ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE DELLA TERRA, DELLA VITA E DELL'AMBIENTE

Ciclo 37

Settore Concorsuale: 05/B1 - ZOOLOGIA E ANTROPOLOGIA

Settore Scientifico Disciplinare: BIO/05 - ZOOLOGIA

CHARACTERIZATION AND EVOLUTION OF MITO-NUCLEAR INTERACTIONS

Presentata da: Alessandro Formaggioni

Coordinatore Dottorato

Barbara Cavalazzi

Supervisore

Marco Passamonti

Co-supervisore

Federico Plazzi

Esame finale anno 2025

Abstract

The emergence of the eukaryotic cell, approximately 2.1 billion years ago, is considered one of the crucial moments in the evolution of life on Earth. Since then, two genomes have coexisted and coevolved within the same cellular environment, resulting in a complex network of interactions. These interactions have been linked to various biological phenomena, including cancer, aging, and speciation, among others. The aim of my thesis is to shed light on some of these interactions and their biological implications.

RNA-RNA interactions are often used by the cell to regulate its homeostasis. In particular, short non-coding RNA (sncRNA) molecules, 19-30 base pairs long, interact with different proteins to target and suppress other RNA transcripts through a mechanism called RNA interference (RNAi). This mechanism has acquired different roles in the cell, such as innate immune response, transposon silencing, and regulation of messenger RNAs. In Lophotrochozoa, one of the main metazoan branches, these pathways are poorly studied. By analysing omics data from 43 species across 9 phyla, I characterized the evolution of two pivotal protein families in the RNAi pathways: the Argonaute and DICER families. The analysis suggested that the common ancestor of Trochozoa lost the endo-siRNA pathway, which is a key pathway in the innate immune response in other metazoan species.

Besides endo-siRNAs, many other RNAi pathways have emerged during eukaryotic evolution. A new class of short non-coding RNAs was recently discovered in the Manila clam *Ruditapes philippinarum*. These short RNAs are transcribed in the mitochondrial genome and target nuclear transcripts. However, how these Small Mitochondrial Highly Transcribed RNAs (smithRNAs) regulate different cellular processes is not yet clear. In my thesis, I focused on identifying proteins that could be involved in the maturation and regulatory pathways of smithRNAs. Analysing publicly available RNA immunoprecipitation (RIP) and cross-linking immunoprecipitation (CLIP) libraries, I observed interactions between proteins involved in the maturation of other sncRNAs (AGO2, Drosha, and DGCR8 in *Homo sapiens*; AGO2 in *Drosophila melanogaster* and *Mus musculus*; and ERGO-1 in *Caenorhabditis elegans*) and mitochondrial tRNAs, where the majority of smithRNAs are located. Moreover, co-immunoprecipitation experiments between the smithRNAs identified in *R. philippinarum* and a protein lysate revealed that part of the proteins interacting with smithRNAs were also interacting with the miRNA let-7. The preliminary structures of smithRNAs interacted with proteins related to the spliceosome, suggesting that their maturational pathway may have co-opted proteins from different nuclear RNAi pathways.

In the second part of my project, I focused on another kind of mito-nuclear interaction. The protein-protein interactions between mitochondrial and nuclear OXPHOS subunits. Twelve subunits are encoded by the mitochondrial genome, while around 70 subunits are encoded by the nuclear genome. The subunits interact closely to ensure the proper functioning of the OPHOS complexes. I was particularly interested in how the co-evolution between subunits could alter the phylogenetic signal retrieved from OXPHOS makers. In literature there are at least two well-known examples of mito-nuclear discordance in deep lineages: in Bivalvia,

Pteriomorphia in sister relationship with Heterodonta. Similarly, in Squamata Serpentes and Agamidae form a monophyletic clade. These two phylogenetic hypotheses are robustly supported by mitochondrial markers, but they were extensively rejected by phylogenomic analyses. I analysed the phylogenetic signal of nuclear OXPHOS genes. For both cases, Bivalvia and Squamata, the close interaction between the OXOPHOS subunits has led the nuclear OXPHOS genes to support the biased mitochondrial topology. In particular, the support for the mitochondrial topology resulted higher for nuclear OXPHOS subunits directly in contact with the mitochondrial counterparts.

The tight interactions between the nuclear and the mitochondrial genome have a clear impact in many biological processes. Co-opting nuclear proteins, the mitochondrial genome has evolved an internal RNAi pathway, which in *R. philippinarum* may be linked to sex determination. While in snakes the adaptive selection on OXPHOS genes is likely linked to their extreme radiation. During my three years project I tried to shed light on these interactions, showing how their importance in main evolutionary processes.

Index

1. Introduction to the Characterization and Evolution of Mito-Nuclear Interactions

A new domain is born

Almost two billion years ago, the endosymbiosis between an archaebacterium and an eubacterium gave rise to one of the most important synapomorphies in life, mitochondria. This single event gave rise to a new domain of life that encompasses highly complex life forms (Williams, 2014, Derelle et al. 2015).

Following the endosymbiotic event, the chimeric cell underwent genomic reorganization, resulting in a significant reduction of the endosymbiont's genome (Gray 2012). Mitochondria are considered close relatives of alpha-proteobacteria (Harrison et al. 2023). Most of alpha-proteobacteria species encode for 2,000-4,000 genes (Koonin and Wolf 2008), whereas the richest mitochondrial genome contains no more than 67 protein coding genes (Burger et al. 2013). In tight symbioses it is common that shared genes are retained in one species and lost in the other one (Moran et al. 2008). Since there are many mitochondria and one nucleus in the eukaryotic cell, it is more energetically efficient to retain the nuclear copy instead of the mitochondrial one (Kelly 2021). Thus, most of the genes in the nucleus have an archaeal origin, whereas a small fraction of genes moved from the mitochondrial genome to the nucleus. To maximize energy efficiency, one might expect mitochondria to have lost all their genes. However, all eukaryotes that produce energy through cellular respiration retain a common set of mitochondrial genes. The CO-location for Redox Regulation (CORR) hypothesis suggests that mitochondria have retained the ability to transcribe and translate the core proteins of the OXidative PHOSphorylation system (OXPHOS) complexes to locally control the number and ratio of these complexes independently in each mitochondrion (Allen et al. 2003; Allen 2015). In bilaterian animals, this gene set is restricted to 13 genes, which encode for the catalytic centres of the complexes. Meanwhile, the remaining nuclear-encoded subunits, which number around 80, assemble around the mitochondrial core of each complex (Lane 2014).

