Svoboda (2015) reported a depth of 10Mb reads are enough to analyse small RNAs (endo-siRNAs). For basic miRNA profiling $<$ 10Mb is enough, assuming more or less homogeneous cell population. Next, an initial quality control filters the raw data based on the assessment of sequencing quality, accessed for each base. Then, a cleaning step removes low quality reads and adapters. A final analysis comprises the mapping of those cleaned reads to reference genome, or to a miRNA databases, such as `miRBase` or `Rfam`. In cases where the homology information did not annotate a close

[1] `https://www.sigmaaldrich.com/DE/de/technical-documents/technical-article/genomics/gene-expression-and-silencing/small-rna-sequencing`, visited 02.01.2022.

candidate, programs as miRDeep2 (Friedländer et al., 2011) are used to annotate new miRNA loci (Hu, Lan, and Miller, 2017).

4.1.2 Computational methods to detect miRNAs

As summarised by Chen, Heikkinen, C. Wang, Y. Yang, Sun, et al. (2019), the miRNA biogenesis derive multiple steps that are susceptible to be transformed into computational rules to detect miRNAs. The interest of this thesis work is focused on the prediction of miRNA genes at genomic level. To do so, at suggested by Hertel, Langenberger, and P. F. Stadler (2013), the prediction of miRNAs relies on *true* miRNA examples and their distinctive features from other ncRNAs. Due to multiple miRNA biogenesis modes (see Chapter 2:Figure 4) it is actually challenging to generalize all features and discover the complete miRNA complement on a specie. In this regard miRNAs with previous characterization or annotation on another species are susceptible to be identified by homology strategies. By other way, when this data is missing, *de novo* approaches could be used, as explained in following sections.

Homology approaches

When annotated miRNAs are available, using sequence and structural comparisons is the simplest way to get a preliminary list of conserved candidates. This analysis is encompassed by homology assumptions and is performed at sequence level using approaches as `BLAST+` (Camacho et al., 2009) or `GotohScan` (Hertel, Jong, et al., 2009). Additionally, comparisons against Hidden Markov Models (HMMs) (based on multiple sequence alignments) can be used by means of `nhmmer` (Wheeler and S. R. Eddy, 2013). After filtering steps obtaining the best candidates, the secondary structure should be assessed (e.g. using `RNAfold` (Lorenz et al., 2011)) to verify a miRNA structure, characterized as a hairpin-loop. In addition, a minimum free energy (MFE) that falls into the reported range for miRNAs, due possible differentiation at energy folding level with respect to other ncRNAs (Ng Kwang Loong and Mishra, 2007). Another alternative is comparing the sequences to a structural model, i.e. ready to use Covariance Model (CM) from `Rfam`, by using the tools in the `INFERNAL` package (Nawrocki and S. R. Eddy, 2013). Early methods explored the conservation from precursors and intragenic regions along multiple species, such as: `srnaloop` (Grad et al., 2003), `MiRscan` (Lim et al., 2003), and `miRseeker` (Lai et al., 2003). In addition, this search can be complemented using multiple structural alignments with collected homologs, identifying conservation into the alignment and validity of its consensus secondary structure Hertel, Langenberger, and P. F. Stadler (2013).

de novo approaches

Since by homology methods, the discovery of new miRNA families is restricted, the development of automatic detection and classification methods are necessary. Hertel, Langenberger, and P. F. Stadler points out the use of machine learning methods, are suited to classify candidate sequences based on inferred miRNA specific characteristics, trained by positive and negative datasets. In the same way, Gomes et al. (2013) split those machine-learning approaches based on their core algorithms, such as: Support vector

machine (SVM), HMM, and näive Bayes (NB) classifiers. This algorithm classification was complemented by Saçar Demirci, Baumbach, and Allmer (2017), who designed a computational framework to evaluate the *ab initio* methods to detect miRNAs (`izMir`), and provided a combined miRNA Decision trees (DT) and NB classification models.

High-throughput expression libraries are suitable to find novel miRNA families. The bulk of expression data can be mapped against a reference genome, using for example `segemehl` (Hoffmann et al., 2009) or `BWA` (H. Li and Durbin, 2009), to devise the current genomic position and collect all mapped reads to identify characteristic read patterns for miRNAs: two stack of reads that coincided with miR and miR* regions and are spaced by the loop region (Hertel, Langenberger, and P. F. Stadler, 2013), assuming a canonical processing. Previous miRNA annotations can be used as well as reference of precursor/mature sequences position and structural patterns that confirm the presence of a hairpin-loop structure, are adequate to validate the annotation.

An updated source of miRNA tools are described in multiple reports see Chen, Heikkinen, C. Wang, Y. Yang, Sun, et al. (2019), Gomes et al. (2013), and Hertel, Langenberger, and P. F. Stadler (2013) and online meta-databases are: `miRToolsGallery` (Chen, Heikkinen, C. Wang, Y. Yang, Knott, et al., 2018) and `tools4miRs` (Lukasik, Wójcikowski, and Zielenkiewicz, 2016).

4.2 Translating canonical rules to computational approaches

4.2.1 `miRNAture` methods

The sequence/structure filters can be grouped by type of evaluation: *Sequence homology*, *Alignment scores*, *Annotation/Structure*, and *Consensus secondary structure*. Table 3 summarises how they are employed in the different modes of the `miRNAture` workflow.

Pairwise comparisons with user-defined query sequences in *Sequence homology* searches are evaluated in terms of E-values, coverage and length of resulting *high scoring pairs* (HSPs) as suggested in Velandia-Huerto, Gittenberger, et al. (2016).

HMM comparisons are evaluated with respect to the default inclusion thresholds of the `nhmmer` models as suggested in the HMMer userguide [2]. Direct comparisons to miRNA CMs make use of the parameters calculated by `cmsearch`: E-value, bitscore and coverage with respect to the length of the CM. A uniform bitscore cutoff of $log_2 2N$ is used, where N denotes the genome size. If a *gathering cutoff* (GA) is available for a CM, for example, in `Rfam` models, `miRNAture` uses a threshold of nGA $= 0.32$ to rescue candidates that potentially represent valid miRNAs. Structural parameters are evaluated with an updated version v2.0.0 of `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019) [3].

The focus of this evaluation step is the correct annotation of mature sequences relative to the precursor. To this end, a precursor length of ≤ 200 nt and a secondary structure with a minimum free energy (MFE) ≤ -10 is required. The additional *Evaluation* stage of `miRNAture` compares the *tree edit* distance (Hofacker et al., 1994) between the consensus structure dot-bracket string of the pre-defined structural alignments of the miRNA family and the re-computed alignment that includes the additional, new precursor sequence.

[2] `http://eddylab.org/software/hmmer/Userguide.pdf`, accessed on 18.12.2020
[3] `https://github.com/Bierinformatik/MIRfix/releases/tag/v2.0.0`, accessed on 27.01.2021

The structural distance is used to determine the confidence level of the new candidate, which passes the validation stage: *High*: Valid consensus secondary structure and tree edit distance ≤ 7 to the consensus secondary structure of the initial family; *Medium*: if fails any of those. In case numerous homologs are found, only a user-specified number of top candidates are processed as described. The remaining putative homologs are reported separately as putative matches.

Table 3: Homology, structure and final filters applied on `miRNAture`. Specific programs used for each mode in parentheses. *Ann.*: Annotation, *SS*: Secondary structure. *CSS*: Consensus secondary structure. *ge*: *gathering cutoff* from `Rfam` family. $nBit = Bitscore/ge$. *ted*: tree edit distance between default miRNA and modified multiple `stockholm` alignments. MFE: Minimum free energy. *HSPs*: high scoring pairs. Pairwise comparisons performed with `blastn`, HMMs with `nhmmer`, and Evaluation by `cmsearch`. Annotation filter, Evaluation of SS, and SS conservation by `MIRfix`. Ann.= Annotation, Str.=Structure., Alig.= Alignment.

Sequence Homology		Alig. Score	Ann./Str. Evaluation		Consensus Evaluation
Pairwise	HMMs	Evaluation	Ann. Filter	SS	SS Conservation
E $\leq$ 0.01 $\geq$ 20 nt HSPs C(f)$\geq$ 70%	E $\leq$ 0.01	E $\leq$ 0.01 C(f)$\geq$ 70% $n_{\text{bitscore}} \geq 0.32 * ge$ B $> \log_2 2N$	Ann. mature seq. Seq. Length $\leq$ 200 nt	MFE < -10	*ted* ≤ 7 Valid CSS

4.2.2 Pilot study: search homology on tunicates

Genomes from *C. robusta*, *C. savignyi* and *O. dioica* were retrieved from sources described in Appendix D: Table 20. Annotated hairpin sequences were retrieved from `miRBase` v.22.1. The homology search was calculated using `miRNAture` v.1.0, as seen in Listing 4.1:

```
miRNAture -stage homology -dataF <Data folder> -speG <specie genome> -speN
<specie name> -speT <specie tag> -w <work dir> -m Blast,Final -pe 0 -str
    4,5,ALL
-blastq <queries folder> -rep relax,150,100
```

Listing 4.1: miRNAture homology parameters.

4.2.3 Comparison of multiple homology-search experiments

The intersected final homology results, calculated using all available `miRNAture` searching modes (`Blast`, `HMM` or `Infernal`), were overlapped using `bedtools`, as follows:

```
bedtools intersect -s -a <MODE1> -b <MODE2> <MODE3> -c > countings
bedtools intersect -s -a <MODE1> -b <MODE2> <MODE3> -wa -wb >
    relation_detail
```

Listing 4.2: Intersection strategy performed with bedtools.

4.2.4 Curation of the *let-7* Family

Let-7 loci from *G. gorilla*, *H. sapiens*, *P. abelii*, *P. troglodytes*, *M. musculus* and *C. savignyi* were retrieved from miRBase v.22 in FASTA and GFF3 format. In addition, the let-7 loci reported in Hertel, Bartschat, et al. (2012) were retrieved and mapped to the genomes listed above. The union of the *let-7* loci from Hertel, Bartschat, et al. (2012) and from miRBase were used as reference for evaluation.

CMs for *let-7* were retrieved from Rfam v.14.4 (RF00027), Hertel, Bartschat, et al. (2012) (17 models from the A, B, C, D, E, F, G, H, I, J, K and L paralogs), and Yazbeck, P. F. Stadler, et al. (2019) (miRBase v.21, curated MIPF0000002). An additional CM was constructed using the bilaterian sequences from miRBase v.22, excluding both, paralogous sequences with 100% identity and sequences from the target species. All models were used as input for Infernal and for and Other_CM homology modes in miRNAture. All *let-7* retrieved sequences from miRBase were used as queries using BLAST mode in miRNAture with strategies 1, 2, 3, 5 and 6.

Both reference loci and the final results of miRNAture are stored in GFF3 format. Comparisons on genomic loci level were performed using bedtools (Quinlan and Hall, 2010) and classified by *Match*: overlaps on the same strand; *Miss*: locus in references without overlap in miRNAture output; *Additional*: candidates detected by miRNAture without overlap in reference.

4.2.5 Curation of Human miRNA Families

miRNA precursor and mature sequences were retrieved from miRBase v.22 and corrected with MIRfix (Yazbeck, P. F. Stadler, et al., 2019) to create a set of representative sequences for each miRNA family with a corrected set of mature positions, corrected precursor sequences and mature-anchored structural miRNA family alignment. From the latter family-specific covariance models where built using Infernal (Nawrocki and S. R. Eddy, 2013).

4.3 miRNAture and its homology assessment

4.3.1 Architecture of miRNAture

The miRNAture pipeline is composed of three modules: (1) *Homology* search operating on miRNA precursors; (2) prediction of the positioning of mature miRNAs within the precursor (*Mature annotation*); and (3) an *Evaluation* scheme designed to identify false positive miRNA annotations. The pipeline is distributed with pre-computed CMs for the miRNAs in Rfam v.14.4 (Kalvari, Nawrocki, Ontiveros-Palacios, et al., 2020), which are used as default for annotation of a target sequence or genome. Users can also add their own CM, and/or a query sequence that will subsequently be annotated. It is also possible to use a combination of built-in and user supplied CMs. miRNAture produces annotation files in GFF3/BED format and FASTA files for validated candidates as well as summary reports that highlight possibly problematic cases, tagging these for manual inspection. The architecture of miRNAture is summarised in Figure 18.

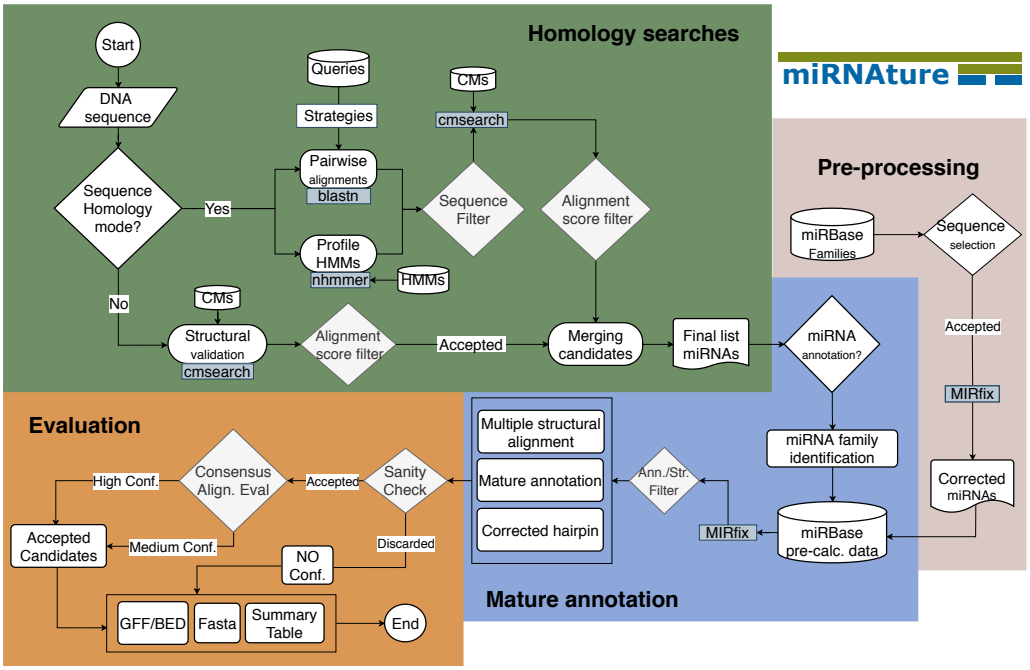


Figure 18: Workflow of `miRNAture`. The starting point is the provided set of target sequence by the user, which is first analyzed in *Homology search* mode to detect miRNA candidates. Specifically, two strategies are available: sequence homology and structural validation. The first one using pairwise alignments, performed with `blastn`, or HMM using `nhmmer`. The second one is based on the use of CM. Each of the described stages has their own filters, accepted candidates being submitted to the next stage while discarded candidates are reported separately for later manual inspection. A merging step produces a final list of homology candidates. After that, *Mature annotation* stage runs on these input sequences and performs a correction of the positioning of mature sequences on the hairpin, generating a correctly anchored family-specific-multiple secondary structure alignment, calculated by `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019). The *Evaluation* stage starts with a sanity check that reviews the mature annotation and performs a comparison of conserved secondary structures with and without the newly annotated candidates. Based on this classification the candidates will be labeled as accepted or discarded. Sequence length and MFE cutoffs are used for further filtering. A final set of candidates is reported in `BED/GFF3` annotation formats and `FASTA` files. A summary file provides overall information about the miRNA candidates and families and contains additional candidates which failed or have not been considered for evaluation due to cutoffs for manual inspection.

In the initial step, `miRNAture` can use either individual miRNA sequences or pre-computed/user provided CMs. In sequence mode, miRNA-specific strategies based on `blastn` (Camacho et al., 2009) are employed. These are discussed in Hertel and P. Stadler (2015) and Velandia-Huerto, Gittenberger, et al. (2016), full details on the parameter choices are given in Appendix D: Table 21. In the following filtering step, overlapping `blastn` hits are aggregated into *extended regions* as described later and in Figure 19.

Extended regions as hotspots possible homology

As mentioned before, when searched annotated miRNA hairpins with typical `blastn` strategies the annotated candidates are expected to be located in the subject genome. An interesting problem were found by Parra-Rincón et al. (2021) and Velandia-Huerto, Gittenberger, et al. (2016), when did a genome-wide annotation of multiple ncRNA families. Using different `blastn` strategies, they found overlapping candidate regions, derived from the same non-coding RNA (ncRNA) query/queries.

In Figure 19 the processing steps performed by `miRNAture` in its `blast` homology mode is summarised, in order to define a *extended region*. The detection of homologous candidates were performed on the scaffold `JH126831.1` from *Latimeria chalumnae* using the `blast` strategy 1[4]. The raw mapped coordinates were labeled as `str1RawBlast` and comprise all the `blast` hits generated in this genomic region without any filtering. After filtered some initial candidates, a reduced number of hits remains (labeled as `str1FilteredBlast`). Then, an iteration of merging was required in order to combine two or more overlapping hits. At this point, comparisons were performed only in terms of genomic coordinates and their correspondent coverage with respect to queries (`str1FusionBlast`). The final product of those comparisons (the *region*) (`str1StrFinal`) was then subject to further evaluation with `cmsearch`.

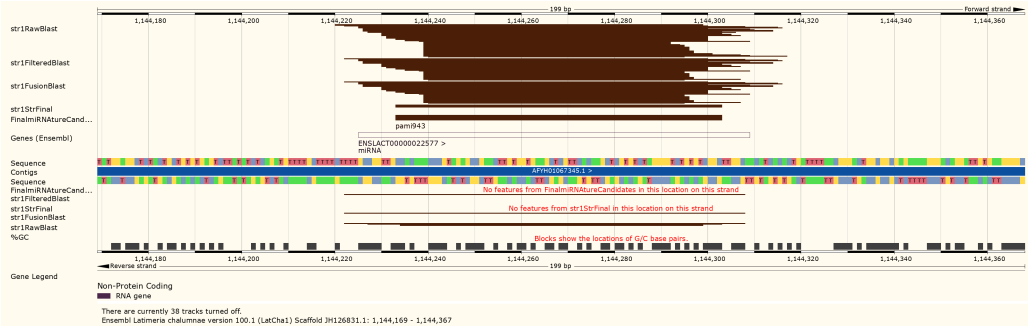


Figure 19: Visualization of merging and annotation process performed by `miRNAture` to generate *extended regions*. `blastn` hits from strategy 1 are coloured as brown tracks.

[4] `blastn dust no -soft_masking false -evalue 0.01 -reward 5 -penalty -4 -gapopen 10 -gapextend 6 -word_size 7 -outfmt 6 -out`

Pilot study: Pure homology miRNA search on solitary tunicates

Taking advantage of the homology module provided in miRNAture, it could be used as a link for quick identification of homologous regions. In that case, chosen tunicate genomes had miRNA annotations deposited on miRBase: *C. robusta* (348), *C. savignyi* (27) and *O. dioica* (66), which were used to identify homologous miRNAs using a reciprocal strategy (see Section 4.2.2). In general, running a pure pairwise homology comparisons yields a set of *extended regions* (as explained above) calculated independently based on the blastn *strategies* included in the pipeline (See Appendix D: Table 21). The subsequent identity of those regions is determined using structural alignments, derived in this example from Rfam v.14.4 miRNA families. The final list of miRNA loci is based on the determination of best-fit family given a genomic region, in terms of their secondary structure alignment scores, calculated with cmsearch (see more details below, in Section Merged structural alignments).

As expected, the number of miRNA loci is higher on *C. robusta*, given their higher number of miRNAs queries annotated on miRBase: $\sim 13\times$ and $\sim 5.3\times$ more than *C. savignyi* and *O. dioica*, respectively. As a comparison, following suggested thresholds from Rfam to annotate a locus, it yielded a reduction of the proportion of annotated hits in all species, see Table 4 column *Loci Default* compared to *Loci* miRNAture. Considering the miRNAture annotations, most of them were identified as *Not shared* loci, see *C. robusta* and *O. dioica*, probably this high number of 'specie-specific' annotated miRNAs is lower because this relation was validated with CMs composed by sequences from other species. On the other hand, this homology strategy located 20 loci reciprocally shared between Ciona spp. A reciprocal relation could be found on *O. dioica* genome, only for a miR-92 (RF00464) locus. At the same time, not all annotated miRNAs were classified into a Rfam family. The main reason stands on the missing of family structural alignments and corresponding CMs, as seen a lot of annotated sequences from *C. robusta* and *O. dioica*.

In detail, the annotated cin-mir-4029 (MI0015580) from miRBase has not been classified into a miRNA family. MirGeneDB reported as a *C. robusta* specific candidate. miRNAture recognized it on *C. robusta* (Chr6, 4033651-4033704, +) and *C. savignyi* (reftig_155, 160374-160426, -) but it did not fit in available miRNA Rfam CMs, resulting in a miss-annotation on both genomes.

Table 4: Reciprocal search of annotated 441 miRNAs on three tunicates genomes (*C. robusta*, *C. savignyi*, and *O. dioica*). Results for suggested method using cmsearch with *Default* and miRNAture. Shared/Not shared sequences are accounted in comparison to other specie. **Fam.**: miRNA families. **Ciro**: *C. robusta*, **Cisa**: *C. savignyi*, **Oidi**: *O. dioica*.

Genome	Loci Default	Loci miRNAture	Fam.	Not Shared	Shared	Ciro	Cisa	Oidi
C. robusta	37	116	50	96	20	-	19	1
C. savignyi	17	33	21	12	21	21	-	1
O. dioica	0	4	4	4	0	0	0	-

Based on this pilot study, it is important to note that despite the availability of Rfam models, the recognition of miRNA families is restricted to specific conserved families that

reported enough *seed* sequences to build a structural alignment and posterior generation of CM. In addition, most of the current annotated precursors on miRBase, have not been categorized into a miRNA family (from Rfam or miRBase), which is a bias when considered those results into genome-wide homology studies. In response to that, miRNAture is flexible allowing the inclusion of additional CMs, relaxing pre-defined filter scores, or even not considering the GA score at all, when new CMs are included (see details in Table 3).

Using HMMs and CMs

Alternatively, Hidden Markov Models (HMMs) of miRNA families pre-computed from Rfam v.14.4 stockholm alignments, or user defined ones, for example, inferred from miRBase (Kozomara, Birgaoanu, and Griffiths-Jones, 2019) can be compared against the target genome using nhmmer (Wheeler and S. R. Eddy, 2013) to determine initial candidate homologs. If CMs for the query families are available, the initial datasets are evaluated with regard to structural alignments using cmsearch Nawrocki and S. R. Eddy, 2013. miRNAture also offers the option to search the target genome with user-defined CMs. These can be obtained, for example, from alignments of miRBase sequences, directly from Rfam, or from the user's own alignments.

Merged structural alignments

Independent of the chosen strategy for the initial step, the candidate sequences are then filtered based on specific threshold values: *E-value*, coverage and if available, bitscore (using family CM threshold value defined by Rfam as GA [5]). In case of overlapping between structured candidates, miRNAture selects the best candidates comparing their scores (following this order: bitscore, *E*-value, and coverage) inferred using cmsearch. It also disambiguates the reading directions in case overlapping candidate loci at both strands. At this point, final candidate lists are merged. For each candidate, coordinates and the supporting initial hits are reported.

All the preliminary lists of candidates generated by each homology search mode, are subject of a merging process to obtain a unique list of non-redundant miRNA candidates on the evaluated sequence. In this case, each of the detected candidates are represented as a vector $\tilde{\mathrm{P}} = (\mathrm{c}, \mathrm{a}, \mathrm{b}, \mathrm{s})$, where c corresponds to the sequence contig, scaffold or chromosome name, a and b are the start and end coordinates with $\mathrm{a} < \mathrm{b}$. At the same time, $\mathrm{s} \in \{+, -\}$, indicate the strand of the detected candidate. Using those definitions, to perform the Merge function, $\tilde{\mathrm{P}}$ is sorted based on: $\tilde{\mathrm{P}}_i \preceq \tilde{\mathrm{P}}_j \iff a_i \leq a_j$, from i, j candidates. Then, to detect overlapping candidates based on their a, b pairs, is defined the Φ value, as:

$$\Phi = \begin{cases} 0, & \text{otherwise} \\ 1, & \text{if } (a_i \leq a_j \wedge b_i \leq b_j \wedge a_j \leq b_i) \vee (a_i \leq a_j \wedge b_i \geq b_j) \end{cases} \tag{4.1}$$

When $\Phi = 1$, the pair $(\tilde{\mathrm{P}}_i, \tilde{\mathrm{P}}_j)$ is subject to score comparisons by their correspondent triplet $\mathrm{Q} = (\mathrm{B}, \mathrm{E}, \mathrm{C(f)})$, where B is the *bitscore*, E their *E-value* and C(f) is the calculated *coverage* respect to the CM model. The pair $(\mathrm{Q}_i, \mathrm{Q}_j)$ is the input parameters to define

[5] https://docs.rfam.org/en/latest/choosing-gathering-threshold.html, accessed on 01.12.2021

the representative homolog between the pair $(\tilde{P}_i, \tilde{P}_j)$ by the use of the `Merge` function, described in detail in Algorithm 1 using both Algorithms 2 and 3.

Algorithm 1: `Merge` function applied for a pair of overlapping candidates $\tilde{P}_i, \tilde{P}_j$ and their correspondent reported scores Q_i, Q_j, respectively. Additional functions are necessary to determine the best candidate, namely: `best_cand` and `fusion` (Algorithms 2, 3).

```
function Merge (Q_i, Q_j);
Input  : Q_i and Q_j candidates
Output : R: reference of the best one.
if b_i == b_j then
    if e_i == e_j then
        if c_i == c_j then
            R = fusion(i, j);
            return R;
        else
            R = best_cand(c_i, c_j, 1);
            return R;
        end
    else
        R = best_cand(e_i, e_j, −1);
        return R;
    end
else
    R = best_cand(b_i, b_j, 1);
    return R;
end
```

Comparison of inferred miRNA homology regions from multiple homology-searches

This three-fold homology search might seem redundant at first glance as, not surprisingly, there is a large overlap between the search results. First, in order to evaluate the supporting scores from the detected set of homology regions, the scaffold `JH126620.1` from the coelacanth genome was selected as target sequence. This scaffold has annotated 8 miRNA hairpins and accounted for a size of 3.03 Mb. Different parameters have been chosen and modified, which internally on `miRNAture` were directly related with the miRNA classification process (Table 5): the $n_{bitscore}$, E and C(f). Experiment A corresponds to the threshold values defined on `miRNAture`, meanwhile B and C, were designed to test those filters with the default threshold values, suggested by `Rfam` to annotate a ncRNA family, or without any thresholds at all.

The number of detected homology regions for each experiment is: $A = 151$, $B = 7$ and $C = 1667$. To evidence the large overlap between homology strategies, the detected

Algorithm 2: `best_cand` identifies the highest, lowest or concatenated text value between a compared pair of scores (i, j), depending on the last parameter p, which could be 1, −1 or *text*, respectively.

```
function best_cand (i, j, m);
Input  : i, j scores and p mode
Output : S
if p == 1 then
    S = max(i, j);
else if p == −1 then
    S = min(i, j);
else
    S = concatenate(i, j);
end
return S;
```

Algorithm 3: `fusion` function combines the reported values for each estimated parameter and generates a new candidate selecting the optimal values between Q_i and Q_j.

```
function fusion (i, j);
Input  : i and j scores
Output : F
while x ≤ length(i) do
    t = best_cand(i[x], j[x]);
    add t to F
end
return F;
```

Table 5: Designed threshold modifications on `miRNAture` applied on the coelacanth scaffold `JH126620.1`. Assigned labels on **Label** column were used for reference over the text.

Label	**Experiment**	$n_{bitscore}$	E	C(f)(%)
A	`miRNAture` default	0.32	0.01	70
B	Default gathering score	1.0	0.01	70
C	No structural filters	0	100	0

A experiment regions (151) are depicted on Figure 20. The structure searches from `infernal` gathered most of the mode-specific hits, $\sim 2.7\times$ the results from the `blast` searches. Meanwhile, reduced number of hits were detected only by means of `hmm`. Only an intersection of 5 conserved regions were detected by all strategies, which corresponds to the families: miR-574 (3) and miR-130 (2). At the same time, one region that apparently did not generate a positive result for the `infernal` strategy (JH126620.1:2652551,2652619, +) was detected using the sequence homology strategies (`blast` and `hmm`). In detail, this region had a corresponding overlapping region on the opposite strand (JH126620.1:2652557, 2652624, -), which had better structure evaluation scores. As a consequence, the *forward* candidate was not considered at the subsequent analysis. As can be seen some miRNAs, are found only with one but not the other search method. The increased sensitivity, thus, justifies the extra effort, in particular when the aim is a comprehensive, high-quality annotation.

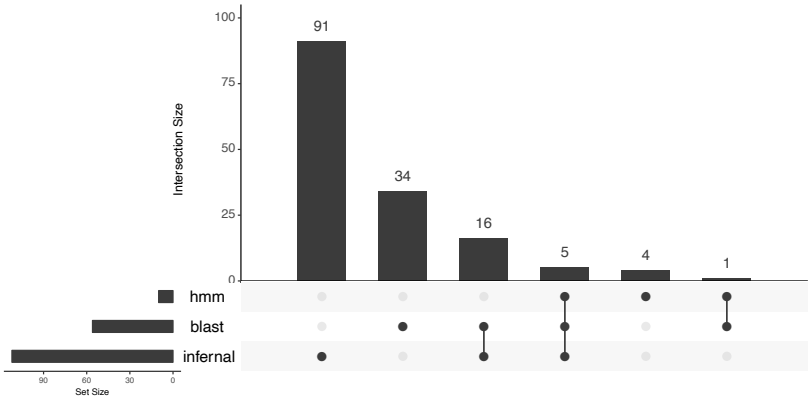


Figure 20: Intersection size of resulting homology regions from each of the available searching modes on `miRNAture` for the *A* experiment.

Those reported *homology regions* were processed and resolved as a unique non-overlapping region, based on the merging ideas described for `miRNAture` on Figure 19, for each the homology modes. In order to identify the supporting homology modes for all of those finally defined regions, Figure 21 summarised the final accounting for the defined experiments.

Figure 21 depicts the relation between homology modes (`blast`, `hmm` and `infernal`) in relation to the mode that yielded the best bitscore (B). In the upper row are those regions detected by multiple structural evaluations (SHARED) with the same B. Meanwhile, on the lower row are grouped those regions recognized by one strategy over other modes (UNIQ). In each panel it is accounted the number of regions detected by each homology modes. As an example, the category `SHARED_BLAST_HMM_Infernal`, encompass those regions with the same bitscore with all homology modes: `Blast`, `HMM` and `infernal`. In this case, were found 5 regions detected by each homology mode. In contrast, the panel `UNIQ_BLAST_NA_NA` reports those regions were the `Blast` bitscore was higher. In spite that `Blast` and `Infernal` modes detected common regions, `Blast` bitscores were higher

than those from `Infernal`. Those comparisons provided evidence to the use of multiple searching modes on `miRNAture`, due the existence of mode-specific regions.

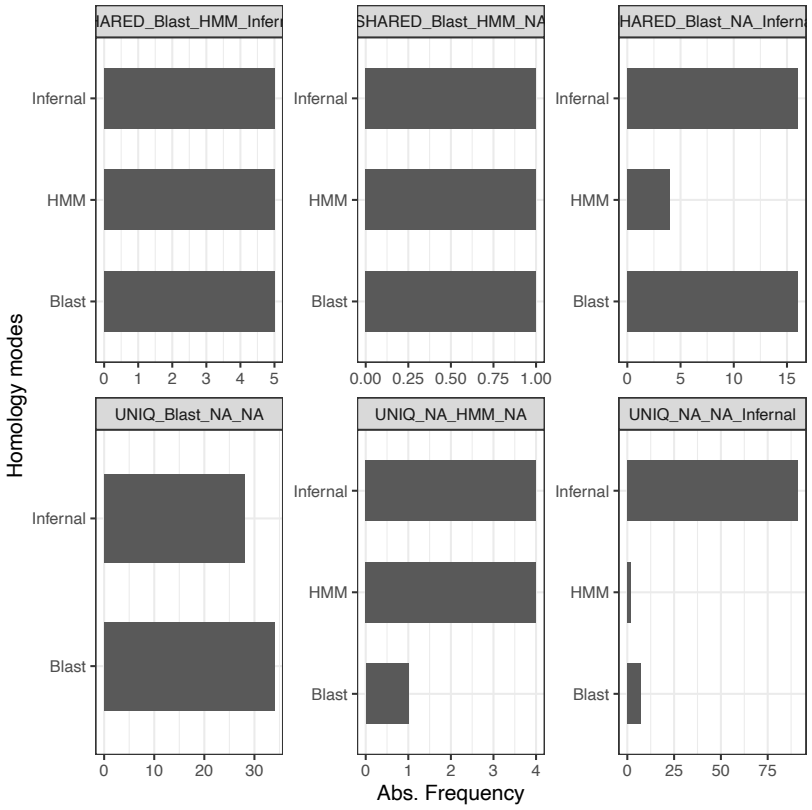


Figure 21: Frequency of detected regions by each homology strategy (`blast`, `hmm` and `infernal`) in regard their structural evaluation accessed by bitscore (B). Each panel defines which mode(s) scored higher (UNIQ) or equal (SHARED).

In addition, to inquire about the structural thresholds based on the $n_{bitscore}$ and E distribution, analysed final regions were depicted on Figure 22. In that representation, each point is a region with its associated mean values of $n_{bitscore}$ and E. Depending on their supporting homology mode(s) (blast, hmm, and/or infernal) the points were coloured. Finally, connected line indicates a that related points belong from same region. As guide, dotted line on x axis is E = 0.01 and dashed line for $n_{bitscore} = 0.32$, both designated as `miRNAture` thresholds. To perform a suitable comparison, each mentioned experiment are represented independently. In one hand, fewer regions were obtained in B experiment respect A and C. This suggests that this parameter combination is able to report the *most conserved regions*, through a high $n_{bitscore}$ numbers and low E scores. Those results were detected with A and C. As noted before, experiments A and C allowed the detection of a high number of hits. In detail, this number is larger on C, with most of the regions reporting $E > 0.01$ and $n_{bitscore} < 0.32$. Essentially, this increment on

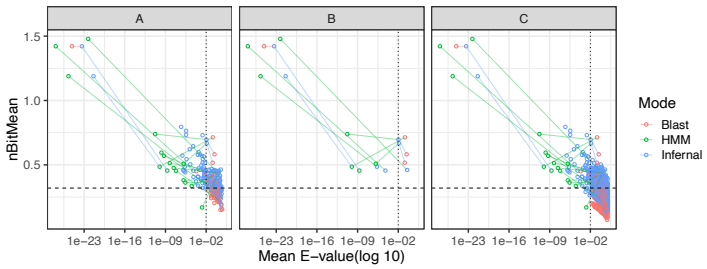


Figure 22: Performance distribution of $n_{bitscore}$ and E for detected hits discriminated by searching mode, on reported 156 homology regions. Intersected lines indicate the designed threshold values from `miRNAture`.

the numbers were not favourable at all, it increases the probability to annotate false positive candidates in genome-wide searches and increases the processing time considerably. Based on those results, a balance should be reached: threshold values should detect high conserved candidates, but at the same time increase the *grey-zone* of acceptance to rescue derived candidates, without being too flexible including false candidates.

Identifying the location of mature miR and miR*

In the next step, `miRNAture` attempts to identify the location of the mature miR and miR* within the preliminary precursor sequences. To this end, it takes advantage of an adapted and updated version of `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019)[6]. Mature miR/miR* sequences were obtained from `miRBase`. For each family, the result of this step is an alignment of corrected and trimmed microRNA (miRNA) precursor sequences annotated with the placement of the mature sequences and finally a structure-annotated sequence alignment. Corrected alignments were pre-calculated for `miRBase` families only and are available together with the corresponding Covariance Models (CMs)[7].

In the final stage, the corrected, structure-annotated alignments are used to evaluate homology search results. Since miRNA hairpins form extremely stable RNA secondary structures (Freyhult, P. P. Gardner, and Moulton, 2005), this can be used in a direct structure prediction and comparison with the consensus structure of the family, to measure how well new candidates structurally conform to a given RNA family. Together with sequence length, folding energy and sequence blocks conforming to the mature miRs, this provides a reliable filtering procedure, summarised in Table 3.

4.3.2 Accessing to `miRNAture` detection performance

Testing the performance of `miRNAture` in terms of measures like sensitivity or precision requires a dataset with a reliable ground truth of positive and negative instances. Such a dataset, however, is currently not available for miRNAs despite extensive efforts of the curators of `miRBase` (Kozomara, Birgaoanu, and Griffiths-Jones, 2019). On the one hand, only positive data, that is, miRNAs with sufficient support, are reported making it impossible to quantify specificity. Spurious annotations (Tarver, Taylor, et al., 2018) and unclear boundaries of what exactly constitutes a miRNA (Velandia-Huerto, Yazbeck, et al.,

[6] `https://github.com/Bierinformatik/MIRfix/releases/tag/v2.0.0`, version 2.0.0
[7] `http://www.bioinf.uni-leipzig.de/publications/supplements/21-001` (accessed on 26.02.2021).

2022), on the other hand, compromise the quantification of specificity. An alternative strategy to evaluate the performance is to use simulated data. This, however, requires an *independent* method to generate artificial data, in our case alignments of miRNA families, which no such tool is available. However, it is possible to use a simple simulation to get some properties of `miRNAture`'s filters. Instead of a quantitative evaluation of `miRNAture`, were therefore considered two scenarios in which a semblance of the ground truth is known from extensive manual curation: the history of the *let-7* family and the human miRNA complement. In both cases the discussion will focus on the differences between current annotation on the findings of `miRNAture`.

Let-7 family on Chordate genomes

The *let-7* family (Reinhart et al., 2000) is one of the most conserved families through metazoan species (Bompfünewerer et al., 2005; Hertel, Lindemeyer, et al., 2006; Pasquinelli et al., 2000; Sempere et al., 2006). It is also one of the largest miRNA families in vertebrates with paralogs appearing both in tightly linked clusters and distributed across several chromosomes (Hertel, Lindemeyer, et al., 2006; Roush and Slack, 2008). Since the evolution of the *let-7* family was studied extensively in the past (Hertel, Bartschat, et al., 2012; T. Liang, C. Yang, et al., 2014; Roush and Slack, 2008; Zhao et al., 2017) it provides probably the best available reference data set. In order to test consistency of the results obtainable with `miRNAture`, re-annotation experiments were performed with several primate genomes, the mouse genome and the Pacific transparent sea squirt, *C. savignyi*, as targets. In each case, all miRNAs annotated for the target genomes were removed from the alignments and CMs of the query, see Section 4.2.4 for details. To consider missing annotations in `miRBase`, `miRNAture` intersected derived *let-7* loci also with the manual annotation of Hertel, Bartschat, et al. (2012). The latter, together with `miRBase` annotation (MIPF0000002), are considered the *gold standard* annotation. Table 6 summarises the results for the *homology* stage and the *final* stage of `miRNAture`, respectively. In summary ≥91% of all annotated *let-7* loci in all species were recovered by `miRNAture`, while in all cases, except the solitary tunicate, one of the annotated loci was not identified. Furthermore, between 1 and 11 additional loci per genome are considered valid, novel *let-7* candidates.

The missing candidates correspond to locus *K-let-7* in all primates, in the nomenclature of Hertel, Bartschat, et al. (2012), which considers homology of paralogs based on synteny. Hertel, Bartschat, et al. (2012) reported those loci as primate-specific novel candidates based on homology. However, the consensus structure generated from a multiple alignment with annotated *K-let-7* sequences shows a multi-loop structure where the typical miRNA hairpin is expected, while *G-let-7-1* was detected in the mouse genome considering only the homology stage, but discarded by structural filters in the evaluation stage for similar reasons (See Figure 23).

Since the *homology* stage of `miRNAture` is optimized for sensitivity and only *let-7* was used as a query, *bona fide* miRNAs that share some similarity with *let-7* are expected to have passed filters. The additional *let-7* loci found by `miRNAture` were therefore compared to the annotation of other miRNA families. Were indeed found overlaps with `miRBase` annotation for the human specific *hsa-mir-4699* (MI0017332), and the families *mir-3596* (MIPF0001194), and *mir-625* (MIPF0000534). A *mir-3596* was annotated in *Rattus*

Table 6: Re-annotation of the *let-7* family. For each species the number of loci annotated by `miRNAture` is shown at the homology stage (Homology) and after the evaluation stage (Final) and compared with the gold standard annotation merged from `miRBase` and Hertel, Bartschat, et al. (2012) (Ann.). We show how often the genomic coordinates from annotation match with candidate region (Match) or are not in the final candidate set (Miss) and the respective ratio over the total number of annotated regions. Candidates which pass the evaluation stage but do not overlap with annotation are counted as Additional. Labels: **Ann.**: Annotation, **Add.**: Additional, **Filt.**: Filtered.

Species	Homology	Final	`MIRfix` Filt.	Ann.	Match	Miss	Ratio Match	Ratio Miss	Add.
Human	26	20	6	14	13	1	0.928	0.07	7
Orang-Utan	27	18	9	14	13	1	0.928	0.07	5
Gorilla	26	20	6	14	13	1	0.928	0.07	7
Chimpanzee	30	24	6	14	13	1	0.928	0.07	11
Mouse	19	14	5	12	11	1	0.916	0.08	3
Sea squirt	7	6	1	5	5	0	1.0	0.0	1

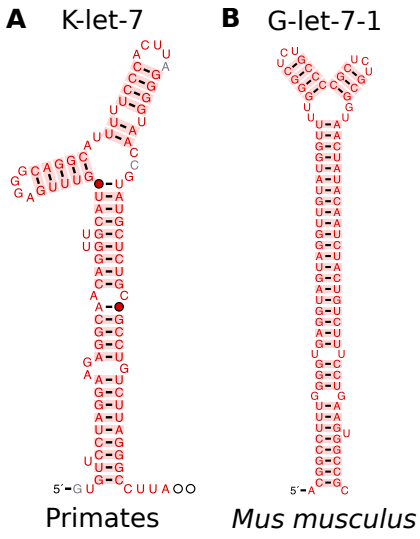


Figure 23: Discarded let-7 sequences. **A.** *K-let-7* consensus structure derived from reported primate sequences in Hertel, Bartschat, et al., 2012. **B.** *G-let-7-1* locus from mouse.

norvegicus and identified by `miRNAture` also in mouse. The *mir-625* family was known in human and macaque only. These cases account for a third of the additional matches. Almost all the remaining loci overlap with regions annotated as repeats. Only three loci (human: chr1:16082685-16082783:+, chimpanzee: AACZ04010697:5895-5965:-, and *C. savignyi*: reftig_41:1114844-1114937,+) do not overlap with available annotation. The similarity of *mir-625* and *let-7* was noted before e.g., by `Rfam`, which includes *mir-625* in their *let-7* miRNA precursor family `RF00027`. In T. Liang, C. Yang, et al. (2014), *mir-3596* is treated as a member of the *let-7* family, highlighting that `miRNAture` presumably classified them correctly as novel candidates.

Simulation of Artificial *let-7* Instances

To check the behavior of `miRNAture` in the presence of large sequence divergence artificially were mutated two of the human *let-7* genes (chr21:16539829-16539913:+ and chr3:52268269-52268368:−) with increasing number of mismatches. For up to 10 point mutations, the loci were recovered at the homology stage, 4 candidates passed the homology filters and 2, overlapping the original loci, also survived the structural filters. At higher artificial mutation rates none of the initial candidates satisfied the structural constraints. As expected, at even higher mutation rates eventually also the initial homology search fails.

Human microRNAs

A typical use case for `miRNAture` is the annotation of a genome of interest with a set of available miRNA family CMs. To simulate such a use case and simultaneously further benchmark `miRNAture`, we used a set of 350 miRNA families with a human entry in `miRBase` v.22 to construct query alignments from which all human sequences were removed (see Section 4.2.5).

At the *homology* stage, `miRNAture` detected miRNA candidates for all but a single family. Considering the annotation of mature sequences and curation at structural level with `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019) in the validation stage, candidates for 323 families (92.23% of initial miRNA CMs) were retained. For 27 families candidates were found on homology level, but later discarded based on the evaluation of structure and localization of mature miRNA regions within the hairpin.

In order to better understand the performance of `miRNAture` at the *homology* stage we distinguish families where all candidates overlap exactly with annotation (337 families, 96.3%) and those where a part of the candidates overlap (12 families). The only family that was not recovered at all is *mir-297* (MIPF0000204), with 100 initial candidates, of which 69 passed the filtering steps. However, none of them matched the annotated loci. In mouse, six precursor loci have read support at mature sequence loci of which 4 are annotated by `miRBase` as *high confidence* miRNAs. Additional homologs at a single locus without read support are annotated in rat and some primate genomes: *Macaca mulatta*, *H. sapiens* (very weak read support) and *P. troglodytes*. Input for `miRNAture` was a CM model built from mouse validated sequences, which in comparison to the known human locus contain 20% more nucleotides and 10% additional consensus positions.

Of the 323 families left after the homology stage, most show a perfect match with current annotation for all accepted candidates (87.9%), see Table 7. The final output of `miRNAture` comprises 284 (81.1%) families with perfect overlaps with the current annotation. Another 28 families show partial matches. Among the remaining 38 families, there are 27 for which no candidate passed the filtering steps. The other 11 families contain additional candidates, but are disjoint from the current annotation (do not show overlap). Table 7 summarises the statistics.

The 11 families with candidates disjoint from human annotation are *mir-1233*, *mir-1291*, *mir-1306*, *mir-140*, *mir-6127*, *mir-645*, *mir-652*, *mir-764*, *mir-873*, *mir-877* and of course *mir-297*. For annotated loci of these miRNAs that were not recovered at least one of the following statements are true: (a) There is no mature sequence alignment available that allows the correct annotation of detected candidates; this is in particular the case for

Table 7: Comparison of Accepted/Filtered `miRNAture` miRNA candidates with respect to the current human miRNA annotation. For a final classification of `miRNAture` miRNA candidates, the latter are intersected with current `miRBase` v.22 annotation on genomic loci level. Candidates were classified as follows: **Accepted**: Candidate passed evaluation stage, **Filtered**: candidate did not pass evaluation. Numbers for all candidates of a specific family overlap (Perfect), some overlap (Partial) and no overlap (Without). Furthermore, we investigate for how many families candidates currently not contained in the annotation of the corresponding family (Additional) are predicted or `Filtered` during evaluation. This set contains families from the `Partial` and `Without` class.

Class	Perfect	Partial	Without	Total	Additional
Accepted	284	28	11	323	178
Filtered	27	0	0	27	5

species-specific families. (b) The location of the mature sequences was determined based on similarity to human loci alone, without additional information. With the artificial removal of the human data this information is unavailable in our benchmark. (c) The `miRNAture` pipeline favours the opposite strand. Details can be found in Section 4.3.2.

For 27 families, all candidates where filtered out even though they show a perfect match with the current annotation. These cases can be traced back to miRNAs which belong to either species specific families, thus lacking homology information, or have only been found in a small set of other species, consequently restricting the available dataset of mature loci, or folding into invalid secondary structure. In total, this lead to rejection of 33 loci, see Section 4.3.2.

In summary, the loci that `miRNAture` did not cover in the human genome fall into two broad classes: (1) members of repetitive families, for which we consider it uncertain whether they should be considered as canonical miRNAs; (2) precursors with deviant secondary structure or unusual placement of the mature sequences within the predicted secondary structure. These families deserve a closer look whether they are canonical miRNAs in the stringent sense used here. If so, they may prompt a future adjustment of the filtering criteria; (3) Families for which the query alignment and secondary structure contains an insufficient number of precursor sequences or contains undetected errors in alignment, positioning of the mature sequences, or consensus structure annotation. Those should be considered as borderline cases that deserve further investigation into the underlying evidence preferably from multiple species.

Additional Candidates

For 178 families (1366 loci) additional candidates were predicted. At the same time, 5 families (6 loci) were removed by filtering steps. A comparison with the current annotation shows that ∼69.0% of those *additional* loci overlap with one or more annotated element(s) (see Table 8). For 12 families we found candidates that overlap with annotation of other miRNA families (different), while for 73 families we find overlaps with repeat regions (repeat), and 31 which overlap with other annotation (other) including, for example, intronic or exonic regions of lncRNAs or coding genes. For the *mir-1233* family, for instance,

20 additional loci were reported. Almost all of them are located on chromosome 15 and overlap retained introns or lncRNAs derived of the palindromic GOLGA8 gene family, described as core duplicons dispersed along $\sim$ 14 kbp, associated to structural variants and genomic instability regions in general Antonacci et al., 2014; Maggiolini et al., 2019.

Table 8: Additionally predicted loci in comparison to current annotation for human (hg38). Reported numbers of families overlapping with each respective annotation category and the number of loci from this families in parentheses. **d**: different miRNA, **r**: repeat and **o**: other non-intergenic region. Numbers were reported keeping a hierarchical comparison to avoid intersections between sets as: $d > r > o$.

	d	***r***	***o***	**Total**
Number	12 (13)	73 (685)	31 (245)	116 (943)
Fraction	0.010 (0.014)	0.629 (0.726)	0.267 (0.26)	

Twelve miRNA families show overlaps with other miRNAs that are annotated as human-specific. An exceptional case is *hsa-mir-499b* (MI0017396). The homology stage of `miRNAture` suggests that it belongs to the *mir-499* family, however, `miRBase` does not include it in this family. We argue that *hsa-mir-499b* is correctly annotated by `miRNAture`.

Additional `miRNAture` Candidates without Annotation Overlaps

For 129 families (423 loci) `miRNAture` predicted candidates that do not overlap with any currently annotated genomic element on the same strand. The miRNA families *mir-544* (50), *mir-548* (42), *mir-1302* (27), *mir-1289* (21), *mir-649* (19), *mir-290* (17), and *mir-297* (11) account for nearly half of them. To further investigate those candidate loci, we intersected available annotation specifically at their 'antisense' strand and found 105 overlaps. Interestingly, more than half (53.9%) of those are found in overlap with repetitive elements, (24.82%) overlap with a miRNA annotated on the opposite strand while in 5.67 a coding gene and in 0.94% a lincRNA is annotated in antisense. Integration of expression patterns derived from a small RNA-Seq dataset from Kuksa et al. (Kuksa et al., 2019) revealed read support for $\sim$8% of these 105 candidates. As example, Figure 24 show a *mir-580* precursor. It was predicted in antisense to the 3'UTR of the protein coding gene *STAM*, a locus well conserved among primates.

Many of the additional loci overlap specific repeat families. All additional *mir-544* loci overlap with the DNA transposon MER (*medium reiterated frequency repeat*). It is interesting to note that the annotated *mir-544* loci are located in the DLK1-DIO3 imprinted region (Edwards et al., 2008). Similarly, 42 *mir-548* loci overlap with Tc1/Mariner. The extensive, repeat-like *mir-548* family has received detailed attention in the past (T. Liang, Guo, and C. Liu, 2012; Piriyapongsa and Jordan, 2007), highlighting its atypical features deriving from Made1 elements, a class of inverted-repeat transposable elements (MITEs). Many *mir-548* loci have been reported to match Made1 elements in both reading directions (Piriyapongsa and Jordan, 2007). The family also features an atypically large divergence among their mature sequences. Some paralogs share the same locus on different strands and generate miRNA:miRNA* duplexes lacking the otherwise typical hairpin loop region (T. Liang, Guo, and C. Liu, 2012).

Furthermore, 12 loci from 11 families passed the validation stage, but the predicted position in the human genome is not overlapping with annotation. A better fit of the mature sequences was found for the opposite strand in 3 cases: *mir-764*, *mir-140* and *mir-1306*.

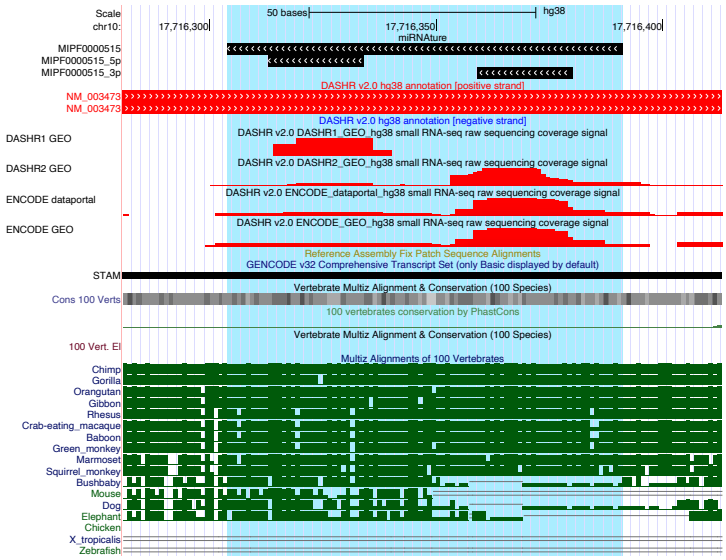


Figure 24: Crossed annotation and expression patterns overlapping in the same region where a locus of *mir-580* was detected by `miRNAture`. Red tracks correspond to sRNA-Seq mapped reads, reported from Kuksa et al., 2019.

Strand-Mismatch Candidates

For six loci that pass all evaluation steps of `miRNAture`, the strand may be mis-annotated. In each case, valid homology regions were detected on both strands, the opposite strand was preferred by `miRNAture` based on the prediction of unusual positions of the mature sequences in relation to the secondary structure for the other strand. Examples are overlap of the mature sequence and the hairpin loop, or a multi-loop structure, see Figure 25. Predictions that match better to the opposite strand than the annotated locus were found for the following families: *mir-101*, *mir-103* (2 loci), *mir-122*, *mir-1245*, *mir-290*, *mir-451*, *mir-4536*, *mir-515*, and *mir-548*. Difficulties with the *mir-451* family are not unexpected due to its atypical biogenesis and a dominant mature product deriving from the loop region (Cifuentes et al., 2010).

Missing Candidates

At the homology stage of `miRNAture`, 90 annotated miRNA loci from 13 families were not recovered. The majority of which were not reported due to the large number of detected homolog loci since `miRNAture` limits candidate lists to the 100 loci with the best

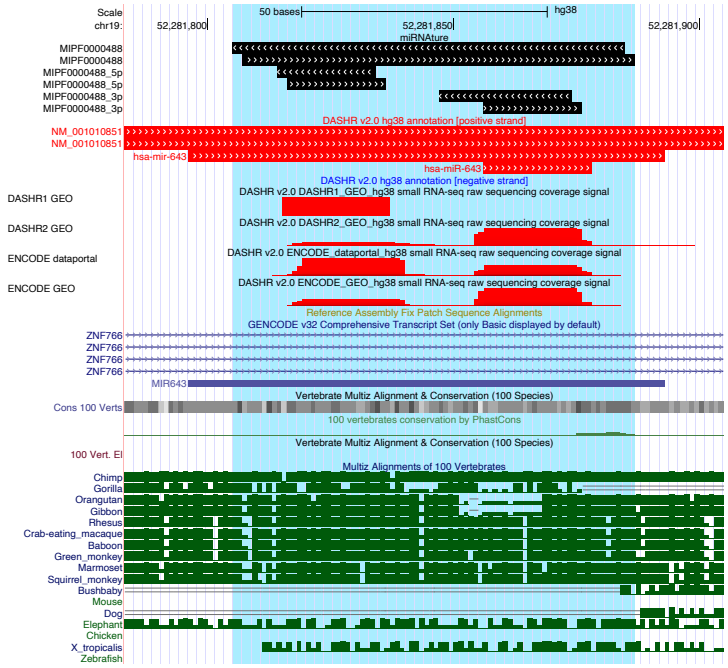


Figure 25: Example of overlappings with current miRNA annotation in human genome. Annotated *mir-643* loci were detected by `miRNAture` on the same strand and additionally an opposite locus from the same family was detected. Supporting expression patterns were detected by both, 5' and 3' miR; however, currently only on the 5' miR is annotated.

bit-scores. This cut-off is intended to exclude candidates associated with highly repetitive sequences for which different, synteny-aware methods have to be used, see, for example, Velandia-Huerto, Berkemer, et al. (2016). Candidates not passing this cut-off are flagged as *potential* candidates for later inspection by the user. 68 missing candidates can be explained in this manner. For example, *mir-548* represents a highly repetitive miRNA (with 74 annotated loci in `miRBase`). `miRNAture` detected in total 6626 candidates by homology searches, highlighting the need for stringent cut-offs. From them, the best 100 bit-score candidates were subject to mature annotation and compared to the annotation, 63 were classified as *potential* and another 2 were predicted on the opposite strand. Among the remaining 22 of the 90 *missing* loci belonging to non-repetitive families are three predictions on the opposite strand (*hsa-mir-103b-1* MI0007261, *hsa-mir-103b-2* MI0007262 and *hsa-mir-371b* MI0017393) and `miRNAture` rejected a total of 19, including five *mir-548* paralogs, four *mir-378* sequences, and three *mir-506* loci, while the remaining 7 undetected loci were not recognized by the corresponding HMM or CM miRNA family models.

In the final result, that is, after the validation stage, 161 loci from 66 families are classified as *missing*, see Table 7. Of these, 45 loci from 38 families were rejected, either at the validation stage (27 families) or did not show genomic loci overlap (11 families) with annotation. For the first group comprising 33 loci, only a limited set of annotated mature sequences was available, and predicted mature sequences were incorrectly placed,

so that the loci were eventually rejected. The *mir-550* family (MIPF0000334), for example, has five loci annotated in human. These were rejected by `miRNAture` because the mouse and chimpanzee loci retained in the input did not pass the secondary structure filters. Similarly, all three *mir-1184* loci predicted at the homology stage were rejected because of the atypical secondary structure of the only remaining input sequences (from chimpanzee).

Another source of *missing* candidates are the 28 families that matched the current annotation only partially, accounting for 116 loci. These include the ten loci assigned to opposite strand and the 66 highly repetitive loci tagged as *potential* that have been discussed above. Of the remaining 40 loci, 19 were not detected by homology and 21 were rejected by `miRNAture` at the evaluation stages. Here, the CM constructed from `miRBase` data after removing the human sequences did not match the annotated human loci. For example, *hsa-let-7g* was rejected at the validation stage because the common *let-7* CM identified a sequence that was shifted relative to the paralog-specific results of Section 4.3.2, resulting in a less stable, shifted MFE structure. While it perfectly matched the mature sequence platypus *oan-let-7g-5p*, human mature sequences overlap the hairpin region, explaining the rejection.

4.3.3 Availability

The `miRNAture` pipeline can be downloaded from `https://github.com/Bierinformatik/miRNAture`. It is provided as `Conda` package for installation, which resolves all dependencies and includes a detailed user manual, a tutorial and extensive example data.

4.4 Discussion

The `miRNAture` pipeline is based on the observation that efficient homology search requires the interplay of fast and ideally loss-less identification of candidate loci in the genome of interest, and subsequent filtering to remove the false-positives. Since it is not difficult to increase the sensitivity of initial search (by simply lowering cut-off values), better and in particular more complete results can be achieved by developing more efficient filters. This is not a new principle, of course. HMMs improve over single-sequence `blast` queries by including patterns of sequence conservation, and covariance models provide another jump in accuracy by incorporating the conservation on secondary structure level. The trade-off, however, is the need for more and more information on the query side. While `blast` requires only a single sequence, `nhmmer` requires a multiple sequence alignment to derive the HMM model, and the CMs used by `cmsearch` need a consensus structure in addition to the sequence alignment. CMs thus are helpful only if the RNA family has an evolutionary well-conserved secondary structure.

`miRNAture` increases the achievable sensitivity by further restricting the scope of queries—its filters are highly specific for *canonical microRNAs*, i.e., those that share all the typical features of miRNA precursors, in particular a secondary structure that resembles a nearly symmetric stem-loop and a sequence conservation pattern governed by the location of the mature products on both sides of the stem region. Therefore, `miRNAture` tends to reject members of atypical families such as those associated with repetitive elements. The main use of `miRNAture` is to reliably process the typical cases

and to limit the need for extensive manual analysis to miRNAs and miRNA-like ncRNAs with atypical features.

The integration of the mature sequences and the evaluation of folding energies reaches beyond the information captured by HMMs and even CMs. This yields more stringent filters that make it feasible to increase the sensitivity of the initial homology search. The cost incurred for this advantage is the restriction of `miRNAture` to canonical microRNAs. While general approaches can be extended to other classes of RNAs, such as box C/D snoRNAs or box H/ACA snoRNAs, class-specific filters need to be developed and tested. This requires extensive domain knowledge and thus makes it difficult to extend the strategy to poorly understood ncRNAs.

The `miRNAture` pipeline is designed specifically to facilitate homology search for canonical miRNAs. The most obvious use case is the annotation of all conserved miRNA families in one or more new genomes. Complementarily, studies into the evolution of specific miRNA families require that (i) distant homologs can be detected reliably and (ii) no spurious apparent homologs are included. Only then is it possible to pinpoint the evolutionary origin of a miRNA family (Hertel, Bartschat, et al., 2012; Hertel, Lindemeyer, et al., 2006; Tarver, Taylor, et al., 2018). Although `miRNAture` usefully assists both tasks, a number of issues remain that will require manual intervention and post-processing. Most importantly, the method relies on correct initial models for each microRNA. We recommend to use models that are specific to individual `miRBase` families, or—in the case of families with divergent paralogs—even paralog specific input alignments. While it is possible to use `Rfam` family models, these turned out to be too promiscuous in many cases, resulting in relatively large fractions of rejected candidates.

The study presented here also highlights the difficulty of benchmarking homology search tools for ncRNAs. The main reason is the lack of a gold standard of sufficient quality and coherence. Databases such as `miRBase` or `Rfam` by design contain entries that satisfy certain levels of evidence. These evidence criteria, however, imply massive ascertainment biases between organisms as a consequence of the large differences in the available empirical evidence. On the other hand, the definition of miRNAs as a class is fuzzy to certain extent as well, implying that not all database entries share all the features that are typical animal miRNAs. In `miRNAture`, very stringent quality criteria are implemented. While the evaluation against the human annotation shows that clear false positive calls are rare and largely confined to repeat-associated families, `miRNAture` fails on `miRBase` families with atypical features. The `miRNAture` pipeline also reports candidates of the homology stage that are later rejected by the automatic curation procedure to enable expert inspection. Such datasets are required to gather enough knowledge about miRNAs with atypical features.

In principle, it would be desirable to benchmark `miRNAture` and similar tools against simulated data with a guaranteed ground truth. The difficulty is that checking the sensitivity and specificity of the filters requires a way of simulating the evolution of artificial miRNAs that is independent of filter rules employed by `miRNAture`. It would be easy of course, to use `miRNAture`, that is, the `MIRfix`-based evaluation to model the selection pressures on miRNAs, but then our filters would be perfect by construction, and no information on the biological correctness of the filters could be gained. On the other hand, it is very simple to construct negative examples, since 10%–20% of randomly placed point mutations is known to almost certainly destroy the secondary structure (Fontana

et al., 1993). We have seen in Section 4.3.2 that this is indeed also the case in our setting and therefore resorted here to comparing `miRNAture` with the known, well-curated miRNA annotation of the human genome. Again a fully quantitative evaluation is difficult, because of the grey-zone between *bona fide* canonical miRNAs and other miRNA-like genes.

The strategy of `miRNAture` may serve as a blueprint for a new generation of homology search tools that rely on class-specific post filters. Here, we have manually constructed the homology and secondary structure filters, making use of explicit knowledge on structure and biogenesis of miRNAs. It seems tempting to use machine learning classifiers for miRNA gene detection, reviewed e.g., in Saçar and Allmer (2014), as an alternative. However, the correlation between miRNAs used for training and their homologs are a concern that will require detailed evaluation before such a strategy can be employed safely. For the time being, explicitly constructed filters thus seem preferable.

— 5 —

From dust to Homology: Genome particularities from the sea vomit *Didemnum vexillum*

Contents

5.1 Current state of animal diversity reflected on genome assembly projects

Current sequencing genome projects have increased due to the fast advances of next generation sequencing technologies, lower associated cost, and the parallel development of efficient computational methods to deal with this astonishing quantity of available genomic information. As shown in Figure 26, the accumulated number of RefSeq genomes (nuclear and mitochondrial) accounts 114,396 stored in NCBI. In case of metazoan species, vertebrates (mammalian + other species) sums up $\sim$ 2026 and invertebrates about 4825 species[1]. In terms of current metazoan taxonomy, those numbers does not reflect the reality, which the artificial invertebrate group represents about 95% of metazoan species (GIGA Community of Scientists, 2013).

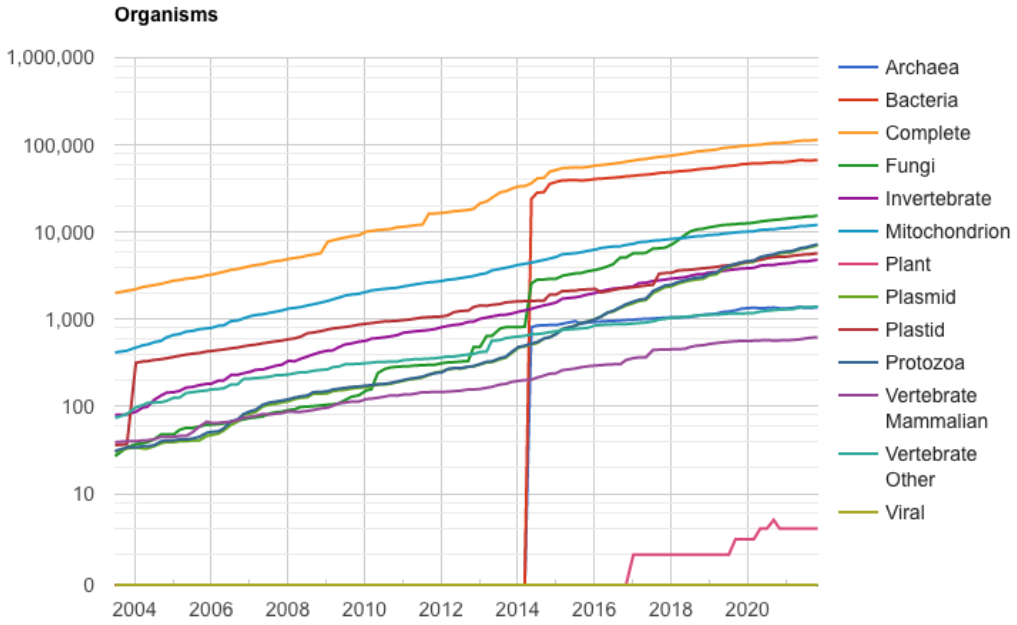


Figure 26: Increment in number RefSeq organism in NCBI. Colours are designated to different phylogenetic or artificial groups. Obtained from `https://www.ncbi.nlm.nih.gov/refseq/statistics/`.

This suggests that genomic biodiversity of animals, in particular invertebrates, is far to be complete being unrepresented in relation to high sampled clades as vertebrates. To tackle this challenge, multiple genome assembly collaborative projects have been organized to increase the sampling of available invertebrate genomic resources. As an example, the Global Invertebrate Genomics Alliance (GIGA) (GIGA Community of Scientists, 2013)[2], is interested in evaluate the broad spectrum of invertebrate phylogenetic diversity, standardize

[1] Data obtained on November 1, 2021
[2] `http://www.gigacos.org/`

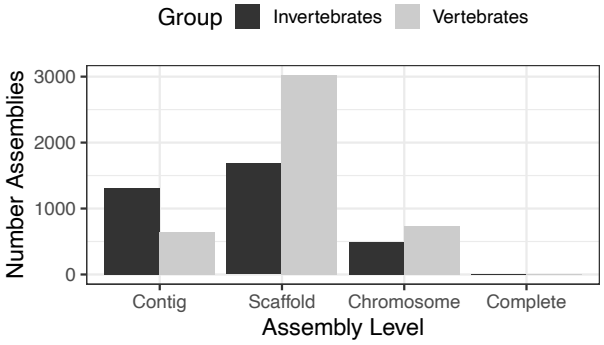


Figure 27: Assembly level of metazoan genomes in NCBI. Data extracted from https://www.ncbi.nlm.nih.gov/genome/.

sequencing and assembly methods to maximize the utility of obtained data, which would serve for comparative studies and the annotation for about 7000 new invertebrates. Other join efforts such as Darwin Tree of Life[3], Earth BioGenome Project[4], European Genome Consortia [5] are expected to sequence, catalogue and describe the broadest number of invertebrate and vertebrate genome assemblies (Redditt, Braund, and Bovon, 1992).

In spite the outstanding efforts to get enough DNA material from the target organisms and the increasing methodologies to generate consensus genome assemblies, additional steps are required to get a meaningful and comparable genome version. After multiple iterations of cleaning and processing the raw material to represent it into a digital representation, this fragmented pieces of information have to be re-assembled to get back a close chromosomal representation for each specie. The current assembly level status of metazoans is revisited in more detail in Figure 27, where an increased contiguity is achieved in this order of assembly categories: Contig $\prec$ Scaffold $\prec$ Chromosome $\equiv$ Complete. The assemblies were binned depending on whether assembly belongs from a vertebrate or invertebrate specie. As a result, in both groups the *scaffold* level is the most common. There are more invertebrate assemblies in *contigs* than in vertebrates. However, at *chromosome* or *complete* levels, vertebrates reported more genomes.

Despite the large number of assemblies at contigs/scaffolds level, those initial assembly stages could be used as significant sources of relevant biological information and annotations. Further considerations should be taken when analysing non-model organisms, which in many aspects, can be challenging to standard protocols or computational pipelines due their biology, sampling conditions, or even sequencing methods. In this chapter, an improvement on the assembly and annotation level to the sea carpet squirt (*Didemnum vexillum*) is assessed. Once the assembly hypothesis are generated, annotation pipelines are pivotal to discover and infer biological relevant relations. Furthermore, the development of annotation workflows in this chapter were done over conserved coding and non-coding RNA elements, as reported in Parra-Rincón et al. (2021).

[3] https://www.darwintreeoflife.org/
[4] https://www.earthbiogenome.org/
[5] https://www.erga-biodiversity.eu/

5.1.1 An invasive and particular non-model species: *Didemnum vexillum*

The marine ascidian *D. vexillum* (family Didemnidae, order Aplousobranchia) were described for the first time by Kott (2002) in the Coromandel Peninsula (New Zealand), as a possible *indigenous* specie. Over time *D. vexillum* was recognized as an aggressive invasive specie that can easily adapt to multiple marine environments as well as in temperate areas. As support of that, their invasive status was changed in New Zealand to *a very recently introduced specie* (G. Lambert, 2009), that once becomes established, it has the capability to grown over organic or inorganic substrates (S. G. Bullard et al., 2007; G. Lambert, 2009; Stefaniak, 2012).

As a colonial organism, it is composed by $\sim$ 1mm zooids, arranged in thin sheets fusing themselves or with external surfaces (Kott, 2002). Additional to its highly plastic life-cycle, it has a high rate of inter-colony fusion (K. F. Smith, Stefaniak, et al., 2012). In fact, each didemnid zooid is hermaphroditic and protandric. As confirmed by Ordóñez et al. (2015), in the same colony coexists multiple sexual stages: 1) Immature, 2) presence of testis, 3) presence of testis and oocytes and 4) presence of oocytes alone. Reproduction can be done asexually by budding or fragmentation, and subsequent reattachment promoting colony growing (S. Bullard et al., 2007; Ordóñez et al., 2015; Valentine et al., 2007) or sexual to promote recombination through the production of new individuals. As highlighted by Ordóñez et al. (2015), the direct impact concerns specially to the aquaculture industry, which paradoxically is the main highway to disperse by shipping these organisms worldwide. In this way, *D. vexillum* has been defined as a *native* species from the Northwest Pacific Ocean, including Japan (G. Lambert, 2009; Stefaniak et al., 2012), and are distributed world-wide (Casso et al., 2019). Studying world-wide biological samples Stefaniak et al. (2012) found the most genetically diversity, the highest number of haplotypes, greater haplotype diversity and specific-haplotypes belonged from Japan populations. Particularly, the recognition of multiple alleles from single-copy *tho2* gene suggested that a single colony should not be composed by only one genome, but chimeras (Stefaniak et al., 2012). The formation of chimeras (via allogenic fusions) resulted in a larger colony with multiple genotypes (Casso et al., 2019).

Chimeras resulted by the fusion of two or more colonies.

In 2016 Velandia-Huerto, Gittenberger, et al. reported the non coding RNAs (ncRNAs) complement on the first draft genome from *D. vexillum* using Illumina paired-end (PE) reads. Despite the draft genome status (with 882,106 contigs and a $N50$ =918 nt) reported at that time, the housekeeping ncRNAs were mostly identified. Specifically for the microRNAs (miRNAs) complement were recognized a substantial restructuring of miRNA families observed as well as in *O. dioica* and Ciona spp. (Fu, Adamski, and E. M. Thompson, 2008; Hendrix, Levine, and Shi, 2010).

5.2 Computational approaches to disentangle *D. vexillum*

5.2.1 Annotation of ncRNAs

Homology searches

Annotated ncRNA candidates from the first assembly of *D. vexillum* were mapped in the new assembly as described in Section 5.2.3. At the same time, homology `blastn` and Hidden Markov Model (HMM) strategies with their corresponding metazoan-specific Covariance Models (CMs) and default CMs evaluation have been applied following the methodology proposed in Velandia-Huerto, Gittenberger, et al. (2016), to annotate candidates that have not been detected with the mapping strategy.

The transfer RNA (tRNA) genes were found using `tRNAscan-SE` v.2.0.3 with default parameters. For other ncRNA families, a final check of candidates was performed to ensure that reported `Rfam` families contain at least one metazoan sequence in their original seed alignment. These last step was performed to report possible false-positive families that could be retrieved applying the default `Rfam` models directly to the genome.

Annotation of mature microRNAs on Rfam models

Public `MySQL` Database from `Rfam` v.14.1 was used to obtain both, accession numbers and annotations of miRNA sequences. `RNAcentral` v.13 (The RNAcentral Consortium, 2018) was accessed to retrieve the stable identifiers between annotated sequences from `Rfam` and `miRBase` v.22.1. Based on these identifiers, the designated *seed* sequences by `Rfam` were classified as: *one to one* if they have available annotated *mature* sequences on `miRBase`. Those sequences that did not have any *mature* annotation were assigned as *one to many* group. For the first group, a validation of the reported mature positions were performed by `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019). Based on those corrected sequences and positions, a new iteration of mature validation were performed including the second group of sequences, which did not report *mature* annotation. From those CMs that all sequences were classified in the *one to many* group, the missing of specific *mature* annotations has been solved inferring the *mature* position based on the reported family-specific `stockholm` alignment, by the identification of conserved correspondent blocks along the alignment, which along the miRNA structure model would correspond to the *stem* region. Based on the previous results, the default `stockholm` alignments were corrected based on the position of the predicted/annotated *mature* sequences. Then, each previously detected miRNA from *D. vexillum* was corrected again with `MIRfix`, based on its previously inferred `Rfam` family. In cases where existed *loci* from the same family, each sequence was compared independently against the corrected set of *seed* sequences from `Rfam`. As a result, those miRNAs from *D. vexillum* that reported mature regions, inside their predicted precursor, and fit into the structural alignment were considered as true candidates. The general methodology resembles an earlier method explained in Chapter 3;Figure 11, without considering the *Evaluation stage*. The *tree edit distance* ($e_{distance}$) calculated with the `Vienna RNA` package (Lorenz et al., 2011), was used to measure the variation between original `stockholm` alignments from `Rfam` and re-calculated ones, that included found loci.

Computational identification of miRNAs

Based on the previously corrected set of `Rfam` *seed* sequences, an evaluation of predicted *D. vexillum* miRNAs was performed using `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019).

Precursors that contain mature annotation and are supported by a correct structural alignment, were considered true candidates (for details see Section 5.2.1). To retrieve phylogenetic distribution of the `Rfam` sequences, taxonomic distribution (annotated as: *kingdom*, *phylum* and *subphylum*) was accessed from NCBI Taxonomy Browser[6] for species in the `Rfam` stockholm alignments.

5.2.2 Study of covariance model thresholds

Reported fasta sequences from precursor miRNAs were retrieved from the *H. roretzi* genome (v.1) (K. Wang et al., 2017). This input file was subject to structural evaluations with `cmsearch`. Control positive sequences were retrieved from `MirGeneDB` v.1.0 (Fromm, Billipp, et al., 2015) and it was composed by all the reported sequences in the database with additional 30 flanking nucleotides[7], obtaining 8656 from 8847 sequences that reported a miRNA family annotation (discarding 'novel' families). Control false sequences were generated from the reported CDS sequences from human genome (v.GRCh38) retrieved from `Ensembl` [8]. Selected sequences reported lengths of 80 and 150 nucleotides. Next, those selected candidates were sampled randomly (with replacement, 95% of confidence, 5% of confidence interval and a total of 4694 sequences) to create random seed groups; this sampling methodology was replicated 10 times. In order to shuffle the nucleotides inside those random seed groups, `shuffleseq` from `EMBOSS:6.6.0.0` (Rice, Longden, and Bleasby, 2000) was applied, generating 100 shuffling steps on the query sequences, as follows:

```
shuffleseq -sequence input.fa -out output.fa -shuffle 100
```

In order to analyse the distribution patterns of bitscore, it was necessary to normalize those values because gathering scores (GA) are covariance model specific. In this case, normalization of `bitscores` (nGA) was performed as referenced in Equation 5.1.

$$\text{nGA} = \frac{\text{B}}{\text{GA}} \tag{5.1}$$

Where, B corresponds to reported `bitscore` from `cmsearch` result and GA is the provided gathering score from `Rfam`.

5.2.3 Mapping previous ncRNA annotation on new assembly

Previous ncRNA annotation was retrieved (Velandia-Huerto, Gittenberger, et al., 2016) in fasta format. All the contigs which reported a ncRNA annotation have been obtained from the reported draft assembly of the *D. vexillum* genome[9]. The resulting was mapped onto the new genome with `lastz`:

```
lastz_32 <NEW_GENOME>[multiple] <OLD_GENOME> --rdotplot=<OUT_DOT-PLOT_FILE>
    --ambiguous=iupac --chain C=0 E=150 H=0 K=4500 L=3000 M=254 O=600
    Q=human_chimp.v2.q T=2 Y=15000 --format=maf+ > <OUTPUT_FILE>
```

[6] https://www.ncbi.nlm.nih.gov/taxonomy
[7] http://mirgenedb.org/static/data/ALL/ALL--pri-30-30.fas
[8] http://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.cds.all.fa
[9] http://tunicata.bioinf.uni-leipzig.de/Download.html

Alignment files were retrieved in `maf` format and were parsed with `Bio::AlignIO Bioperl` library. The criteria to obtain the best genome coordinates was chosen based on the relation R_{mn} between the length of the mapped region into the new genome (m) and the original size of the query contig in the old genome (n). The relation was defined as noted in Equation 5.2:

$$R_{mn} = \frac{m}{n} \tag{5.2}$$

In this case, the best mapping candidates are those which reported $R_{mn} = 1$, but to retrieve the maximum number of mapping between the two genome versions, $R_{mn} \geq 0.90$ was also considered.

From 247 contigs, was possible to map 213 in the raw results after the mapping stage with *lastz*, which generated 153892 relations R_{mn}. After filter R_{mn}, resulted in: 1.09% ($R_{mn} = 1$), 1.37% ($0.95 \leq R_{mn} < 1$), 0.35% ($0.90 \leq R_{mn} < 0.95$) and 0.27% ($0.85 \leq R_{mn} < 0.90$), the remaining percentage of candidates (96.15%) reported $R_{mn} < 0.85$, which in this strategy were considered as low mapping score. In the other hand, 111 contigs reported at least one high mapping score ($R_{mn} \geq 0.85$).

At the same time, previously ncRNAs were obtained and mapped against the new *D. vexillum* assembly with `blastn`, as follows:

```
blastall -p blastn -d <DB> -i <QUERY> -F F -e 10e-5 -m 8 -o <OUT>
```

If one contig reported more than one candidate in the new genome, we chose the one with the highest `blastn bitscore`. After mapping all the candidates with `blastn`, the true locations were obtained after applying the following filters:

- Identity $\geq 85\%$.
- E-value $\leq 10^{-10}$.
- Size relation between homology region of query h_m and its calculated size h_n have to be $\frac{h_m}{h_n} \geq 0.9$.

An additional confirmation step was performed using the covariance models from `Rfam`v.14.1 onto the retrieved fasta sequences, using `infernal` package:

```
cmsearch -g -Z <NT number (Mb)> --toponly --tblout <OUT_TABULAR> -o
<OUT_FILE> <FASTA> <CM>
```

5.2.4 Annotation of *homeobox* proteins

A collection of reported *homeobox* proteins from human (of the family *Homeoboxes*, 516)[10], *C. robusta*, *C. savignyi*[11], *B. leachii* (Blanchoud et al., 2018), *H. roretzi* (Sekigami et al., 2017, 2019) and a variety of species from the `HomeoDB` (Zhong, Butts, and P. W. H. Holland, 2008) were retrieved from the corresponding references. This set was used to search along

[10] `https://www.genenames.org/cgi-bin/genegroup/download?id=516&type=branch` retrieved from `HGNC` database on October 10, 2019 (B. Yates et al., 2016)

[11] From `Ensembl v100` (A. D. Yates et al., 2019)

the annotated transcriptome and protein sequences from *D. vexillum* using `tblastn` and `blastp`, respectively. The best candidates were obtained with an identity percent of ≥35, E-value $\leq 10^{-5}$ and a query coverage of 70%.

As a complement, pairwise genome alignments with the new assembly from *D. vexillum* and close species that reported annotations of *homeobox* genes: *B. floridae*, *B. leachii*, *B. schlosseri*, *C. savignyi*, *C. robusta*, *H. roretzi* and *O. dioica*, were performed with `LASTZ` (Harris, 2007). References from *homeobox* genes were obtained from `Aniseed` (Brozovic et al., 2017) using the Gene Builder with the term *hox*, except from *B. floridae* where updated annotations (for v.2) were searched and retrieved from `LanceletDB` (You et al., 2019). Cross-matching of shared regions and reported genes and homology searches were performed to support the identification of *homeobox* candidates.

5.2.5 Detection of orthologous proteins involved in skeletogenesis

The RUNX, SOX, and Hh homologs were searched in the output of `eggNOG-Mapper` for all studied chordate species. The corresponding orthology groups have the accession numbers: `KOG3982`, `KOG0527` and `KOG3638`, respectively. Due to the lack of true RUNX orthologs on *D. vexillum*, we performed an additional analysis to confirm the presence of some homology signal. We retrieved the *RUNX* sequences reported on Hecht et al. (2008), from available 16 chordates from `NCBI`: AN08565.1, AAN08567.1, AAQ88389.1, AAS02047.1, AAS21356.1, BAA03485.1, BAF36001.1, BAF36011.1, EAX04278.1, EDL03777.1, EDL29993.1, ENSCINT00000004611.3, NP_001001890.1, NP_001092121.1, NP_004341.1 and NP_571678.1. Those sequences were searched with `blastp` in the proteome of *D. vexillum* and the following 10 species: *B. floridae*, *B. leachii*, *B. schlosseri*, *C. robusta*, *C. savignyi*, *M. oculata*, *M. occidentalis*, *O. dioica*, *P. marinus*, and *L. chalumnae*. On the other hand, the `PFAM` domain *Runt* (PF00853) was searched along all the reported proteomes of the described species using `hmmscan` (`HMMER` v.3.1b1) (S. R. Eddy, 2011). Filtering was based on the *gathering score* reported by `PFAM` and a low E-value <0.001.