The coevolution between OXPHOS subunits

Mitochondrial and nuclear subunits interact closely to assemble OXPHOS complexes. In animals, the mitochondrial genome accumulates mutations faster than the nuclear genome (Lynch et al. 2006). Moreover, it has been a common belief that the mitochondrial genome lacks recombination (Birky 2001). In this scenario, adaptive variants cannot be selected over deleterious ones; instead, selection acts on the entire mitochondrial genome. The accumulation of deleterious variants in non-recombinant lineages is known as the Hill-Robertson effect (Hill and Robertson 1966). According to this hypothesis, the mitochondrial genome would face mutational erosion, although the lack of recombination in mitochondria cannot be completely ruled out

(Fragkoulis et al. 2024). To maintain the efficiency of OXPHOS, it has been proposed that nuclear-encoded OXPHOS subunits and other nuclear-encoded mitochondrial-interacting (Nuc-mt) proteins accumulate variants that compensate for deleterious mutations in mitochondrial subunits. Thus, mito-nuclear compensation could maintain respiratory function and prevent a continuous decline in fitness in eukaryotes (Levin et al. 2014; Havird et al. 2017). Consistently, Nuc-mt genes (whether they are OXPHOS subunits or ribosomal RNAs) show higher evolutionary rates and evidence of positive selection than mitochondrial genes and other nuclear genes (Sloan et al. 2014; Barreto et al. 2018). However, some clades show a reversed pattern, where the evolutionary rates of mitochondrial genes are generally higher than those of Nuc-mt genes (Piccinini et al. 2021). According to the "nuclear compensation" hypothesis, a single detrimental mutation in mitochondrial genes is compensated by several positive mutations in the nuclear counterparts. Simply measuring higher evolutionary rates in Nuc-mt genes is not sufficient to confirm the hypothesis. Nuclear compensatory mutations must temporally follow and be physically associated with deleterious mitochondrial variants. In this context, data on primates failed to confirm the hypothesis, as the majority of mitochondrial substitutions did not occur before Nuc-mt substitutions between contact site pairs (Weaver et al. 2022).

Mitochondrial and Nuc-mt genes are in continuous coevolution to maintain cellular functions. Thus, as two populations diverge, the mitochondria of one population may not be able to cooperate with the nuclear genes of the other populations (Rand et al. 2004). Experiments that produced cybrid lines (i.e., cell lines where the nuclear genome from one species is combined with an enucleated cell from a second species) using human nuclei and cells from different primates (e.g., chimpanzees, gorillas, orangutans) showed that the older the common ancestor between humans and the primate species, the less effective the respiration. This ranged from no measurable respiration in lemur cybrids to diminished respiration in chimpanzee cybrids (Kenyon and Moraes 1997). Mito-nuclear interactions have been shown to interfere with hybridization. When crossing two species of swordtail fish, *Xiphophorus birchmanni* (father) and *Xiphophorus malinche* (mother), the frequency of two Nuc-mt loci was biased toward the *X. malinche* allele: hybrids homozygous for the *X. malinche* allele had normal development, while heterozygous hybrids or those homozygous for the *X. birchmanni* allele exhibited reduced complex I function or incomplete embryonic development (Moran et al. 2024). Thus, reproductive barriers may arise primarily due to mito-nuclear incompatibilities. Many other studies showed that mito-nuclear interactions coevolve along population-specific trajectories (Wolff et al. 2014; Hill 2019). These trajectories are also shaped by environmental factors; for instance, a mutation in the cytochrome c oxidase subunit 3 (COX3) has allowed the Bar-headed Goose to adapt to high altitudes, shortening its migration route (Scott et al. 2015). Another example is the convergent adaptation of different lineages of the Atlantic molly to hydrogen sulphide tolerance, which has been linked to mutations in the COX1 and COX3 genes (Greenway et al. 2020).

Mitochondrial signalling towards the nucleus

A complex network of signals is established to control energy production. Most of the signals are related to OXPHOS (e.g., ATP, NADH, reactive oxygen species), and they act as proxies for the state of the respiratory chain (Woodson and Chory 2008). Thus, signals from the mitochondria to the nucleus were initially thought to be passive. However, in recent years, different types of signals encoded by the mitochondrial genome that affect nuclear and cellular functions have been discovered. Aside from the 37 genes encoded in bilaterian mitochondrial genomes, there is limited space for additional genes. Nevertheless, many other small peptides could be encoded through alternative open reading frames. Some of these peptides have been characterized; they influence mitochondrial bioenergetics (Kienzle et al. 2023), attenuate pathologies such as Alzheimer's disease, prostate cancer, and macular degeneration (Miller et al. 2020; Miller et al. 2022), trigger the immune responses (Rice et al. 2023), and are secreted outside the cells, acting as signals in circulation (Cobb et al. 2016; Kienzle et al. 2023). These mitochondrial peptides have been mostly studied in mammals. However, we would expect to find them in all metazoans. Contrastingly, mitochondrial protein-coding genes that are not linked to OXPHOS have been uniquely found in Bivalvia and Brachiopoda among bilaterians (Breton et al. 2014; Niaison et al. 2021). These proteins have no homology with other known proteins, and their function remains uncertain; hence, they are referred to as ORFans (Breton et al. 2014). In bivalves, ORFans are commonly found in species that exhibit a unique mode of mitochondrial genome inheritance, called doubly uniparental inheritance (DUI) (Zouros 2013). In DUI systems, mitochondria are inherited in a sex-specific manner: males inherit mitochondria from both parents, but transmit only male-type mitochondria, while females inherit and transmit female-type mitochondria (Zouros et al. 1994; Ghiselli et al. 2011; Zouros and Rodakis 2019; Passamonti and Plazzi 2020). Consequently, in DUI species, there are two separate mitochondrial lineages, whose amino acid sequences can diverge by up to 30% (Passamonti and Ghiselli 2009). In DUI species, ORFans are sex-specific. Male-specific and female-specific ORFans have been detected in clams and freshwater mussels, and they have been linked to processes such as spermatogenesis, embryo development, and the maintenance of gonochorism (Faure et al. 2011; Milani et al. 2013; Milani et al. 2014). In this context, mitochondrially encoded proteins may play a role in determining the sex of these species. However, the exact relationship between mitochondrial ORFans, DUI, and sex determination remains unclear (Breton et al. 2014).

Mitochondria control cellular functions through ncRNA mechanisms

A wide range of ncRNA types is transcribed by the nucleus to control cellular functions, and some of these ncRNAs target mitochondrial transcripts, affecting mitochondrial functions (Roiz-Valle et al. 2023). However, communication by means of RNA molecules occurs in both directions. Some ncRNAs have been shown to be transported from the mitochondria to the nucleus, where they associate with chromatin and affect nuclear RNA transcription (Sriram et al. 2024). Mitochondria also encode short RNAs (sRNAs). sRNAs are molecules ranging from 18 to 30 nucleotides that are processed from longer RNAs with specific secondary structures through the action of endonucleases. sRNAs regulate the cellular levels of many transcripts (e.g., messenger RNAs, mobile elements, viral RNAs) by pairing with these transcripts and interacting with proteins involved in RNA interference (Grimson et al. 2008; Bartel 2018; Shi et al. 2022). Initial evidence suggested that mitochondrial sRNAs are predominantly encoded from sense transcripts and target antisense transcripts, primarily affecting mitochondrial functions (Ro et al. 2013). Further analyses have identified a new class of mitochondrial sRNAs that target nuclear transcripts, known as small mitochondrial highly transcribed RNAs (smithRNAs; Pozzi et al., 2017). SmithRNAs were first discovered in the Manila clam, *Ruditapes philippinarum*. According to analyses of small RNA sequencing libraries, the two mitochondrial lineages of *R. philippinarum*, male and female, transcribe small RNAs that, through *in silico* analyses, were found to target the 3' untranslated region (UTR) of messenger RNAs transcribed in the nucleus. Some of these messenger RNAs encode proteins related to sex determination in other metazoan species (Pozzi et al. 2017). Further analysis confirmed the *in vivo* functionality of some smithRNAs. For instance, when clam specimens were injected with a solution containing the 106t smithRNA, which was predicted to target the clam homolog of a human histone-lysine N-methyltransferase (Pozzi et al. 2017), a higher methylation rate on histone H3 was observed compared to the methylation rate in specimens injected with pure water (Passamonti et al. 2020). Moreover, smithRNAs are likely produced in a broad range of Metazoa, since an *in silico* analysis detected their presence in *Mus musculus, Danio rerio* and *Drosophila melanogaster* (Passamonti et al. 2020). The mitochondrial genome appears to readily evolve small RNA structures: its circular chromosome is transcribed as a single polycistronic transcript, with coding genes separated by secondary structures, which are processed by ribonucleases. Thus, small RNAs are likely generated from the processing of the polycistronic transcript. These RNAs can easily acquire biological functions, as their likelihood of targeting messenger RNAs, as estimated through *in silico* simulations, is quite high (Plazzi et al. 2024). Indeed, smithRNAs have also been linked to various molecular processes, such as recovery from COVID-19 (Pozzi 2022). Additionally, some smithRNAs have been conserved over time, such as the smithRNA produced from mt-tRNA-Met, which has been demonstrated to be conserved across Chordata (Pozzi and Dowling 2022).

Overall, smithRNAs appear to be an effective tool for mitochondria to control cellular functions. However, many aspects of their biology remain unclear. If these molecules are indeed conserved, as current data suggests,

the pathway responsible for smithRNA production and activity should also be under selective pressure. It has been reported that human AGO2 likely interacts with smithRNAs (Pozzi and Dowling 2022), but the entire processing mechanism is still unknown. Future research focused on identifying smithRNA-interacting proteins may clarify whether smithRNAs are part of a peculiar pathway or if they rely on different proteins co-opted from other RNA interference pathways.

1.1 The Scope of "Characterization of Mito-Nuclear Interactions"

The aim of this thesis is to better characterize different types of nuc-mt interactions. I first analysed how the coevolution between mitochondrial and nuclear subunits can affect their phylogenetic signal. By retrieving the protein sequences of OXPHOS subunits from 31 Bivalvia species, I inferred the phylogenetic tree of the Bivalvia class. The results showed that the coevolution between mitochondrial and nuclear subunits deeply affects their phylogenetic signal. This effect is more pronounced in the nuclear subunits that are in closer contact with their mitochondrial counterparts. Studying the phylogenetic signal of OXPHOS genes can shed light on the mechanisms of mito-nuclear compensation, but also help identify phylogenetic artifacts that deviate from the true evolutionary history of the species.

As I previously mentioned, mitochondria rely on RNAi mechanisms to control cellular functions. The most parsimonious hypothesis is that these small RNAs are processed by proteins co-opted from cellular RNAi pathways. In some cases, these pathways are conserved across Metazoa, as is the case for microRNAs (miRNAs) and PIWI-interacting RNAs (piRNAs) (Lim et al. 2014; Moran et al. 2017). However, the mode of maturation of endogenous small-interfering RNAs (endo-siRNAs) varies significantly among different metazoan lineages (Czech et al. 2008; Watanabe et al. 2008; Billi et al. 2014; Fridrich et al. 2020). The evolution of these pathways has never been fully explored in Lophotrochozoa. Therefore, I annotated the proteomes of 43 lophotrochozoan species, focusing on proteins belonging to the DICER and Argonaute families, which are key components of RNAi pathways. According to my findings, the miRNA and piRNA pathways are conserved across all Lophotrochozoa phyla, whereas the endo-siRNA pathway was lost in the most recent common ancestor of Trochozoa.