RMST annotation

Ten RMST covariance models (RF01962-RF01971) were retrieved from `Rfam` v.14 with `cmfetch`. Using `cmsearch` on selected genomes of chordates (*B. floridae* (Brfl), *B. belcheri* (Brbe), *O. dioica* (Oidi), *M. occidentalis* (Mlis), *M. oculata* (Mata), *M. occulta* (Mlta), *B. schlosseri* (Bosc), *H. roretzi* (Haro), *S. thompsoni* (Sath), *B. leachii* (Bole), *D. vexillum* (Dive), *C. robusta* (Ciro), *C. savignyi* (Cisa), *P. marinus* (Pema), *D. rerio* (Dare), *L. chalumnae* (Lach), *M. musculus* (Mumu) and *H. sapiens* (Hosa)), echinoderms (*S. purpuratus* (Stpu) and *P. miniata* (Pami)) and hemichordata (*S. kowalevskii* (Sako)). True candidates were retrieved if reported an E-value $\leq 10^{-3}$ and the 32% of the GA.

5.2.6 Mitochondrial genes

Mitochondrial complete genome from isolated clade A (NC_026107) and isolated clade B (KM259617.1) of *D. vexillum* were retrieved from `GenBank` as reported by K. F. Smith, Abbott, et al. (2015). Both sets of sequences where mapped with `blastn` against the new *D. vexillum* genome. The best candidates were retrieved adjusting identity ≤95%, E-value

≤0.001 and coverage 100% cutoffs. Final coordinates files are available in `GFF3` format. Filtering of the intergenic coordinates was performed by a `Perl` script and this output was depicted with LuaTEX package `pgfmolbio`. Annotated Tunicata mitochondrial genomes were collected from `NCBI`. Multiple mitochondrial genome alignments were calculated using `progressiveMauve` (Darling, Mau, and Perna, 2010).

5.2.7 Genome Browser construction

`GFF3` annotation files for coding genes, ncRNAs and mtDNA were processed using `MakeHub` (Hoff, 2019) as preprocessing step to generate the input files of the hub. The input files were used to create a genome Hub hosted on the UCSC hub site (Raney et al., 2013).

5.3 Non-model assembly and annotation hypotheses

5.3.1 Generalities of *D. vexillum* genome

Despite the availability of a previous *D. vexillum* draft assembly, reported contiguity is not suitable to predict larger coding/non-coding genes, where > 80% contigs with 1 Kb length (Velandia-Huerto, Gittenberger, et al., 2016). To extend and improve those initial efforts, Parra-Rincón et al. (2021) performed a *de novo* assembly and annotation using an *hybrid* assembly: combining 28.5Gb of Illumina and 12.5Gb of PacBio data. In comparison to previous draft assembly, the contig length was increased $\sim 8\times$, the $N50 = 6539$nt was extended, the number of *scaffolds* (109,769) lowered, and reporting a genome size of 517.55 Mb, including the assembly of the mitochondrial genome (with 16.13Kb). At the same time, through a transcriptome assembly from Illumina paired-end (PE) reads, the annotation of 62,194 coding-genes were reported, with their corresponding 64,424 protein transcription products. In terms of non-coding genes, using a homology approach were detected 4877 loci, including numerous families from transfer RNA (tRNA) and miRNAs (see in detail Table 10).

In comparison to other metazoan assemblies reported on NCBI, *D. vexillum* reported a genome size and GC content expected for invertebrates species (Table 9). However, the accounted genes and proteins are higher, even when compared with mammals ($\sim 2\times$) and close species as *C. robusta* ($\sim 4.4\times$) or *B. schlosseri* ($\sim 1.6\times$). Those numbers suggest that current draft genome contains redundancies at scaffold or/and annotation level. In the following sections, the developed approach to detect and quantify them are described in more detail.

5.3.2 *D. vexillum* genome assembly: running out of road to get sequences

About 27 extractions of genomic material collected from the sea carpet squirt (*Didemnum vexillum*) resulted on a notorious DNA degradation, presumably due an action of acidic environment from the tunic bladder cells (restricted to some groups of ascidians, including the Didemnidae). The bulk of their cytoplasm comprises a large vacuole containing

Table 9: Comparison of *D. vexillum* assembly respect NCBI metazoan assemblies. *Other Animals* represents a collection of metazoan species that are not in the mentioned groups. **Bold** names are tunicate species.

Group	Assemblies	Size	GC	Genes	Proteins
Mammals	213	2478.97	41.53	30378	42753.0
Birds	371	1070.84	41.50	15613	14146.0
Reptiles	35	1901.84	43.77	23178	38712.0
Amphibians	15	3779.43	43.25	27589	41969.0
Fishes	179	848.83	41.00	28534	41965.0
Other Animals	154	367.19	34.50	27296	32086.5
Insects	275	280.20	36.87	15338	21191.0
Flatworms	33	539.42	40.70	13309	12795.0
Roundworms	73	78.27	35.10	14737	16380.0
C. robusta	Satou et al.	122.99	36.20	14072	61667.0
B. schlosseri	Voskoboynik et al.	580	41.00	38730	46519.0
D. vexillum	Parra-Rincón et al.	517.55	36.20	62194	64424.0

sulphuric acid, which accounts for a tunic pH < 3.0 in didemnids (Hirose, 2001) that may be involved in chemical defence.

This material were used to obtain long libraries of PacBio, obtaining about 5 millions of subreads, with an N50 = 2.3 Kbp (Parra-Rincón et al., 2021). Taking advantage of a previous draft genome assembled with Illumina paired end reads in Velandia-Huerto, Gittenberger, et al. (2016), a hybrid assembly approach combining both libraries were done. This strategy was devised given the difficulty to assembly the genome by standard methodologies, including short read only assembly (`ABySS` (Jackman et al., 2017)), hybrid assemblers (`DBG2OLC` (Ye et al., 2016), `Wengan` (Di Genova et al., 2020)), and `LazyB` (Gatter et al., 2020), and a long read only assembly method (`wtdbg2` (J. Ruan and H. Li, 2019)), which were able to assemble $\leq 20\%$ of the reported final assembly. Single-copy elements annotated on other genomes were found, called in average 2, and in some cases < 11 variants, which in general made up on average 41.5% of reads in each site (Parra-Rincón et al., 2021).

This evidenced the presence of multiple haplotypes on the sequenced data. For that reason, the hybrid approach included three rounds of *chimeric* corrections, such as: one run with PacBio subreads using the `SMRT` suite to reduce spurious contigs, a second one was calculated including the SMRT reads together with Illumina data (as implemented in `Proovread-2.13.13` (Hackl et al., 2014)) and finally, another one with the Celera assembler, using the parameter `doChimeraDetection`.

Comparisons to close homologs

To discover the existence of redundant annotations, *D. vexillum* genes were represented by their longest protein product and were compared to multiple annotated databases,

such as: *C. robusta* proteins, ortholog clusters from the `eggNOG` database (Huerta-Cepas et al., 2018), and non-redundant (nr) database from `NCBI`. Through this strategy, were discovered that $\sim$ 42% of annotated proteins and their associated genes have at least one close homolog on described databases. Most of them, showing homology relations to close metazoan proteins and/or *C. robusta* annotations (26,005).

In the other way, other 58% of proteins did not report close homologs, even when relaxed homology search parameters. Looking in detail the support to annotate those genes, 24,762 were not annotated using the available transcriptome data, but were annotated by the *ab initio* prediction models, using `Maker` (Cantarel et al., 2007). At the same time, from the same dataset, at least one UTR region is missing for 28,431 and $\sim$ 10.6% of those genes are annotated as isolated elements in their host scaffolds. Finally, 5977 protein-coding transcripts did not have a clear close homolog.

Evidence of gene fragmentation

In a detailed inspection of genes that reported close homologs in at least one of the query databases ($\sim$ 42% of *D. vexillum* annotation) led to a characterization of relations between the annotated gene (*model*) and its close homolog (represented as n_g) in *D. vexillum*. By this approach, 2482 *D. vexillum* genes with close homologs in *C. robusta*, served to categorize the nature of homology relations, as follows: $1 : n_g$, with $n_g > 1$ genes, or $1 : 1$. For $1 : n_g$ relations, the n_g *D. vexilum* genes could assemble a complete or incomplete coverage respect to their *C. robusta model*. In terms of *model* coverage, closer related genes can be aligned over a long overlapping region (*split*) or be composed by short overlappings that cover almost all the *model* (*fragmented*). In both cases, *D. vexillum* genes are found along multiple scaffolds. In other way, $1 : 1$ relations can display complete or incomplete coverage. Multiple examples of explained categories are depicted in Figure 28A and B, respectively.

In terms of redundancy, *fragmented* case contributes directly to additional annotations, given their span over multiple scaffolds but matching for the same annotated model (in this case *C. robusta*). Those classified as *split* could result a gene duplication event or product of an assembly artefact that duplicated the same region. Additional evidence is required to differentiate between both cases. Meanwhile, *Incomplete* ones could account for a real non-functional gene or an assembly-interrupted model, as seen in Figure 28B.

In detail, Figure 28 A depicts the $1 : n_g$ relations: the protein `KY.Chr14.999.v1.SL1-1`, was split into 4 mappings on *D. vexillum* over multiple scaffolds. At the same time, protein `KY.Chr1.580.v1.ND1-1`, is *fragmented* over 3 *D. vexillum* proteins, located in three scaffolds. The main difference is the way how mapped candidates assemble the original model. In Figure 28B, the protein `KY.Chr11.363.v1.SL1-1` was identified as a *1:1* example on *D. vexillum*, but incomplete.

5.3.3 Mapping previously detected non-coding RNAs

A multi-step methodology to get the correct mapping of previous candidates was designed to get a final candidate genome coordinates for each ncRNA element (see Section 5.2.3). In detail, some candidates reported more than one position on the genome or even, the same positions shared with another candidate(s). In this case, the final reported mapped

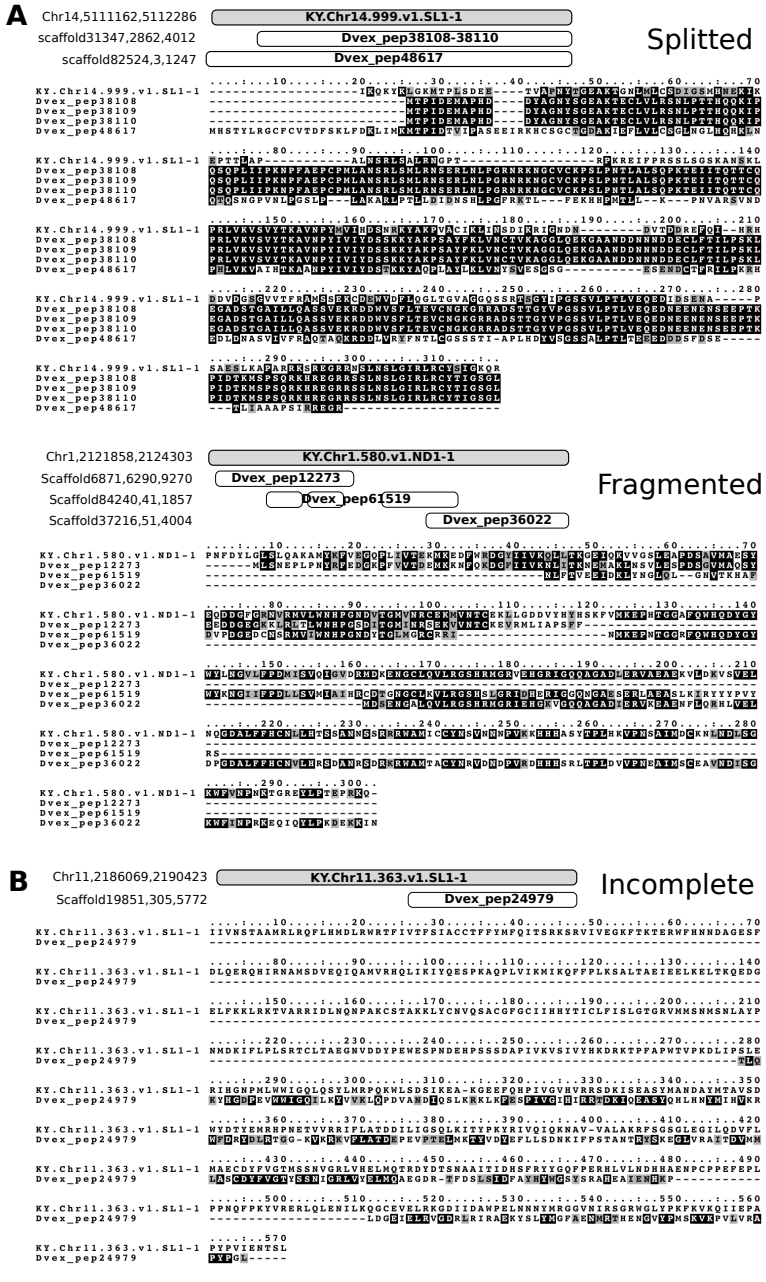


Figure 28: Multiple protein alignments showing found relations between *C. robusta* and *D. vexillum* proteins: **A.** Complete (Fragmented and Split) and **B.** Incomplete. Alignment representation for each case is represented, *Ciona* protein coloured in grey. Protein-coding gene coordinates are described on figure left side.

candidates were those that mapped 1 : 1 to the new assembly and does not share the same positions with an overlapping candidate. In this way, 77 loci were retrieved and had reported additional support from genome alignments, 36 have been identified on the new genome in another location that is different to the correspondent new region of the old contig. At the end it was possible to map 105 previously annotated candidates by this strategy which were included in the final set of candidates with the tag `MAPPED`. From those candidates, 8 reported an additional mapping position which were also included in the final results, due those candidates in the new assembly reported high homology scores, description of those families are in Appendix C: Table 18.

5.3.4 Annotation of conserved coding genes

Given this challenging genome as described above, additional support should be considered at the annotation of conserved genes. A combination of pairwise homology, multiple genome-wide alignments, comparisons to clustered datasets, and a final manual refinement was required to annotate the conserved *homeobox* genes and a comprehensive set of genes related to modulate skeletogenesis, as described in the following Sections: Homeobox transcription factors and Skeletogenesis proteins, respectively.

Homeobox transcription factors

In a preliminary scan, specifically searching for *homeobox* transcription factors, a combined `blastp/tblastn` strategy identified 48 coding sequences with their corresponding number of genes located in 47 scaffolds. The most frequent found proteins are homologs from the families: *ZEB2*, *LHX2* and *Irx* transcription factors. Additionally, a genome-wide alignments approach was used to compare existing annotations of homeobox genes in six tunicate and one cephalochordate genomes to the *D. vexillum* assembly. Only one of the 48 *homeobox loci* had annotated homologs in four of the six query species, which corresponds to a *Hox2* gene, located on the *scaffold16549-size8805*. Several other Hox genes, however, were not recognized by the default homology annotation pipeline because of incomplete overlaps, and in some cases, no gene was recognized for *D. vexillum* (Figure 29).

Taking into account the reported organization of the Hox genes in other tunicate genomes, as: *H. roretzi* and *Ciona* spp., it is expected to find three anterior, three middle-group, and three posterior Hox genes (Sekigami et al., 2017, 2019). Based on found candidates and a more detailed manual search with genome alignments as support, were found evidence for two anterior genes (*Hox2* and *Hox3*), two central genes (*Hox4* and *Hox6/7*-like), and the three expected posterior genes in *D. vexillum*, as referred on Figure 29E. The assembly of the HOX gene region unfortunately is too fragmented to conclusively rule out the presence of *Hox1* and *Hox5* or to provide any linkage information of the reported Hox genes.

Skeletogenesis proteins

Because the Didemnidae can mineralize calcium to form spicules in their tunics, it is worth to search for key proteins involved in skeletogenesis, as described in Wagner and Aspenberg (2011): *Sox*, *Hedgehog* (Hh), and *RUNX*, which corresponded to the ortholog groups:

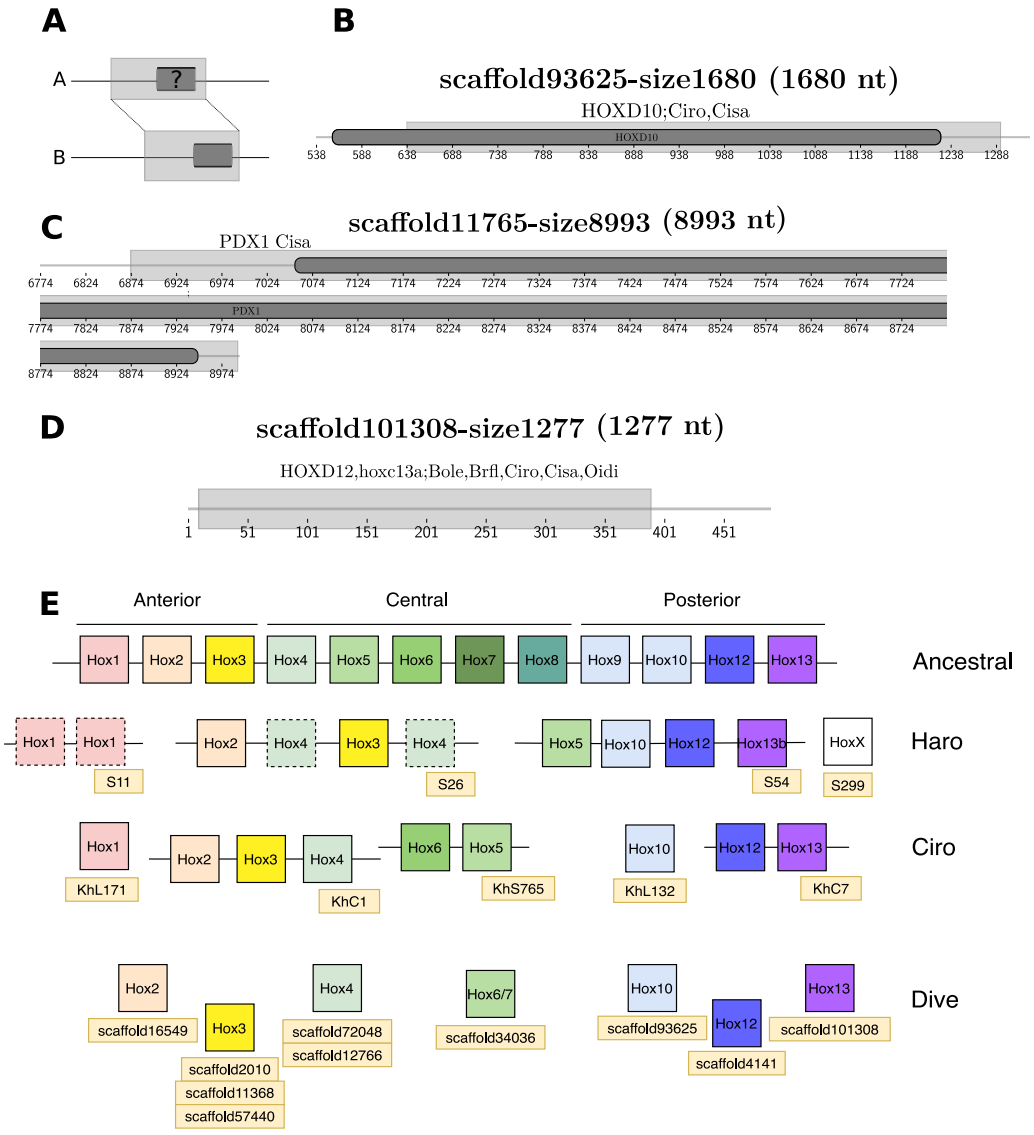


Figure 29: Detection of *Homeobox* genes on *D. vexillum*. **A.** Model of detection, a shared region between genomes *A* and *B* is detected and referenced as grey boxes. Correspondence is denoted by dotted lines between genomes. The dark grey box in genome *B* represents an annotated gene whereas the dark grey box mark represents the putative orthologous region. **B-D** show examples of putative orthologous Hox gene assignment in *D. vexillum*. Specific details are explained in the main text. **E** summarize the complete *Homeobox* genes annotation in *D. vexillum* (**Dive**) in comparison to reported genes on *C. robusta* (**Ciro**) and *H. roretzi* (**Haro**). Genomic locations were retrieved from `ANISEED`, Hox cluster of the chordate ancestor is depicted (Sekigami et al., 2017, 2019). Uncertain positions of some genes are represented as a dotted box, e.g. *Hox1* and *Hox4* in *H. roretzi*. For specific genome coordinates see Appendix C: Table 19.

KOG0527 (SOX), *KOG3638* (Hh), and *KOG3982* (RUNX) on the `eggNOG` database. Gene phylogenies for these ortholog groups (including the chordate sequences used as reference and the orthologs annotated in the `eggNOG` database) are shown in Figure 30. In *D. vexillum*, were found seven members of the SOX family belonging to SoxB1, SoxB2, SoxC, SoxD and SoxE subgroups as defined in Guth and Wegner (2008). Overall, two paralogs for the SOXC (*SOX4/SoxC#*32 and *SOX4/SoxC#*33) and SoxB2 (*SOX14/SoxB2#*5 and *SOX14/SoxB2#*6) were found in the *D. vexillum* annotation, see Figure 30A and Appendix C: Figure 50 for the complete tree.

All tunicates except *O. dioica* reported members of the Hh families (Figure 30B). The basal Hh family, previously reported in Ciona (Takatori, Satou, and Satoh, 2002) and in amphioxus Shimeld (1999), was detected in all ascidians. In the vertebrates, were confirmed the presence of the three Hh genes: Desert (DHh), Indian (IHh) and Sonic-hedgehog (SHh) (Ingham and McMahon, 2001; Shimeld, 1999). In ascidians, several clades of Hh genes were found. There are at least three Hh families in the ascidians: Hh clade A (with medium bootstrap support of 61), Hh clade B (with full bootstrap support in *Ciona*) and Hh clade C (with full bootstrap support in the botryllids). The *D. vexillum* Hh does not group with any of the other clades. The resulted analysis supports an independent diversification of the Hh family in ascidians.

The key regulators of skeletogenesis RUNX-related transcription factor (RUNX) proteins were not found in *D. vexillum*. This does not necessarily indicate a true loss, however, because in a detailed domain-based homology search, were found parts of the *Runt* domain (PF00853) among 15 proteins from *D. vexillum*, albeit with truncated sequences. The phylogenetic distribution of the orthologs found (Appendix C: Figure 49), shows a defined clade of tunicate sequences that belong to the ancestral RUNX family, which has been detected in this study in amphioxus and is known to be expressed in *Ciona* and *Oikopleura* (Nah et al., 2014). This suggests that RUNX proteins may not be truly absent in *D. vexillum*. Was observed in passing that the RUNX family has undergone additional duplications in the lampreys (Appendix C: Figure 49).

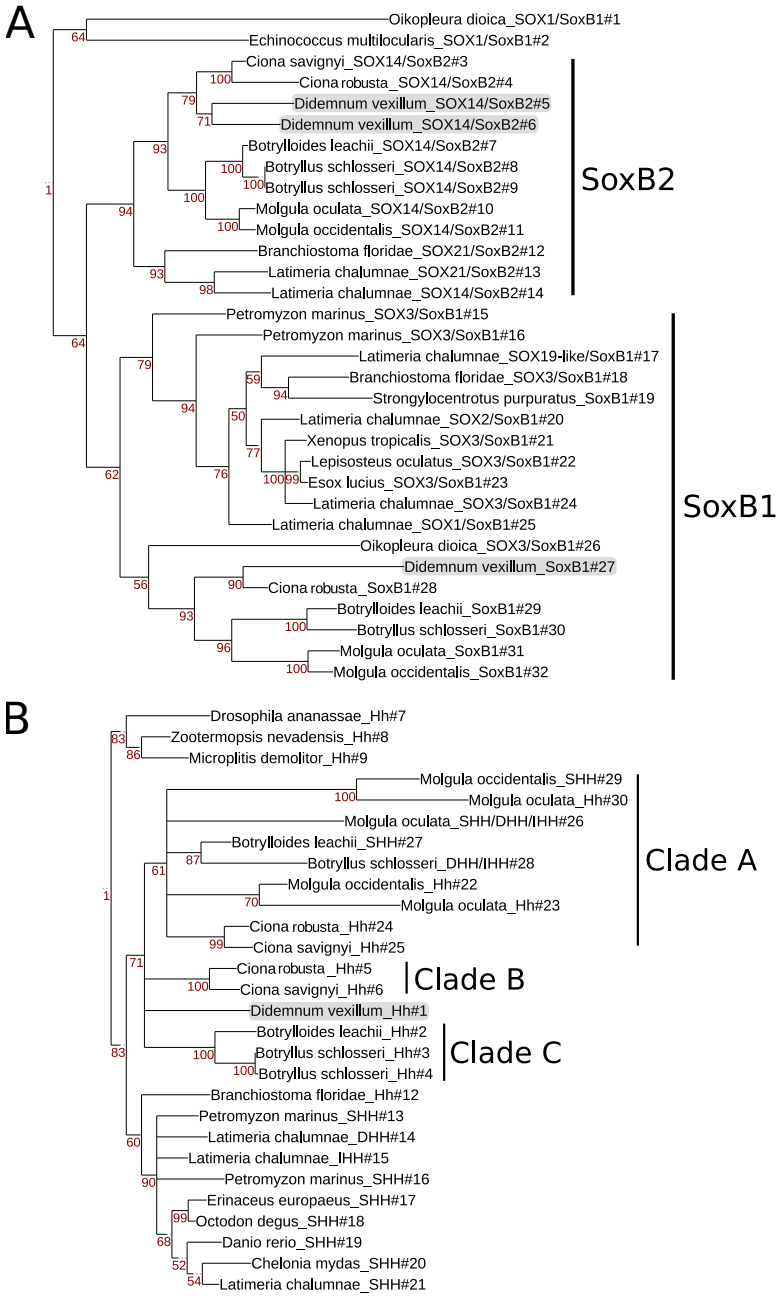


Figure 30: Phylogenetic analysis of skeletogenesis proteins found in *D. vexillum*. **A.** SoxB1/B2 family, **B** Hh family. The sea vomit is highlighted in grey. A tree of the complete SOX family can be found in Appendix C: Figure 50. Trees were built using Maximum Likelihood (ML) with the JTT+G+I substitution model generating 100 bootstrap replicates.

5.3.5 Annotation of Non-coding RNAs

A different approach respect to coding genes, was performed to annotate non coding RNAs (ncRNAs). In this way, sequence patterns and secondary structure comparisons were accessed via homology-based methods, combining `blastn` searches, Hidden Markov Model (HMM) profiles, and Covariance Models (CMs) as described in Velandia-Huerto, Gittenberger, et al. (2016) with some modifications, as detailed in Methods. Not counting transfer RNAs (tRNAs), were identified 2153 ncRNA *loci* corresponding to 271 distinct families. A search with `tRNAscan-SE` (Chan and Lowe, 2019) resulted in 18,343 tRNA predicted loci, including pseudogenes and undetermined isotype candidates. In addition, were mapped the 206 families of ncRNAs identified in a preliminary draft of the *D. vexillum* genome Velandia-Huerto, Gittenberger, et al. (2016) to new assembly (see Section 5.2.3). As in other genomes, in particular the pol-III transcribed RNAs including 5S ribosomal RNA (rRNA), tRNAs, and U6 RNA, as well as the small nuclear RNAs (snRNAs) transcribed by pol-II appear in multiples copies (Marz, Kirsten, and P. F. Stadler, 2008). The data are summarized in Table 10 and described in detail in the following Sections: House-keeping ncRNA families, A conserved long non-coding RNA and microRNA complement.

Table 10: Annotated ncRNAs families and *loci* (in parentheses) in the *D. vexillum* genome. *Homology* corresponds to previously reported numbers of ncRNAs by homology (Velandia-Huerto, Gittenberger, et al., 2016), *Mapped* corresponds to the number of ncRNAs that were mapped in the first genome draft (Velandia-Huerto, Gittenberger, et al., 2016). *Final* corresponds to the current list of candidate ncRNAs. *NA:* Not available.

ncRNA Family	Homology	Mapped	Final
Cis-Reg	3 (333)	0	3 (333)
microRNAs (miRNAs)	248 (2065)	17 (20)	235 (1582)
misc RNAs	1 (1)	1 (1)	2 (2)
lncRNAs	2 (8)	0	2 (8)
Ribozyme	3 (11)	0	3 (11)
rRNAs	4 (84)	0	4 (84)
snoRNAs	6 (9)	6 (9)	12 (18)
snRNAs	9 (87)	2 (34)	9 (115)
tRNAs	23 (2724)	NA	23 (2724)
mt-tRNAs	0	21	21
mt-rRNAs	0	2	2
Total	277 (5322)	26 (64)	271 (4877)

House-keeping ncRNA families

Here are grouped ncRNA families that are well conserved through metazoan species: tRNAs, rRNAs, snRNAs and small nucleolar RNAs (snoRNAs). Usually, those families

are identified using pairwise comparisons or automated searches, easily detectable due high conservation patterns at sequence and secondary structure, and early divergence detected over ancestral species.

Transfer RNAs A number of 2724 tRNAs and 15,619 tRNA pseudogenes or with undetermined isotype (23) were found. The most abundant tRNA is $tRNA_{Thr}$ with 1395 copies, while only a single copy of $tRNA_{SeC}$ was observed. Surprisingly, `tRNAscan-SE` reported numerous suppressor tRNAs: 153 ($tRNA_{Suppressor\text{-}TCA}$: 145, $tRNA_{Suppressor\text{-}TTA}$: 7, and $tRNA_{Suppressor\text{-}CTA}$: 1).

Ribosomal RNAs As in most eukaryotes, the small and large subunit (SSU 18S and LSU 28S) rRNAs are organized in repetitive units of the rRNA operon (see in Dyomin et al. (2016) conserved examples in chicken). It also contains the 5.8S rRNAs. In this case, *D. vexillum* reported 6 clusters of rRNAs: two clusters are composed of repetitions of 5S rRNA (*scaffold1545-size16374* and *scaffold22447-size6833*), two clusters contain small subunit ribosomal RNA (SSU) 18S, 5.8S, and large subunit ribosomal RNA (LSU) 28S rRNA elements within (*scaffold4839-size12187* and *scaffold9164-size12300*) see Figure 31, and one cluster contains repetitions of 5.8 rRNAs with a locus of LSU 28S rRNA (*scaffold4349-size12561*). At the same time for the subunit 5S rRNA 71 loci were detected and from them 52 are located on the *scaffoldUncertain*. For the other rRNAs elements, in total were found 6 5.8S, 3 SSU, and 4 LSU rRNAs.

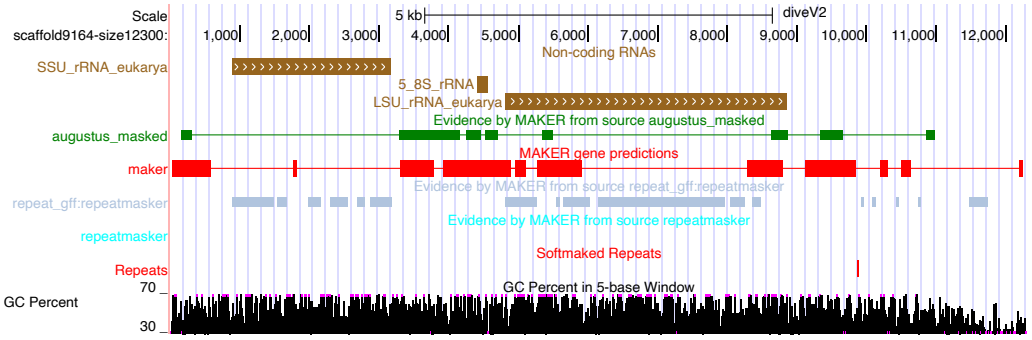


Figure 31: rRNA cluster in *scaffold9164-size12300* with SSU, 5.8S and LSU rRNA, colored as gold regions. This cluster resembled the same conserved architecture reported by Dyomin et al. (2016).

Small Nucleolar RNAs. Conserved snoRNA families were detected: 3 U3, 2 copies for SNORD14, SNORD18, snoZ39, and SNORA36, as well as a single copy of SNORD29, SNORD33, SNORD35, SNORD36, SNORD52, SNORD63, and SNORD83.

Spliceosomal RNAs All RNA components of the spliceosome machinery were found in the new genome assembly. As usual, the snRNAs of the major spliceosome appear in multiple copies U6 (46), U5 (9), U1 (21), U2 (27), U4 (3). Among the snRNAs of the minor spliceosome, U12 appear once, while there are 2 loci coding for U4atac, U6atac, and 4 U11 genes.

Other small nuclear RNAs Not all snRNAs are located by a simple homology strategy, others like the expected genes for the RNA component of the signal recognition particle as well as the RNase P RNA, RNase MRP RNA, and 7SK RNA, are more

challenging. These groups are notoriously difficult to be detected by homology search without the benefit of known homologs in closely related species (Menzel, Gorodkin, and P. F. Stadler, 2009). For tunicates, no homologs were found for the telomerase RNA, U7 snRNA, and Y RNAs, although their presence in the genome is expected. A thorough search along reported Tunicata genomes successfully reported vault snRNA *loci*, except for *D. vexillum* (see Table 11 for details).

Table 11: Presence/absence of housekeeping snRNA candidates on tunicate genomes. Tags are used to report the values after evaluation by `cmsearch`: **B**: bitscore, **E**: E-value and **N**:Number of true candidates. Average of bitscore or E-value are shown if more than one candidate was found.

Species	**snRNAs**			
	U7	**vault**	**Y**	**Telomerase**
B. leachii	NA	B:57.6, E:3^{-10}, N:3	NA	NA
B. schlosseri	NA	B:46.9, E:9.8^{-7}, N:1	NA	NA
C. robusta	NA	B:48.6, E:7^{-8}, N:3	NA	NA
C. savignyi	NA	B:47.1, E:3.49^{-8}, N:4	NA	NA
H. roretzi	NA	B:57.85, E:1.65^{-10}, N:2	NA	NA
M. oculata	NA	B:54.9, E:1.75^{-9}, N:4	NA	NA
M. occulta	NA	B:59.8, E:8.30^{-11}, N:3	NA	NA
M. occidentalis	NA	B:59.8, E:3.05^{-8}, N:4	NA	NA
O. dioica	NA	NA	NA	NA
S. thompsoni	NA	B:44.7, E:3^{-8}, N:7	NA	NA
D. vexillum	NA	NA	NA	NA

A conserved long non-coding RNA

RMST lncRNA: extended annotation along deuterostomes

According to L. Wang et al. (2016) multiple long non-coding RNAs (lncRNAs) involved in nervous system mechanisms are annotated without specific functionality, for example the conserved family *Rhabdomyosarcoma 2-associated transcript conserved region* (RMST). Confirmed conservation of complete RMST families is depicted in Figure 32. There, human RMST cover a region of about 100 kb in chromosome 12. Predicted RMST families using family specific CM are contained in this region, matching with previous annotations that are high conserved in selected species of primates, mouse and rat. A reduced RMST families are conserved in lizard, clawed frog and coelacanth. The number of conserved RMST families are dramatically reduced in zebrafish and lamprey.

Extending the same approach to tunicate genomes, the homology searches with suggested `Rfam` threshold scores, detected the most conserved RMST families in all selected species but not in tunicates (Figure 33A). Relaxing those scores and considering the *E*-value distribution, two structured lncRNAs were found in this clade: the RMST8 and RMST9, the latter one has already been previously annotated by (Velandia-Huerto, Gittenberger, et al., 2016). As a result of the iteration and re-building of the correspondent CM with newly detected tunicate sequences, the occurrence of the complete RMST family

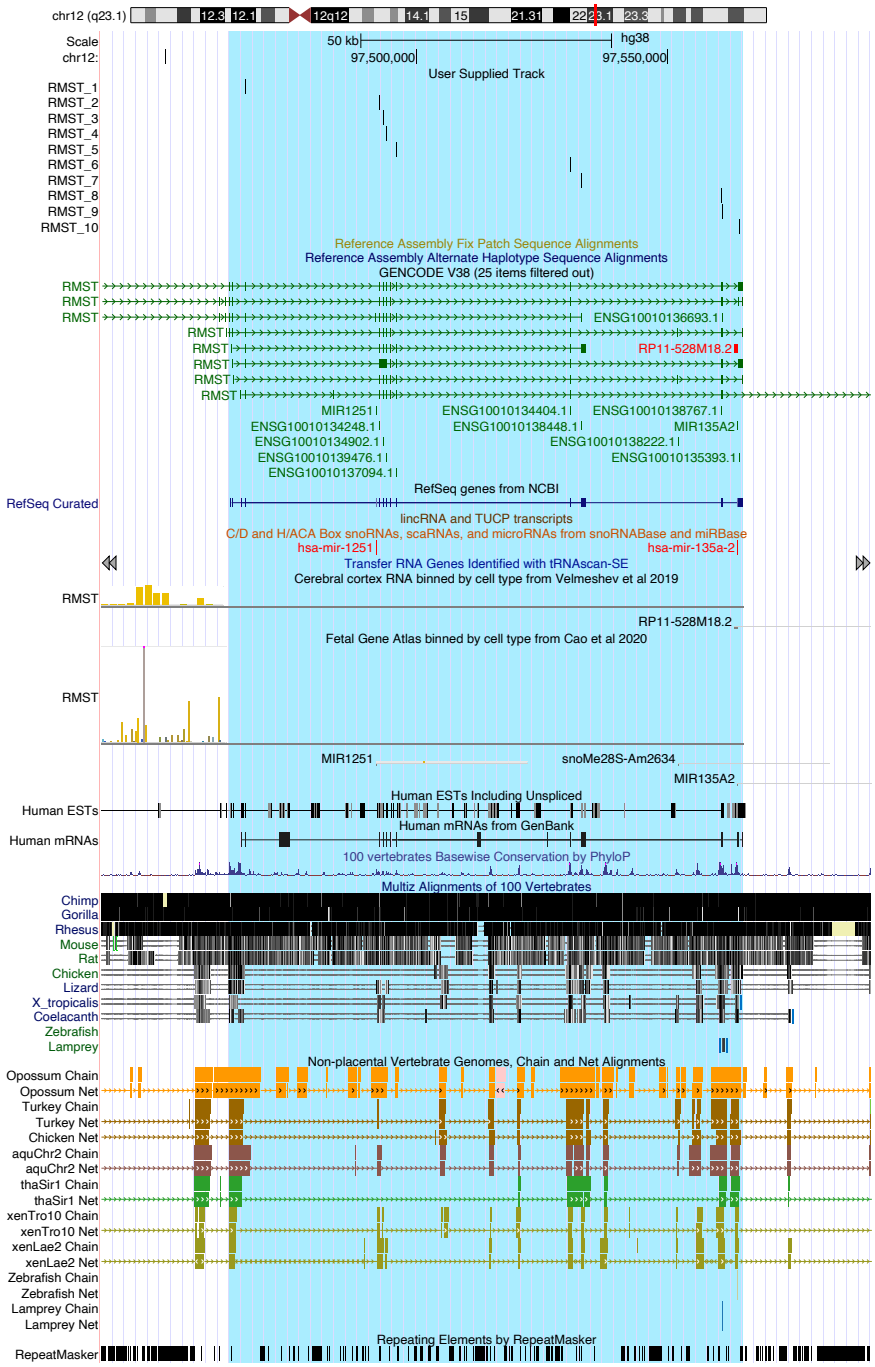


Figure 32: Conservation of lncRNA RMST and current annotated families. The conservation level of the complete transcript is depicted by Multiz alignments from selected vertebrates and chained alignments, showing blocks of conservation. Computational prediction is depicted in the upper track.

in deuterostomes is shown in Figure 33B. RMST 8 and 9 were detected in all deuterostomes. Were found two additional RMST families (RMST 6 and 7) in the coelacanth suggesting an initial expansion in the ancestor of lobe finned fishes (Sarcopterygii). The complete set of RMST 1, 2, 3, 4, 5, 6, 7, and 10 were detected in mammals. Because of their relevance in neural development (Chodroff et al., 2010; Ng et al., 2013), it would be interesting to study the evolution of RMSTs in the tetrapods, and the ancestral role of RMST 8 and 9 in the deuterostomes, the tetrapods and mammals.

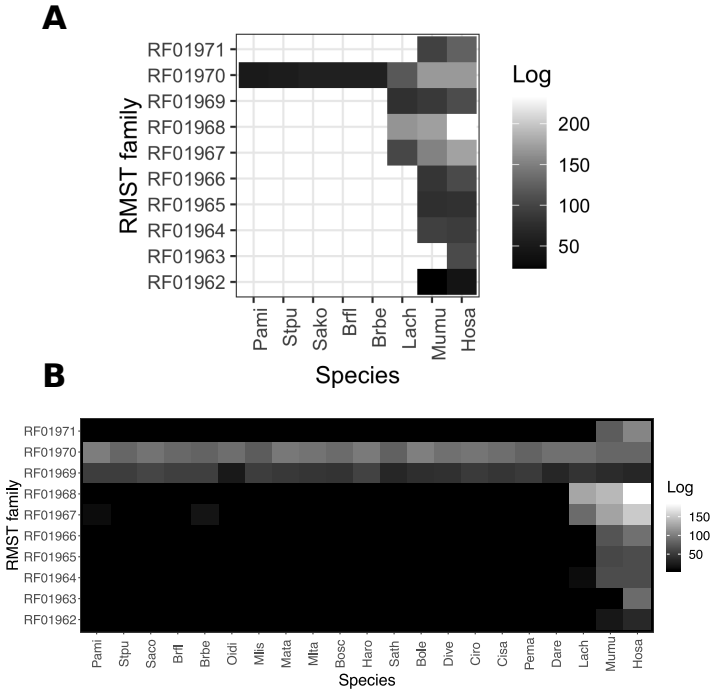


Figure 33: Evolution of the RMST lncRNA. **A.** Suggested RMST annotation by `Rfam` (Kalvari, Nawrocki, Argasinska, et al., 2018). **B.** Expanded annotation including all candidates. Gray intensity represents the $\max(-\log_2 E)$, with $E = E$-value, compared by specie and RMST family. Not available candidates are represented without colour. **Rfam accessions:** RMST1 (RF01962), 2(RF01963), 3 (RF01964), 4 (RF01965), 5 (RF01966), 6 (RF01967), 7 (RF01968), 8 (RF01969), 9 (RF01970) and 10 (RF01971). Species tags are described on Section 5.2.5.