With a clearer understanding of RNAi pathways in Lophotrochozoa, I proceeded to analyze potential interactions between smithRNAs and RNAi-related proteins in *R. philippinarum*. By immunoprecipitating two *R. philippinarum* smithRNAs (and their putative immature forms) exposed to the clam's protein fraction, I was able to identify proteins that interact with smithRNAs. Moreover, I analyzed publicly available enhanced Cross-Linking and ImmunoPrecipitation (eCLIP) and RNA Immunoprecipitation (RIP) sequencing libraries of a wide range of Argonaute proteins, DROSHA and DGCR8 in different animals. It resulted that some Argonaute proteins, such as AGO2 in *Homo sapiens, M. musculus* and *D. melanogaster* and ERGO-1 in *Caenorhabditis elegans,* and DROSHA may be involved in the smithRNA pathway. In particular, the smithRNAs located in the mt-tRNA-Met showed the strongest interaction with most of these proteins. Hence, I propose that initiation and elongation factors, which have been reported to interact with Argonaute proteins, might be involved in the loading of smithRNAs on AGO proteins.

Overall, the results of my thesis highlight the close connection between the nucleus and mitochondria, as well as the evolutionary implications of this interaction.

1.2 References

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proc Natl Acad Sci U S A* 112:10231–10238.

Allen JF, Horner DS, Cavalier-Smith T, Willison K, Leaver CJ, Martin W. 2003. The function of genomes in bioenergetic organelles. *Philos Trans R Soc Lond B Biol Sci* [Internet] 358:19–38. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2002.1191

Barreto FS, Watson ET, Lima TG, Willett CS, Edmands S, Li W, Burton RS. 2018. Genomic signatures of mitonuclear coevolution across populations of Tigriopus californicus. *Nat Ecol Evol* 2:1250–1257.

Bartel DP. 2018. Metazoan MicroRNAs. *Cell* 173:20–51.

Billi AC, Fischer SEJ, Kim JK. 2014. Endogenous RNAi pathways in C. elegans. *WormBook*:1–49.

Birky CW. 2001. The Inheritance of Genes in Mitochondria and Chloroplasts: Laws, Mechanisms, and Models. *Annu Rev Genet* 35:125–148.

Breton S, Milani L, Ghiselli F, Guerra D, Stewart DT, Passamonti M. 2014. A resourceful genome: Updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. *Trends in Genetics* [Internet] 30:555–564. Available from: http://www.cell.com.ezproxy.unibo.it/article/S0168952514001383/fulltext

Burger G, Gray MW, Forget L, Lang BF. 2013. Strikingly Bacteria-Like and Gene-Rich Mitochondrial Genomes throughout Jakobid Protists. *Genome Biol Evol* 5:418–438.

Cobb LJ, Lee C, Xiao J, Yen K, Wong RG, Nakamura HK, Mehta HH, Gao Q, Ashur C, Huffman DM, et al. 2016. Naturally occurring mitochondrial-derived peptides are agedependent regulators of apoptosis, insulin sensitivity, and inflammatory markers. *Aging* 8:796–808.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, et al. 2008. An endogenous small interfering RNA pathway in Drosophila. *Nature* 453:798–802.

Faure E, Delaye L, Tribolo S, Levasseur A, Seligmann H, Barthélémy R-M. 2011. Probable presence of an ubiquitous cryptic mitochondrial gene on the antisense strand of the cytochrome oxidase I gene. *Biol Direct* 6:56.

Fragkoulis G, Hangas A, Fekete Z, et al. Linear DNA-driven recombination in mammalian mitochondria. *Nucleic Acids Res*. 2024;52(6):3088-3105. doi:10.1093/nar/gkae040

Fridrich A, Modepalli V, Lewandowska M, Aharoni R, Moran Y. 2020. Unravelling the developmental and functional significance of an ancient Argonaute duplication. *Nat Commun* 11:6187.

Ghiselli F, Milani L, Passamonti M. 2011. Strict Sex-Specific mtDNA Segregation in the Germ line of the DUI Species Venerupis philippinarum (Bivalvia: Veneridae). *Mol Biol Evol* 28:949–961.

Gray MW. 2012. Mitochondrial Evolution. *Cold Spring Harb Perspect Biol* 4:a011403–a011403.

Greenway R, Barts N, Henpita C, Brown AP, Arias Rodriguez L, Rodríguez Peña CM, Arndt S, Lau GY, Murphy MP, Wu L, et al. 2020. Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proceedings of the National Academy of Sciences* 117:16424–16430.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455:1193–1197.

Harrison SA, Rammu H, Liu F, Halpern A, Nunes Palmeira R, Lane N. 2023. Life as a Guide to Its Own Origins. *Annu Rev Ecol Evol Syst* 54:327–350.

Havird JC, Trapp P, Miller CM, Bazos I, Sloan DB. 2017. Causes and Consequences of Rapidly Evolving mtDNA in a Plant Lineage. *Genome Biol Evol* 9:323–336.

Hill GE. 2019. Mitonuclear Ecology. Oxoford University Press

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* 8:269–294.

Kelly S. 2021. The economics of organellar gene loss and endosymbiotic gene transfer. *Genome Biol* 22:345.

Kenyon L, Moraes CT. 1997. Expanding the functional human mitochondrial DNA database by the establishment of primate xenomitochondrial cybrids. *Proceedings of the National Academy of Sciences* 94:9131–9135.