5.3.6 microRNA complement

A critical assessment of miRNA-CMs threshold values

An initial microRNA (miRNA) annotation approach on *D. vexillum* was performed by Velandia-Huerto, Gittenberger, et al. (2016). This strategy made use of `Rfam` CMs with

pre-defined threshold scores, required to distinguish between true or false candidates. In order to test the correctness of this idea, the miRNA CMs from `Rfam` were used to perform a re-validation of 37 miRNAs reported for *Halocynthia roretzi* by K. Wang et al. (2017). Eleven miRNA loci were annotated, 9 previously reported in K. Wang et al. (2017) from the families let-7, miR-33, miR-124, miR-133 and miR-219, and two additional from miR-10. A manual inspection on the results, led to determine that missing candidates were actually detected, but filtered due their bitscore is lower than suggested family bitscore threshold, gathering score (GA), calculated by the `Rfam` (Kalvari, Nawrocki, Argasinska, et al., 2018).

To extend this evaluation, 4 experimental treatments were set up including two controls: a positive one composed by precursor sequences extracted from `MirGeneDB` (*true*) (Fromm, Billipp, et al., 2015) and other negative, constituted by swapped sequences derived from human CDS (*rand*). Two additional experiments were included: sequence homology candidates found in *C. robusta* (*exp*) and previous described sequences from *H. roretzi* (*haro*). The reported normalized bitscore ($n_{bitscore}$) density distribution for all the proposed treatments is described in Figure 34. The intercept in (nGA = 1.0) specify the normalized threshold value nGA that had been selected on Velandia-Huerto, Gittenberger, et al. (2016) as filter as well as by the `Rfam`.

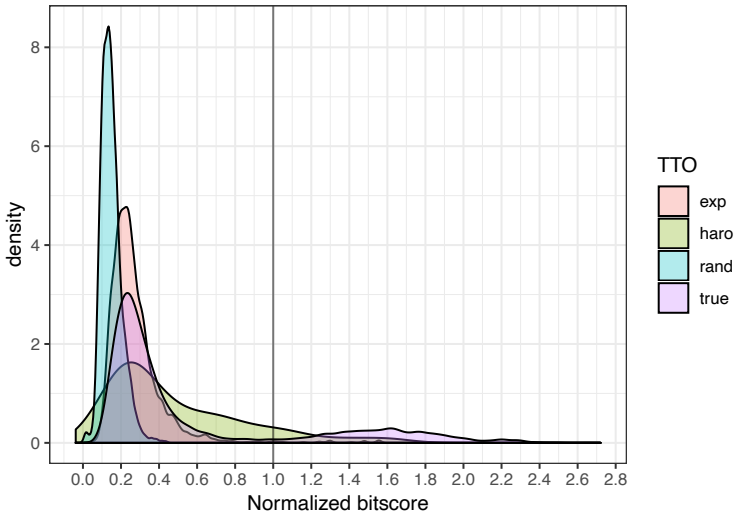


Figure 34: Density distribution of normalized bitscore ($n_{bitscore}$). **exp** corresponds to results from structural alignments on *C. robusta* candidates with the `blastn` number 4, described in this study, it constitutes an 'external' group of candidates, **haro** represents evaluation of annotated candidates on *H. roretzi* genome (K. Wang et al., 2017), **rand** is the generated negative control and **true** the results from sequence retrieved from `MirGeneDB` (Fromm, Billipp, et al., 2015).

Interestingly, a lot of candidates annotated in the *true* dataset reported $n_{bitscore} < nGA$, showing a multimodal distribution with one peak in ~ 0.25 and ~ 1.835. Specifically for this positive control group, Figure 35A, the distribution density of *true* $n_{bitscore}$ shows two groups when considered the family classification given by the CM. The candidates

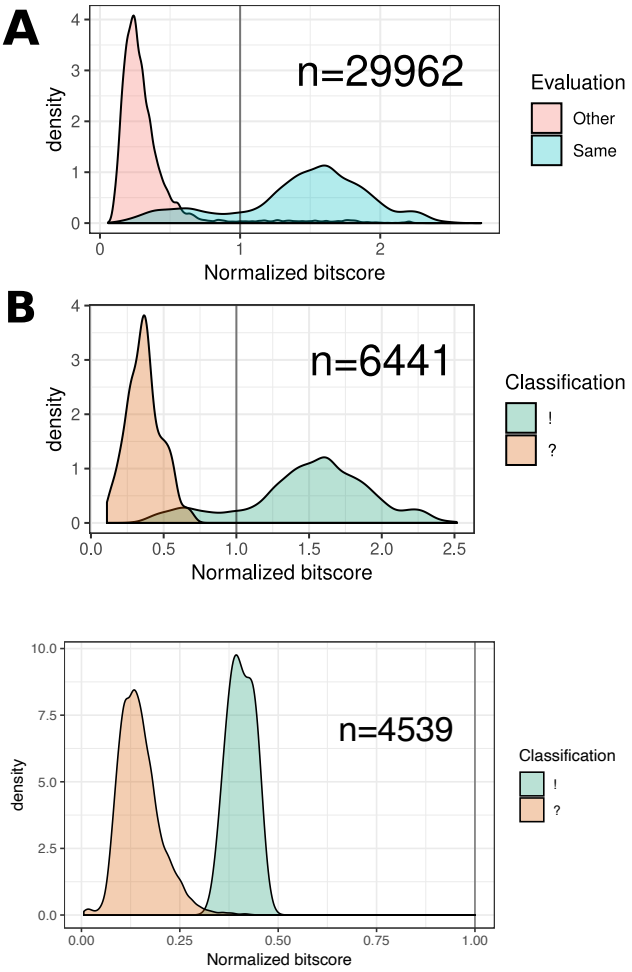


Figure 35: Density distribution of Control Positive sequences. **A.** Depicts the distribution of $n_{bitscore}$from control positive sequences. Previous structural evaluation in this control set was possible to know the annotated miRNA family, in this case 'Same' corresponds to the evaluation with the same CM with the annotated family. From final 29,962 evaluated results. 6441 were evaluated with the same miRNA CM. The remaining 23,521 miss this validation. **B.** From the 'Same' group, the colours split the data in two groups corresponding to the `cmsearch` classification, based on the default 'inclusion threshold' (E-value ≤ 0.01).

Figure 36: Density distribution of Negative control sequences. Classification criteria that defined '!' and '?' groups are defined by the E-value threshold suggested by `Rfam`.

are classified as 'Same' when the family was correctly predicted or 'Other' the opposite case. Again, it is noted that even families classified as miRNAs by the same CM family, reported $n_{bitscore} < nGA$. Additionally, considering the suggested criteria from `infernal` based on the *E*-value threshold, splits the 'Same' candidates into two groups: '!' with *E*-value≤ 0.01 or '?' *E*-value> 0.01. As shown in Figure 35B, a subset of *true* candidates reported a range of $n_{bitscore} = (0.298, 2.516)$, suggesting that $nGA \geq 0.29$.

The negative control reported a leptokurtic ($kurtosis = 5.24$) and positive skewed ($skewness = 1.054$) distribution, with an unique peak about ~ 0.15 and a maximum in 0.438 (data not shown). When considered a *E*-value discrimination criteria as before, two different distributions are evident: candidates that reported an $E - value \leq 0.01$ (4539) distribute with a $\mu = 0.1492 \pm 0.054$. Few candidates (5), reported greater E-values, but those candidates did not exceed the default threshold value nGA, as shown in Figure 36.

Based on the last results, the nGA threshold required to be modified in order to

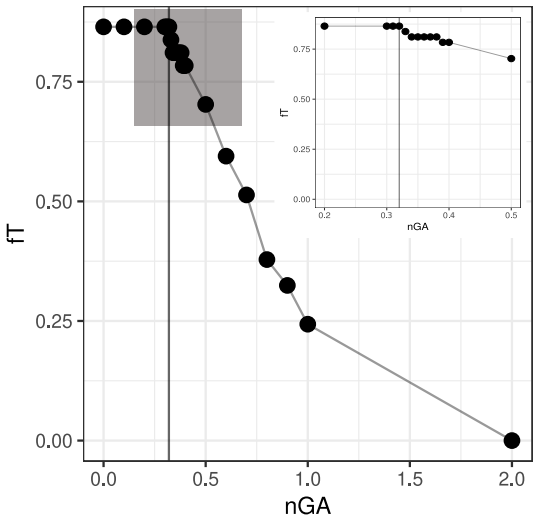


Figure 37: Evaluation of nGA on reported candidates from *H. roretzi* (K. Wang et al., 2017). fT corresponds to the absolute frequency of the classified true candidates with the same annotated miRNA family and covariance model. nGA represents the threshold nGA value applied to classify the candidates. The frequency was calculated on all the reported conserved candidates from *H. roretzi* (n=37). From this set, 4 families does not have a correspondent name on Rfam miRNA families (miR-3876, miR-3182, miR-3598 and miR-1502). Selected new threshold (nGA = 0.32) is depicted as an intercept on x axis.

consider a broader range of possible true candidates in a classification process, specifically for miRNAs. In this way, it is required that new candidates should be fitted by both, a lower $E - value \leq 0.01$ and a nGA greater than true negative control distribution as shown in Figure 36 as '?'. According to the distribution of this negative control data set, the true negative ones reported a confidence Interval ($CI = 0.012$, $\alpha = 0.05$, $\mu = 0.367$) and in general, the Control Negative ($CI = 0.002$, $\alpha = 0.05$, $\mu = 0.149$). Assuming that the reported miRNA on *H. roretzi* are effectively true candidates, the value of nGA was re-defined in comparison to the absolute frequency of true candidates. The threshold was defined as nGA = 0.32, based on the negative control distribution and the number of successfully annotated candidates (Figure 37).

miRNA annotation on *D. vexillum*

The miRNA annotation pipeline, described in the Section 5.2.1, identified 2065 *loci* encoding members of 248 distinct miRNA families. An additional 20 *loci*, which harbour two additional families, correspond to previously reported miRNAs (Velandia-Huerto, Gittenberger, et al., 2016) which successfully mapped into the new assembly. To avoid the annotation of false positives due to the modification of the threshold values (see Figure 34), the position of the *mature* sequence was evaluated using `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019) using both, the `Rfam` database for the miRNA families alignments and `miRBase` as source for the annotated *mature* sequences (as explained in more detail in Chapter 3:Figure 11). As a result, the definition of a true miRNA candidate relies not only on the homology results given by the sequence/secondary structure comparison, but also in the annotation of their *mature* sequence(s). In addition, a conserved position of the mature products within the defined miRNA family was required as additional evaluation step. To this end, candidates that reported homologous *mature* regions were compared against their original corrected `stockholm` alignments, by the calculation of the *tree edit*

distance ($e_{distance}$) between generated consensus secondary structures, as described on Section 5.2.1.

Due to the last classification, the final list of homologous miRNA on *D. vexillum* were categorized based on the secondary structure variation respect to the correspondent stockholm alignment family as follows: *high* ($e_{distance} = 0, 2$), *medium* ($e_{distance} = 3, 5$), and *low* ($e_{distance} = 6, 7$) and *no fitting* ($e_{distance} > 7$). As an example, Figure 38 shows examples for each category to illustrate the conservation degree respect to the current corrected multiple alignments from `Rfam`.

Following this way, a number of 1582 *loci* were found, from which 1394 fulfilled all the designed filters and reported a set of *mature* sequences harboured at the predicted hairpin structure. The other 188 have broken the conservation block in the defined family alignment, despite having shown a high conservation at hairpin level. Taking into account those detected miRNAs with *mature* annotation, the distribution of *loci* shows that 75% of miRNA families have less than 6 *loci*. The corresponding 25% of miRNA families have a higher median of ∼11.5 *loci*. Within these miRNA families, miR-544 (65), miR-578 (70), and miR-944 (97), had the highest number of *loci*.

The phylogenetic distribution of the miRNAs in the `Rfam` *seed* alignment, were retrieved along with their annotated *kingdom*, *phylum* and *subphylum*. The annotated miRNA families and their *loci* in *D. vexillum* were compared as shown in Appendix: C: Figure 47. Were found 18 miRNA families shared in more than 2 *phyla*: **mir-124**, mir-598, miR-7, **let-7**, **miR-1, miR-133**, miR-33, *lin-4*, **miR-137**, **miR-153**, miR-2, **miR-31**, miR-449, **miR-183**, **miR-190, miR-210, miR-219**, and **miR-8**. Families highlighted in bold showed a conserved structure (panel labelled as *VALID_STR*), even when the *D. vexillum* sequences were included into the alignment. In this analysis, were uncovered two potential additional tunicate-specific families: **ciona-mir-92** (RF01117) and **mir-281** (RF00967) to the previously reported **mir-1497** (RF00953) (K. Wang et al., 2017). In contrast, a subset of 13 miRNA family candidates did not fit into the corrected `stockholm` alignment (classified as *NO_VALID_STR*), despite a previous homology validation.

In a previous study of the miRNA complement in the solitary species *H. roretzi* (K. Wang et al., 2017) a more extensive list of tunicate-specific miRNAs was reported (21). From these list, only **miR-1497** (RF00953) was detected because the availability of corresponding CM. In *D. vexillum* a 21 from the 25 *conserved* miRNA families in metazoans were identified. Other families, including **mir-9**, **mir-182**, **mir-184**, **mir-200**, and **mir-218**, were not found. These families (except **mir-200**) were also found to be absent in other tunicates such as *C. savignyi* and *O. dioica* (K. Wang et al., 2017).

From previously reported set of miRNAs in Velandia-Huerto, Gittenberger, et al. (2016), 16 families were detected only in *D. vexillum* and not in other tunicates. From this set, 10 families were annotated in the new assembly and 4 were discarded because their mature sequences could not be annotated (**mir-130**, **mir-460**, **mir-185**, and **mir-233**), does not have a CM (**mir-4068**), and another was not found in the new assembly (**mir-9**). From the set of shared families in colonial tunicates, all were annotated and validated by this strategy, except **mir-340** (RF00761). The latter showed a good homology but did not pass the conditions of the current structural alignment, which used only vertebrate sequences to assign homology. In this study, **mir-31** was reported as the sole miRNA candidate that passed all filtering criteria to be exclusively found in solitary ascidian species. At the same time, were excluded 502 candidates based on their lack of conserved

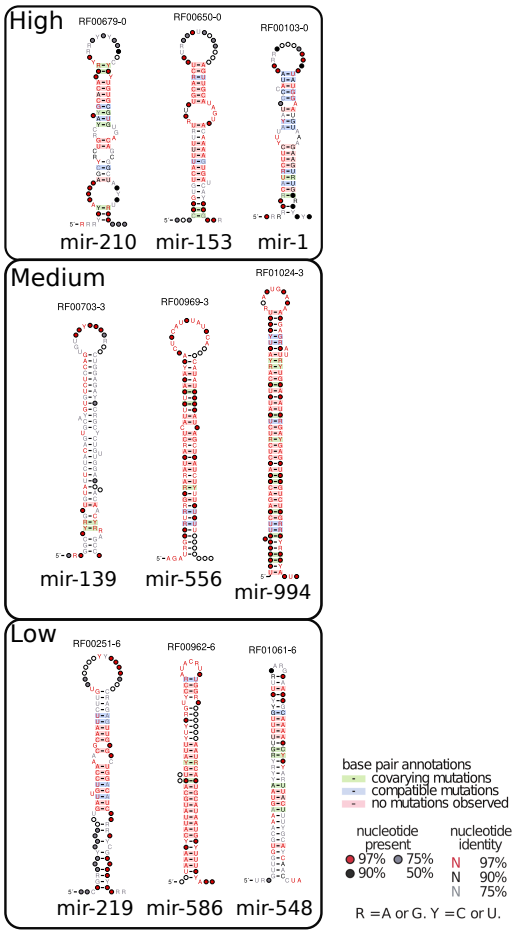


Figure 38: Examples of miRNA candidates detected and validated on *D. vexillum*. Labels (*High*, *Medium* and *Low*) refer to the *tree edit distance* ($e_{distance}$) calculated for the reported secondary structure from reported `Rfam` alignments and the same structure including the detected miRNA of the same family as explained on Chapter 3:Figure 11. Secondary structures were generated with `R2R` (Weinberg and Breaker, 2011), which provides annotations for sequence and structure conservation, as detailed in the legend.

mature sequences inside annotated precursors.

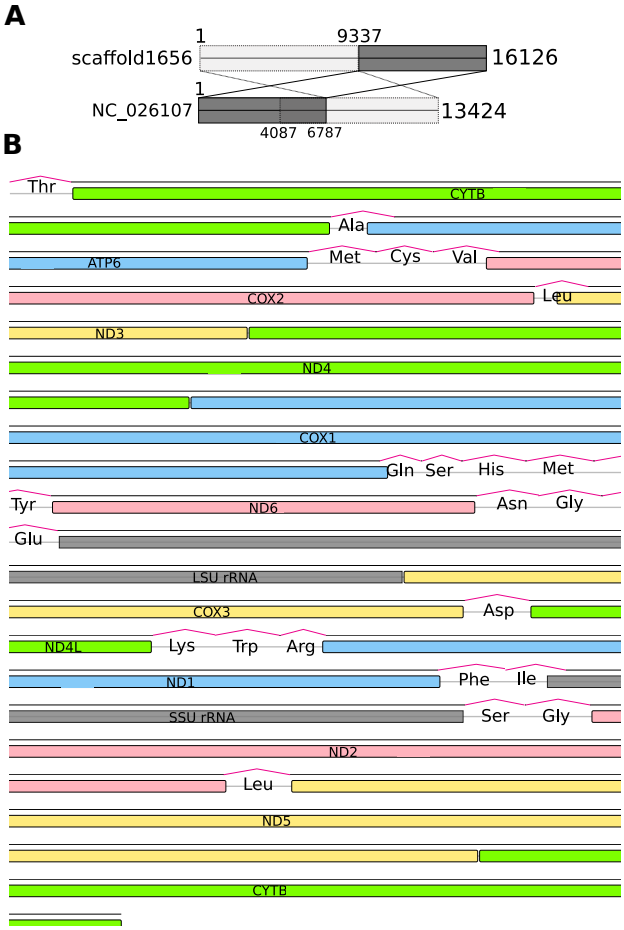


Figure 39: Mitochondrial genome from *D. vexillum*. **A.** Pairwise alignment between the newly assembled mtDNA (located on scaffold scaffold1656) and the reported mtDNA (Accession number:NC_026107). **B.** Distribution of reported sequences on the newly assembled mtDNA. For practical means, inter-genic regions were not considered. Genes sizes and order is shown, including the tRNA elements (red bends) and rRNA (gray boxes).

5.3.7 Annotation of mithochondrial DNA

The mitochondrial genome of *D. vexillum* maps to a single scaffold *scaffold1656-size16126* and very closely matches the two previously reported mitogenomic sequences (K. F. Smith, Abbott, et al., 2015), known as Clade A and Clade B. The mt-LSU is 99.9% identical to Clade A, and diverges about 3.6% from Clade B, confirming that the collected organisms belongs to clade A, see also Velandia-Huerto, Gittenberger, et al. (2016). Mapping the currently reported elements from mtDNA, resulted in the gene order depicted on Figure 39. In this case, intergenic distances were reduced, but the size and the order of the genes in the new assembly were conserved. The 37 expected elements of mtDNA were mapped to the new assembly. The gene order of the mitogenome matches that of clade A, but differs from other tunicate species, as shown in the multiple alignment of the mitogenomes in Appendix C: Figure 48.

5.4 Discussion

The genome assembly reported here pertains to a specimen of *Didemnum vexillum* Clade A, determined by the mt-LSU RNA. *D. vexillum* has a similar genome size and GC content as other deuterostome genomes, including ten tunicate genomes. Among tunicates, solitary organisms appear to have smaller genomes ($\leq$250 Mb) than colonial ones (with range from 160 to 723 Mb). The *D. vexillum* genome thus appears in the typical size range for colonial tunicates, and in terms of its size, it is comparable to the amphioxus genome (see Appendix C: Table 18).

At the same time, the contiguity of the assembly still falls short of those available for other ascidians. Despite considerable efforts, a partial degradation of the genomic DNA detected in all field samples, presumably due to the unusually acidic milieu of the tunic bladder cells (restricted to some groups of ascidians, including the Didemnidae). The bulk of their cytoplasm comprises a large vacuole containing sulfuric acid, which accounts for a tunic pH < 3.0 in didemnids (Hirose, 2001) that may be involved in chemical defense. In contrast, tunic pH > 6.0 was measured for *Perophora* and *Clavelina* species. The acidic pH may account for the observed gDNA degradation, possibly due to increased deamination rates (Lindahl and Nyberg, 1974; Shapiro and Klein, 1966). The partial degradation of gDNA is a confounding factor for genome assembly, particularly limiting the achievable PacBio read lengths. As a consequence, to avoid DNA shearing during extraction for long read sequencing in this species, extraction methods for complex genomes should be considered, including extraction methods based on pulsed field gradient gel electrophoresis (Schwartz and Cantor, 1984), or low-melting agarose microbeads or plugs, as well as other agarose based methods used previously for plant tissues and cells for shearing avoidance (Vogelstein and Gillespie, 1979; H.-B. Zhang et al., 1995). In addition, long-term EtOH storage of *D. vexillum* tissues should be avoided, and tissues should be deep-frozen with liquid nitrogen immediately after collection. Although we believe that the latter alone may not resolve the problem, it certainly provides an additional step of caution for extractions on this species.

The natural genetic diversity of *D. vexillum*, furthermore, is too large for standard genome assembly tools to produce satisfactory assemblies from pooled sequencing of multiple individuals. We therefore resorted to a strategy that reduces the impact of variation, possibly at the expense of contiguity. This genetic diversity is likely associated with *chimerism* of the sampled colony, a phenomenon reported both for *D. vexillum* (Casso et al., 2019; Rinkevich and Fidler, 2014; Watts, Hopkins, and Goldstien, 2019) and other colonial tunicates (Rinkevich, 2005). Chimeric colonies appear to be a natural strategy to potentiate the invasiveness behavior, e.g., enhancing the colony survival having multiple genotypes inside the colony that would respond to a broader set of environmental conditions (Casso et al., 2019).

As a consequence, the assembly is far from perfect. Its contiguity is sufficient to provide exome-level information supporting detailed insights into the gene content of *D. vexillum*. It can be used for phylogenetic purposes, to study the gene structure of the majority of the coding genes, or the evolution of non-coding RNAs. It is insufficient, however, for investigations that involve large-scale synteny, e.g. an assessment of genome rearrangements, and it likely does not represent accurate copy numbers of repetitive elements.

The construction of a reference genome for *D. vexillum* that is on par with better

understood tunicates such as *Ciona robusta* will mostly likely require the creation of an inbred line, as has been the case with other tunicate assemblies (Satou et al., 2019). The high level of diversity observed here may also help to shed light on the fast spread and adaptation of *D. vexillum* to diverse biomes around the globe. It is reminiscent of the increased mutation rate observed for *C. robusta* which is linked to high diversity and adaptive evolution (Tsagkogeorga, Cahais, and Galtier, 2012).

Functional annotation of the predicted *D. vexillum* proteome by comparison with 11 chordates resulted in 8349 orthology groups. The vast majority is shared among chordates. We identified 292 orthology groups in tunicates only (present in more than one tunicate). Among them five functional groups shared by all tunicates, including *lytic polysaccharide monooxygenase and cellulose-degrading processes* (ENOG5028N9R). Other shared orthology groups did not have a specific annotation, however in some cases protein domains (e.g., *sulfotransferase and pleckstrin families* and some transmembrane domains) were recognizable. From all the available chordate orthology groups, 1737 groups were not recovered in our *D. vexillum* assembly. Most notably, we did not find any member of the *RUNX* family, which correspond to key regulators of skeletogenesis together with *HH* and *SOX* family members. We observed that tunicates, except *Oikopleura dioica*, showed a tunicate specific expansion of Hh members. We found seven members of the *SOX* family. A phylogenetic analysis revealed duplication events for SoxC and SoxB2 in *D. vexillum*. We also identified seven of nine tunicate *homeobox* transcription factors of HOX family, the contiguity of the assembly is insufficient to conclusively rule out the absence of the remaining two genes (*Hox1* and *Hox5*) or to determine the genomic organization of the HOX gene cluster. However, a much more extensive annotation effort will be necessary not only for *D. vexillum* but also for tunicate genomes in general, in order to produce a more complete picture of the functional landscape.

The new assembly increased the number of detected ncRNA families to 4877 genomic loci corresponding to 271 families. From these, most of the detected *loci* were *housekeeping* ncRNAs (rRNAs, tRNAs, snRNAs, and snoRNAs) and those *loci* were found in a conserved cluster organization, as seen on tRNAs, rRNAs, and snRNAs. At the same time, a new set of regulatory ncRNAs (miRNAs, Cis-regulatory RNAs and lncRNAs) were detected. As expected, the conserved set of miRNAs were annotated: mir-124, mir-598, mir-7, let-7, mir-1, mir-133, mir-33, lin-4, mir-137, mir-153, mir-2, mir-31, mir-449, mir-183, mir-190, mir-210, mir-219, and mir-8. In comparison to previous miRNA tunicate surveys (Velandia-Huerto, Brown, et al., 2018; K. Wang et al., 2017), we validated previous reports of tunicate-specific mir-1497 (RF00953), and also reported additional specific families, such as ciona-mir-92 (RF01117) and mir-281 (RF00967), by detecting their mature position and evaluating them along a secondary family specific structural multiple alignment. Further studies will allow us to continue to refine the complete miRNA complement in *D. vexillum* and reconstruct the evolutionary history of miRNAs in the tunicates. We were not able to identify homologs of other expected ncRNA families, as: *vault*, *U7* and *Y* RNA and *Telomerase* RNA.

The new assembly of the *D. vexillum* genome described here provides an integrated effort to contribute to the ongoing Tunicata genome projects and constitutes the first annotation dataset for a species in the Aplousobranchia. We hope that the new *D. vexillum* genome annotation presented here triggers more biological studies in a representative of a highly invasive species with a colonial life history.

Part III

Applications

— 6 —

Evolutionary analysis of miRNA over tunicate genomes

Contents

6.1 Chordata miRNA evolution

The chordates, besides containing well-known vertebrates including ourselves, are composed by easily recognizable body plans that have been characterized as monophyletic groups: cephalochordates, tunicates and vertebrates (Dai et al., 2009; Delsuc, Brinkmann, et al., 2006; Stach, 2008; Swalla and A. B. Smith, 2008). Among them, morphologically cephalochordates and vertebrates are more similar, meanwhile tunicates are recognized as simplified and highly derived organisms. Authors as Stach (2008) supported the Notochordata hypothesis (Cephalochordata + Craniata), and recognized the pivotal role of tunicates in the understanding of chordates evolution. However, Delsuc, Brinkmann, et al. (2006), Delsuc, Tsagkogeorga, et al. (2008), and Putnam et al. (2008) evidenced the validity of the Olfactores hypothesis (Tunicata + Craniata), suggested by Jefferies (1991) based on proposed autapomorphies detected on the olfactor stem lineage when compared with mitrates fossils.

This phylogenetic relation has an impact on further evolutionary inferences including the gains and losses of microRNAs (miRNAs). As an example the deep analysis reported by Candiani (2012) studying miRNAs in cephalochordates and their conservation in regard to vertebrates, compared previous annotations in tunicates. In brief, this study found that miRNA complement in tunicates are very divergent in comparison to the chordata dataset, reflected in studies from Hendrix, Levine, and Shi (2010) that identified about 331 miRNAs in *C. robusta* (as detailed in Chapter 3) where most of them were recognized as Ciona-specific families. In the same sense, Fu, Adamski, and E. M. Thompson (2008) reported few deuterostomes conserved miRNAs in *O. dioca*. By a comparison to the annotated miRNA complement in cephalochordates, see references in Candiani (2012). Seems that in general, tunicates shares similar conserved miRNAs with vertebrates. This conclusion was based on the comparisons of *C. robusta* and *O. dioca* and both available amphioxus genomes (*B. belcherei* and *B. floridae*) at that time.

Evidence of origin of ancient miRNA families in tunicates have been reported by Z. Yang et al. (2014) on miR-181. Despite they found evidence of homologs in arctic lamprey (*Lampetra japonica*) and fruit fly (*Drosophila melanogaster*), the putative mature sequence was not recognized on the pre-miRNA. At the same time, homologs on amphioxus (*B. floridae*) and mosquito (*Anopheles gambiae*) the candidate pre-miRNA did not fold as a hairpin. The unique candidate was detected in *C. robusta* and multiple detected copies in vertebrates suggested an evolution by multiple replication of the ancestral gene (Z. Yang et al., 2014). Surprisingly, in the most recent version of `MirGeneDB` v.2.1 (Fromm, Høye, et al., 2021), did not include this sequence on *C. robusta* complement and further evaluation, using `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019), this miRNA failed to be correctly positioned on the reported pre-miRNA (MI0007171). Available miRNA model for this family in `Rfam` v.14.4 (RF00076) are composed by 19 vertebrate sequences, meanwhile their `miRBase` family (MIPF0000007) contained 163 sequences annotated on 37 vertebrates, without the inclusion of *C. robusta* candidate. A challenge is to decide whether this family emerged on the base of olfactores or at the divergence of vertebrates, based on current available miRNA annotation databases/projects.

The recent availability of tunicate genomes over the past 10 years represents a unique opportunity to update the current miRNA annotation on this clade and complement previous studies, that approached the miRNA annotation, over the chordates. In this

chapter, a computational annotation of the miRNA complement was performed over 16 tunicate species and additional 12 deuterostome genomes, using the recently developed `miRNAture` (as detailed in Chapter 4). Described methods and results are part of a manuscript *in preparation* by Velandia-Huerto, Fallmann, and P. F. Stadler.

6.2 Tunicates as targets to find miRNA signals

6.2.1 Collection of genomic information for studied genomes

Genome sequences and gene annotations from 29 deuterostomes: 2 hemichordates, 5 echinoderms, 3 cephalochordates, 16 tunicates, and 3 vertebrates, retrieved from `NCBI` (Sayers et al., 2009), `Echinobase` (Cary, R. A. Cameron, and Hinman, 2018), `ANISEED` (Dardaillon et al., 2019), `Ensembl` and independent databases as listed on Appendix A: Table 20.

Based on the subset of selected species corresponding phylogenetic tree, used to account miRNA evolutionary history, was based on recent phylogenies as a reference from the Tunicata clade, described by Braun, Leubner, and Stach (2019), Delsuc, Philippe, et al. (2018), Giribet (2018), and Kocot et al. (2018).

6.2.2 Preliminary assessment of genome quality

A principal component analysis (PCA) was performed over the following normalized parameters on genome assemblies: GC content, genome size, number of contigs, and percentage of *complete* BUSCO orthologs (Simão et al., 2015). All analyses were calculated using `R` package `factoextra` (Kassambara and Mundt, 2020).

6.2.3 Validation and curation of `miRBase` annotated miRNAs

Studied species that previously reported annotated miRNAs in `miRBase` or another databases, were subject to an evaluation step to validate the correct position of reported mature sequences respect their precursor sequence. Reported precursor and mature sequences annotated on those species in `miRBase` v.22.1, were used to build anchored structural alignments using `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019). To extend the precursor sequences, same genome versions as reported in `miRBase` were used. Specifically, miRNA annotation (including hairpin, mature and genome sequences) for the *sea pineapple* (*H. roretzi*) were retrieved from K. Wang et al. (2017).

Then, input files and necessary processing steps were executed using in-house `Perl` scripts using `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019). Valid miRNAs were identified as precursors with a successfully positioned mature sequences, and with a precursor hairpin-like structure.

6.2.4 Automated construction of miRNA structural alignments, Hidden Markov and Covariance Models

As described by Velandia-Huerto, Fallmann, and P. F. Stadler (2021), a set of quality-filtering steps could be used to build family structural alignments and their corresponding

Covariance Models (CMs). In this case, to build new structural alignments from `miRBase` sequences, all sequences from metazoan species were chosen, removing all of those from studied organisms (listed in Appendix D: Table 20). Given that curated subset, a genetic algorithm was implemented and used to maximize the quality the final structural alignment. To do so, filtering miRNA sequences was done in function of selected parameters, such as: Identity percentage (I), phylogenetic distribution of sequences (D) and quality (Q)[1], where: $\mathrm{I} = (70, 80, 90, 100)$, $\mathrm{D} = (Metazoa, Vertebrata, Mammalia, Primates)$ and $\mathrm{Q} = (normal, high)$. An individual is defined as a vector $\tilde{\mathrm{A}_{\mathrm{l}}} = \begin{pmatrix} \mathrm{I} \\ \mathrm{D} \\ \mathrm{Q} \end{pmatrix}$, which return a structural alignment using `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019), using selected sequences that fit into selected parameters. The *fitness* function (F) to be maximized was defined through empirical observations over structural alignment features (see Equation 6.1).

$$\mathrm{F} = [\mathrm{N_{seq}} + (-\mathrm{F_{energy}}\,\mathrm{N_{spe}}) + (10\,\mathrm{N_{parts}})] \tag{6.1}$$

Where $\mathrm{N_{seq}}$ is the final number of sequences, $\mathrm{N_{spe}}$ is the number of species, $\mathrm{F_{energy}}$ corresponds to folding energy calculated using `RNAalifold` (Lorenz et al., 2011) and $\mathrm{N_{parts}}$ accounts the number of additional (> 1) stem-loops on the reported consensus structure.

To set up the experiments, used operators were:

- Initial population $n = 40$.
- *Selection* = Tournament, $n = 39$.
- *Crossover* = Single point, probability=0.7;
- *Mutation* = Displacement mutation, probability=0.1.

The implementation were performed in `Python` v3.7.9 using `deap` package (De Rainville et al., 2014). Finally, Hidden Markov Models (HMMs) and CMs were build as described in Velandia-Huerto, Fallmann, and P. F. Stadler (2021) using `RNAalifold` (Lorenz et al., 2011) and `Infernal` package v.1.1.2 (Nawrocki and S. R. Eddy, 2013) (see Listing 6.1).

```
clustalo -i <multifasta file> --outfmt clu -o <output file>
RNAalifold --aln-stk=<fasta file> <align file>
cmbuild <Covariance Model> <STO file>
cmcalibrate --cpu=20 <CM>
cmsearch --cpu 4 --tblout <TABULAR OUT> -o <OUT> <CM> <GENOME>
```

Listing 6.1: Modification of Covariance Model

[1] Confidence of the annotation assigned by `miRBase`, see https://www.mirbase.org/blog/2014/03/high-confidence-microrna/

6.2.5 Build CMs from available miRNAs

A comprehensive set of curated miRNA families derived from `miRBase` v.22.1 were build following the proposed classification described on Figure 40. First, those families that were grouped on one of the detected metazoan 1415 `miRBase` families (14,713 loci), were subject to a construction of their CMs and HMMs, removing sequences from studied species. This distinction yields two groups: one with all sequences from tunicate species (*Tunicate-specific*) and the other that included ≥ 0 tunicate sequences with additional sequences from other species (*Non-tunicate specific*). After removing specific target sequences, the first dataset resulted with 986 CMs (labelled as **A**). Removed sequences were prone to a structural evaluation using `MIRfix`, by positioning the reported mature sequence. This method structurally checked and built 391 CMs from annotated loci from 136 families (label **B**). On the other side, tunicate specific families generated 13 CMs (label **C**). In other way, miRNAs without a family classification (15302) were subject to a validation using `MIRfix`, from this group we focused on 562, that contained at least one of the studied species. Through `blastn` iterations using `miRNAture` (strategies 3 and 9), were identified 511 loci that reported close homologs in ≥ 1 studied species or not generated any homologs at all (39). Inside the first group, 216 were evaluated with `MIRfix`, generating 70 valid CMs (label **D**). Correspondingly, those that reported specie-specific results accounted for 190 CMs (label **E**). Finally, those 39 annotated sequences without close-homologs (*No Homology* branch), resulted in 21 CMs (label **F**). The *H. roretzi* miRNA annotation (provided by K. Wang et al. (2017)) accounted 329 miRNAs **G**. Those candidates were processed as the *No classified* dataset, resulting in 257 valid miRNAs by `MIRfix` means.

Tunicate specific models

MiRNA families composed exclusively by studied species (i.e *Ciona* spp.) were subject to an alternative structural alignment construction, adding trusted-homologs in additional tunicates as detailed in Figure 41. Reported hairpin and mature sequences were collected and aligned using `MIRfix` to get the anchored structural alignment and subsequently get their HMM and CM, as described in Section 6.2.4. Next, using `miRNAture` on the first iteration, each miRNA family was searched on additional tunicate genomes. Then, the best candidate for each family and additional genome, defined in terms of global bitscore (B), was concatenated and re-aligned using `RNAalifold` (Lorenz et al., 2011). If this process is part of the first iteration, original *seed* sequences were removed. Resulting HMMs and CMs from first iteration were searched on their original annotated species. The process conclude once all studied species were evaluated. In case additional iterations, same steps can be done without removing sequences. This workflow resembled proposed iteration explained in Chapter 3:Figure 10.

6.2.6 Detection of miRNA homologs by `miRNAture`

An extensive homology search was performed over studied species using `miRNAture` (Velandia-Huerto, Fallmann, and P. F. Stadler, 2021). First, a set of 3395 annotated miRNAs in `miRBase` were searched using `BLAST` mode, which used `blastn` through two previous search strategies reported for miRNAs (Hertel and P. Stadler, 2015; Velandia-Huerto, Gittenberger, et al., 2016). Second, HMMs and CMs (derived from constructed miRNA

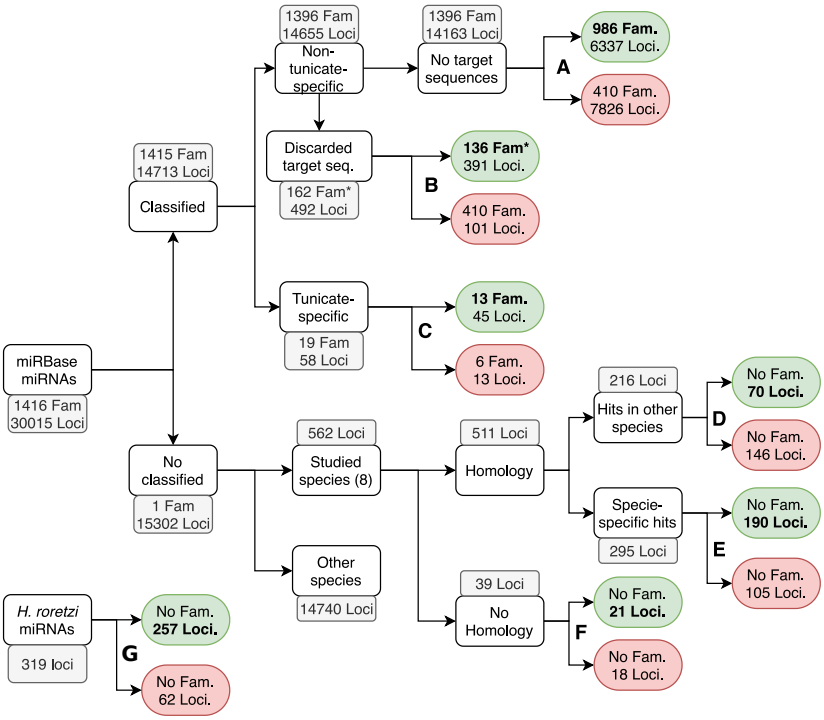


Figure 40: Structural validation of miRNA from `miRBase` and *H. roretzi*. Accepted and filtered families/loci are depicted in green and red, respectively. Bold numbers show the number of families/loci used to build final 2492 CMs.

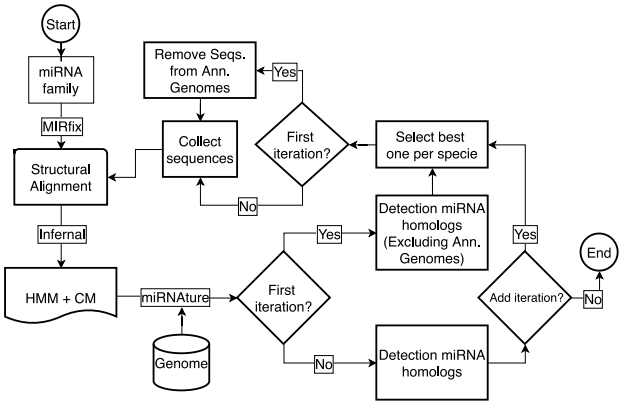


Figure 41: General workflow to increase the number of trusted homologs on a miRNA family using selection of the best candidates into growing alignment.

structural alignments, as described previously in Section 6.2.5, and from 821 `Rfam` v.14.1 miRNA families), were used to validate sequence homology regions by direct searches to the subject genomes (using `Other` and `Infernal` modes in `miRNAture`). In Table 12, a total number of 2492 miRNA CMs were used through 7 experiments, depending on the miRNAs family, as described in Figure 40. Command line parameters are described on Listing 6.2. Multiple generated miRNA `GFF3` files were merged using `bedtools` (Quinlan and Hall, 2010), in case were detected overlapping on the miRNA annotations, global `cmsearch` scores were compared to report the highest score candidate. Families with $r_p > 1$ (see Equation 6.2), served to infer the distribution of miRNAs. Visualization of miRNA loci along with phylogenetic tree of species was done using `ggtree` (Yu, 2020) and `treeio` (L.-G. Wang et al., 2019) `R` packages.

```
miRNAture -stage complete -sublist <LIST_CMs> -dataF <DATA_FOLDER> -speG
<specie_genome> -speN <specie_name> -speT <specie_tag> -w <WORK_PATH> -m
Blast,Infernal,HMM,Final -pe 1 -str 3,9,ALL -blastq <BLAST_QUERIES> -rep
relax,150,100 -usrM <USER_CM_MODELS>
```

Listing 6.2: Annotation of deuterostome miRNAs using `miRNAture`

6.2.7 Comparison to previous miRNA annotations

Annotated miRNAs for the species *B. floridae*, *C. robusta*, *P. miniata*, *S. kowalevskii* and *S. purpuratus*, were obtained from `MirGeneDB` (Fromm, Domanska, et al., 2019). For the species: *O. dioica*, *C. savignyi*, and *P. marinus* were retrieved from `miRBase` release 22.1 (Kozomara and Griffiths-Jones, 2010). *H. roretzi* annotation were retrieved from K. Wang et al. (2017). In cases where an updated version of genome assembly was found, reported sequences were mapped to the current genome version using `blastn`, see updated genomes in Appendix D Table 20. Final candidate miRNAs annotations in `GFF3` format were obtained using `miRNAture`. Comparisons between annotated and predicted loci were calculated using `bedtools` (Quinlan and Hall, 2010). Annotated elements were classified as *Match*, if both elements reported an overlap in genome location level, otherwise were counted as a *Miss* (see Table 14). *Additional* candidates accounted for those predicted elements that missed an overlap.