Kienzle L, Bettinazzi S, Choquette T, Brunet M, Khorami HH, Jacques JF, Moreau M, Roucou X, Landry CR, Angers A, et al. 2023. A small protein coded within the mitochondrial canonical gene nd4 regulates mitochondrial bioenergetics. *BMC Biol* [Internet] 21:111. Available from: /pmc/articles/PMC10193809/

Koonin E V., Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* [Internet] 36:6688–6719. Available from: https://dx.doi.org/10.1093/nar/gkn668

Lane N. 2014. Bioenergetic Constraints on the Evolution of Complex Life. *Cold Spring Harb Perspect Biol* [Internet] 6. Available from: /pmc/articles/PMC3996473/

Levin L, Blumberg A, Barshad G, Mishmar D. 2014. Mito-nuclear co-evolution: the positive and negative sides of functional ancient mutations. *Front Genet* 5.

Lim RSM, Anand A, Nishimiya-Fujisawa C, Kobayashi S, Kai T. 2014. Analysis of Hydra PIWI proteins and piRNAs uncover early evolutionary origins of the piRNA pathway. *Dev Biol* 386:237–251.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. *Science (1979)* 311:1727–1730.

Milani L, Ghiselli F, Guerra D, Breton S, Passamonti M. 2013. A Comparative Analysis of Mitochondrial ORFans: New Clues on Their Origin and Role in Species with Doubly Uniparental Inheritance of Mitochondria. *Genome Biol Evol* 5:1408–1434.

Milani L, Ghiselli F, Maurizii MG, Nuzhdin S V., Passamonti M. 2014. Paternally Transmitted Mitochondria Express a New Gene of Potential Viral Origin. *Genome Biol Evol* 6:391–405.

Miller B, Kim S-J, Kumagai H, Mehta HH, Xiang W, Liu J, Yen K, Cohen P. 2020. Peptides derived from small mitochondrial open reading frames: Genomic, biological, and therapeutic implications. *Exp Cell Res* 393:112056.

Miller B, Kim S-J, Mehta HH, Cao K, Kumagai H, Thumaty N, Leelaprachakul N, Braniff RG, Jiao H, Vaughan J, et al. 2022. Mitochondrial DNA variation in Alzheimer's disease reveals a unique microprotein called SHMOOSE. *Mol Psychiatry*.

Moran BM, Payne CY, Powell DL, Iverson ENK, Donny AE, Banerjee SM, Langdon QK, Gunn TR, Rodriguez-Soto RA, Madero A, et al. 2024. A lethal mitonuclear incompatibility in complex I of natural hybrids. *Nature* 626:119–127.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and Evolution of Heritable Bacterial Symbionts. *Annu Rev Genet* 42:165–190.

Moran Y, Agron M, Praher D, Technau U. 2017. The evolutionary origin of plant and animal microRNAs. *Nat Ecol Evol* 1:0027.

Niaison T, Guerra D, Breton S. 2021. The complete mitogenome of the inarticulate brachiopod *Glottidia pyramidata* reveals insights into gene order variation, deviant ATP8 and mtORFans in the Brachiopoda. *Mitochondrial DNA Part B* 6:2701–2703.

Passamonti M, Calderone M, Delpero M, Plazzi F. 2020. Clues of in vivo nuclear gene regulation by mitochondrial short non-coding RNAs. *Sci Rep* 10:8219.

Passamonti M, Ghiselli F. 2009. Doubly Uniparental Inheritance: Two Mitochondrial Genomes, One Precious Model for Organelle DNA Inheritance and Evolution. *DNA Cell Biol* 28:79–89.

Passamonti M, Plazzi F. 2020. Doubly Uniparental Inheritance and beyond: The contribution of the Manila clam *Ruditapes philippinarum*. *Journal of Zoological Systematics and Evolutionary Research* 58:529–540.

Piccinini G, Iannello M, Puccio G, Plazzi F, Havird JC, Ghiselli F. 2021. Mitonuclear Coevolution, but not Nuclear Compensation, Drives Evolution of OXPHOS Complexes in Bivalves. *Mol Biol Evol* 38:2597–2614.

Plazzi F, Le Cras Y, Formaggioni A, Passamonti M. 2024. Mitochondrially mediated RNA interference, a retrograde signaling system affecting nuclear gene expression. *Heredity (Edinb)* 132:156–161.

Pozzi A. 2022. COVID-19 and Mitochondrial Non-Coding RNAs: New Insights From Published Data. *Front Physiol* 12.

Pozzi A, Dowling DK. 2022. New Insights into Mitochondrial–Nuclear Interactions Revealed through Analysis of Small RNAs. *Genome Biol Evol* 14.

Pozzi A, Plazzi F, Milani L, Ghiselli F, Passamonti M. 2017. SmithRNAs: Could Mitochondria “Bend” Nuclear Regulation? *Mol Biol Evol* 34:1960–1973.

Rand DM, Haney RA, Fry AJ. 2004. Cytonuclear coevolution: the genomics of cooperation. *Trends Ecol Evol* 19:645–653.

Rice M, Kim J, Immun M, Park CY, Lai R, Barr C, Son J, Tor K, Kim E, Lu R, et al. 2023. The Human Mitochondrial Genome Encodes for an Interferon-Responsive Host Defense Peptide. *Elife*.

Ro S, Ma H-Y, Park C, Ortogero N, Song R, Hennig GW, Zheng H, Lin Y-M, Moro L, Hsieh J-T, et al. 2013. The mitochondrial genome encodes abundant small noncoding RNAs. *Cell Res* 23:759–774.

Roiz-Valle D, Caravia XM, López-Otín C. 2023. Mechanisms of mitochondrial microRNA regulation in cardiovascular diseases. *Mech Ageing Dev* 212:111822.

Scott GR, Hawkes LA, Frappell PB, Butler PJ, Bishop CM, Milsom WK. 2015. How Bar-Headed Geese Fly Over the Himalayas. *Physiology* 30:107–115.

Shi J, Zhou T, Chen Q. 2022. Exploring the expanding universe of small RNAs. *Nat Cell Biol* 24:415–423.

Sloan DB, Triant DA, Wu M, Taylor DR. 2014. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. *Mol Biol Evol* 31:673–682.

Sriram K, Qi Z, Yuan D, Malhi NK, Liu X, Calandrelli R, Luo Y, Tapia A, Jin S, Shi J, et al. 2024. Regulation of nuclear transcription by mitochondrial RNA in endothelial cells. *Elife* 13.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453:539–543.

Weaver RJ, Rabinowitz S, Thueson K, Havird JC. 2022. Genomic Signatures of Mitonuclear Coevolution in Mammals. *Mol Biol Evol* [Internet] 39. Available from: https://dx.doi.org/10.1093/molbev/msac233

Wolff JN, Ladoukakis ED, Enríquez JA, Dowling DK. 2014. Mitonuclear interactions: evolutionary consequences over multiple biological scales. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369:20130443.

Woodson JD, Chory J. 2008. Coordination of gene expression between organellar and nuclear genomes. *Nat Rev Genet* 9:383–395.

Zouros E. 2013. Biparental Inheritance Through Uniparental Transmission: The Doubly Uniparental Inheritance (DUI) of Mitochondrial DNA. *Evol Biol* 40:1–31.

Zouros E, Ball AO, Saavedra C, Freeman KR. 1994. Mitochondrial DNA inheritance. *Nature* 368:818–818.

Zouros E, Rodakis GC. 2019. Doubly Uniparental Inheritance of mtDNA: An Unappreciated Defiance of a General Rule. *Adv Anat Embryol Cell Biol* 231:25–49.

2. Mito-Nuclear Coevolution and Phylogenetic Artifacts: the Case of Bivalve Mollusks

This chapter was written in collaboration with Federico Plazzi and Marco Passamonti and has been published in the journal *Scientific Reports*. Supplementary materials are available via the link to the original article (https://doi.org/10.1038/s41598-022-15076-y).

2.1 Introduction

Deep bivalve phylogeny: state-of-art

Bivalves are an extremely diverse group with about 50,000 living species (Gosling 2003). Deep evolutionary relationships among major clades within the molluscan class Bivalvia are only recently coming to a shared figure. The class is split into two main subgroups, Protobranchia and Autobranchia, whose origins root deep in the middle Ordovician periods (Morton 1996; Cope and Babin 1999; Cope 2002; Fang 2006; Tsubaki et al. 2011). Most likely, extant protobranchs resemble the Cambrian forerunners the most, for many molluscan symplesiomorphies are present, like a well-developed foot and true molluscan ctenidia devoted to gas exchange (Yonge 1939; Stasek 1963); moreover, food is brought to the mouth by palp proboscides. Two sister groups are usually acknowledged within Protobranchia, Nuculida and Solemyida, which are given an ordinal status (Starobogatov 1992; Morton 1996; von Salvini-Plawen and Steiner 1996; Waller 1998; Steiner and Hammer 2000; Passamaneck et al. 2004); analyses mainly based on molecular markers proposed to exclude the protobranch superfamily Nuculanoidea from Protobranchia and to better place it within Autobranchia (Giribet and Wheeler 2002; Giribet and Distel 2003; Bieler and Mikkelsen 2006; Plazzi and Passamonti 2010); the name Opponobranchia was proposed for remaining protobranchs (Giribet 2008). On the other hand, the clade Protobranchia has been recovered by most of large-scale datasets (González et al. 2015; Lemer et al. 2019), but with some exceptions (Lemer et al. 2019). Therefore, the monophyly of this clade still needs to be assessed.

The way of feeding is radically different in Autobranchia (=Autolamellibranchiata *sensu*; Giribet, 2008), whose common ancestor developed a feeding gill, one of the main drivers of the Ordovician bivalve radiation (Cope and Babin 1999) and led most groups to the key ecological shift towards infaunalization (Cope 2002; Fang 2006; Plazzi et al. 2017). Autobranchia is comprised by three major clades (subclasses; Newell, 1965): Heterodonta (clams, cockles, razor clams, and their kin), Palaeoheterodonta (freshwater mussels and their kin), and Pteriomorphia (mytilids, oysters, scallops, and their kin; Combosch et al., 2017; González et al., 2015; Plazzi et al., 2016). Moreover, the former subclass Anomalodesmata (Myra 1963; Newell 1965; J. Carter et al. 2011; Morton and Machado 2019) has been found to be nested within Heterodonta (Harper et al. 2000; Giribet and Wheeler 2002; Dreyer et al. 2003; Giribet and Distel 2003; Harper et al. 2006; Taylor, Williams, Glover, et al. 2007; Giribet 2008; Lemer et al. 2019). Currently, Archiheterodonta (order Carditida) are considered sister group to other Euheterodonta, which are further split into Anomalodesmata itself and Imparidentia (Giribet and Distel 2003; Taylor, Williams, Glover, et al. 2007; Taylor, Williams, and Glover 2007; J. Carter et al. 2011; Bieler et al. 2014; Combosch et al. 2017; Lemer et al. 2019; Morton and Machado 2019).

Relationships among the main bivalve sub-lineages remained unresolved or uncertain until recently. With minor issues linked to the position of Nuculanida and Anomalodesmata, two main hypotheses have been put forward: the Heteroconchia hypothesis, which involves a sister group relationship between Heterodonta and

Palaeoheterodonta (Fig. 1a), and the Amarsipobranchia hypothesis, which involves the sister group relationship between Heterodonta and Pteriomorphia instead (Fig. 1b).

The traditional taxonomic view and morphological analyses of Autobranchia heralded the Heteroconchia hypothesis (Waller 1990; Waller 1998; Cope 2002; Giribet and Distel 2003; J. Carter et al. 2011; Bieler et al. 2014); however, a closer relationship between Heterodonta and Pteriomorphia has been suggested following palaeontological evidence (Morris 1980; Cope 1996; Sánchez and Babin 2003; Sánchez 2006; Fang and Sanchez 2012; Cope and Kříž 2013). The Amarsipobranchia hypothesis was also highly supported by molecular phylogenetics, using mitochondrial markers (Giribet and Distel 2003; Doucet-Beaupré et al. 2010; Plazzi and Passamonti 2010; Plazzi et al. 2011; Plazzi et al. 2013; Plazzi et al. 2016). Contrastingly, the Heteroconchia hypothesis is always supported when nuclear markers are used (either combined with morphological data or not), as well as by means of transcriptomics (Giribet and Distel 2003; Kocot et al. 2011; Smith et al. 2011; Sharma et al. 2012; Bieler et al. 2014; González et al. 2015; Lemer et al. 2019). This is a clear example of mito-nuclear phylogenetic discordance (Toews and Brelsford 2012).