6.2.8 Inference of conserved miRNA families

Conserved miRNA families were identified by a presence ratio r_p score, calculated by the relation between species with annotated miRNA (p_{spe}) and the total species in the specific clade (T_{spe}), see Equation 6.2.

$$r_p = \frac{p_{spe}}{T_{spe}} \tag{6.2}$$

The set of species were clustered based on correspondent lineage, such as: Hemichordata, Echinodermata, Cephalochordata, Tunicata, and Vertebrata. Additional clades were inferred from `miRBase stockholm` files and were included in the final matrix, such as: Vertebrata, Ecdysozoa, Lophotrochozoa, Cnidaria, Hemichordata, Echinodermata, and Deuterostomia. Families annotated exclusively in one species were filtered, then conserved

Table 12: Description of homology experiments to search miRNAs using `miRNAture`.*: Models calculated from loci and not from families.

Exp.	Target miRNAs	miRNAture mode	Final CM
A	miRNAs from `miRBase`, with family classification. Target sequences were removed.	Blast, HMM, Infernal, OTHER_CM, Final	986 + 821 `Rfam`
B	miRNAs from `miRBase`, with family classification. That belongs from target species.	Blast, Final	136 (selected sequences)
C	miRNAs from `miRBase`, with family classification. All sequences belong from tunicate species.	Blast, HMM, OTHER_CM, Final	13
D	miRNAs from `miRBase`, without family classification. Annotated in target species with additional close-homologs in another species.	Blast,Final	70*
E	miRNAs from `miRBase`, without family classification. Annotated in target species with specie-specific hits.	Blast, Final	190*
F	miRNAs from `miRBase`, without family classification. Did not report close-homologs in other species	Blast, Final	21*
G	*H. roretzi* miRNAs annotated by K. Wang et al. (2017).	Blast, Final	255*
Total			2492

families must report $r_p > 0$ on selected clades, e.g. to define the conserved miRNA in Olfactores: Tunicata (r_p^t) + Vertebrata (r_p^v), and not in other species (r_p^o) it follows: $r_p^t > 0 \wedge r_p^v > 0 \wedge r_p^o = 0$.

6.2.9 Synteny analysis

Obtained miRNA annotation files in `GFF3` format by `miRNAture` (Velandia-Huerto, Fallmann, and P. F. Stadler, 2021) were compared with available genome annotation to get corresponding 3 upstream/downstream coding-genes by each strand, using `bedtools` (Quinlan and Hall, 2010). To optimize speed and manage the cross-references required to use the annotated data, a `MYSQL` database was build for each specie for annotated data and their relation to adjacent miRNA elements. Once adjacent elements to annotated miRNAs were detected, a *vector* representation of those relations were build, taking into account

the genomic order of adjacent coding-elements and their associated miRNA. On this, a further step was done to detect miRNA clusters based on common adjacent elements. In detail, a miRNA cluster M_c is defined as a set of miRNA loci $M_c = (M_1, M_2, \ldots, M_n)$ that are ordered in forward or reverse strand and did not have any element z in the ordered sequence of annotations and $z \notin M_c$. Elements M were re-labelled to avoid redundancy on subsequent comparisons.

To detect groups of homologous miRNAs, their adjacent coding genes were compared using an all-vs-all strategy, using `DIAMOND` (Buchfink, C. Xie, and Huson, 2014) (as detailed in Listing 6.3). Based on those results, close homologous genes were detected if reported:

- E-value $\leq 10e^{-5}$
- Identity $\geq 30\%$.
- Subject coverage $\geq 20\%$.

Finally, a weighted graph was generated taking into account those relations with score > 0. Final groups were detected using a Chinese whispers algorithm, implemented in `python` (Ustalov et al., 2019) over the graph structure.

```
diamond blastp -d <query> --very-sensitive -p 16 --matrix BLOSUM45 -q
    <subject.fa> --evalue 0.0001 --id 25 --query-cover 30 -o out.tab
```

Listing 6.3: Detection close homolog genes using `DIAMOND`

6.3 Tunicates as source of unexplored miRNA annotations

6.3.1 Genome status of analysed species

To assembly a complete dataset to study the miRNA complement over the tunicata clade, all their available genomes were retrieved. It accounts 23 species, which covered all defined classes from the *subphylum Tunicata* (according to phylogeny proposed by Braun, Leubner, and Stach (2019), Delsuc, Philippe, et al. (2018), Giribet (2018), and Kocot et al. (2018)): Ascidiacea (14), Appendicularia (8), and Thaliacea (1). At the same time, additional sister species were included: 3 representatives from vertebrates, 4 cephalochordates, and as an outgroup, were considered 5 species from echinoderms and 2 hemichordates (see detailed genome sources in Appendix D; Table 20). In order to assess to the heterogeneity and identify inherent patterns from selected genomes, a Principal Component Analysis (PCA) together with a Hierarchical Clustering on Principal Components (HCPC) were calculated from the following inferred parameters: number of contigs/scaffolds/chromosomes, nucleotides frequency, GC content, standard deviation of GC and completeness of metazoan universal orthologs (calculated with `BUSCO`).

In general terms, were identified 5 defined clusters, inferred from 3 principal components that explained 84.4% of detected variance along studied genomes (see Figure 42). Cluster 1 and 2, are composed by vertebrate species, one branch dominated by lamprey genomes and another by coelacanth + human; cluster 3 is composed by a combination between

Branchiostoma sp. + hemichordates + colonial Stolidobranchia and another Copelata-specific branch. Cluster 4 comprises a large group of solitary tunicates, divided in three subtrees: the largest one dominated by solitary tunicate species from *Phlebobranchia* (Ciona sp + Phallussia sp.) and *Stolidobranchia* (*M. oculta* + *M. occulata* + Halocynthia sp. + *S. clava*) and the echinoderm *Asterias rubens*. Additionally, a subtree composed by Ambulacraria species and a more dispersed set composed by a draft genome tunicates, such as *Oikopleuridae* + *D. vexillum* (Aplousobranchia). Finally, the cluster 5 grouped another set of genomes from *Oikopleuridae* + *S. thompsoni* (Thaliacea) genomes, which in comparison to studied sequences, were identified as an outgroup due their high number of contigs (mean = 563,859.7) and low percentage of `BUSCO` completeness (mean = 27.3%).

6.3.2 State and structural assessment of miRNA annotation

In terms of reported miRNA annotations, four tunicates were detected with `miRBase` annotations: *C. robusta*, *C. savignyi*, *O. dioica*, and *H. roretzi* from K. Wang et al. (2017). At the same time, the ambulacrarians: *S. kowalevskii*, *S. purpuratus*, *P. miniata*; the cephalochordates: *B. floridae* and *B. belcherei* and the vertebrate *P. marinus*, have annotations on `miRBase` v.22.1. A preliminary validation step over those miRNA annotations was performed due previous reports, such as Fromm, Domanska, et al. (2019) and Velandia-Huerto, Fallmann, and P. F. Stadler (2021), who detected a considerable number of annotated false positives on `miRBase`. The way to validate those loci is based on their secondary structural folding and the (corrected) position of their reported mature sequences. Through this validation method, generated gold standard dataset was used for further comparisons.

As seen in Table 13, annotated miRNA loci over 11 studied species reported on `miRBase` v.22.1, and for *H. roretzi*) (K. Wang et al., 2017), were re-validated in terms of their secondary structural folding and position of reported mature(s) sequences. Surprisingly, the sea squirt *C. robusta* accounted the highest number of miRNA annotations, 13× to the closest related specie: *C. savingyi* and even a larger number than the coelacanth (244) and lancelet (162). Additional support was confirmed including the validation performed by the `MirGeneDB` v.2.0 (Fromm, Domanska, et al., 2019) for 5 species, accounted in the column **Supp.**. This validation reduced in lancelet and sea squirt $\sim 55\%$ and $\sim 40.2\%$ of loci, respectively. Additional annotations in this database were identified on sea urchin (6) and sea squirt (14). Next, this set of loci were mapped into the genomic sequence used in this study (Appendix D:Table 20) which in sea urchin, lancelet, sea squirt and coelacanth were updated with respect to `miRBase` genomes. As a last filter, accounted in the column **Filt.**, annotated miRNAs were evaluated in terms of the position of reported mature sequence and the secondary structure using `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019). In this step, the species with most filtered annotations were: *C. robusta* (60%), *H. roretzi* (20%) and *P. marinus* (18.5%).

At the end of this filtering, as described in **Acc.** column, $> 70\%$ of candidates from almost all species were validated for further comparisons, except for lancelet (35.8) and the sea squirt (14.8), were most of the loci were filtered because they failed the validation in `MirGeneDB` and failed the structural evaluation performed in this study, respectively.

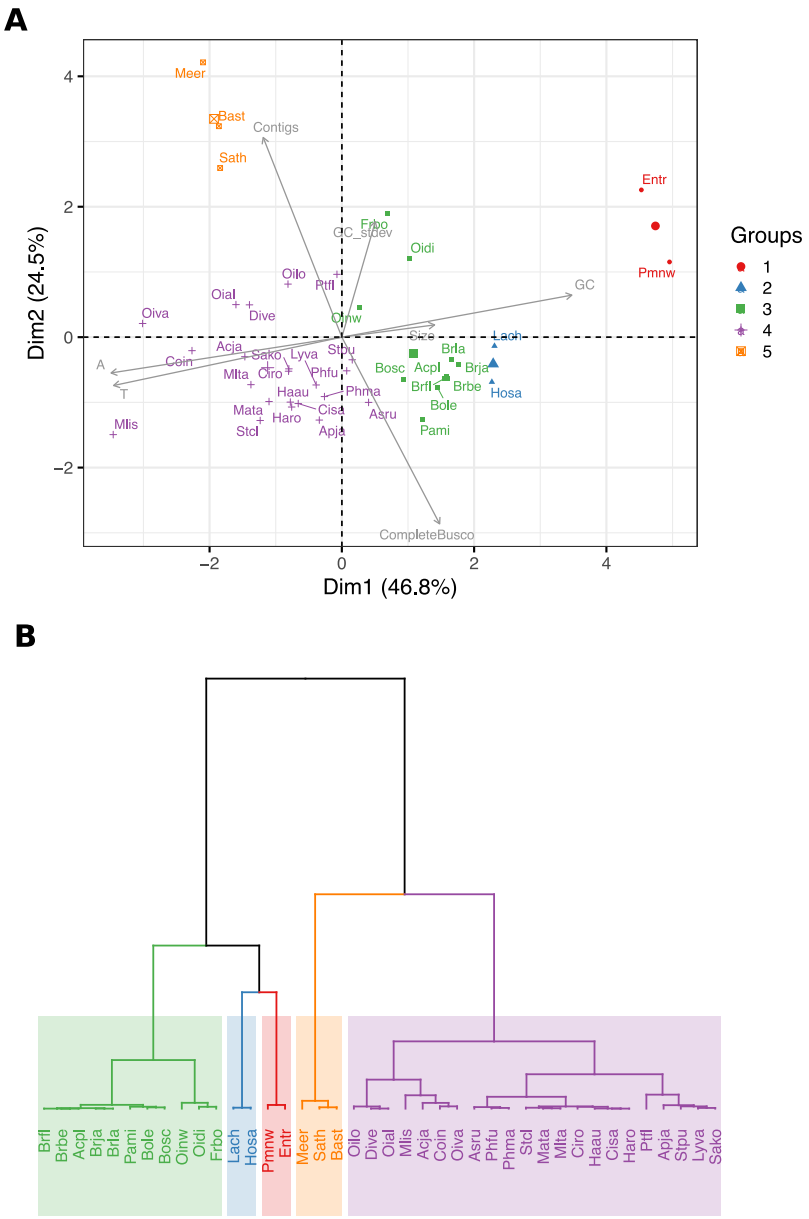


Figure 42: A. Principal Component Analysis (PCA) of genomic features from studied genomes. Selected variables that contributed to the variance are depicted as grey vectors. Colours refer to inferred clusters using HCPC. **B.** HCPC cluster cladogram inferred from calculated PCA.

Table 13: Evaluation of current `miRBase` annotations on studied species. *Supp.* accounts miRNAs validated by `MirGeneDB`. *Filt.* for filtered loci with structural evaluation. *: Updated genomes to last reported assembly. **NA**: Not available.

Source	Specie	Ann.	Supp.	Novel	Mapped	Filt.	Accepted
`miRBase`	*S. kowalevskii*	91	83	0	72	8	64 (70.3%)
	*S. purpuratus**	64	53	6	54	6	48 (75.0%)
	*P. miniata**	49	58	0	51	6	45 (91.8%)
	*B. floridae**	162	90	0	67	9	58 (35.8%)
	*C. robusta**	351	141	14	130	78	52 (14.8%)
	L. variegatus	50	NA	NA	50	1	49 (98.0%)
	B. belcherei	118	NA	NA	110	10	100 (84.7%)
	C. savignyi	27	NA	NA	19	0	19 (70.4%)
	O. dioica	66	NA	NA	47	0	47 (71.2%)
	*P. marinus**	244	NA	NA	238	44	194 (79.5%)
K. Wang et al.	*H. roretzi*	319	NA	NA	319	64	255 (79.9%)

6.3.3 microRNA repertory on tunicate species

As explained in Methods, a comprehensive miRNA homology search was done over studied genomes using `miRNAture` (Velandia-Huerto, Fallmann, and P. F. Stadler, 2021), looking for a set of 2492 miRNA families, derived from `miRBase`. In this way, for species described in Section 6.3.2, the dataset of annotated candidates was compared to the predicted miRNAs (see Section 6.3.3). Finally, a complete characterization for all species is reported taking into account all studied species, as described in Section 6.3.3.

Species with annotation

A comprehensive homology search allowed to detect a predicted set of miRNA loci over described species, see in detail Section 6.2.6. As previously described, the set of annotated miRNAs were validated (as described in Section 6.3.2) and compared to the predicted miRNAs by this study (see in Table 14). At a first glance, over all species the predicted number of miRNAs (column `miRNAture` **predicted**) exceeds by $\sim 25.6\times$ annotated miRNAs (column **Valid Ann.**). About 4.8% of this difference is explained by the presence of miRNAs detected as *truncated* in regard to the family model or with a short loop region < 8nt, which were filtered from the final predicted dataset (column *Final Pred.*). Then, as a product of the comparison between final predicted (columnn **Final Pred.**) and the validated annotated (column **Valid Ann.**) different categories were created: matching (Match) and missing (Miss) annotations, in terms of genomic positions considering the strand.

In one way, the prediction of miRNAs using `miRNAture` was able to recover in average 84.2% of annotated miRNAs. At the same time, the proportion of missing loci in average is 15.7%, with an outlier by *C. robusta* (29 loci, 55.7%). Remaining set of predicted

candidates outside described categories were classified as *additional* ones, which in average, represented a 89.3% in relation to final predicted dataset over all species. A detailed inspection of those categories is presented in the following sections.

Table 14: Comparison annotated predicted candidates to annotated miRNAs. *:Annotations were mapped to current update genome assembly. **Pred.**: Predicted, **Filt.**: Filtered.

Species	`miRNAture` **Pred.**	**Final Pred.**	**Filt.**	**Valid Ann.**	**Match**	**Miss**	**Ratio Match**	**Ratio Miss**	**Add.**
S. kowalevskii	571	493	78	64	59	5	.921	.078	434
L. variegatus	3728	2977	751	49	41	8	.836	.163	2936
*S. purpuratus**	4008	3176	832	48	41	7	.854	.145	3135
*P. miniata**	4088	3347	741	45	42	3	.933	.066	3305
B. belcherei	2738	2183	555	100	82	18	.820	.180	2101
*B. floridae**	2600	2212	388	58	53	5	.913	.086	2159
O. dioica	1090	784	306	47	45	2	.957	.042	740
*C. robusta**	391	341	50	52	23	29	.442	.557	318
C. savignyi	171	140	31	19	16	3	.842	.157	124
H. roretzi	352	340	12	255	226	29	.886	.113	114
*P. marinus**	4089	3348	741	194	166	28	.855	.144	3182

Detected missing annotations were explained by different reasons, such as: *close* predictions, predicted elements that have a shift 200 nt from original annotations, that accounted for 6.89%. At the same time the 46.65% of the candidates were annotated originally by their source database as *novel* candidates that did not fit to previous annotations and fell in this category by database-specific criteria. Another option that did not throw results was the position on the opposite strand.

In the same way, a larger number of *additional* elements, in comparison to the match loci, were detected as filtered candidates with a short loop region or a truncated in the *seed* region (inferred from annotated miRNA mature sequences), as seen in Appendix D: Figure 51. Through this classification, were identified some conserved families, found in 10 species with > 5.4 loci per specie: mir-3149 (MIPF0001935), mir-4536 (MIPF0001319), mir-297 (MIPF0000204), and mir-2513 (MIPF0000980). Short-loop candidates were conserved > 3 species, and accounted for the following families: mir-297 (MIPF0000204) and mir-1277 (MIPF0001937), with a ratio of > 3.7 loci per specie.

On the other hand, accepted families detected as *additional* respect current annotation were located in all species (see column **Add.** in Table 14), accounting for 854 families. Inside that, were detected 33%(281) as species-specific and 67.1% (573) as shared > 2 species. In the first category, we identified 5 families: mir-4526 (MIPF0001545; 66 loci, *S. purpuratus*), MIPFNEW0702 (55), mir-9198 (MIPF0001887; 24) and MIPFNEW0733 (19) from *P. marinus*, and mir-511 (MIPF0000807, 33, *B. belcherei*), that contributed on their host genomes with > 19 loci. On the second category, were possible to detect the family mir-3149 (MIPF0001935) over all species and 7 families over ten species: mir-1277, mir-4703, mir-4679, mir-2513, mir-944, mir-466, and mir-297.

General miRNA annotation

Current miRNA repertory detected along all studied species is depicted on Figure 43. In panel A, phylogenetic distribution is labeled with gain/losses inferred by Dollo parsimony (Farris, 1977; Quesne, 1974). The final miRNA family number with corresponding specie is reported in the leaves and each branch reported their gains (in green) and losses (negative red numbers). At the Deuterostomia ancestor 610 miRNAs were identified. This number increased at the base of chordates by 34 and in ambulacrarians by 8. Next, a high number of losses (191) were detected at the base of cephalochordates, in comparison to the olfactores ancestor, which gained/lost almost the same proportion of families (+27/-23).

Overall, when compared the ancestor in vertebrates and tunicates, a higher loss proportion was detected at the divergence of tunicates (−116) concerning the −65 lost families on the divergence of vertebrates. Specifically, the number of conserved families at the root of tunicates (533) were reduced to < 100 families along almost all tunicates except: *B. schlosseri*, *D. vexillum*, *S. thompsoni* and *O. dioica*. At the same time, a high reduced number of families were identified at Appendicularians (−395) and at the divergence of Thaliaceans (−272). The diversification from Stolidobranchia and Phlebobranchia shows 488 shared miRNA, and a higher loss proportion at Stolidobranchia (−158), and presenting high loss rate at the base of Phalusia sp + Corella spp. (−237), and at the divergence of Ciona spp. (−286). Gains were more consistently identified along other clades, but not in tunicates.

In detail, miRNA loci number is reported in Figure 43B. Identification of truncated loci (based on annotated mature sequences in comparison to the structural model) and miRNAs with short loop region (≤ 8 nt, based on Bartel (2018) and Fromm, Domanska, et al. (2019)) was done based on a posterior analysis of `miRNAture` results (highlighted in red colours). Loci classified as *Accepted* (coloured with green) passed described filters and were considered the final miRNA dataset obtained in this homology strategy. As previously described, a reduction of miRNA loci is evident along the Tunicata clade, showing in average 374 miRNA loci, in comparison to other clades: Cephalochordata (2119), Vertebrata (3638), Hemichordata (1092) and Echinodermata (2737).

6.3.4 Phylogenetic assessment to conserved miRNA families

The distribution of miRNAs over single/multiple clade(s) was assessed by the calculation of detected miRNA families on each defined clade (see Section 6.2.8). In general the number of shared miRNA families (presence on at least > 2 species) is higher (805) than the specie-specific (500) miRNAs. Only on tunicates and hemichordates the number of specie-specific miRNA is greater than the shared one. In one way, shared miRNAs were detected considering their presence on monophyletic groups, identified on the phylogeny depicted in Figure 43. In this way, selected comparisons are particularly relevant to describe the evolution of miRNAs in tunicates, such as: Tunicate-/Cephalochordate-/Vertebrate-specific families, Olfactores, Notochordata, Chordata, Deuterostomes, and Ancestral miRNAs.

As described in Table 15, the list of shared families on described clades are reported with their clades and total number of miRNAs. In detail, the identified set of 22 *ancestral*

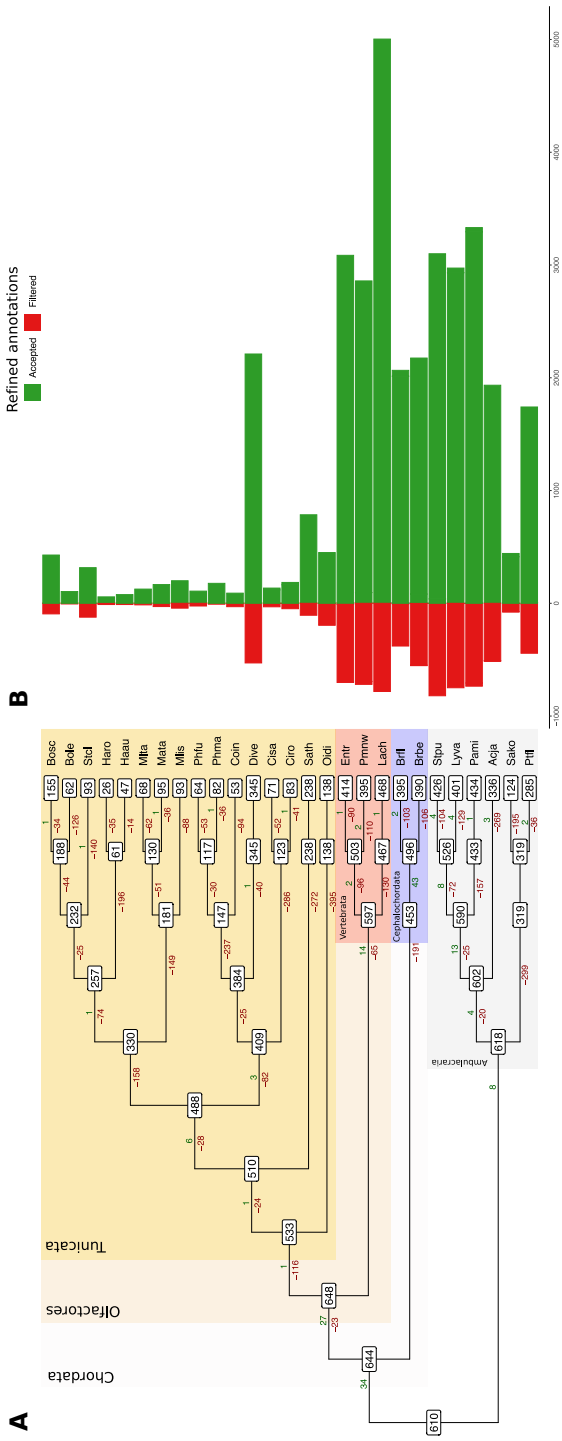


Figure 43: A. Dollo parsimony of miRNA families from 30 Deuterostomia species: Hemichordata (2), Echinodermata (5), Cephalochordata (2), Tunicata (17) and Vertebrata (3). Phylogenetic distribution was inferred from previous reports (Braun, Leubner, and Stach, 2019; Delsuc, Philippe, et al., 2018; Giribet, 2018; Kocot et al., 2018)). **B.** Total number of miRNA loci for described species.

miRNA families covered species from all monophyletic clades and at least one of the following outgroup clades: Ecdysozoa, Lophotrochozoa, Cnidaria, or Deuterostomia (inferred from `miRBase` structural alignments). At the emergence of deuterotomes, were found only three families: mir-135, mir-1497, and mir-4520. The `miRBase` annotated mir-135 and mir-4520 on vertebrates and mir-1497 a tunicate-specific family. By this homology approach, those families originated on the divergence of Deuterostomia, being conserved on the base of Hemichordata, but lost on Echinodermata. On chordates, cephalochordates and Olfactores displayed complete conservation. Specifically, there is a lost of mir-4520 on Petromyzontidae. On tunicates, those families were lost in the Copelata. On Thaliacea, were lost mir-1497 and mir-135. Interestingly, mir-4520 were lost at the base of Stolidobranchia and by its way mir-135 were lost in Phlebobranchia. Those elements were not identified on the genomes from *S. clava*, *M. occulata*, *M. occidentalis* and *C. savignyi*, but identified on their close relatives. Inside chordates, three possible hypothesis of the origin of vertebrates have been proposed: Atriozoa [(Cephalochordata + Tunicata) + Vertebrata], Notochordata [(Tunicata + (Cephalochordata + Vertebrata)], and Olfactores [Cephalochordata (Tunicata + Vertebrata)] (see Stach (2008) for details). In this thesis work, as seen in Figure 43, the Olfactores hypothesis is adopted. As seen in Table 15, mentioned hypothesis were calculated: Atriozoa did not generate shared candidates, Notochordata accounted for 6 conserved miRNAs, and by Olfactores it reached a higher 24 shared miRNA families. Independently, the chordata clades displayed the presence of specific-miRNAs. Both, Cephalochordates and Vertebrates have almost similar number of specific miRNAs: 40 and 44, respectively. Meanwhile, tunicates reported a set of 14 specific families, 10 of them annotated as *de novo* families due their lack `miRBase` classification. Interestingly, the sequence bfl-let-7b, originally reported in `miRBase` for *B. floridae* without family classification, were detected over 2 tunicates: *H. aurantium* and *M. occulta*, and in *P. marinus* but not in additional lampreys and neither in the updated lancelet genome. Detected specific families on vertebrates and cephalochordates are reported in Appendix D; Table 22.

6.3.5 Synteny as rich source of conserved miRNA relations

An additional layer of information can be deduced from a synteny analysis that included annotated miRNAs and their adjacent protein-coding annotations, used as conservation signals (as described in Section 6.2.9). By this method, the reconstruction of the genomic context allow the identification of miRNA clusters or singletons (based on described definition in Section 6.2.9). In this way, the reconstruction of family' evolutionary history was approached at loci level. As evidence of that, an example is described: the tunicate specific miR-1473 family.

miR-1497

The family miR-1497 has been annotated in `miRBase` on Ciona spp. and *O. dioica* (MIPF0000458 family). In fact, the construction of the family Covariance Model (CM) (RF00953) was done by `Rfam` including only the Ciona spp. sequences, which coincided with the family alignment calculated in this work, using a genetic algorithm (see Section 6.2.5). After the iteration with this model and the structural evaluation by `miRNAture`, detected

Table 15: Conserved miRNA families over deuterostomes, including ancestral families detected when compared to outgroup clades. Deuterostomia and Invertebrate chordata sets did not report specific candidates. Olfactores is composed by Tunicata + Vertebrata, meanwhile Notochordata is composed by Cephalochordates + Vertebrata. *: New families without `miRBase` family classification. Chord.: Chordata, Ceph.:Cephalochordata, Not.: Notochordata, Tun.: Tunicata. ‡:Appendix D: Table 22.

	Ancestral	**Chord.**	**Olfactores**	**Not.**	**Ceph.**	**Tun.**	**Vert.**
			mir-19				
	let-7		mir-515				
	mir-29		mir-221				
	mir-25		mir-140				
	mir-9		mir-338				
	mir-8		mir-95			mir-1473	
	mir-133		mir-126			mir-92	
	mir-10		mir-325			mir-1490	
	mir-1		mir-448			mir-4079	
	mir-219		mir-337	mir-129		cin-mir-135*	
	mir-153	mir-135	mir-485	mir-331		cin-mir-4220*	
	mir-31	mir-1497	mir-671	mir-1298	‡	cin-mir-4035*	‡
	mir-33	mir-4520	mir-742	mir-1296		cin-mir-4049*	
	mir-190		mir-744	mir-1949		cin-mir-4072*	
	mir-137		mir-281	mir-1467		cin-mir-5601*	
	lin-4		mir-1905			csa-mir-216b*	
	mir-2831		mir-2450			csa-mir-217*	
	mir-2513		mir-1388			odi-let-7a*	
	mir-2574		mir-2355			bfl-let-7b*	
	mir-3747		mir-3544				
	mir-3718		mir-4423				
	mir-5879		mir-2131				
	mir-8186		mir-1789				
			mir-9209				
Total	22	3	24	6	40	14	44

loci were identified on: 1 hemichordate, 2 (2) cephalochordates, 10 (9) tunicates, and 2 (1) vertebrates, numbers in parentheses indicate those genomes with available gene annotation. Based on the phylogenetic distribution and the inference of shared families (see definition in Section 6.2.8) the origin of this family was assigned on the divergence of chordates. Evidenced syntenic support was detected overall tunicate species as shown in Figure 44A, by means of their adjacent genes. Homology relations between adjacent elements are represented by grey lines. The reconstructed phylogeny included 9 species from the potential 12. Some of them did not report adjacent elements to detected miR-1497 locus/loci, such as *C. inflata*. Other regions did not report homologous relations, even with a larger number of elements as Branchiostoma spp. and *L. chalumnae*. The relation was successfully calculated on tunicate genomes, that have shown their miR-1497 locus lying on

the intronic region of a protein coding gene with annotated *Zona pellucida-like* domains[2]. In most cases, this domain is found in secreted glycoproteins[3]. This relation is restricted to one gene, which in Halocynthia spp. and *D. vexillum* was split. In addition, a strand switch of those regions was detected in the Stolidobranchia representatives (Botryllus spp. + Halocynthia spp.).

The final consensus secondary structures reported in `Rfam` and its augmented version, including detected tunicate sequences, are reported in Figure 44B. The identity of the original `Rfam` alignment was produced by the high similarity of the sequences. However, on the calculated alignment this identity on the consensus structure is not shared when included additional tunicates. High conservation is displayed on the mature sequence block at 3' end (see adjacent black line). The same region is not correctly positioned and even truncated on the `Rfam` model. In regard to previous annotations, in this approach were not identified the reported candidates on the appendicularian (*O. dioica*) reported by Fu, Adamski, and E. M. Thompson (2008) as a cluster. The homology strategy, previous to the validation, did not recognize a possible candidate on the updated genome and even the annotation on `miRBase` did not reported coordinates on previous assembly. In the same way, the candidate in *C. savignyi* failed the structural evaluations to position the reported mature sequence, which indicated that more mature sequences are required to generate the validated hairpin because current annotated mature sequence belongs from *O. dioica*. However, homology detected miRNA candidate was included with their adjacent genes, showing a conserved inverted region with *C. robusta*.

6.4 Discussion

The assessment of animal miRNA evolution has been discussed since their full characterization in 2001 by Lagos-Quintana, Rauhut, Lendeckel, et al.; Lau et al.; Lee and Ambros. In terms of results and evolutionary trends, has been noted that the number of miRNA families and their origin has been correlated with an increased morphology and developmental complexity (Heimberg et al., 2008). As an initial comparison, the miRNA complement has not been detected on the ancient specie *Trichoplax adhaerens* (Hertel, Jong, et al., 2009), but the complete processing machinery has been traced back in unicellular organisms (Bråte et al., 2018). By the other way, on those organisms with complex morphological patterns, it was recognized the burst of new families and annotations, as the case for vertebrates or mammalian species (Hertel and P. Stadler, 2015). However, as explained in this work, the distribution of annotations and database miRNA registers are biased towards model organisms, such as human, mouse, fruit fly and roundworm. It is interesting to ascertain whether this increased annotations are an evolutionary consequence, or are a systematic bias.

Particularly, tunicates can be used to tackle this question, since the miRNA annotation progress have been done using few tunicate models. Furthermore, those marine invertebrates with huge morphological and developmental heterogeneities, have been reported at genomic level as organisms with a clear tendency to reduce, cluster and/or reshape the genome architecture acquired on the divergence of chordates or at the Olfactores ancestor.

[2] `Pfam` accession number: PF00100.
[3] Interpro accession number: IPR042235

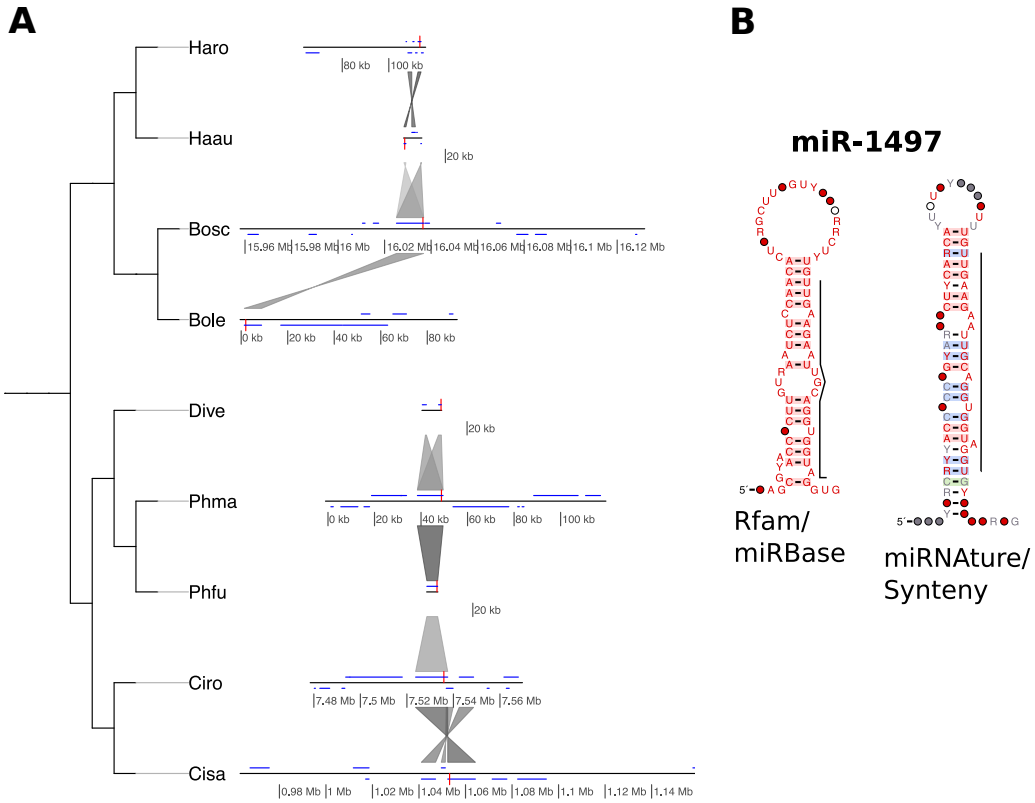


Figure 44: A. Synteny of miR-1497 identified in tunicate species. Red and blue lines represent miRNAs and adjacent coding genes, respectively. Candidate for *C. savignyi* was found by the `homology` mode in `miRNAture`. **B.** Consensus secondary structure of reported and calculated miR-1497 after synteny evaluation. Left structure is reported by the `Rfam` (RF00953) using sequences from `miRBase`. Right structure represents the augmented alignment including tunicate sequences supported by synteny.

Their annotated miRNA complement, have been characterized by the identification of well-conserved families and specie-specific families (Fu, Adamski, and E. M. Thompson, 2008; Hendrix, Levine, and Shi, 2010; Norden-Krichmar et al., 2007; K. Wang et al., 2017) or even the presence of miRNA-offset RNA (moR), detected lying close to miRNAs (Shi et al., 2009). In this context, as reported by Velandia-Huerto, Brown, et al. (2018), in the last 10 years the availability of new tunicate genomes have increased, opening the door to fill the miRNA annotation gap on this clade and in general, in chordates.

Nevertheless, the quality of genome assemblies are heterogeneous. Tunicate genomes were recognized in multiple similarity clusters, clearly distinguished from vertebrates and assembling a cluster with colonial species, *B. schlosseri* and *B. leachii*, and cephalochordates. A bigger group, composed by solitary species, displaying close similarities between *Phlebobranchia* and *Stolidobranchia* genomes. The position of two Oikopleuridae and the *D. vexillum* is uncertain in this group, but they were recognized as a dispersed, due

their current draft genome assemblies. In the same way, an *outgroup* set was identified, assembled with species from Oikopleuridae and Thaliacea, that reported in common the largest number of contigs and low `BUSCO` completeness.

As detailed in Chapter 3, a consensus database that use all available miRNA mature and precursor annotations to build structural alignments is missing. As a response, to use the complete potential of `miRNAture` (Velandia-Huerto, Fallmann, and P. F. Stadler, 2021), multiple annotation sources were used to search miRNAs on tunicates, such as `miRBase` annotations, `Rfam` models, and curated alignments derived from `miRBase` families/sequences. The structural ascertainment translated the evaluation into a numerical score, which allowed the implementation of a genomic algorithm that selected a set of sequences to get the best alignment under designed score assumptions. As a result, the automated strategy generated 1671 metazoan Covariance Models (CMs), derived from `miRBase` sequences. Similar approaches were described and developed by Yazbeck, Tout, et al. (2017) and Yazbeck, P. F. Stadler, et al. (2019) using the `MIRfix` pipeline, but using all available sequences from `miRBase` families, without considering further filtering rules or variations on the alignment construction. Another approach to create iterative alignments and CMs was reported by Eggenhofer, Hofacker, and Höner zu Siederdissen (2016), but the curation steps performed by `MIRfix` are not included in this strategy such as the construction of mature anchored-structural alignments. In general, these methods can be extended improving the score function(s) that could consider another a more general way to classify and evaluate miRNA alignments.

Using the correct position of the mature region by means of `MIRfix` (Yazbeck, P. F. Stadler, et al., 2019), apparent high available miRNA annotations on the sea squirt genome were filtered. This high number of false positives are supported by a previous validation, using RNA-seq data, by the `MirGeneDB` database which supported only the 40% of previous `miRBase` annotations. However, in `MirGeneDB` the genome versions should be updated towards data consistency. For example, in the last version of `MirGeneDB` (Fromm, Høye, et al., 2021), the scientific name of the sea squirt is incorrect and genome assembly are not updated, impacting further evolutionary references and the species sampling for tunicates is focused only in this specie. This calls for a validation by multiple computational and experimental criteria that should be used as strategy to produce a *gold standard* miRNA annotation.

As approached by Velandia-Huerto, Fallmann, and P. F. Stadler (2021), a true set of miRNAs to quantify the performance of a computational tool to discover miRNAs is not easy to define, given the current miRNA classifications/definitions and ongoing new annotations. In comparisons to annotated miRNAs, the ratio of matches reported by `miRNAture` was about 89.3% over studied genomes (see Table 14). However, the number of additional loci is high and needs to be further validated using experimental data. Also, this approach led to quantify `miRNAture` performance by discovering possible false positives in the miRNA annotation from human, as previously highlighted by Velandia-Huerto, Fallmann, and P. F. Stadler (2021). In this sense, a *gold standard* miRNA dataset, despite current shortcomings, should be generated based on intersected accumulated evidence from miRNAs.

The complete annotation landscape of miRNA loci and families supported a reduction at the divergence of tunicates, which lost about 116 families in regard to the 65 lost in vertebrates in comparison to the ancestral Olfactores families (648). The conserved

533 families on the tunicate divergence was dramatically reduced in almost all species of ascidians, appendicularians and salps, with specific cases with high number annotations identified on colonial tunicates, which could posses multiple haplotypes in their assembled draft genomes, increasing systematically the number of detected loci (as reported in the *D. vexillum* genome (Parra-Rincón et al., 2021)). To further evolutionary comparisons, the phylogenetic tree distribution (Figure 43) allowed the inference of conserved miRNAs along specific divergence points. Next, those results were suitable to be compared to earlier approaches that studied lancelet and tunicate common families in regard to the vertebrate miRNA complement (Candiani, 2012). In this study the Olfactores clade generated 24 shared families, in comparison to the 6 conserved by the Notochordata clade. Evenly, the number of families derived at the base of vertebrates (44) is high respect to the tunicate ones (14), and similar to cephalochordates (40). More broadly, the conserved deuterostome conserved list was expanded, reaching a set of ancestral 22 families in accordance to earlier evolutionary reports and their families (Hertel and P. Stadler, 2015; Velandia-Huerto, Brown, et al., 2018).

The distribution of miR-1473 was extended in tunicates. In complement of this approach, an automatic strategy based on close-homolog adjacent anchors allowed the detection of miRNA regions that share a common divergence. As a product of this extension, an updated miR-1473 structural alignment, including sequences from tunicates except appendicularians and thaliaceans, was build using the results of the syntenic analysis. Loci reported by Fu, Adamski, and E. M. Thompson (2008) in *O. dioica* were not found by the homology strategy and did not report genomic coordinates on previous assembly, as reported in `miRBase` (see family MIPF0000458). This suggests a high derivate sequence, that could escape from current homology methods, and a required further validation of the tentatively diverged mature sequences supported by experimental evidence. In general, the use of this approach could be applied to other miRNA families to obtain mature anchored-structural alignments, derived from sequences grouped by a common syntenic context. An enhanced structural alignment that could be used as an iterative model can be used to expand the annotation over additional close-related species.

Part IV

Conclusions and Perspectives

— 7 —

Conclusions and Perspectives

Contents

7.1 Conclusions

The understanding of the regulatory networks behind the scenes of the cell functionality is a central topic that have been occupied the molecular biology from its origins. The cellular machinery, viewed as an open high organized system, has converged to assign those regulatory tasks to molecular *artefacts*, found along all living organism. The current model of *canonical* microRNA (miRNA), has opened the door to comprehend the tightly controlled and efficient biogenesis mechanism to generate miRNA precursors and their mature counterparts. Interestingly, this process is affected by the exact expression of their components, but inevitably further *epigenetic* factors, such as: the availability and correct selection of the generated mature sequences by Argonaute (AGO), being a crucial step to the subsequent formation of miRNA-induced silencing complex (miRISC). By its own, the non-canonical biogenesis challenged the consensus canonical model, due their alternative sequence, structural and transcriptional patterns, and reported source of precursors (small/long-RNAs, intronic regions or even random generated hairpins). The distinction of those miRNA classes is crucial to delimit correctly the miRNA definitions and clarify current elements that are on a grey classification-zone.