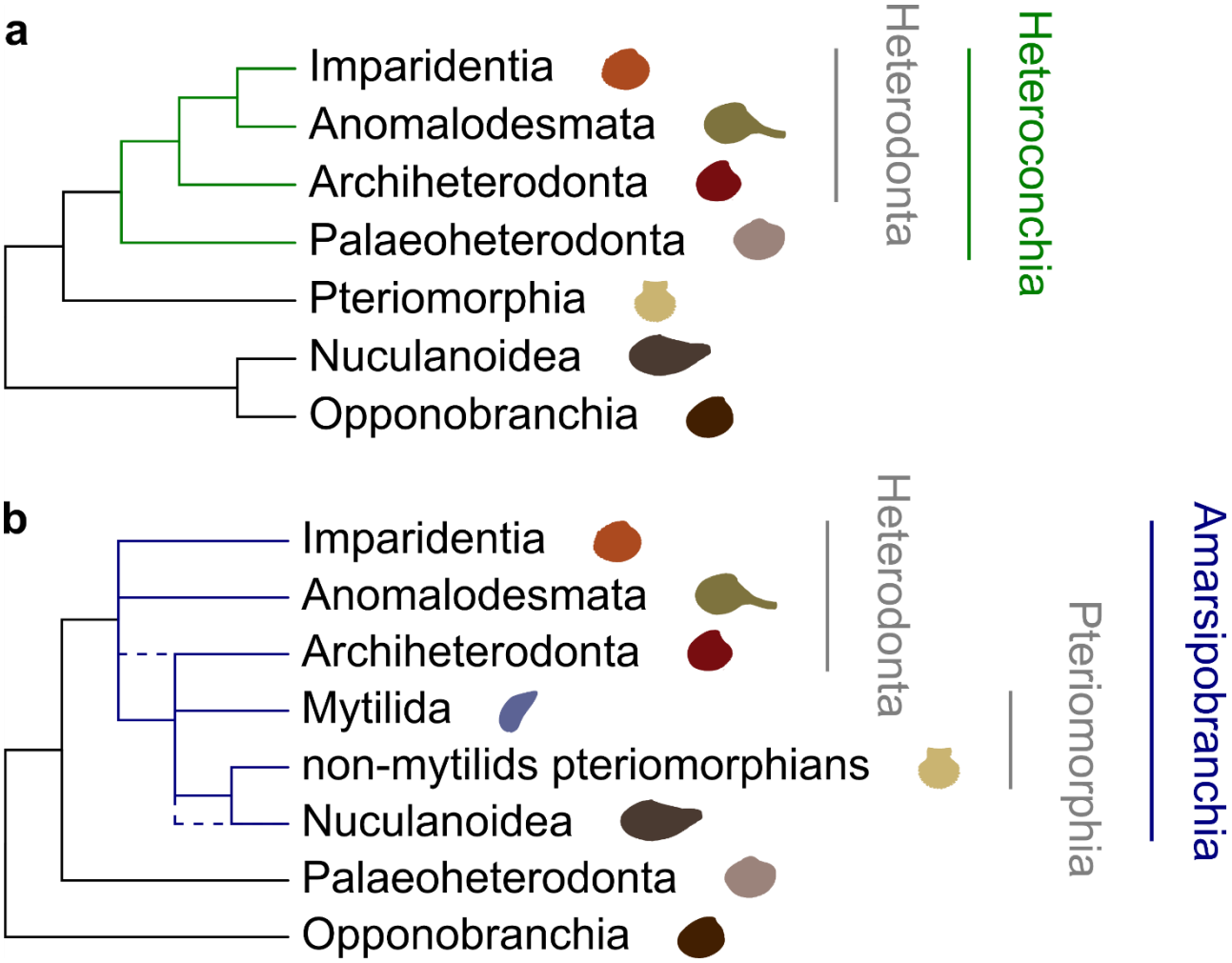


Figure 1. The two main alternative resolutions of the Bivalvia phylogenetic tree. (a) The Heteroconchia hypothesis. (b) The Amarsipobranchia hypothesis.

The OXPHOS genes and mito-nuclear coevolution

The massive ATP production of aerobic respiration in eukaryotes is mostly made possible through the oxidative phosphorylation (OXPHOS) pathway, which takes place across the inner mitochondrial membrane. OXPHOS pathway is carried out by five enzymatic complexes (CI-V). The genes encoding for the subunits are mostly located in the nuclear genome (around 70 genes), while 13 genes are typically harbored in the mitochondrial genome (mtDNA), at least in most bilaterians. All the complexes but Complex II (CII) have cooperating

subunits that are encoded by genes that are located on two different genomes, which show different mutation rate, population size and way of inheritance (Sloan et al. 2018).

In particular, the low recombination rate of mtDNA leads to the accumulation of slightly deleterious mutations (Lynch 1996). This process would affect the efficiency of OXPHOS, but slightly negative mutations can be counterbalanced by compensatory mutations in the nuclear genes (Osada and Akashi 2012) or even by new nuclear subunits added to the OXPHOS complexes (van der Sluis et al. 2015). According to this model of mito-nuclear coevolution, the process is driven by the accumulation of slightly deleterious mitochondrial mutations, which affects the selective pressure on the interacting nuclear subunits. Indeed, a correlation between the amino acid substitution rate of mitochondrial genes and their interacting nuclear counterparts was shown (Havird et al. 2015; Rockenbach et al. 2016; Weng et al. 2016). The evolutionary rate correlation (ERC; Wolfe & Clark, 2015) analysis is considered highly reliable to detect signals of mito-nuclear coevolution (Yan et al. 2019) and bivalves are among the clades where a positive ERC has been identified (Yan et al. 2019; Forsythe et al. 2021; Piccinini et al. 2021).