In terms of their functionality, once miRISC is loaded, the targeting function of a miRNA could affect many mRNAs, as reported in animals. To make more complex this relation, target sites located on UTR regions of messenger RNAs (mRNAs), could possess multiple sites for different miRNA seeds. Finally, the complementary relation once the miRNA reaches its target(s), affects and directs the intensity of the control mechanism over targeted transcripts.

In terms of animal evolution, miRNAs have been identified along all metazoans, laying into multiple genomic locations with a tendency to be organized in clusters, generated very often by tandem or non-local duplications. Since their location is not a random process, it has been noted a correlation between their locations in conserved old genes with higher and broadly expression, subjects of a strong purifying selection. However, the generation of new miRNAs is not a static process, which can be initiated from random local formation of stem-loops that most of the time are located on a specie-specific gene, affected by the presence of a fast evolution rates. Detected conservation patterns of miRNAs along metazoans helped to recognize an expansion of the miRNAs families. This expansion is strongly correlated with a broad interaction with additional control networks, that are evidenced, for example on an increased morphological complexity. As evidence for that, multiple burst of new families have been detected over the metazoan tree. As a contraposition of this trend, an extreme morphological simplification observed in tunicates has yielded an overall lost of the miRNA complement detected at the divergence of the clade inside Olfactores. In spite this reduction it is still possible to detect chordate conserved clusters, indicating a high pressure to conserve those regions together, despite the genomic restructuring mechanisms detected on tunicates. At the same time, current literature has reported large clusters of specie-specific families, as explained before, containing *young* miRNA families.

The central role played by miRNA databases is reflected in the growing number of research that used them to add support to their observations for a particular set of miRNAs. As described in this work, most of the miRNAs are annotated in `miRBase`, being an exclusive database to standardize the miRNA annotations, it does not escape from

annotation artefacts or false positives. Despite growing literature has highlighted particular erroneous cases, still those annotations are present and are being used e.g. as training set of Machine Learning approaches (see (Ben Or and Veksler-Lublinsky, 2021)). Recently, as an indirect complement of that, the `RNACentral` database is available to organize and compare multiple RNA annotations. Particularly on miRNAs, this data aggregation should consider their multiple biogenesis entities, such as: pri-miRNA, pre-miRNAs, and mature products. At the same time, those entities should be included in the construction of homology predictive models, based on multiple structural alignments, as Hidden Markov Models (HMMs) or Covariance Models (CMs). The use of this information improves the quality of the alignments. This was demonstrated by the curation of `Rfam` miRNA-families, using their additional information derived from `miRBase`. The inclusion of the correct position of mature sequences inside the precursor, led to the generation of *anchored* (by mature region) multiple structural alignments. Through this alternative, it was identified 33 families with recognized pitfalls, such as: invalid long hairpin structures, insufficient number of correct sequences to build an alignment, or with structural issues. Moreover, the multiple structural alignment could be composed by sequences that did not belong from the same family, as recognized in the miR-154 (RF00641). In addition, a recognition of the non-canonical miRNA families could be useful to be included on the databases. This would help to avoid making assumptions about a canonical mature position, Dicer or Drosha processing cleavages, or the inference of the precursor sequence based on an incorrect mature position. This complete effort to curate current data on available `miRBase` or `Rfam` annotations, benefit the generation of further automatic pipelines to predict the miRNA complement on genome sequencing projects or query nucleotide sequences.

In addition to provide a curated set of miRNA families and build correct structural alignments, a detailed evolutionary history of miRNA paralogs could be addressed by the projection of current annotated/validated miRNAs onto new assembled species or species on the first steps of annotation, using homology strategies. Common searching strategies that combine sequence searches (using `blast` or HMMs) or structural alignments (by CMs) to predict those sequences/models on the subject genome/sequence.

The identification of merged homology regions that combine all of those comparison methods and the posterior structural evaluation with specific miRNA-features, has been developed in `miRNAture`, as proposed to limit the need of extensive manual analysis of miRNA specific features. Further to the sequence analysis provided by homology search tools, the integration of the correct position of the mature sequence and the structural evaluation of corrected pre-miRNA, led `miRNAture` ascertain a stringent way to increase sensitivity when performed a homology search. By this way, `miRNAture` used as a benchmark the human annotation, identifying 87.9% of human annotated miRNAs and tagged 27 families as false positives, on the `miRBase`. Interestingly, reporting additional candidates from 178 families recognized to be overlapping into long non-coding RNAs (lncRNAs) genes or intronic/exonic gene regions, repeat families (as inverted repeat transposable elements (MITEs)) or being antisense elements of other miRNAs or coding genes. Some of those elements were found actually being transcribed, but in some cases not previously annotated. In the same way, miss annotations were recognized as members of repetitive families, precursors with deviant secondary structure or unusual placement of mature sequence respect to reported precursor, or families with insufficient number of mature sequences or precursor sequences to build an alignment, or incorrect initial

structural alignments which affected the loci prediction.

The previous assessment performed in human have evidenced that a clear definition of a miRNA and the limits between canonical/non-canonical are still loosely defined. In one hand, this is significant since human contains large number of data, used to support or filter those annotations, and still there are gaps on its miRNA classification. In the other hand, same kind of results could be faced on a new genome assembly projects. Due to lower associated cost and efficient generation of large quantity of data, the refinement of homology and annotation techniques should be considered. In the same way to miRNA annotations, currently high number of vertebrate genome assemblies did not represent the reality in terms of diversity in comparison to invertebrates, which are unrepresented at genomic level.

The significance of sequencing an unrepresented animal, as *Didemnum vexillum* (Aplousobranchia), has been evidenced at the beginning of the collection of the raw DNA material. The challenge started with the extraction of genomic material, degraded by a low pH reported in bladder cells of didemnids. Surprisingly, due the presence high levels of genomic variance, standard assembly pipelines assembled $< 20\%$ of achieved genome size by a hybrid approach, using available PacBio and Illumina reads. This could be explained by multiple events of chimerism/*multichimerism* between colonies, a strategy reported increasing the genetic variability, generate synergistic complementation, or increase the probability of mate location (see a complete analysis in Rinkevich (2005)). The significance of this high variability was evidenced on the protein coding annotation, where suspicious cases were isolated and classified. At the same time, relevant conserved genes were identified despite their fragmentation, such as: HOX and skeletogenesis-related genes, by means of homology comparisons with close related candidates in another tunicate species. In case of non-coding elements, homology strategies successfully annotated house-keeping families and traced the evolution of the RMST lncRNA. In particular, for the comprehensive annotation of the miRNA complement, a methodological analysis of threshold values of CMs and specific annotations filters, founded the basis to further being implemented in `miRNAture`. This obtained relevant biological data, despite the contiguity challenges that represented the sequencing and assembly, supported the use of available draft genomes as rich source of biological information. In case of *D. vexillum*, constitutes an opportunity to improve existing ways to generate expected contiguity with PacBio. Meanwhile, a public genome browser condense current calculated information, prone to be updated and further validated.

In a broad miRNA exploration, the miRNA complement for 16 tunicates was predicted using multiple annotation sources, such as individual sequences and structural models for miRNA families, derived from `Rfam` and `miRBase`. To this end, the automatized construction of mature-anchored structural alignments using `miRBase` sequences constituted the first approach to generate corresponding sequence and profiles. In total 2492 models were designed to search systematically 7 sets of annotated miRNAs into the target species. The use of an automatized strategy as `miRNAture`, enables this kind of homology comparisons, gathering multiple sources of evidence to annotated a unique, non-overlapping miRNA annotations.

This approach is useful as first strategy to get a potential number of canonical miRNAs on a subject sequence or genome, reducing time and throwing a list of miRNA loci that could be used to further analyses. In terms of predicted miRNA families and loci, tunicates

evidenced a reduction in comparison to other chordates and at the divergence of the clade, 533 miRNA families were identified, but all tunicate species reduced between 34 and 91% of this set. Inferences of evolutionary conserved families indicated a larger number of families shared in Olfactores (24) than in Notochordates 6. The synteny analysis of conserved adjacent genes added a rich layer to complement those evolutionary analyses. By this strategy, the phylogenetic distribution of miR-1497 was extended and curated in tunicate clade and the structural alignment was increased. The proposed idea can be extrapolated to create synteny-supported structural alignments, that consider the position of annotated/reported mature sequences as an anchor. However, as seen in tunicates, the availability of gene annotations and the contiguity of the genome could restrict the quality of those comparisons.

7.2 Perspectives and open questions

The most pressing open problems, however, concern practical aspects of data analysis: What is the level of completeness of miRNA annotation that can be achieved by homology search? In other words, how good are the available tools in identifying the last common ancestors of miRNA families, and how good are our estimates of the gain and loss of individual ortholog groups and entire families of homologous miRNAs? It would appear that quantitative statements on patterns and regularities of miRNA evolution are at present limited by technical (computational) and biological (definitional) issues. The available evidence seems to support clear differences between (sub)types of miRNAs, implying that quantitative studies required clear distinctions of miRNA types. Additional data will certainly help to clarify these issues. High coverage small-RNA-seq data would be of particular value for species that do not have close relatives for which such data are already available.

Several aspects of miRNA evolution remain poorly understood. How prevalent are anti-sense miRNAs, i.e., those produced from a common locus in both reading directions such as iab-4/iab-8 (Hui et al., 2013)? Are evolutionary transitions from miR to miR* common? Examples of this kind of "arm-switching" have been reported e.g. in the mir-10 and mir-100 families (Griffiths-Jones, Hui, et al., 2011). Are there cases in which the position of the precursor hairpin shifts relative to the mature product, i.e., new miR or miR* are introduced? Finally, there is of course the broad topic of conservation of miRNA function and thus of their target sites. Answers to all these topics require a reliable, accurate, and (reasonably) complete miRNA annotation. Hence, it is more than worthwhile to address the many technical issues addressed in this contribution in future research and to invest in tools and pipelines to process the ever-increasing wealth of miRNA data with much less user intervention and expert curation.

Appendices

— A —

Clusters in tunicate genomes

A.1 Largest miRNA clusters in some chordate species

Table 16: Details of biggest miRNA cluster for chordate species. **No.**: Number loci.

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. robusta	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

— B —

Curation of miRNA databases

B.1 Correspondence between `Rfam` and `miRBase` sequences

The described family label annotation performed over 4849 `Rfam` sequences generated a 65.4% of the assignments to a `miRBase` gene family, 25.6% did not reported a matching hit on `miRBase` and the remaining 9.3% mapped into a sequence without any `miRBase` family annotation (NoFam).

The complete overview is depicted on Figure 45, which last described categories (NOHIT, NOFam and miRBaseFam) binned the number of miRNA families. In detail Figure 45A depicted those models that reported an unique label: 154 families have been built with sequences that have all representatives on the same `miRBase` family, 5: (RF00800, RF00872, RF01942, RF02002 and RF02013) did not generated any `miRBase` hit, and the family *mir-1251* (RF01938) mapped into an unclassified set of `miRBase` precursors. At the same time, those families that contained 2 or 3 family labels reported a high frequency of families that included into their alignments sequences without hits (*NOHIT*) and/or without assigned family (*NoFam*) (see Figure 45B).

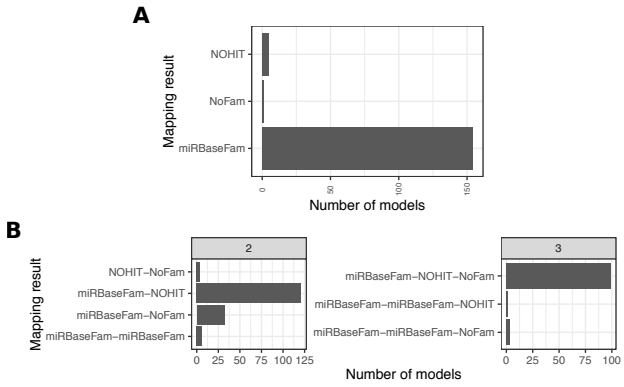


Figure 45: Number of models that mapped 1 (A), 2 or 3 (B) miRNA gene families from `miRBase`. Detail of labels: **NOHIT**: did not have a mapping representative on the `miRBase` precursors. **NoFam**: Models that mapped onto a precursor without a family classification provided from `miRBase`

Finally, those models that reported $>$ 3 mapped gene families were described in

more detail on Figure 46. It was detected a set of 10 Rfam families which contained a heterogeneous set of miRNA families and were potential candidates to be revisited in case that their consensus secondary structure fails the evaluation, as explained in Figure 11.

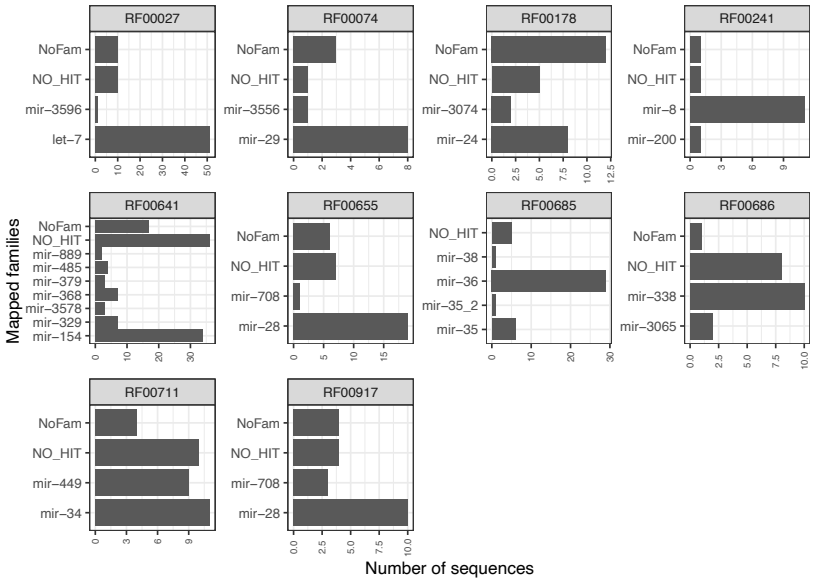


Figure 46: Mapping results of the sequences from Rfam models which have reported > 3 miRBase miRNA families.

B.2 Discarded Rfam models

Table 17: Discarded miRNA families in the first annotation round on the Rfam models. The **Ref.** referred to annotation on Rfam database. **Rmvd.** = Number of removed sequences. **NA**: Not available.

miRNA Family	Ref.	Seed	Full	Input	Rmvd.	Comments
mir-31	RF00661	28	163	28	NA	The inference of mature region failed, because the stem matching region reported a 15 bp size, which is less than the minimum defined to report a mature region ($\geq$ 19 bp). Consequently, the generation of required files for MIRfix failed. See consensus secondary structure for the family on Rfam.

mir-198	RF00681	3	9	4	NA	Have been detected 4 sequences with `miRBase` mature annotation, which failed the correction process due their short length < 20 nt: MI0002918, MI0002919, MI0002920, and MI0002921. The annotated evidence for those miRNA mature sequences were annotated by similarity to the human miRNA hsa-mir-198 (MI0000240) supported by 39 reads in `miRBase`[1]. After the first curation process, the remaining sequences from the model did not have a mature set to be annotated.
mir-458	RF00750	7	71	7	6	For the sequence `CAAE01013759.1/80040-80116` was recognized the sequence annotated on `miRBase`: MI0003253. After the curation process, the predicted mir and mir* sequences did not fit for 6 of the 7 sequences. Need better mature sequences.
mir-287	RF00788	7	9	7	6	For the sequence `AE014134.6/17574610-17574702` was found in `miRBase` the precursor MI0000381 with MIMAT0000360 as mature, supported by 7 reads. In this approach this precursor sequence was corrected and predicted the mir. The remaining sequences from the family were discarded, because this only one set of predicted mir and annotated mir* were not able to detect the corresponding regions on the other sequences with a correct folding.

[1] `http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000240`

mir-42	RF00794	5	4	6	6	The family FR847113.2/8891453-8891358 has been annotated in miRBase as: MI0000495, which reported homology based predicted mir sequence (MIMAT0000467). The annotated mature sequence did not fit into the stem structure for the remaining sequences. This family needs better mature support.
mir-BART3	RF00866	1	0	7	1	Only one sequence from human was detected as valid in this miRNA family. The other sequences corresponded to the Epstein-Barr virus, which were out of the scope of this computational strategy.
mir-1302	RF00951	24	3551	24	NA	Did not found mature annotations on miRBase and their secondary structure resulted on predicted mature regions with lengths of ≤ 19 bp.
mir-604	RF001041	2	14	3	2	The sequence CM000324.3/30534620-30534527 was identified on miRBase as which reported the mature MIMAT0012809. Based on the posterior correction, two of three sequences were removed because the mature prediction did not fit into the folding sequence.
mir-1419	RF01919	5	13	5	2	The predicted mature region was insufficient to perform the correction. The structure reported by Rfam alignment was misleading and did not corresponded for a hairpin-like one.
mir-2518	RF01944	4	57	4	2	The predicted mature region was insufficient to perform the correction. The structure reported by Rfam alignment was misleading and did not corresponded for a hairpin-like one.

mir-1803	RF02094	6	859	6	2	The sequence `CM000094.4/92880219-92880307` was identified as MI0007548 in `miRBase` with the validated mature: MIMAT0007720. This sequence was not enough to perform the annotation of the mature region for the other sequences, for that reason 5 precursors were removed. This family needs better mature support.
mir-56	RF02214	4	4	4	NA	The predicted mature region was insufficient to perform the correction. The structure reported by `Rfam` alignment was misleading and did not corresponded for a hairpin-like one.

— C —

Didemnum vexillum annotation

C.1 ncRNA mapping from previous draft *D. vexillum* assembly

Table 18: Annotated *loci* in the draft version of *D. vexillum* that reported more than one mapping position on the new alignment. Old coordinates are reported as: Name, Chromosome, Strand, Start, End. New coordinates are reported as: Chromosome, Start-End, Strand, (Bitscore, E-Value)

Candidate	Supported by Alignment	Other position
mir-276.dvex159218.+.706.786	scaffold9268-size9828, 4294-4374, Reverse, (38.8, 2.5e-10)	scaffold7042-size10846, 1855-1935, Forward, (34.0, 3e-09)
SNORD18.dvex152227.+.1372.1432	scaffold225-size23884, 4646-4706, Reverse, (26.8, 9.4e-07)	scaffold76090-size2378, 1649-1710, Reverse, (18.7, 2.8e-05)
U4atac.dvex622135.-.311.415	scaffold23895-size6623, 2192-2306, Forward, (37.5, 2.5e-08)	scaffold78418-size2279, 1339-1453, Forward, (59.1, 8.7e-12)
U6.dvex134697.+.19.112	scaffold66798-size2802, 2097-2193, Reverse, (40.5, 3.8e-09)	scaffold68538-size2716, 1005-1098, Forward, (41.1, 3e-09)
U6.dvex152032.+.83.175	scaffold30047-size5748, 717-809, Forward, (35.2, 3.2e-08)	scaffold97836-size1478, 229-325, Forward, (29.7, 3e-07)
U6.dvex435452.+.1786.1884	scaffold925-size18266, 4693-4791, Forward, (43.2, 1.3e-09)	scaffold17919-size7584, 1862-1964, Forward, (33.8, 5.8e-08)
U6.dvex595726.-.285.383	scaffold7916-size11956, 1129-1227, Forward, (26.8, 9.6e-07)	scaffold85789-size1993, 494-591, Forward, (31.7, 1.4e-07)
U6.dvex619958.+.139.235	scaffold28367-size5973, 674-770, Reverse, (36.4, 2e-08)	scaffold60312-size3152, 619-715, Forward, (36.4, 2e-08)

C.2 Phylogenetic distribution of Rfam miRNA alignments

Distribution axis corresponds to the taxonomic classification for the species reported on the `stockholm` seed alignment by `Rfam`. Panels *VALID_STR* and *NO_VALID_STR* which contains those *loci* that fit into the alignment or not, respectively (Figure 47).

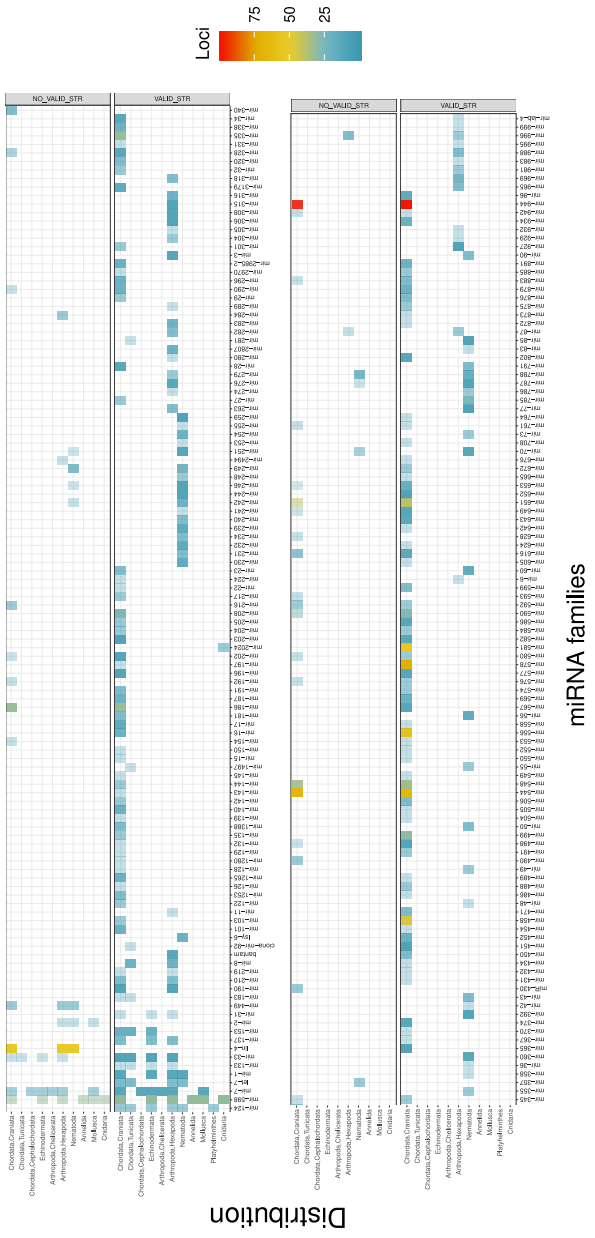


Figure 47: Distribution of final set of miRNA families annotated on *D. vexillum*.

C.3 Mitochondrial genome alignment

Tunicata mt-DNA is represented along a coordinate system, which start from 0 to the length of the mt-DNA. The resulting alignment is centered on the conserved block along all genomes, which overlaps with the position of mt-LSU. Other conserved blocks have been detected and are highlighted by the same color and the same corresponding joining line. Negative numbers along the coordinates are useless in terms of distances, otherwise these refer to a translocated region, i.e. *Cionas* comparison (see Figure 48).

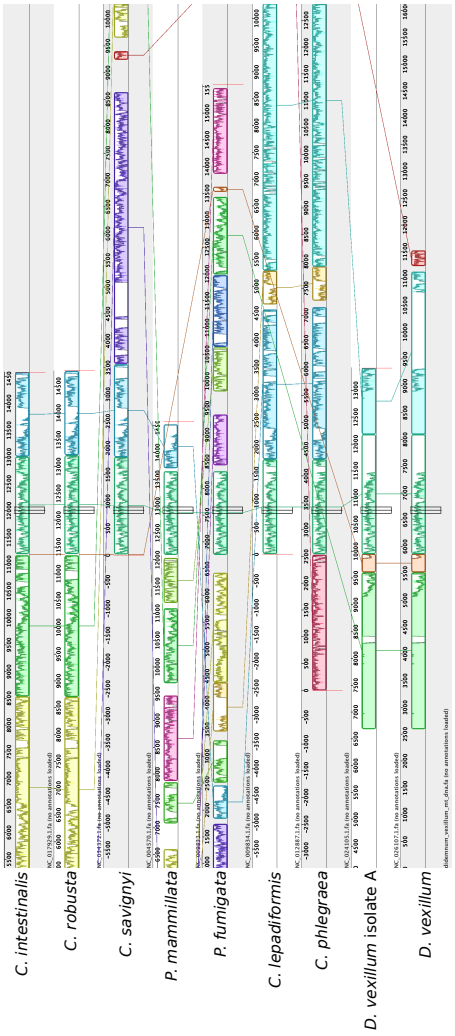


Figure 48: Graphic representation of a mitochondrial genome multiple alignment.

C.4 Hox genes

C.5 RUNX family phylogeny

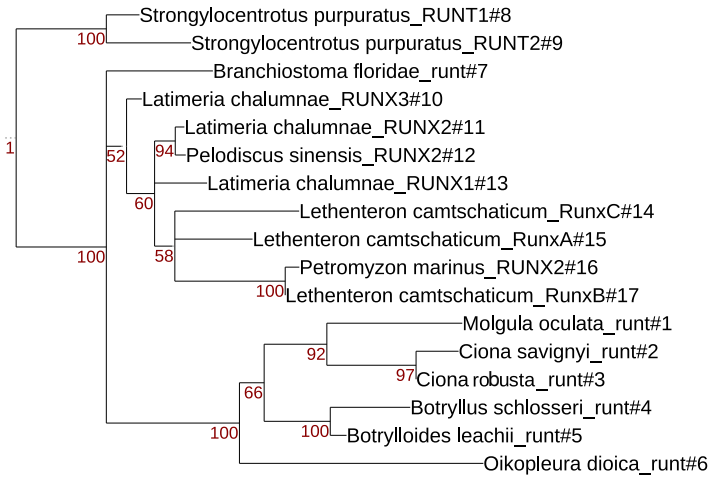


Figure 49: Phylogenetic analisis of RUNX family.

C.6 SOX family phylogeny

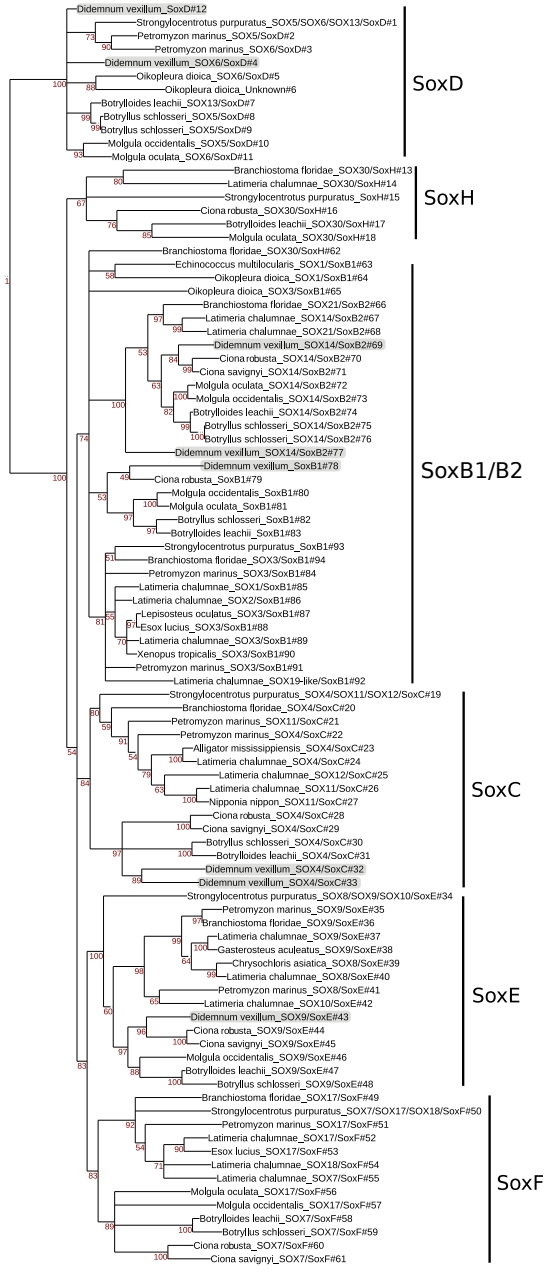


Figure 50: Complete phylogenetic tree of the SOX family.

Table 19: Final results from annotation of candidates of Hox gene family in the *D. vexillum* genome. **A**: Anterior, **C**: Central, **AP**: Ancestral Posterior. NA: not available. Species labels: **Brla**: *Branchiostoma lanceolatum*, **Ciin**: *C. intestinalis*, **Haro**: *H. roretzi*, **Hefr**: *Heterodontus francisci*, **Lame**: *Latimeria menadoensis*, **Bosc**: *B. schlosseri*, **Ciro**: *C. robusta*, **Cisa**: *C. savignyi*, **Bole**:*B. leachii*, **Oidi**: *O. dioica*.

Class	Hox Gene	Manual Curation	Homology Support	Alignment Support
A	Hox2	scaffold16549, 5363-5551,-	Dvex_pep49845,2876-6340,-, Brla, Ciin, Haro, Hefr, Lame	4226-4827, +, Bole, Bosc, Ciro, Cisa
		scaffold2010, 9838-9972, -	NA	9755-10025, +, Bole, Cisa, Haro
A	Hox3	scaffold11368, 1079-1225, -	NA 1072-1253, +, Bosc, Ciro, Cisa	
		scaffold57440, 880-1005, -	NA	877-1131, +, Ciro, Cisa, Haro
C	Hox4	scaffold72048, 2108-2254, +	NA	1980-2259, +, Bole, Bosc, Ciro, Haro
		scaffold12766, 1763-1909, -	NA	1267-2089, +, Ciro, Cisa
C	Hox6/7-like	scaffold34036, 585-779, -	NA	633-807, +, Haro
AP	Hox10	scaffold93625, 1125-1223, -	(Dvex_pep30911, Dvex_pep30912), 556-1226, -, Ciro, Cisa[a]	638-1292, +, Ciro,Cisa
AP	Hox12	scaffold4141, 8233-8496, +	NA	8146-8486, +, Bole, Ciro, Cisa, Haro
AP	Hox11/12/13.a	scaffold101308, 4-243, +	NA	10-389, +, Bole, Brfl, Ciro, Cisa, Oidi

[a] Not found by standard homology searches. Reported region has shown a high homology with a corresponding harbouring-Hox gene region in *Ciona* sp. Reported gene (Divexi.CG.Dive2019.scaffold93625.g29940) in *D. vexillum* was annotated on that region.

— D —

Data sources

D.1 Studied Deuterostome genomes

Table 20, retrieved 28 deuterostomes were tagged based on their phylogenetic classification, as follows: H= Hemichordata, E= Echinodermata, C=Cephalochordata, T= Tunicata and V=Vertebrata.

D.2 Blast strategies used for miRNA homology

`Blastn` Camacho et al. (2009) strategies integrated in `miRNAture` for miRNAs detection on homology level. Strategies 1-4 are based on Velandia-Huerto, Gittenberger, et al. (2016), strategy 5 by (Hertel, Bartschat, et al., 2012) and `blastn` default strategy (6).

D.3 Structural consistency evaluation

In terms of the evaluation provided by the use of `miRNAture`, each miRNA loci was evaluated in terms of structural features. In regard to the inferred position of mature sequences, the precursor sequence were subdivided into 5 regions, which 2 corresponds to mature regions (5' and 3'), two additional regions (5'-arm, 3'-arm) and 1 loop region. By this subdivision, specific checks were evaluated:

- The length of loop region should be > 8 nt.
- Position of mature miRNAs should be located inside the global alignment respect their family covariance model, specifically the *seed* region: nucleotides 2-8 from 5'-arm and 13-16 from 3'-arm.

D.4 Conserved miRNAs

Table 20: Analyzed chordate genomes.

Clade	Specie	Label	Size (Mb)	Assembly State	Reported Fragments	Version
H	*Ptychodera flava*	Ptfl	1228.7	Scaffolds	218,255	1.0.14
	Saccoglossus kowalevskii	Sako	775.8	Scaffolds	54,120	Skow_1.1
E	*Strongylocentrotus purpuratus*	Stpu	921.9	Scaffolds	871	Spur_5.0
	Patiria miniata	Pami	608.3	Scaffolds	30	Pmin_3.0
	Anneissia japonica	Apja	589.63	Scaffolds	76,727	Ajap1.0
	Asterias rubens	Asru	417.6	Scaffolds	150	Arub1.3
	Lytechinus variegatus	Lyva	869.6	Scaffolds	33	Lvar3.0
C	*Branchiostoma floridae*	Brfl	513.5	Chromosomes	432	Bfl_VNyyK
	Branchiostoma belcheri	Brbe	426.1	Scaffolds	2308	Haploidv18h27
T	*Botrylloides leachii*	Bole	159.1	Scaffolds	1778	ANISEED v.1
	Botryllus schlosseri	Bosc	580.4	Chromosomes	14	ANISEED v.1
	Corella inflata	Coin	120.7	Contigs	67,285	v.1
	Ciona robusta	Ciro	123.0	Chromosomes	67	KY
	Ciona savignyi	Cisa	177.0	Reftigs	375	ENSEMBL CSAV2.0
	Didemnum vexillum	Dive	519.3	Scaffolds	109,770	v.2
	Halocynthia aurantium	Haau	128.1	Contigs	11,610	ANISEED v.1
	Halocynthia roretzi	Haro	119.6	Contigs	3047	ANISEED v.1
	Molgula occulta	Mlta	174.8	Contigs	23,663	ANISEED v.1
	Molgula occidentalis	Mlis	249.6	Contigs	21,251	ANISEED v.1
	Molgula oculata	Mata	154.4	Contigs	10,554	ANISEED v.1
	Oikopleura dioica	Oidi	64.3	Chromosomes	19	OKI2018_I69
	Phallusia fumigata	Phfu	231.9	Contigs	34,699	ANISEED* v.1
	Phallusia mammillata	Phma	233.9	Contigs	11,004	ANISEED* v.1
	Salpa thompsoni	Sath	318.7	Contigs	478,281	v.1
	Styela clava	Stcl	340.5	Chromosomes	211	v.1
V	*Entosphenus tridentatus*	Entr	927.7	Chromosomes	25,006	NCBI v7.0
	Latimeria chalumnae	Lach	2861	Scaffolds	22,819	ENSEMBL LatCha1
	Petromyzon marinus	Pmnw	1089.05	Chromosomes	1434	NCBI kPetMar1.pri

Table 21: `Blastn` strategies integrated in `miRNAture`.

	Blastn strategies					
Flag	1	2	3	4	5	6
-dust	no					D
-soft_masking	false					D
-reward	5	4	5	4	D	D
-penalty	-4	-5	-4	-5	D	D
-gapopen	10	3	25	12	D	D
-gapextend	6	5	10	8	D	D
-word_size	7				D	D
-evalue	0.01				$10e^{-10}$	D
-outfmt	6					

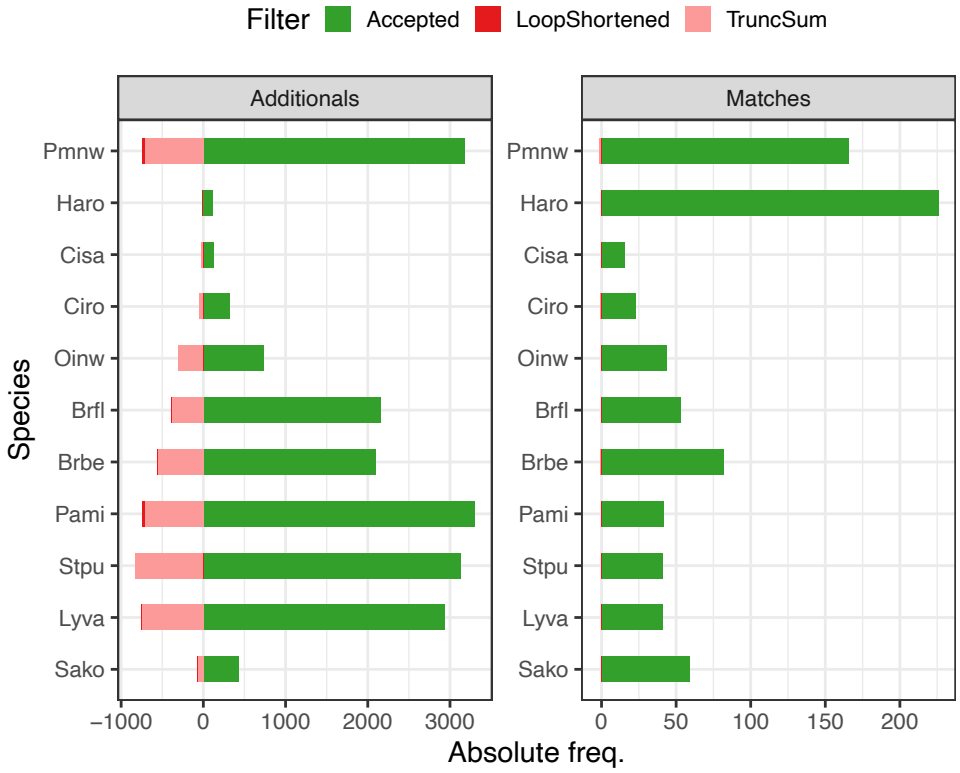


Figure 51: Final number of filtered miRNAs over 11 annotated species. Labels: *P. marinus* (**Pmwn**),*H. roretzi* (**Haro**), *C. savignyi* (**Cisa**),*C. robusta* (**Ciro**),*O. dioica* (**Oinw**),*B. floridae* (**Brfl**),*B. belcherei* (**Brbe**),*P. miniata* (**Pami**),*S. purpuratus* (**Stpu**),*L. variegatus* (**Lyva**), and *S. kowalevskii* (**Sako**).

Table 22: Conserved miRNA families over cephalochordates and vertebrates.

Cephalochordata	Vertebrata
	mir-17
	mir-430
mir-4872	mir-181
mir-4875	mir-103
mir-4868	mir-23
mir-4057	mir-27
mir-2071	mir-24
mir-2059	mir-26
mir-4889	mir-128
mir-2072	mir-148
mir-2068	mir-205
mir-4864	mir-21
mir-4890	mir-192
mir-4888	mir-132
mir-2056	mir-143
mir-2057	mir-136
mir-4865	mir-146
mir-2063	mir-147
mir-2062	mir-455
mir-4866	mir-431
mir-4860	mir-491
mir-4876	mir-684
mir-4863	mir-456
mir-4879	mir-551
mir-4859	mir-878
mir-2061	mir-939
mir-2058	mir-1306
mir-2076	mir-767
mir-2070	mir-935
mir-4873	mir-622
mir-4928	mir-769
mir-2066	mir-1538
mir-4861	mir-1451
mir-4878	mir-3604
mir-4869	mir-3074
mir-4880	mir-3065
mir-4899	mir-3190
mir-4857	mir-1260b
mir-4874	mir-3154
mir-4856	mir-3072
mir-2067	mir-3934
mir-4891	mir-3192
	mir-4677
	mir-7143
40	44

List of Symbols

F	*Fitness* score evaluated in all individuals
Q	`miRBase` quality classification of a loci
B	Calculated bitscore from `cmsearch`
C(f)	Calculated coverage respect to corresponding covariance model.
E	Calculated E-value from `cmsearch`
p	Comparison parameter: 1, -1, or *text*
$\mathcal{B}$	Emission probabilities or observation likelihoods
F_{energy}	Folding energy of the consensus secondary structure
GA	gathering threshold defined for each `Rfam` family
b	Genome coordinate end
a	Genome coordinate start
z	Genomic element different to miRNAs.
i, j	Homology detected candidates
$\tilde{A}_l$	Individual with specific values of *identity*, *taxonomic distribution* and `sequence quality`
π	Initial probability distribution
m	Length of the mapped region into the new genome
h_m	Length of the mapped region into the subject
M_c	microRNA cluster
M	microRNA locus
r_p	miRNA presence ratio over a defined clade
r_p^o	miRNA presence ratio over other clades
r_p^t	miRNA presence ratio over tunicates
r_p^v	miRNA presence ratio over vertebrates
nGA	Normalized *gathering score* suggested by `Rfam`
$n_{bitscore}$	Normalized bitscore

n_g	Number of close related genes in target genome
N_{seq}	Number of sequences that compose a final structural alignment
p_{spe}	Number of species in a clade with a miRNA family
T_{spe}	Number of species in a clade
N_{spe}	Number of species that compose a final structural alignment
N_{parts}	Number of stem-loops on the consensus secondary structure
n	Original size of the query contig
h_n	Original size of the query
Φ	Overlapping variable
I	Percentage of identity between two sequences
R_{mn}	Relation of original and mapped query on new genome
Q	Score triplet, composed by bitscore, E-value and coverage
c	Sequence contig/scaffold identification
λ	Sequence of observations
$\mathcal{O}$	Sequence of observations
Q_s	Set of N states in a Markov chain
s	Strand where element is located
D	Taxonomical distribution of a family
A_{ij}	Transition probability matrix
$e_{distance}$	Tree edit distance
$\tilde{P}$	Vector representing detected structural candidates

List of Abbreviations

T_m melting temperature.
m^7G 7-methylguanosine.
3'UTR 3'untranslated region.
CM Covariance Model.
HMM Hidden Markov Model.
MITE inverted repeat transposable element.
MSA multiple sequence alignment.
dsRNA double-stranded RNA.
lncRNA long non-coding RNA.
mRNA messenger RNA.
miRNA microRNA.
moR miRNA-offset RNA.
ncRNA non-coding RNA.
piRNA piwi-interacting RNA.
rRNA ribosomal RNA.
sdRNA small nucleolar RNA-derived RNA.
shRNA short-hairpin RNA.
siRNA short-interfering RNA.
snRNA small nuclear RNA.
snoRNA small nucleolar RNA.
ssRNA single-stranded RNA.
stRNA small temporal RNA.
tRNA transfer RNA.

AGO Argonaute.

CRT cyclic reversible termination.

DT Decision trees.

EXP1 exportin-1.
EXP5 exportin-5.

GA gathering score.
GIGA Global Invertebrate Genomics Alliance.

HCPC Hierarchical Clustering on Principal Components.
Hh *Hedgehog*.

ISH *in-situ* hybridization.

LNA locked nucleic acid.
Loqs Loquacious.
LSU large subunit ribosomal RNA.

MFE minimum free energy.
miRISC miRNA-induced silencing complex.
MSP maximal-scoring segment pair.

NB näive Bayes.
NGS next generation sequencing.
nr non-redundant.
nt nucleotides.

PAZ PIWI-AGO-ZWILLE.
PCA Principal Component Analysis.
PE paired-end.
PIWI *P*-element *i*nduced *wi*mpy testis.
pre-miRNA precursor miRNA.
pri-miRNA primary miRNA.

qRT-PCR Quantitative Real-time PCR.

RACE rapid amplification of cDNA ends.
RISC RNA-induced silencing complex.
RMST Rhabdomyosarcoma 2-associated transcript.
RNAi RNA interference.

SBL Sequencing by ligation.
SBS Sequencing by synthesis.
SCFG stochastic context-free grammar.
SMRT Single Molecule Real Time.

SNA single-nucleotide addition.

SSU small subunit ribosomal RNA.

SVM Support vector machine.

TRBP TAR RNA-binding protein.

URS unique identifier.

ZMW Zero Mode Waveguide.

Definition Index

List of Figures

List of Tables

Bibliography

Albà, M. M. and J. Castresana (Nov. 2004). „Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes“. In: *Mol. Biol. Evol.* 22.3, pp. 598–606. DOI: 10.1093/molbev/msi045.

Aldridge, S. and J. Hadfield (2012). „Introduction to miRNA profiling technologies and cross-platform comparison“. In: *Methods Mol. Biol.* 822, pp. 19–31.

Alles, J. et al. (Mar. 2019). „An estimate of the total number of true human miRNAs“. In: *Nucleic Acids Res.* 47.7, pp. 3353–3364. DOI: 10.1093/nar/gkz097.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (Oct. 1990). „Basic local alignment search tool“. In: *J. Mol. Biol.* 215.3, pp. 403–410. DOI: 10.1016/s0022-2836(05)80360-2.

Ambros, V. et al. (Mar. 2003). „A uniform system for microRNA annotation“. In: *RNA* 9.3, pp. 277–279. DOI: 10.1261/rna.2183803.

Antonacci, F. et al. (Oct. 2014). „Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability“. In: *Nat. Genet.* 46.12, pp. 1293–1302. DOI: 10.1038/ng.3120.

Anzelon, T. A., S. Chowdhury, S. M. Hughes, Y. Xiao, G. C. Lander, and I. J. MacRae (Sept. 2021). „Structural basis for piRNA targeting“. In: *Nature* 597.7875, pp. 285–289. DOI: 10.1038/s41586-021-03856-x.

Aravin, A. A., N. M. Naumova, A. V. Tulin, V. V. Vagin, Y. M. Rozovsky, and V. A. Gvozdev (July 2001). „Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline“. In: *Curr. Biol.* 11.13, pp. 1017–1027. DOI: 10.1016/s0960-9822(01)00299-8.

Auyeung, V. C., I. Ulitsky, S. E. McGeary, and D. P. Bartel (Feb. 2013). „Beyond Secondary Structure: Primary-sequence Determinants License Pri-miRNA Hairpins for Processing“. In: *Cell* 152.4, pp. 844–858. DOI: 10.1016/j.cell.2013.01.031.

Backes, C. et al. (Dec. 2015). „Prioritizing and selecting likely novel miRNAs from NGS data“. In: *Nucleic Acids Res.* 44.6, e53–e53. DOI: 10.1093/nar/gkv1335.

Baker, M. (Sept. 2010). „MicroRNA profiling: Separating signal from noise“. In: *Nat. Methods* 7.9, pp. 687–692. DOI: 10.1038/nmeth0910-687.

Barbieri, M. (Mar. 2016). „What is information?“ In: *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 374.2063, p. 20150060. DOI: 10.1098/rsta.2015.0060.

Bartel, D. P. (Jan. 2009). „MicroRNAs: Target Recognition and Regulatory Functions“. In: *Cell* 136.2, pp. 215–233. DOI: 10.1016/j.cell.2009.01.002.

Bartel, D. P. (Mar. 2018). „Metazoan MicroRNAs“. In: *Cell* 173.1, pp. 20–51. DOI: 10.1016/j.cell.2018.03.006.

Ben Or, G. and I. Veksler-Lublinsky (May 2021). „Comprehensive machine-learning-based analysis of microRNA–target interactions reveals variable transferability of interaction rules across species“. In: *BMC Bioinf.* 22.1, p. 264. DOI: 10.1186/s12859-021-04164-x.

Berezikov, E. (Nov. 2011). „Evolution of microRNA diversity and regulation in animals“. In: *Nat. Rev. Genet.* 12.12, pp. 846–860. DOI: 10.1038/nrg3079.

Berezikov, E., F. Thuemmler, L. W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, and R. H. A. Plasterk (Oct. 2006). „Diversity of microRNAs in human and chimpanzee brain“. In: *Nat. Genet.* 38.12, pp. 1375–1377. DOI: 10.1038/ng1914.

Berná, L. and F. Alvarez-Valin (July 2014). „Evolutionary Genomics of Fast Evolving Tunicates“. In: *Genome Biol. Evol.* 6.7, pp. 1724–1738. DOI: 10.1093/gbe/evu122.

Bernhart, S. H., I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler (Nov. 2008). „RNAalifold: Improved consensus structure prediction for RNA alignments“. In: *BMC Bioinf.* 9.1, p. 474. DOI: 10.1186/1471-2105-9-474.

Bernhart, S., I. L. Hofacker, and P. F. Stadler (2006). „Local RNA Base Pairing Probabilities in Large Sequences“. In: *Method. Biochem. Anal.* 22, pp. 614–615.

Bernstein, E., A. A. Caudy, S. M. Hammond, and G. J. Hannon (Jan. 2001). „Role for a bidentate ribonuclease in the initiation step of RNA interference“. In: *Nature* 409.6818, pp. 363–366. DOI: 10.1038/35053110.

Blanchoud, S., K. Rutherford, L. Zondag, N. J. Gemmell, and M. J. Wilson (Apr. 2018). „De novo draft assembly of the Botrylloides leachii genome provides further insight into tunicate evolution“. In: *Sci. Rep.* 8.1, p. 5518. DOI: 10.1038/s41598-018-23749-w.

Bohmert, K. (Jan. 1998). „AGO1 defines a novel locus of Arabidopsis controlling leaf development“. In: *EMBO J* 17.1, pp. 170–180. DOI: 10.1093/emboj/17.1.170.

Bompfünewerer, A. F. et al. (2005). „Evolutionary Patterns of Non-Coding RNAs“. In: *Th.\ Biosci.* 123, pp. 301–369.

Bråte, J. et al. (Oct. 2018). „Unicellular Origin of the Animal MicroRNA Machinery“. In: *Curr. Biol.* 28.20, 3288–3295.e5. DOI: 10.1016/j.cub.2018.08.018.

Braun, K., F. Leubner, and T. Stach (Nov. 2019). „Phylogenetic analysis of phenotypic characters of Tunicata supports basal Appendicularia and monophyletic Ascidiacea“. In: *Cladistics* 36.3, pp. 259–300. DOI: 10.1111/cla.12405.

Brozovic, M. et al. (Nov. 2017). „ANISEED 2017: Extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets“. In: *Nucleic Acids Res.* 46.D1, pp. D718–D725. DOI: 10.1093/nar/gkx1108.

Brudno, M., S. Malde, A. Poliakov, C. B. Do, O. Couronne, I. Dubchak, and S. Batzoglou (July 2003). „Glocal alignment: Finding rearrangements during alignment“. In: *Method. Biochem. Anal.* 19.Suppl 1, pp. i54–i62. DOI: 10.1093/bioinformatics/btg1005.

Brunetti, R., C. Gissi, R. Pennati, F. Caicci, F. Gasparini, and L. Manni (2015). „Morphological evidence that the molecularly determined Ciona intestinalis type A and type B are different species: Ciona robusta and Ciona intestinalis“. In: *J. Zool. Syst. Evol. Res.* 53.3, pp. 186–193.

Buchfink, B., C. Xie, and D. H. Huson (Nov. 2014). „Fast and sensitive protein alignment using DIAMOND“. In: *Nat. Methods* 12.1, pp. 59–60. DOI: 10.1038/nmeth.3176.

Bullard, S. et al. (Mar. 2007). „The colonial ascidian Didemnum sp. A: Current distribution, basic biology and potential threat to marine communities of the northeast and west coasts of North America“. In: *J. Exp. Mar. Biol. Ecol.* 342.1, pp. 99–108. DOI: 10.1016/j.jembe.2006.10.020.

Bullard, S. G., B. Sedlack, J. F. Reinhardt, C. Litty, K. Gareau, and R. B. Whitlatch (Mar. 2007). „Fragmentation of colonial ascidians: Differences in reattachment capability among species“. In: *J. Exp. Mar. Biol. Ecol.* 342.1, pp. 166–168. DOI: 10.1016/j.jembe.2006.10.034.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden (Dec. 2009). „BLAST+: Architecture and applications“. In: *BMC Bioinf.* 10.1, p. 421. DOI: 10.1186/1471-2105-10-421.

Campo-Paysaa, F., M. Sémon, R. A. Cameron, K. J. Peterson, and M. Schubert (Jan. 2011). „microRNA complements in deuterostomes: Origin and evolution of microRNAs. miRNA origins and evolution“. In: *Evol. Dev.* 13.1, pp. 15–27. DOI: 10.1111/j.1525-142x.2010.00452.x.

Candiani, S. (Feb. 2012). „Focus on miRNAs evolution: A perspective from amphioxus“. In: *Brief. Funct. Genomics* 11.2, pp. 107–117. DOI: 10.1093/bfgp/els004.

Cantarel, B. L., I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado, and M. Yandell (Nov. 2007). „MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes“. In: *Genome Res.* 18.1, pp. 188–196. DOI: 10.1101/gr.6743907.

Cary, G. A., R. A. Cameron, and V. F. Hinman (2018). „EchinoBase: Tools for Echinoderm Genome Analyses“. In: *Methods Mol. Biol.* 1757, pp. 349–369.

Casso, M., D. Tagliapietra, X. Turon, and M. Pascual (Oct. 2019). „High fusibility and chimera prevalence in an invasive colonial ascidian“. In: *Sci. Rep.* 9.1, p. 15673. DOI: 10.1038/s41598-019-51950-y.

Cech, T. R. and J. A. Steitz (Mar. 2014). „The Noncoding RNA Revolution–Trashing Old Rules to Forge New Ones“. In: *Cell* 157.1, pp. 77–94. DOI: 10.1016/j.cell.2014.03.008.

Chan, P. P. and T. M. Lowe (2019). „tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences“. In: *Methods Mol. Biol.* 1962, pp. 1–14.

Cheloufi, S., C. O. Dos Santos, M. M. W. Chong, and G. J. Hannon (Apr. 2010). „A dicer-independent miRNA biogenesis pathway that requires Ago catalysis“. In: *Nature* 465.7298, pp. 584–589. DOI: 10.1038/nature09092.

Chen, L., L. Heikkinen, C. Wang, Y. Yang, K. E. Knott, and G. Wong (Jan. 2018). „miRToolsGallery: A tag-based and rankable microRNA bioinformatics resources database portal“. In: *Database* 2018. DOI: 10.1093/database/bay004.

Chen, L., L. Heikkinen, C. Wang, Y. Yang, H. Sun, and G. Wong (June 2019). „Trends in the development of miRNA bioinformatics tools“. In: *Brief. Bioinform.* 20.5, pp. 1836–1852. DOI: 10.1093/bib/bby054.

Chodroff, R. A., L. Goodstadt, T. M. Sirey, P. L. Oliver, K. E. Davies, E. D. Green, Z. Molnár, and C. P. Ponting (2010). „Long noncoding RNA genes: Conservation of sequence and brain expression among diverse amniotes“. In: *Genome Biol.* 11.7, R72. DOI: 10.1186/gb-2010-11-7-r72.

Cifuentes, D. et al. (June 2010). „A Novel miRNA Processing Pathway Independent of Dicer Requires Argonaute2 Catalytic Activity“. In: *Science* 328.5986, pp. 1694–1698. DOI: 10.1126/science.1190809.

Clote, P., F. Ferré, E. Kranakis, and D. Krizanc (Apr. 2005). „Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency“. In: *RNA* 11.5, pp. 578–591. DOI: 10.1261/rna.7220505.

Consortium, R. et al. (Oct. 2020). „RNAcentral 2021: Secondary structure integration, improved sequence search and new member databases“. In: *Nucleic Acids Res.* 49.D1, pp. D212–D220. DOI: 10.1093/nar/gkaa921.

Creugny, A., A. Fender, and S. Pfeffer (2018). „Regulation of primary microRNA processing“. In: *FEBS Lett* 592.12, pp. 1980–1996.

Csurös, M. (June 2010). „Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood“. In: *Method. Biochem. Anal.* 26.15, pp. 1910–1912. DOI: 10.1093/bioinformatics/btq315.

Dai, Z., Z. Chen, H. Ye, L. Zhou, L. Cao, Y. Wang, S. Peng, and L. Chen (Jan. 2009). „Characterization of microRNAs in cephalochordates reveals a correlation between microRNA repertoire homology and morphological similarity in chordate evolution“. In: *Evol. Dev.* 11.1, pp. 41–49. DOI: 10.1111/j.1525-142x.2008.00301.x.

Dambal, S., M. Shah, B. Mihelich, and L. Nonn (July 2015). „The microRNA-183 cluster: The family that plays together stays together“. In: *Nucleic Acids Res.* 43.15, pp. 7173–7188. DOI: 10.1093/nar/gkv703.

Dardaillon, J. et al. (Nov. 2019). „ANISEED 2019: 4d exploration of genetic data for an extended range of tunicates“. In: *Nucleic Acids Res.* 48.D1, pp. D668–D675. DOI: 10.1093/nar/gkz955.

Darling, A. E., B. Mau, and N. T. Perna (June 2010). „progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement“. In: *PLoS ONE* 5.6, e11147. DOI: 10.1371/journal.pone.0011147.

Dayhoff, M. O. and R. S. Ledley (Dec. 1962). „Comprotein. a computer program to aid primary protein structure determination“. In: *Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall)*. AFIPS '62 (Fall). Philadelphia, Pennsylvania: ACM Press, pp. 262–274. DOI: 10.1145/1461518.1461546.

De Rainville, F.-M., F.-A. Fortin, M.-A. Gardner, M. Parizeau, and C. Gagné (Feb. 2014). „Deap. enabling nimbler evolutions“. In: *ACM SIGEVOlution* 6.2, pp. 17–26. DOI: 10.1145/2597453.2597455.

Dehal, P., Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D. M. Goodstein, et al. (2002). „The draft genome of Ciona intestinalis: Insights into chordate and vertebrate origins“. In: *Science* 298.5601, pp. 2157–2167.

Delsuc, F., H. Brinkmann, D. Chourrout, and H. Philippe (Feb. 2006). „Tunicates and not cephalochordates are the closest living relatives of vertebrates“. In: *Nature* 439.7079, pp. 965–968. DOI: 10.1038/nature04336.

Delsuc, F., H. Philippe, G. Tsagkogeorga, P. Simion, M.-K. Tilak, X. Turon, S. López-Legentil, J. Piette, P. Lemaire, and E. J. P. Douzery (Apr. 2018). „A phylogenomic framework and timescale for comparative studies of tunicates“. In: *BMC Biol.* 16.1, p. 39. DOI: 10.1186/s12915-018-0499-2.

Delsuc, F., G. Tsagkogeorga, N. Lartillot, and H. Philippe (Nov. 2008). „Additional molecular support for the new chordate phylogeny“. In: *Genesis* 46.11, pp. 592–604. DOI: 10.1002/dvg.20450.

Denoeud, F. et al. (Dec. 2010). „Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate“. In: *Science* 330.6009, pp. 1381–1385. DOI: 10.1126/science.1194167.

Dexheimer, P. J. and L. Cochella (June 2020). „MicroRNAs: From Mechanism to Organism“. In: *Front. Cell Dev. Biol.* 8, p. 409. DOI: 10.3389/fcell.2020.00409.

Di Genova, A., E. Buena-Atienza, S. Ossowski, and M.-F. Sagot (Dec. 2020). „Efficient hybrid de novo assembly of human genomes with WENGAN“. In: *Nat. Biotechnol.* 39.4, pp. 422–430. DOI: 10.1038/s41587-020-00747-w.

Dong, H., J. Lei, L. Ding, Y. Wen, H. Ju, and X. Zhang (May 2013). „MicroRNA: Function, Detection, and Bioanalysis“. In: *Chem. Rev.* 113.8, pp. 6207–6233. DOI: 10.1021/cr300362f.

Dunham, I. (Sept. 2005). *Genome Sequencing*. Accessed: 14.01.2021. DOI: 10.1038/npg.els.0005378.

Dyomin, A. G., E. I. Koshel, A. M. Kiselev, A. F. Saifitdinova, S. A. Galkina, T. Fukagawa, A. A. Kostareva, and E. R. Gaginskaya (June 2016). „Chicken rRNA Gene Cluster Structure“. In: *PLoS One* 11.6, e0157464. DOI: 10.1371/journal.pone.0157464.

Eddy, S. R. (Oct. 1998). „Profile hidden Markov models“. In: *Method. Biochem. Anal.* 14.9, pp. 755–763. DOI: 10.1093/bioinformatics/14.9.755.

Eddy, S. R. (Oct. 2011). „Accelerated Profile HMM Searches“. In: *PLoS Comput. Biol.* 7.10, e1002195. DOI: 10.1371/journal.pcbi.1002195.

Eddy, S. R. and R. Durbin (1994). „RNA sequence analysis using covariance models“. In: *Nucleic Acids Res.* 22.11, pp. 2079–2088. DOI: 10.1093/nar/22.11.2079.

Edgar, R. C. (Aug. 2004). „MUSCLE: A multiple sequence alignment method with reduced time and space complexity“. In: *BMC Bioinf.* 5, p. 113.

Edwards, C. A. et al. (June 2008). „The Evolution of the DLK1-DIO3 Imprinted Domain in Mammals“. In: *PLoS Biol.* 6.6, e135. DOI: 10.1371/journal.pbio.0060135.

Eggenhofer, F., I. L. Hofacker, and C. Höner zu Siederdissen (May 2013). „CMCompare webserver: Comparing RNA families via covariance models“. In: *Nucleic Acids Res.* 41.W1, W499–W503. DOI: 10.1093/nar/gkt329.

Eggenhofer, F., I. L. Hofacker, and C. Höner zu Siederdissen (June 2016). „RNAlien – Unsupervised RNA family model construction“. In: *Nucleic Acids Res.* 44.17, pp. 8433–8441. DOI: 10.1093/nar/gkw558.

Eid, J. et al. (Jan. 2009). „Real-Time DNA Sequencing from Single Polymerase Molecules“. In: *Science* 323.5910, pp. 133–138. DOI: 10.1126/science.1162986.

Ellwanger, D. C., F. A. Büttner, H.-W. Mewes, and V. Stümpflen (Mar. 2011). „The sufficient minimal set of miRNA seed types“. In: *Method. Biochem. Anal.* 27.10, pp. 1346–1350. DOI: 10.1093/bioinformatics/btr149.

ENCODE Project Consortium (Sept. 2012). „An integrated encyclopedia of DNA elements in the human genome“. In: *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247.

Ender, C., A. Krek, M. R. Friedländer, M. Beitzinger, L. Weinmann, W. Chen, S. Pfeffer, N. Rajewsky, and G. Meister (Nov. 2008). „A Human snoRNA with MicroRNA-Like Functions“. In: *Mol. Cell* 32.4, pp. 519–528. DOI: 10.1016/j.molcel.2008.10.017.

Fang, W. and D. P. Bartel (Oct. 2015). „The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes“. In: *Mol. Cell* 60.1, pp. 131–145. DOI: 10.1016/j.molcel.2015.08.015.

Farris, J. S. (Mar. 1977). „Phylogenetic Analysis Under Dollo's Law“. In: *Syst. Zool.* 26.1, p. 77. DOI: 10.2307/2412867.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello (Feb. 1998). „Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans“. In: *Nature* 391.6669, pp. 806–811. DOI: 10.1038/35888.

Fitch, W. M. (June 1970). „Distinguishing Homologous from Analogous Proteins“. In: *Syst. Zool.* 19.2, p. 99. DOI: 10.2307/2412448.

Fitch, W. M. (May 2000). „Homology“. In: *Trends Genet.* 16.5, pp. 227–231. DOI: 10.1016/s0168-9525(00)02005-9.

Fontana, W., P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster (Mar. 1993). „RNA folding and combinatory landscapes“. In: *Phys. Rev. E* 47.3, pp. 2083–2099. DOI: 10.1103/physreve.47.2083.

França, G. S., L. C. Hinske, P. A. F. Galante, and M. D. Vibranovski (Mar. 2017). „Unveiling the Impact of the Genomic Architecture on the Evolution of Vertebrate microRNAs“. In: *Front. Genet.* 8, p. 34. DOI: 10.3389/fgene.2017.00034.

França, G. S., M. D. Vibranovski, and P. A. F. Galante (Apr. 2016). „Host gene constraints and genomic context impact the expression and evolution of human microRNAs“. In: *Nat. Commun.* 7.1, p. 11438. DOI: 10.1038/ncomms11438.

Freyhult, E., P. P. Gardner, and V. Moulton (2005). „A comparison of RNA folding measures“. In: *BMC Bioinf.* 6, p. 241.

Friedländer, M. R., S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky (Sept. 2011). „miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades“. In: *Nucleic Acids Res.* 40.1, pp. 37–52. DOI: 10.1093/nar/gkr688.

Friedman, R. C., K. K.-H. Farh, C. B. Burge, and D. P. Bartel (Oct. 2008). „Most mammalian mRNAs are conserved targets of microRNAs“. In: *Genome Res.* 19.1, pp. 92–105. DOI: 10.1101/gr.082701.108.

Fromm, B., T. Billipp, et al. (Nov. 2015). „A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome“. In: *Annu. Rev. Genet.* 49.1, pp. 213–242. DOI: 10.1146/annurev-genet-120213-092023.

Fromm, B., D. Domanska, et al. (Jan. 2019). „MirGeneDB 2.0: The metazoan microRNA complement“. In: *Nucleic Acids Res.* 48.D1, pp. D132–D141.

Fromm, B., E. Høye, et al. (Nov. 2021). „MirGeneDB 2.1: toward a complete sampling of all major animal phyla“. en. In: *Nucleic Acids Res.*

Fu, X., M. Adamski, and E. M. Thompson (Jan. 2008). „Altered miRNA Repertoire in the Simplified Chordate, Oikopleura dioica“. In: *Mol. Biol. Evol.* 25.6, pp. 1067–1080. DOI: 10.1093/molbev/msn060.

Gatter, T., S. von Loehneysen, P. Drozdova, T. Hartmann, and P. F. Stadler (2020). „Economic Genome Assembly from Low Coverage Illumina and Nanopore Data“. In: *20th International Workshop on Algorithms in Bioinformatics (WABI 2020)*. Ed. by C. Kingsford and N. P. Pisanti. Leibniz International Proceedings in Informatics. bioRxiv: 939454. Schloss Dagstuhl: Dagstuhl Publishing, German, p. 10.

Gauthier, J., A. T. Vincent, S. J. Charette, and N. Derome (Aug. 2018). „A brief history of bioinformatics“. In: *Brief. Bioinform.* 20.6, pp. 1981–1996. DOI: 10.1093/bib/bby063.

Gebert, L. F. R. and I. J. MacRae (Aug. 2018). „Regulation of microRNA function in animals“. In: *Nat. Rev. Mol. Cell Bio.* 20.1, pp. 21–37. DOI: 10.1038/s41580-018-0045-7.

Ghildiyal, M. and P. D. Zamore (Feb. 2009). „Small silencing RNAs: An expanding universe“. In: *Nat. Rev. Genet.* 10.2, pp. 94–108. DOI: 10.1038/nrg2504.

GIGA Community of Scientists (Dec. 2013). „The Global Invertebrate Genomics Alliance (GIGA): Developing Community Resources to Study Diverse Invertebrate Genomes“. In: *J. Hered.* 105.1, pp. 1–18. DOI: 10.1093/jhered/est084.

Giribet, G. (Apr. 2018). „Phylogenomics resolves the evolutionary chronicle of our squirting closest relatives“. In: *BMC Biol.* 16.1, p. 49. DOI: 10.1186/s12915-018-0517-4.

Godfrey-Smith, P. and K. Sterelny (2016). „Biological Information“. In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2016. Metaphysics Research Lab, Stanford University.

Gomes, C. P. C., J.-H. Cho, L. Hood, O. L. Franco, R. W. Pereira, and K. Wang (May 2013). „A Review of Computational Tools in microRNA Discovery“. In: *Front. Genet.* 4, p. 81. DOI: 10.3389/fgene.2013.00081.

Goodwin, S., J. D. McPherson, and W. R. McCombie (May 2016). „Coming of age: Ten years of next-generation sequencing technologies“. In: *Nat. Rev. Genet.* 17.6, pp. 333–351. DOI: 10.1038/nrg.2016.49.

Grad, Y., J. Aach, G. D. Hayes, B. J. Reinhart, G. M. Church, G. Ruvkun, and J. Kim (May 2003). „Computational and Experimental Identification of C. elegans microRNAs“. In: *Mol. Cell* 11.5, pp. 1253–1263. DOI: 10.1016/s1097-2765(03)00153-9.

Gregory, R. I., K.-p. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar (Nov. 2004). „The Microprocessor complex mediates the genesis of microRNAs“. In: *Nature* 432.7014, pp. 235–240. DOI: 10.1038/nature03120.

Griffiths-Jones, S. (Jan. 2003). „Rfam: An RNA family database“. In: *Nucleic Acids Res.* 31.1, pp. 439–441. DOI: 10.1093/nar/gkg006.

Griffiths-Jones, S. (Jan. 2004). „The microRNA Registry“. In: *Nucleic Acids Res.* 32.Database issue, pp. D109–11.

Griffiths-Jones, S., J. H. L. Hui, A. Marco, and M. Ronshaugen (Feb. 2011). „MicroRNA evolution by arm switching“. In: *EMBO Rep.* 12.2, pp. 172–177.

Griffiths-Jones, S., H. K. Saini, S. van Dongen, and A. J. Enright (Jan. 2008). „miRBase: Tools for microRNA genomics“. In: *Nucleic Acids Res.* 36.Database issue, pp. D154–8.

Guth, S. I. E. and M. Wegner (May 2008). „Having it both ways: Sox protein function between conservation and innovation“. In: *Cell. Mol. Life Sci.* 65.19, pp. 3000–3018. DOI: 10.1007/s00018-008-8138-7.

Ha, M. and V. N. Kim (July 2014). „Regulation of microRNA biogenesis“. In: *Nat. Rev. Mol. Cell Bio.* 15.8, pp. 509–524. DOI: 10.1038/nrm3838.

Hackl, T., R. Hedrich, J. Schultz, and F. Förster (July 2014). „proovread : Large-scale high-accuracy PacBio correction through iterative short read consensus“. In: *Method. Biochem. Anal.* 30.21, pp. 3004–3011. DOI: 10.1093/bioinformatics/btu392.

Hamilton, A. J. and D. C. Baulcombe (Oct. 1999). „A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants“. In: *Science* 286.5441, pp. 950–952. DOI: 10.1126/science.286.5441.950.

Hammond, S. M. (Jan. 2006). „microRNA detection comes of age“. In: *Nat. Methods* 3.1, pp. 12–13. DOI: 10.1038/nmeth0106-12.

Hammond, S. M., E. Bernstein, D. Beach, and G. J. Hannon (Mar. 2000). „An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells“. In: *Nature* 404.6775, pp. 293–296. DOI: 10.1038/35005107.

Harris, R. S. (2007). *Improved Pairwise Alignment of Genomic DNA*. University Park, PA, USA.

Hecht, J. et al. (Mar. 2008). „Evolution of a Core Gene Network for Skeletogenesis in Chordates“. In: *PLoS Genet.* 4.3, e1000025. DOI: 10.1371/journal.pgen.1000025.

Heid, C. A., J. Stevens, K. J. Livak, and P. M. Williams (Oct. 1996). „Real time quantitative PCR“. In: *Genome Res.* 6.10, pp. 986–994.

Heimberg, A. M., L. F. Sempere, V. N. Moy, P. C. J. Donoghue, and K. J. Peterson (Feb. 2008). „MicroRNAs and the advent of vertebrate morphological complexity“. In: *Proc. Natl. Acad. Sci. U. S. A.* 105.8, pp. 2946–2950.

Hendrix, D., M. Levine, and W. Shi (Apr. 2010). „miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data“. In: *Genome Biol.* 11.4, R39. DOI: 10.1186/gb-2010-11-4-r39.

Hertel, J., D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater, and P. F. Stadler (Jan. 2009). „Non-coding RNA annotation of the genome of Trichoplax adhaerens“. In: *Nucleic Acids Res.* 37.5, pp. 1602–1615. DOI: 10.1093/nar/gkn1084.

Hertel, J., S. Bartschat, A. Wintsche, C. Otto, T. S. of the Bioinformatics Computer Lab, and P. F. Stadler (Mar. 2012). „Evolution of the let-7 microRNA Family“. In: *RNA Biol.* 9.3, pp. 231–241. DOI: 10.4161/rna.18974.

Hertel, J., D. Langenberger, and P. F. Stadler (Dec. 2013). „Computational Prediction of MicroRNA Genes“. In: *Methods in Molecular Biology*. Ed. by J. Gorodkin and W. L. Ruzzo. Totowa, NJ: Humana Press, pp. 437–456. DOI: 10.1007/978-1-62703-709-9_20.

Hertel, J., M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I. L. Hofacker, P. F. Stadler, and S. of Bioinformatics Computer Labs 2004 and 2005 (Feb. 2006). „The expansion of the metazoan microRNA repertoire“. In: *BMC Genomics* 7.1, p. 25. DOI: 10.1186/1471-2164-7-25.

Hertel, J. and P. Stadler (Mar. 2015). „The Expansion of Animal MicroRNA Families Revisited“. In: *Life* 5.1, pp. 905–920. DOI: 10.3390/life5010905.

Hirose, E. (July 2001). „Acid Containers and Cellular Networks in the Ascidian Tunic with Special Remarks on Ascidian Phylogeny“. In: *Zool. Sci.* 18.5, pp. 723–731. DOI: 10.2108/zsj.18.723.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster (1994). „Fast Folding and Comparison of RNA Secondary Structures“. In: *Monatsh.\ Chem.* 125, pp. 167–188.

Hoff, K. J. (Oct. 2019). „MakeHub: Fully Automated Generation of UCSC Genome Browser Assembly Hubs“. In: *Genomics, Proteomics & Bioinformatics* 17.5, pp. 546–549. DOI: 10.1016/j.gpb.2019.05.003.

Hoffmann, S., C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller (Sept. 2009). „Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures“. In: *PLoS Comput. Biol.* 5.9, e1000502. DOI: 10.1371/journal.pcbi.1000502.

Holland, L. Z. (Mar. 2014). „Genomics, evolution and development of amphioxus and tunicates: The Goldilocks principle. AMPHIOXUS AND TUNICATES“. In: *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 324.4, pp. 342–352. DOI: 10.1002/jez.b.22569.

Holland, L. Z. (Feb. 2016). „Tunicates“. In: *Curr. Biol.* 26.4, R146–R152. DOI: 10.1016/j.cub.2015.12.024.

Hu, Y., W. Lan, and D. Miller (2017). „Next-Generation Sequencing for MicroRNA Expression Profile“. en. In: *Methods Mol. Biol.* 1617, pp. 169–177.

Huerta-Cepas, J. et al. (Nov. 2018). „eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses“. In: *Nucleic Acids Res.* 47.D1, pp. D309–D314. DOI: 10.1093/nar/gky1085.

Hui, J. H. L., A. Marco, S. Hunt, J. Melling, S. Griffiths-Jones, and M. Ronshaugen (Jan. 2013). „Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods“. In: *Nucleic Acids Res.* 41.5, pp. 3352–3361. DOI: 10.1093/nar/gks1445.

Ingham, P. W. and A. P. McMahon (Dec. 2001). „Hedgehog signaling in animal development: Paradigms and principles“. In: *Gene. Dev.* 15.23, pp. 3059–3087. DOI: 10.1101/gad.938601.

International Human Genome Sequencing Consortium (Oct. 2004). „Finishing the euchromatic sequence of the human genome“. In: *Nature* 431.7011, pp. 931–945. DOI: 10.1038/nature03001.

Jackman, S. D. et al. (2017). „ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter“. In: *Genome Res.* 27, pp. 768–777.

Jefferies, R. P. (1991). „Two types of bilateral symmetry in the Metazoa: Chordate and bilaterian“. In: *Ciba Found. Symp.* 162, 94–120, discussion 121–7.

Jue, N. K., P. G. Batta-Lona, S. Trusiak, C. Obergfell, A. Bucklin, M. J. O'Neill, and R. J. O'Neill (Sept. 2016). „Rapid Evolutionary Rates and Unique Genomic Signatures Discovered in the First Reference Genome for the Southern Ocean Salp,Salpa thompsoni(Urochordata, Thaliacea)“. In: *Genome Biol. Evol.* 8.10, pp. 3171–3186. DOI: 10.1093/gbe/evw215.

Jurafsky, D. and J. H. Martin (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3ed. Draft. USA: Prentice Hall PTR.

Kalvari, I., J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, and A. I. Petrov (Jan. 2018). „Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families“. In: *Nucleic Acids Res.* 46.D1, pp. D335–D342.

Kalvari, I., E. P. Nawrocki, J. Argasinska, N. Quinones-Olvera, R. D. Finn, A. Bateman, and A. I. Petrov (June 2018). „Non-Coding RNA Analysis Using the Rfam Database“. In: *Current Protocols in Bioinformatics* 62.1, e51. DOI: 10.1002/cpbi.51.

Kalvari, I., E. P. Nawrocki, N. Ontiveros-Palacios, et al. (Nov. 2020). „Rfam 14: Expanded coverage of metagenomic, viral and microRNA families“. In: *Nucleic Acids Res.* 49.D1, pp. D192–D200. DOI: 10.1093/nar/gkaa1047.

Kassambara, A. and F. Mundt (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Katoh, K. (July 2002). „MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform“. In: *Nucleic Acids Res.* 30.14, pp. 3059–3066. DOI: 10.1093/nar/gkf436.

Kim, V. N., J. Han, and M. C. Siomi (Feb. 2009). „Biogenesis of small RNAs in animals“. In: *Nat. Rev. Mol. Cell Bio.* 10.2, pp. 126–139. DOI: 10.1038/nrm2632.

Kocot, K. M., M. G. Tassia, K. M. Halanych, and B. J. Swalla (Apr. 2018). „Phylogenomics offers resolution of major tunicate relationships“. In: *Mol. Phylogenet. Evol.* 121, pp. 166–173. DOI: 10.1016/j.ympev.2018.01.005.

Koonin, E. V. (Dec. 2005). „Orthologs, Paralogs, and Evolutionary Genomics“. In: *Annu. Rev. Genet.* 39.1, pp. 309–338. DOI: 10.1146/annurev.genet.39.073003.114725.

Korlach, J. (2015). *Understanding Accuracy in SMRT Sequencing*. https://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf. Accessed: 2021-12-30.

Koscianska, E., J. Starega-Roslan, L. J. Sznajder, M. Olejniczak, P. Galka-Marciniak, and W. J. Krzyzosiak (Apr. 2011). „Northern blotting analysis of microRNAs, their precursors and RNA interference triggers“. In: *BMC Mol. Biol.* 12.1, p. 14. DOI: 10.1186/1471-2199-12-14.

Kott, P. (Aug. 2002). „A complex didemnid ascidian from Whangamata, New Zealand“. In: *J. Mar. Biol. Assoc. Uk.* 82.4, pp. 625–628. DOI: 10.1017/s0025315402005970.

Kozomara, A. and S. Griffiths-Jones (Oct. 2010). „miRBase: Integrating microRNA annotation and deep-sequencing data“. In: *Nucleic Acids Res.* 39.Database, pp. D152–D157. DOI: 10.1093/nar/gkq1027.

Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones (Jan. 2019). „miRBase: From microRNA sequences to function“. In: *Nucleic Acids Res.* 47.D1, pp. D155–D162.

Kozomara, A. and S. Griffiths-Jones (Nov. 2013). „miRBase: Annotating high confidence microRNAs using deep sequencing data“. In: *Nucleic Acids Res.* 42.D1, pp. D68–D73. DOI: 10.1093/nar/gkt1181.

Krogh, A. (1998). „An introduction to hidden Markov models for biological sequences“. In: *Computational Methods in Molecular Biology*. Ed. by S. L. Salzberg, D. B. Searls, and S. Kasif. Vol. 32. New Comprehensive Biochemistry. Elsevier, pp. 45–63. DOI: 10.1016/s0167-7306(08)60461-5.

Kuksa, P. P., A. Amlie-Wolf, Ž. Katanić, O. Valladares, L.-S. Wang, and Y. Y. Leung (Mar. 2019). „DASHR 2.0: integrated database of human small non-coding RNA genes and mature products“. en. In: *Bioinformatics* 35.6, pp. 1033–1039.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl (Oct. 2001). „Identification of Novel Genes Coding for Small Expressed RNAs“. In: *Science* 294.5543, pp. 853–858. DOI: 10.1126/science.1064921.

Lagos-Quintana, M., R. Rauhut, J. Meyer, A. Borkhardt, and T. Tuschl (Feb. 2003). „New microRNAs from mouse and human“. In: *RNA* 9.2, pp. 175–179.

Lai, E. C., P. Tomancak, R. W. Williams, and G. M. Rubin (June 2003). „Computational identification of Drosophila microRNA genes“. In: *Genome Biol.* 4.7, R42. DOI: 10.1186/gb-2003-4-7-r42.

Lam, J. K. W., M. Y. T. Chow, Y. Zhang, and S. W. S. Leung (Sept. 2015). „siRNA Versus miRNA as Therapeutics for Gene Silencing“. In: *Molecular Therapy - Nucleic Acids* 4, e252. DOI: 10.1038/mtna.2015.23.

Lambert, A., J.-F. Fontaine, M. Legendre, F. Leclerc, E. Permal, F. Major, H. Putzer, O. Delfour, B. Michot, and D. Gautheret (July 2004). „The ERPIN server: An interface to profile-based RNA motif identification“. In: *Nucleic Acids Res.* 32.Web Server, W160–W165. DOI: 10.1093/nar/gkh418.

Lambert, G. (2009). „Adventures of a sea squirt sleuth: Unraveling the identity of Didemnum vexillum, a global ascidian invader“. In: *Aquat. Invasions* 4.1, pp. 5–28. DOI: 10.3391/ai.2009.4.1.2.

Langenberger, D., M. V. Çakir, S. Hoffmann, and P. F. Stadler (Nov. 2012). „Dicer-Processed Small RNAs: Rules and Exceptions. DICER-PROCESSED SMALL RNAs“. In: *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 320.1, pp. 35–46. DOI: 10.1002/jez.b.22481.

Lau, N. C., L. P. Lim, E. G. Weinstein, and D. P. Bartel (Oct. 2001). „An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans“. In: *Science* 294.5543, pp. 858–862. DOI: 10.1126/science.1065062.

Lee, R. C. and V. Ambros (Oct. 2001). „An Extensive Class of Small RNAs in Caenorhabditis elegans“. In: *Science* 294.5543, pp. 862–864. DOI: 10.1126/science.1065329.

Lee, R. C., R. L. Feinbaum, and V. Ambros (Dec. 1993). „The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14“. In: *Cell* 75.5, pp. 843–854. DOI: 10.1016/0092-8674(93)90529-y.

Legendre, M., A. Lambert, and D. Gautheret (Oct. 2004). „Profile-based detection of microRNA precursors in animal genomes“. In: *Method. Biochem. Anal.* 21.7, pp. 841–845. DOI: 10.1093/bioinformatics/bti073.

Lemaire, P. and J. Piette (June 2015). „Tunicates: Exploring the sea shores and roaming the open ocean. A tribute to Thomas Huxley“. In: *Open Biology* 5.6, p. 150053. DOI: 10.1098/rsob.150053.

Li, H. and R. Durbin (May 2009). „Fast and accurate short read alignment with Burrows–Wheeler transform“. In: *Method. Biochem. Anal.* 25.14, pp. 1754–1760.

Li, L. and Y. Liu (2011). „Diverse small non-coding RNAs in RNA interference pathways“. In: *Methods Mol Biol.* 764, pp. 169–182.

Li, W. and K. Ruan (Jan. 2009). „MicroRNA detection by microarray“. In: *Anal. Bioanal. Chem.* 394.4, pp. 1117–1124. DOI: 10.1007/s00216-008-2570-2.

Li, Y., Z. Zhang, F. Liu, W. Vongsangnak, Q. Jing, and B. Shen (Jan. 2012). „Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis“. In: *Nucleic Acids Res.* 40.10, pp. 4298–4305. DOI: 10.1093/nar/gks043.

Liang, H. and W.-H. Li (Mar. 2009). „Lowly Expressed Human MicroRNA Genes Evolve Rapidly“. In: *Mol. Biol. Evol.* 26.6, pp. 1195–1198. DOI: 10.1093/molbev/msp053.

Liang, T., L. Guo, and C. Liu (2012). „Genome-Wide Analysis of mir-548 Gene Family Reveals Evolutionary and Functional Implications“. In: *J. Biomed. Biotechnol.* 2012, pp. 1–8. DOI: 10.1155/2012/679563.

Liang, T., C. Yang, P. Li, C. Liu, and L. Guo (Nov. 2014). „Genetic Analysis of Loop Sequences in the Let-7 Gene Family Reveal a Relationship between Loop Evolution and Multiple IsomiRs“. In: *PLoS ONE* 9.11, e113042. DOI: 10.1371/journal.pone.0113042.

Lim, L. P., N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel (Apr. 2003). „The microRNAs ofCaenorhabditis elegans“. In: *Gene. Dev.* 17.8, pp. 991–1008. DOI: 10.1101/gad.1074403.

Lin, H. and A. Spradling (June 1997). „A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary“. In: *Development* 124.12, pp. 2463–2476. DOI: 10.1242/dev.124.12.2463.

Lindahl, T. and B. Nyberg (July 1974). „Heat-induced deamination of cytosine residues in deoxyribonucleic acid“. In: *Biochemistry-us.* 13.16, pp. 3405–3410. DOI: 10.1021/bi00713a035.

Liu, C.-G., G. A. Calin, B. Meloon, et al. (June 2004). „An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues“. In: *Proc. Natl. Acad. Sci. U. S. A.* 101.26, pp. 9740–9744.

Liu, C.-G., G. A. Calin, S. Volinia, and C. M. Croce (Mar. 2008). „MicroRNA expression profiling using microarrays“. In: *Nat. Protoc.* 3.4, pp. 563–578. DOI: 10.1038/nprot.2008.14.

Lorenz, R., S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker (Nov. 2011). „ViennaRNA Package 2.0“. In: *Algorithm. Mol. Biol.* 6.1, p. 26. DOI: 10.1186/1748-7188-6-26.

Lott, S. C., R. A. Schäfer, M. Mann, R. Backofen, W. R. Hess, B. Voß, and J. Georg (Apr. 2018). „GLASSgo – Automated and Reliable Detection of sRNA Homologs From a Single Input Sequence“. In: *Front. Genet.* 9, p. 124. DOI: 10.3389/fgene.2018.00124.

Lukasik, A., M. Wójcikowski, and P. Zielenkiewicz (Apr. 2016). „Tools4miRs – one place to gather all the tools for miRNA analysis“. In: *Method. Biochem. Anal.* 32.17, pp. 2722–2724. DOI: 10.1093/bioinformatics/btw189.

Mack, G. S. (June 2007). „MicroRNA gets down to business“. In: *Nat. Biotechnol.* 25.6, pp. 631–638.

Maggiolini, F. A. M. et al. (Mar. 2019). „Genomic inversions and GOLGA core duplicons underlie disease instability at the 15q25 locus“. In: *PLOS Genet.* 15.3, e1008075. DOI: 10.1371/journal.pgen.1008075.

Marco, A., M. Ninova, M. Ronshaugen, and S. Griffiths-Jones (June 2013). „Clusters of microRNAs emerge by new hairpins in existing transcripts“. In: *Nucleic Acids Res.* 41.16, pp. 7745–7752. DOI: 10.1093/nar/gkt534.

Margulies, M. et al. (July 2005). „Genome sequencing in microfabricated high-density picolitre reactors“. In: *Nature* 437.7057, pp. 376–380. DOI: 10.1038/nature03959.

Marz, M., T. Kirsten, and P. F. Stadler (Nov. 2008). „Evolution of Spliceosomal snRNA Genes in Metazoan Animals“. In: *J. Mol. Evol.* 67.6, pp. 594–607. DOI: 10.1007/s00239-008-9149-6.

Menzel, P., J. Gorodkin, and P. F. Stadler (Oct. 2009). „The tedious task of finding homologous noncoding RNA genes“. In: *RNA* 15.12, pp. 2075–2082. DOI: 10.1261/rna.1556009.

Michel, J.-B. et al. (Jan. 2011). „Quantitative Analysis of Culture Using Millions of Digitized Books“. In: *Science* 331.6014, pp. 176–182. DOI: 10.1126/science.1199644.

Michlewski, G. and J. F. Cáceres (Jan. 2019). „Post-transcriptional control of miRNA biogenesis“. en. In: *RNA* 25.1, pp. 1–16.

Missal, K., D. Rose, and P. F. Stadler (Sept. 2005). „Non-coding RNAs in Ciona intestinalis“. In: *Method. Biochem. Anal.* 21.Suppl 2, pp. ii77–ii78. DOI: 10.1093/bioinformatics/bti1113.

Monteys, A. M., R. M. Spengler, J. Wan, L. Tecedor, K. A. Lennox, Y. Xing, and B. L. Davidson (Jan. 2010). „Structure and activity of putative intronic miRNA promoters“. In: *RNA* 16.3, pp. 495–505. DOI: 10.1261/rna.1731910.

Morgenstern, B. (July 2004). „DIALIGN: Multiple DNA and protein sequence alignment at BiBiServ“. In: *Nucleic Acids Res.* 32.Web Server issue, W33–6.

Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. CSHL Press. ISBN: 9780879697129.

Moussian, B., H. Schoof, A. Haecker, G. Jürgens, and T. Laux (Mar. 1998). „Role of the ZWILLE gene in the regulation of central shoot meristem cell fate during Arabidopsis embryogenesis“. en. In: *EMBO J.* 17.6, pp. 1799–1809.

Nah, G. S. S., B.-H. Tay, S. Brenner, M. Osato, and B. Venkatesh (Nov. 2014). „Characterization of the Runx Gene Family in a Jawless Vertebrate, the Japanese Lamprey (Lethenteron japonicum)“. In: *PLOS ONE* 9.11, pp. 1–13. DOI: 10.1371/journal.pone.0113445.

Nakashima, K., L. Yamada, Y. Satou, J.-i. Azuma, and N. Satoh (Feb. 2004). „The evolutionary origin of animal cellulose synthase“. In: *Dev. Genes Evol.* 214.2, pp. 81–88. DOI: 10.1007/s00427-003-0379-8.

Nawrocki, E. P. and S. R. Eddy (Sept. 2013). „Infernal 1.1: 100-fold faster RNA homology searches“. In: *Method. Biochem. Anal.* 29.22, pp. 2933–2935. DOI: 10.1093/bioinformatics/btt509.

Nawrocki, E. P. and S. Eddy (2005). „Query-Dependent Banding (QDB) for Faster RNA Similarity Searches“. In: *PLoS Comput. Biol.* preprint.2007, e56. DOI: 10.1371/journal.pcbi.0030056.eor.

Needleman, S. B. and C. D. Wunsch (Mar. 1970). „A general method applicable to the search for similarities in the amino acid sequence of two proteins“. In: *J. Mol. Biol.* 48.3, pp. 443–453. DOI: 10.1016/0022-2836(70)90057-4.

Ng Kwang Loong, S. and S. K. Mishra (Feb. 2007). „Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification“. In: *RNA* 13.2, pp. 170–187.

Ng, S.-Y., G. K. Bogu, B. S. Soh, and L. W. Stanton (Aug. 2013). „The Long Noncoding RNA RMST Interacts with SOX2 to Regulate Neurogenesis“. In: *Mol. Cell* 51.3, pp. 349–359. DOI: 10.1016/j.molcel.2013.07.017.

Nguyen, T. A., M. H. Jo, Y.-G. Choi, J. Park, S. C. Kwon, S. Hohng, V. N. Kim, and J.-S. Woo (June 2015). „Functional Anatomy of the Human Microprocessor“. In: *Cell* 161.6, pp. 1374–1387. DOI: 10.1016/j.cell.2015.05.010.

Nobel Prize Outreach AB (2021a). *The Nobel Prize in Chemistry*. https://www.nobelprize.org/prizes/medicine/1980/summary. Accessed: 2021-12-01.

Nobel Prize Outreach AB (2021b). *The Nobel Prize in Physiology or Medicine*. https://www.nobelprize.org/prizes/medicine/2006/summary. Accessed: 2021-11-01.

Norden-Krichmar, T. M., J. Holtz, A. E. Pasquinelli, and T. Gaasterland (Nov. 2007). „Computational prediction and experimental validation of Ciona intestinalis microRNA genes“. In: *Bmc Genomics* 8.1, p. 445. DOI: 10.1186/1471-2164-8-445.

Notredame, C., D. G. Higgins, and J. Heringa (Sept. 2000). „T-coffee: A novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton“. In: *J. Mol. Biol.* 302.1, pp. 205–217. DOI: 10.1006/jmbi.2000.4042.

Nozawa, M., S. Miura, and M. Nei (2010). „Origins and Evolution of MicroRNA Genes in Drosophila Species“. In: *Genome Biol. Evol.* 2, pp. 180–189.

Okamura, K. (Nov. 2011). „Diversity of animal small RNA pathways and their biological utility. Diversity of animal small RNA pathways“. In: *Wiley Interdisciplinary Reviews: RNA* 3.3, pp. 351–368. DOI: 10.1002/wrna.113.

Okamura, K., E. Ladewig, L. Zhou, and E. C. Lai (Mar. 2013). „Functional small RNAs are generated from select miRNA hairpin loops in flies and mammals“. In: *Gene. Dev.* 27.7, pp. 778–792. DOI: 10.1101/gad.211698.112.

Ordóñez, V., M. Pascual, M. Fernández-Tejedor, M. C. Pineda, D. Tagliapietra, and X. Turon (Mar. 2015). „Ongoing expansion of the worldwide invader Didemnum vexillum (Ascidiacea) in the Mediterranean Sea: High plasticity of its biological cycle promotes establishment in warm waters“. In: *Biol. Invasions* 17.7, pp. 2075–2085. DOI: 10.1007/s10530-015-0861-z.

Oxford Nanopore Technologies (2022). *How nanopore sequencing works.* https://nanoporetech.com/how-it-works. Accessed: 2022-01-15.

Ozata, D. M., I. Gainetdinov, A. Zoch, D. O'Carroll, and P. D. Zamore (Nov. 2018). „PIWI-interacting RNAs: Small RNAs with big functions“. In: *Nat. Rev. Genet.* 20.2, pp. 89–108. DOI: 10.1038/s41576-018-0073-3.

Park, J., K. Xu, T. Park, and S. V. Yi (Sept. 2011). „What are the determinants of gene expression levels and breadths in the human genome?“ In: *Hum. Mol. Genet.* 21.1, pp. 46–56. DOI: 10.1093/hmg/ddr436.

Parra-Rincón, E., C. A. Velandia-Huerto, A. Gittenberger, J. Fallmann, T. Gatter, F. D. Brown, P. F. Stadler, and C. I. Bermúdez-Santana (Dec. 2021). „The Genome of the "Sea Vomit" Didemnum vexillum“. In: *Life* 11.12, p. 1377. DOI: 10.3390/life11121377.

Pasquinelli, A. E. et al. (Nov. 2000). „Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA“. In: *Nature* 408.6808, pp. 86–89. DOI: 10.1038/35040556.

Pearson, W. R. and D. J. Lipman (Apr. 1988). „Improved tools for biological sequence comparison.“ In: *Proc. Natl. Acad. Sci.* 85.8, pp. 2444–2448. DOI: 10.1073/pnas.85.8.2444.

Pearson, W. R. (June 2013). „An introduction to sequence similarity ("homology") searching“. In: *Curr. Protoc. Bioinformatics* Chapter 3, Unit3.1.

Pertsemlidis, A. and J. W. Fondon (Sept. 2001). „Having a BLAST with bioinformatics (and avoiding BLASTphemy)“. In: *Genome Biol.* 2.10, pp. 1–10. DOI: 10.1186/gb-2001-2-10-reviews2002.

Piriyapongsa, J. and I. K. Jordan (Feb. 2007). „A Family of Human MicroRNA Genes from Miniature Inverted-Repeat Transposable Elements“. In: *PLoS ONE* 2.2, e203. DOI: 10.1371/journal.pone.0000203.

Pritchard, C. C., H. H. Cheng, and M. Tewari (Apr. 2012). „MicroRNA profiling: Approaches and considerations“. In: *Nat. Rev. Genet.* 13.5, pp. 358–369. DOI: 10.1038/nrg3198.

Putnam, N. H. et al. (June 2008). „The amphioxus genome and the evolution of the chordate karyotype“. In: *Nature* 453.7198, pp. 1064–1071. DOI: 10.1038/nature06967.

Quesne, W. J. L. (Dec. 1974). „The Uniquely Evolved Character Concept and its Cladistic Application“. In: *Syst. Zool.* 23.4, p. 513. DOI: 10.2307/2412469.

Quinlan, A. R. and I. M. Hall (Jan. 2010). „BEDTools: A flexible suite of utilities for comparing genomic features“. In: *Method. Biochem. Anal.* 26.6, pp. 841–842. DOI: 10.1093/bioinformatics/btq033.

Rabiner, L. (Feb. 1989). „A tutorial on hidden Markov models and selected applications in speech recognition“. In: *Proc. IEEE* 77.2, pp. 257–286. DOI: 10.1109/5.18626.

Raney, B. J. et al. (Nov. 2013). „Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser“. In: *Method. Biochem. Anal.* 30.7, pp. 1003–1005. DOI: 10.1093/bioinformatics/btt637.

Redditt, P. L., D. C. Braund, and F.-o. Bovon (Nov. 1992). *Philip (Person)*. DOI: 10.5040/9780300261912-0282.

Reiche, K. and P. F. Stadler (May 2007). „RNAstrand: Reading direction of structured RNAs in multiple sequence alignments“. In: *Algorithm. Mol. Biol.* 2.1, p. 6. DOI: 10.1186/1748-7188-2-6.

Reichholf, B., V. A. Herzog, N. Fasching, R. A. Manzenreither, I. Sowemimo, and S. L. Ameres (Aug. 2019). „Time-Resolved Small RNA Sequencing Unravels the Molecular Principles of MicroRNA Homeostasis“. In: *Mol. Cell* 75.4, 756–768.e7. DOI: 10.1016/j.molcel.2019.06.018.

Reinhart, B. J., F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun (Feb. 2000). „The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans“. In: *Nature* 403.6772, pp. 901–906. DOI: 10.1038/35002607.

Rhoads, A. and K. F. Au (Oct. 2015). „PacBio Sequencing and Its Applications“. In: *Genomics, Proteomics & Bioinformatics* 13.5, pp. 278–289. DOI: 10.1016/j.gpb.2015.08.002.

Rice, P., I. Longden, and A. Bleasby (June 2000). „EMBOSS: The European Molecular Biology Open Software Suite“. In: *Trends Genet.* 16.6, pp. 276–277. DOI: 10.1016/s0168-9525(00)02024-2.

Rinkevich, B. (Sept. 2005). „Natural chimerism in colonial urochordates“. In: *J. Exp. Mar. Biol. Ecol.* 322.2, pp. 93–109. DOI: 10.1016/j.jembe.2005.02.020.

Rinkevich, B. and A. Fidler (Mar. 2014). „Initiating laboratory culturing of the invasive ascidian Didemnum vexillum“. In: *Management of Biological Invasions* 5.1, pp. 55–62. DOI: 10.3391/mbi.2014.5.1.05.

RNAcentral Consortium (Jan. 2019). „RNAcentral: A hub of information for non-coding RNA sequences“. In: *Nucleic Acids Res.* 47.D1, pp. D1250–D1251.

Roden, C. et al. (Jan. 2017). „Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation“. In: *Genome Res.* 27.3, pp. 374–384. DOI: 10.1101/gr.208900.116.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén (Nov. 1996). „Real-Time DNA Sequencing Using Detection of Pyrophosphate Release“. In: *Anal. Biochem.* 242.1, pp. 84–89. DOI: 10.1006/abio.1996.0432.

Rougvie, A. E. (Sept. 2001). „Control of developmental timing in animals“. In: *Nat. Rev. Genet.* 2.9, pp. 690–701. DOI: 10.1038/35088566.

Roush, S. and F. J. Slack (Oct. 2008). „The let-7 family of microRNAs“. In: *Trends Cell Biol.* 18.10, pp. 505–516. DOI: 10.1016/j.tcb.2008.07.007.

Ruan, J. and H. Li (Dec. 2019). „Fast and accurate long-read assembly with wtdbg2“. In: *Nat. Methods* 17.2, pp. 155–158. DOI: 10.1038/s41592-019-0669-3.

Ruby, J. G., C. H. Jan, and D. P. Bartel (June 2007). „Intronic microRNA precursors that bypass Drosha processing“. In: *Nature* 448.7149, pp. 83–86. DOI: 10.1038/nature05983.

Rüegger, S. and H. Großhans (Oct. 2012). „MicroRNA turnover: When, how, and why“. In: *Trends Biochem. Sci.* 37.10, pp. 436–446. DOI: 10.1016/j.tibs.2012.07.002.

Ruppert, E. E. (Jan. 2005). „Key characters uniting hemichordates and chordates: Homologies or homoplasies?“ In: *Can. J. Zoolog.* 83.1, pp. 8–23. DOI: 10.1139/z04-158.

Saçar Demirci, M. D., J. Baumbach, and J. Allmer (Aug. 2017). „On the performance of pre-microRNA detection algorithms“. In: *Nat. Commun.* 8.1, p. 330. DOI: 10.1038/s41467-017-00403-z.

Saçar, M. D. and J. Allmer (2014). „Machine learning methods for microRNA gene prediction“. In: *Methods Mol Biol* 1107, pp. 177–187.

Sanger, F., S. Nicklen, and A. R. Coulson (Dec. 1977). „DNA sequencing with chain-terminating inhibitors“. In: *Proc. Natl. Acad. Sci.* 74.12, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

Satoh, N. and M. Levine (2005). „Surfing with the tunicates into the post-genome era“. In: *Genes and Development*.

Satou, Y., R. Nakamura, D. Yu, R. Yoshida, M. Hamada, M. Fujie, K. Hisata, H. Takeda, and N. Satoh (Oct. 2019). „A Nearly Complete Genome of Ciona intestinalis Type A (C. robusta) Reveals the Contribution of Inversion to Chromosomal Evolution in the Genus Ciona“. In: *Genome Biol. Evol.* 11.11, pp. 3144–3157. DOI: 10.1093/gbe/evz228.

Sayers, E. W. et al. (Nov. 2009). „Database resources of the National Center for Biotechnology Information“. In: *Nucleic Acids Res.* 38.suppl_1, pp. D5–D16. DOI: 10.1093/nar/gkp967.

Schuster, S. C. (Dec. 2007). „Next-generation sequencing transforms today's biology“. In: *Nat. Methods* 5.1, pp. 16–18. DOI: 10.1038/nmeth1156.

Schwartz, D. C. and C. R. Cantor (May 1984). „Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis“. In: *Cell* 37.1, pp. 67–75. DOI: 10.1016/0092-8674(84)90301-5.

Seitz, H., H. Royo, M.-L. Bortolin, S.-P. Lin, A. C. Ferguson-Smith, and J. Cavaillé (Sept. 2004). „A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain“. en. In: *Genome Res.* 14.9, pp. 1741–1748.

Seitz, H., N. Youngson, S.-P. Lin, S. Dalbert, M. Paulsen, J.-P. Bachellerie, A. C. Ferguson-Smith, and J. Cavaillé (June 2003). „Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene“. In: *Nat. Genet.* 34.3, pp. 261–262. DOI: 10.1038/ng1171.

Sekigami, Y., T. Kobayashi, A. Omi, K. Nishitsuji, T. Ikuta, A. Fujiyama, N. Satoh, and H. Saiga (2017). „Hox gene cluster of the ascidian, *Halocynthia roretzi*, reveals multiple ancient steps of cluster disintegration during ascidian evolution“. In: *Zoological Lett.* 3, p. 17.

Sekigami, Y., T. Kobayashi, A. Omi, K. Nishitsuji, T. Ikuta, A. Fujiyama, N. Satoh, and H. Saiga (2019). „Note to: Hox gene cluster of the ascidian, *Halocynthia roretzi*, reveals multiple ancient steps of cluster disintegration during ascidian evolution“. In: *Zoological Lett.* 5, p. 8.

Sempere, L. F., C. N. Cole, M. A. Mcpeek, and K. J. Peterson (Nov. 2006). „The phylogenetic distribution of metazoan microRNAs: Insights into evolutionary complexity and constraint“. In: *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 306B.6, pp. 575–588. DOI: 10.1002/jez.b.21118.

Shapiro, R. and R. S. Klein (July 1966). „The Deamination of Cytidine and Cytosine by Acidic Buffer Solutions. Mutagenic Implications*“. In: *Biochemistry-us.* 5.7, pp. 2358–2362. DOI: 10.1021/bi00871a026.

Shendure, J. and H. Ji (Oct. 2008). „Next-generation DNA sequencing“. In: *Nat. Biotechnol.* 26.10, pp. 1135–1145. DOI: 10.1038/nbt1486.

Shi, W., D. Hendrix, M. Levine, and B. Haley (Jan. 2009). „A distinct class of small RNAs arises from pre-miRNA–proximal regions in a simple chordate“. In: *Nat. Struct. Mol. Biol.* 16.2, pp. 183–189. DOI: 10.1038/nsmb.1536.

Shimeld, S. M. (Jan. 1999). „The evolution of the hedgehog gene family in chordates: Insights from amphioxus hedgehog“. In: *Dev. Genes Evol.* 209.1, pp. 40–47. DOI: 10.1007/s004270050225.

Siebert, S. and R. Backofen (June 2005). „MARNA: Multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons“. In: *Method. Biochem. Anal.* 21.16, pp. 3352–3359. DOI: 10.1093/bioinformatics/bti550.

Siederdissen, C. H. z. and I. L. Hofacker (Sept. 2010). „Discriminatory power of RNA family models“. In: *Method. Biochem. Anal.* 26.18, pp. i453–i459. DOI: 10.1093/bioinformatics/btq370.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov (June 2015). „BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs“. In: *Method. Biochem. Anal.* 31.19, pp. 3210–3212. DOI: 10.1093/bioinformatics/btv351.

Slatko, B. E., A. F. Gardner, and F. M. Ausubel (Apr. 2018). „Overview of Next-Generation Sequencing Technologies“. In: *Current Protocols in Molecular Biology* 122.1, e59. DOI: 10.1002/cpmb.59.

Smith, J. M. (June 2000). „The Concept of Information in Biology“. In: *Roy. I. Philos. Suppl.* 67.2, pp. 177–194. DOI: 10.1086/392768.

Smith, K. F., C. L. Abbott, Y. Saito, and A. E. Fidler (Feb. 2015). „Comparison of whole mitochondrial genome sequences from two clades of the invasive ascidian, Didemnum vexillum“. In: *Mar. Genom.* 19, pp. 75–83. DOI: 10.1016/j.margen.2014.11.007.

Smith, K. F., L. Stefaniak, Y. Saito, C. E. C. Gemmill, S. C. Cary, and A. E. Fidler (Jan. 2012). „Increased Inter-Colony Fusion Rates Are Associated with Reduced COI Haplotype Diversity in an Invasive Colonial Ascidian Didemnum vexillum“. In: *PLoS ONE* 7.1, e30473. DOI: 10.1371/journal.pone.0030473.

Smith, T. and M. Waterman (Mar. 1981). „Identification of common molecular subsequences“. In: *J. Mol. Biol.* 147.1, pp. 195–197. DOI: 10.1016/0022-2836(81)90087-5.

Song, R., S. Ro, and W. Yan (2010). „In situ hybridization detection of microRNAs“. In: *Methods Mol. Biol.* 629, pp. 287–294.

Stach, T. (Oct. 2008). „Chordate phylogeny and evolution: A not so simple three-taxon problem“. In: *J. Zool.* 276.2, pp. 117–141. DOI: 10.1111/j.1469-7998.2008.00497.x.

Stark, A., N. Bushati, C. H. Jan, P. Kheradpour, E. Hodges, J. Brennecke, D. P. Bartel, S. M. Cohen, and M. Kellis (Jan. 2008). „A single Hox locus in Drosophila produces

functional microRNAs from opposite DNA strands“. In: *Gene. Dev.* 22.1, pp. 8–13. DOI: 10.1101/gad.1613108.

Stefaniak, L. M. (2012). „Didemnum vexillum: Identity, origin, and life history of an invasive ascidian“. PhD thesis. University of Connecticut.

Stefaniak, L., H. Zhang, A. Gittenberger, K. Smith, K. Holsinger, S. Lin, and R. B. Whitlatch (July 2012). „Determining the native region of the putatively invasive ascidian Didemnum vexillum Kott, 2002“. In: *J. Exp. Mar. Biol. Ecol.* 422-423, pp. 64–71. DOI: 10.1016/j.jembe.2012.04.012.

Sung, W.-K. (Nov. 2009). *Algorithms in Bioinformatics. A Practical Introduction*. Chapman and Hall/CRC. ISBN: 9780429141492. DOI: 10.1201/9781420070347.

Svoboda, P. (May 2015). „A toolbox for miRNA analysis“. In: *FEBS Lett* 589.14, pp. 1694–1701. DOI: 10.1016/j.febslet.2015.04.054.

Swalla, B. J., C. B. Cameron, L. S. Corley, and J. R. Garey (Jan. 2000). „Urochordates Are Monophyletic Within the Deuterostomes“. In: *Systematic Biol.* 49.1, pp. 52–64. DOI: 10.1080/10635150050207384.

Swalla, B. J. and A. B. Smith (Jan. 2008). „Deciphering deuterostome phylogeny: Molecular, morphological and palaeontological perspectives“. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1496, pp. 1557–1568. DOI: 10.1098/rstb.2007.2246.

Takatori, N., Y. Satou, and N. Satoh (Aug. 2002). „Expression of hedgehog genes in Ciona intestinalis embryos“. In: *Mech. Develop.* 116.1-2, pp. 235–238. DOI: 10.1016/s0925-4773(02)00150-8.

Tanzer, A. and P. F. Stadler (May 2004). „Molecular Evolution of a MicroRNA Cluster“. In: *J. Mol. Biol.* 339.2, pp. 327–335. DOI: 10.1016/j.jmb.2004.03.065.

Tarver, J. E., E. A. Sperling, A. Nailor, A. M. Heimberg, J. M. Robinson, B. L. King, D. Pisani, P. C. Donoghue, and K. J. Peterson (Aug. 2013). „miRNAs: Small Genes with Big Potential in Metazoan Phylogenetics“. In: *Mol. Biol. Evol.* 30.11, pp. 2369–2382. DOI: 10.1093/molbev/mst133.

Tarver, J. E., R. S. Taylor, M. N. Puttick, G. T. Lloyd, W. Pett, B. Fromm, B. E. Schirrmeister, D. Pisani, K. J. Peterson, and P. C. J. Donoghue (May 2018). „Well-Annotated microRNAomes Do Not Evidence Pervasive miRNA Loss“. In: *Genome Biol. Evol.* 10.6, pp. 1457–1470. DOI: 10.1093/gbe/evy096.

The RNAcentral Consortium (Nov. 2018). „RNAcentral: A hub of information for non-coding RNA sequences“. In: *Nucleic Acids Res.* 47.D1, pp. D221–D229. DOI: 10.1093/nar/gky1034.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (Nov. 1994). „CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice“. In: *Nucleic Acids Res.* 22.22, pp. 4673–4680. DOI: 10.1093/nar/22.22.4673.

Thomson, R. C., D. C. Plachetzki, D. L. Mahler, and B. R. Moore (July 2014). „A critical appraisal of the use of microRNA data in phylogenetics“. In: *Proc. Natl. Acad. Sci.* 111.35, E3659–E3668. DOI: 10.1073/pnas.1407207111.

Treiber, T., N. Treiber, and G. Meister (Sept. 2018). „Regulation of microRNA biogenesis and its crosstalk with other cellular pathways“. In: *Nat. Rev. Mol. Cell Bio.* 20.1, pp. 5–20. DOI: 10.1038/s41580-018-0059-1.

Tsagkogeorga, G., V. Cahais, and N. Galtier (2012). „The Population Genomics of a Fast Evolver: High Levels of Diversity, Functional Constraint, and Molecular Adaptation in the Tunicate Ciona intestinalis“. In: *Genome Biol. Evol.* 4.8, pp. 852–861. DOI: 10.1093/gbe/evs054.

Tsagkogeorga, G., X. Turon, N. Galtier, E. J. P. Douzery, and F. Delsuc (Aug. 2010). „Accelerated Evolutionary Rate of Housekeeping Genes in Tunicates“. In: *J. Mol. Evol.* 71.2, pp. 153–167. DOI: 10.1007/s00239-010-9372-9.

Ustalov, D., A. Panchenko, C. Biemann, and S. P. Ponzetto (Sept. 2019). „Watset: Local-global Graph Clustering with Applications in Sense and Frame Induction“. In: *Comput. Linguist.* 45.3, pp. 423–479. DOI: 10.1162/coli_a_00354.

Valentine, P. C., M. R. Carman, D. S. Blackwood, and E. J. Heffron (Mar. 2007). „Ecological observations on the colonial ascidian Didemnum sp. in a New England tide pool habitat“. In: *J. Exp. Mar. Biol. Ecol.* 342.1. Proceedings of the 1st International Invasive Sea Squirt Conference, pp. 109–121. DOI: 10.1016/j.jembe.2006.10.021.

Velandia-Huerto, C. A., F. D. Brown, A. Gittenberger, P. F. Stadler, and C. I. Bermúdez-Santana (July 2018). „Nonprotein-Coding RNAs as Regulators of Development in Tunicates“. In: *Results and Problems in Cell Differentiation*. Ed. by M. Kloc and J. Kubiak. Vol. 65. Results Probl Cell Differ. Cham: Springer International Publishing, pp. 197–225. DOI: 10.1007/978-3-319-92486-1_11.

Velandia-Huerto, C. A., J. Fallmann, and P. F. Stadler (Feb. 2021). „miRNAture–Computational Detection of microRNA Candidates“. In: *Genes* 12.3, p. 348. DOI: 10.3390/genes12030348.

Velandia-Huerto, C. A., J. Fallmann, and P. F. Stadler (in prep.). „The bona fide miRNA complement on tunicates, based on an automatizated homology approach“.

Velandia-Huerto, C. A., A. M. Yazbeck, J. Schor, and P. F. Stadler (2022). „Evolution and Phylogeny of MicroRNAs — Protocols, Pitfalls, and Problems“. In: *miRNomics: MicroRNA Biology and Computational Analysis*. Ed. by J. Allmer and M. Yousef. New York, NY: Springer US, pp. 211–233. ISBN: 978-1-0716-1170-8. DOI: 10.1007/978-1-0716-1170-8_11.

Velandia-Huerto, C. A., S. J. Berkemer, A. Hoffmann, N. Retzlaff, L. C. Romero Marroquín, M. Hernández-Rosales, P. F. Stadler, and C. I. Bermúdez-Santana (Aug. 2016). „Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies“. In: *Bmc Genomics* 17.1, p. 617. DOI: 10.1186/s12864-016-2927-4.

Velandia-Huerto, C. A., A. A. Gittenberger, F. D. Brown, P. F. Stadler, and C. I. Bermúdez-Santana (Aug. 2016). „Automated detection of ncRNAs in the draft genome sequence of a colonial tunicate: The carpet sea squirt Didemnum vexillum“. In: *Bmc Genomics* 17.1, p. 691. DOI: 10.1186/s12864-016-2934-5.

Vogelstein, B. and D. Gillespie (Feb. 1979). „Preparative and analytical purification of DNA from agarose.“ In: *Proc. Natl. Acad. Sci.* 76.2, pp. 615–619. DOI: 10.1073/pnas.76.2.615.

Voskoboynik, A. et al. (July 2013). „The genome sequence of the colonial chordate, Botryllus schlosseri“. In: *eLife* 2, e00569. DOI: 10.7554/elife.00569.

Wagner, D. O. and P. Aspenberg (June 2011). „Where did bone come from? An overview of its evolution“. In: *Acta Orthop.* 82.4. PMID: 21657973, pp. 393–398. DOI: 10.3109/17453674.2011.588861.

Wang, L.-G. et al. (Oct. 2019). „Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data“. In: *Mol. Biol. Evol.* 37.2, pp. 599–603. DOI: 10.1093/molbev/msz240.

Wang, K., C. Dantec, P. Lemaire, T. A. Onuma, and H. Nishida (Apr. 2017). „Genome-wide survey of miRNAs and their evolutionary history in the ascidian, Halocynthia roretzi“. In: *Bmc Genomics* 18.1, p. 314. DOI: 10.1186/s12864-017-3707-5.

Wang, L., Y. Liu, S. Sun, M. Lu, and Y. Xia (July 2016). „Regulation of neuronal-glial fate specification by long non-coding RNAs“. In: *Rev. Neurosci.* 27.5, pp. 491–499. DOI: 10.1515/revneuro-2015-0061.

Wang, Y., J. Luo, H. Zhang, and J. Lu (Apr. 2016). „microRNAs in the Same Clusters Evolve to Coordinately Regulate Functionally Related Genes“. In: *Mol. Biol. Evol.* 33.9, pp. 2232–2247. DOI: 10.1093/molbev/msw089.

Watts, A. M., G. A. Hopkins, and S. J. Goldstien (Feb. 2019). „Chimerism and population dieback alter genetic inference related to invasion pathways and connectivity of biofouling populations on artificial substrata“. In: *Ecology and Evolution* 9.6, pp. 3089–3104. DOI: 10.1002/ece3.4817.

Weinberg, Z. and R. R. Breaker (Jan. 2011). „R2R - software to speed the depiction of aesthetic consensus RNA secondary structures“. In: *BMC Bioinf.* 12.1, p. 3. DOI: 10.1186/1471-2105-12-3.

Wen, J., E. Ladewig, S. Shenker, J. Mohammed, and E. C. Lai (Sept. 2015). „Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates“. In: *PLOS Comput. Biol.* 11.9, e1004441. DOI: 10.1371/journal.pcbi.1004441.

Wheeler, T. J. and S. R. Eddy (July 2013). „nhmmer: Dna homology search with profile HMMs“. In: *Method. Biochem. Anal.* 29.19, pp. 2487–2489. DOI: 10.1093/bioinformatics/btt403.

Wightman, B., I. Ha, and G. Ruvkun (Dec. 1993). „Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans“. In: *Cell* 75.5, pp. 855–862. DOI: 10.1016/0092-8674(93)90530-4.

Will, S., T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen (Mar. 2012). „LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs“. In: *RNA* 18.5, pp. 900–914. DOI: 10.1261/rna.029041.111.

Wilson, R. C. and J. A. Doudna (May 2013). „Molecular Mechanisms of RNA Interference“. In: *Annu. Rev. Biophys.* 42.1, pp. 217–239. DOI: 10.1146/annurev-biophys-083012-130404.

Winter, J., S. Jung, S. Keller, R. I. Gregory, and S. Diederichs (Mar. 2009). „Many roads to maturity: Microrna biogenesis pathways and their regulation“. In: *Nat. Cell Biol.* 11.3, pp. 228–234. DOI: 10.1038/ncb0309-228.

Winter, J., S. Link, D. Witzigmann, C. Hildenbrand, C. Previti, and S. Diederichs (Apr. 2013). „Loop-miRs: Active microRNAs generated from single-stranded loop regions“. In: *Nucleic Acids Res.* 41.10, pp. 5503–5512. DOI: 10.1093/nar/gkt251.

Wolf, Y. I., P. S. Novichkov, G. P. Karev, E. V. Koonin, and D. J. Lipman (Apr. 2009). „The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages“. In: *Proc. Natl. Acad. Sci.* 106.18, pp. 7273–7280. DOI: 10.1073/pnas.0901808106.

Xie, M., M. Li, A. Vilborg, N. Lee, M.-D. Shu, V. Yartseva, N. Šestan, and J. A. Steitz (Dec. 2013). „Mammalian 5′-Capped MicroRNA Precursors that Generate a Single MicroRNA“. In: *Cell* 155.7, pp. 1568–1580. DOI: 10.1016/j.cell.2013.11.027.

Xu, S., P. D. Witmer, S. Lumayag, B. Kovacs, and D. Valle (Aug. 2007). „MicroRNA (miRNA) Transcriptome of Mouse Retina and Identification of a Sensory Organ-specific miRNA Cluster“. In: *J. Biol. Chem.* 282.34, pp. 25053–25066. DOI: 10.1074/jbc.m700501200.

Yang, Z., X. Wan, Z. Gu, H. Zhang, X. Yang, L. He, R. Miao, Y. Zhong, and H. Zhao (Sept. 2014). „Evolution of the mir-181 microRNA family“. In: *Comput. Biol. Med.* 52, pp. 82–87.

Yates, A. D. et al. (Nov. 2019). „Ensembl 2020“. In: *Nucleic Acids Res.* 48.D1, pp. D682–D688. DOI: 10.1093/nar/gkz966.

Yates, B., B. Braschi, K. A. Gray, R. L. Seal, S. Tweedie, and E. A. Bruford (Oct. 2016). „Genenames.org: The HGNC and VGNC resources in 2017“. In: *Nucleic Acids Res.* 45.D1, pp. D619–D625. DOI: 10.1093/nar/gkw1033.

Yaylak, B. and B. Akgül (2022). „Experimental MicroRNA Detection Methods“. en. In: *Methods Mol. Biol.* 2257, pp. 33–55.

Yazbeck, A. M., K. R. Tout, P. F. Stadler, and J. Hertel (Mar. 2017). „Towards a Consistent, Quantitative Evaluation of MicroRNA Evolution“. In: *Journal of Integrative Bioinformatics* 14.1, p. 20160013. DOI: 10.1515/jib-2016-0013.

Yazbeck, A. M., P. F. Stadler, K. Tout, and J. Fallmann (Apr. 2019). „Automatic curation of large comparative animal MicroRNA datasets“. In: *Method. Biochem. Anal.* 35.22. btz271, pp. 4553–4559. DOI: 10.1093/bioinformatics/btz271.

Ye, C., C. M. Hill, S. Wu, J. Ruan, and Z. (Ma (Aug. 2016). „DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies“. In: *Sci. Rep.* 6.1, p. 31900. DOI: 10.1038/srep31900.

You, L. et al. (Jan. 2019). „LanceletDB: An integrated genome database for lancelet, comparing domain types and combination in orthologues among lancelet and other species“. In: *Database* 2019. DOI: 10.1093/database/baz056.

Yu, G. (Mar. 2020). „Using ggtree to Visualize Data on Tree-Like Structures“. In: *Current Protocols in Bioinformatics* 69.1, e96. DOI: 10.1002/cpbi.96.

Zeng, Y. and B. R. Cullen (Jan. 2003). „Sequence requirements for micro RNA processing and function in human cells“. In: *RNA* 9.1, pp. 112–123. DOI: 10.1261/rna.2780503.

Zhang, B. H., X. P. Pan, S. B. Cox, G. P. Cobb, and T. A. Anderson (Jan. 2006). „Evidence that miRNAs are different from other RNAs“. In: *Cell. Mol. Life Sci.* 63.2, pp. 246–254. DOI: 10.1007/s00018-005-5467-7.

Zhang, H.-B., X. Zhao, X. Ding, A. H. Paterson, and R. A. Wing (1995). „Preparation of megabase-size DNA from plant nuclei“. In: *The Plant Journal* 7.1, pp. 175–184.

Zhao, B.-W., L.-F. Zhou, Y.-L. Liu, S.-M. Wan, and Z.-X. Gao (Mar. 2017). „Evolution of Fish Let-7 MicroRNAs and Their Expression Correlated to Growth Development in Blunt Snout Bream“. In: *Int. J. Mol. Sci.* 18.3, p. 646. DOI: 10.3390/ijms18030646.

Zhong, Y.-F., T. Butts, and P. W. H. Holland (Sept. 2008). „HomeoDB: A database of homeobox gene diversity“. In: *Evol. Dev.* 10.5, pp. 516–518. DOI: 10.1111/j.1525-142x.2008.00266.x.

Zuker, M. (July 2003). „Mfold web server for nucleic acid folding and hybridization prediction“. In: *Nucleic Acids Res.* 31.13, pp. 3406–3415. DOI: 10.1093/nar/gkg595.

Curriculum Scientiae

Personal Information

Name	Cristian Arley Velandia Huerto
Birth	September 21, 1990
Birthplace	Bogotá D.C
E-mail	cristian@bioinf.uni-leipzig.de
Webpage	`https://cavelandiah.github.io/`

Education

since October 2017	PhD Student Bioinformatics Group Universität Leipzig
2014–2016	Master Student Bioinformatics at Universidad Nacional de Colombia • Thesis: Estudio computacional del recambio de genes y pseudogenes de miRNAs en genomas del subphylum Tunicata. Grade: 4.5/5.0
2008–2013	Bachelor Student Biology at Universidad Nacional de Colombia • Thesis: Búsqueda y validación computacional de RNAs no codificantes homólogos en el genoma del tunicado *Didemnum vexillum*. Grade: 4.0/5.0

Research Experience

2017-2021	DAAD Research Fellow Doctoral Studies, Universität Leipzig, Germany
2015-2018	Research Fellow Estudio de la organización genómica de regiones adyacentes a dominios ancestrales del sistema inmune de cordados inferiores. Colciencias, Project:110165843196.
2013-2015	Research Fellow Regiones estructurales de ARNs no codificantes del genoma del tunicado *Didemmun vexillum*. FPIT, Banco de la Republica, Project: 3256.

Teaching Experience (384 hours)

2014, 2017, 2018 Teaching Assistant

- 1st Programming for Evolutionary Biology goes to the Americas, Universidad Nacional de Colombia.
- 6st-7st Programming for Evolutionary Biology, (Universität Leipzig, Freie Universität Berlin)

Conferences

October, 2016-2021	Bioinformatik Herbstseminar Universität Leipzig Bioinformatik, Leipzig,Germany.
September, 2019	German Conference of Bioinformatics 2019 DKFZ, Heidelberg, Germany. Talk: Key Structural Patterns for miRNA Family Reconstruction
February, 2019	34th TBI Winterseminar Institute for Theoretical Chemistry, Universität Wien, Bled, Slovenia. Talk: How to improve the detection of miRNA homologs? An outlook from tunicate genomes.
October, 2018	GIGA III: Global Invertebrate Genomics Alliance Nova Southeastern University, Willemstad, Curaçao. Talk: Accessing to the miRNA complement in tunicate genomes, an *in silico* approach.

Computer skills

Operating Systems:	MacOs, Linux, Windows
Programming:	Bash, Perl, Python, R, Java, Bash, MySQL
Others	Conda, Git, VIM, MS Office suite, TEX, HTML, CSS, Markdown

Languages

Spanish:	native speaker
English:	fluent
German:	intermediate level

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbstständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, 21.02.2022

(Ort, Datum)

Cristian A. Velandia M.

(Unterschrift)