Quite surprisingly, the Amarsipobranchia clade is also supported by nuclear genes encoding for the OXPHOS subunits (Piccinini et al. 2021). Moreover, nuclear and mitochondrial OXPHOS genes show significant ERC and a similar dN/dS ratio (Piccinini et al. 2021) (the ratio between nonsynonymous substitution rate and the synonymous substitution rate; Nielsen, 2005).

The mtDNA of bivalves has a highly variable architecture, showing features that are unique among metazoans. Gene order is not conserved inside the class and the high frequency of rearrangements prevents to infer an ancestral gene order for Autobranchia (Ghiselli et al. 2021). Among Protobranchia, in the mitochondrial sequence of *Solemya velum* the leading strand, which is also the AC-rich one, harbours the genes *co1, co2, co3, nadh1, nadh2, nadh4, nadh4L and nadh5*, whereas the other strand harbours the genes *atp8, atp6, cytb, nadh1* and *nadh6* (Plazzi et al. 2013). Among Bivalvia, this is likely the most ancestral gene arrangement (Plazzi et al. 2013).

In Palaeoheterodonta the genome organization is highly conserved, and notable rearrangements were never detected within this subclass. Most of the protein coding genes are retained on the GT-rich strand (*atp6, atp8, co1, co2, co3, nadh3, nadh4, nadh4L* and *nadh5*), whereas the other strand harbours *cytb, nadh1, nadh2* and *nadh6* (Guerra et al. 2017).

Heterodonta and Pteriomorphia show a high degree of rearrangement. Few blocks of genes are shared between different orders of the same clade, and sometimes even among the same family (Ren et al. 2010). It is however worth noting that in Heterodonta and Pteriomorphia all genes are retained on one strand, which is rich in G+T.

As a matter of fact, the unidirectional replication of the mitochondrial genome leads to an asymmetric nucleotide composition of the two strands, increasing the G+T content in the heavy strand (Saccone et al. 1999). Most metazoans harbour most of the genes on the light strand, which is rich in A+C, but mollusks show an inverted pattern, in that in these species most of the genes are located on the GT-rich strand (Sun et al. 2018; Formaggioni et al. 2021). The position of mitochondrial genes on different strands has already been reported as a source of phylogenetic artifacts (Hassanin et al. 2005; Sun et al. 2018). Thus, there could be a relationship between the diverging phylogenetic signal of the mitochondrial markers and the location of some genes in Palaeoheterodonta compared to Heterodonta and Pteriomorphia.

In this study, I performed a phylogenetic analysis using mitochondrial (mt-OXPHOS) and nuclear OXPHOS (nu-OXPHOS) markers, exploiting different phylogenetic approaches. For the sake of comparison, I added two more datasets: genes related to the glycolytic pathway and the genes related to the biogenesis of regulative small noncoding RNAs (sncRNAs). I also analysed different features of markers selected for phylogenies: how the phylogenetic signal is distributed along the genes, codon usage, amino acid composition and strand location of the markers. I tested possible relationships between these features and the retrieved phylogenetic signals,

Regardless of the phylogenetic method, the Amarsipobranchia are supported only by the OXPHOS markers, both nuclear and mitochondrial. This phylogenetic signal is mostly retained in the organellar markers; among nuclear genes, subunits in direct contact with the mitochondrial counterparts lend most support to this topology. Moreover, I report an unbalanced nucleotide and amino acid composition between Amarsipobranchia and the Palaeoheterodonta, with a higher guanine and thymine content in the latter clade. I suggest that this pattern might be related to a different transcriptional mechanism, which has driven the mitochondrial phylogenetic signal to support Amarsipobranchia.

2.2 Results

The phylogenetic analysis on the four datasets

The datasets were comprised by 35 species, for four species two mitochondrial haplotypes were sampled (i.e., the female and male mitochondrial haplotypes; see below) (Table 1). All four datasets were incomplete, glycolysis being the most incomplete matrix (Supplementary table S1). Conversely, the mt-OXPHOS dataset was the most complete. Species showed a different range of completeness as well: *Myzuhopecten yessoensis* was the most complete species, while the outgroup *Graptacme eborea* was the least complete species (Supplementary figure S1). After the masking step, the mt-OXPHOS dataset was the shortest but also that with the lowest number of discarded sites. The longest dataset was the glycolysis one; the sncRNAs dataset was that with the highest number of discarded sites (Supplementary table S1).

Clade	**Order**	**Family**	**Species**
Protobranchia	Nuculida	Nuculanidae	*Ennucula tenuis*
	Solemyida	Solemyidae	*Solemya velum*
	Nuculanida	Sareptidae	*Aequiyoldia eightsii*
Pteriomorphia	Pectinida	Pectinidae	*Amusium pleuronectes*
	Pectinida	Pectinidae	*Mizuhopecten yessoensis*
	Arcida	Arcidae	*Tegillarca granosa*
	Ostreida	Ostreidae	*Magallana angulata*
	Ostreida	Ostreidae	*Saccostrea glomerata*
	Ostreida	Pinnidae	*Pinna atropurpurea*
	Ostreida	Margaritidae	*Pinctada margaritifera*
	Mytilida	Mytilidae	*Bathymodiolus azoricus*
	Mytilida	Mytilidae	*Mytilus edulis* (F and M)
	Mytilida	Mytilidae	*Perna viridis*
Palaeoheterodonta	Unionida	Unionidae	*Cristaria plicata* (F and M)
	Unionida	Unionidae	*Lampsilis cardium*
	Unionida	Unionidae	*Sinohyriopsis cumingii* (F and M)
	Unionida	Maragaritiferidae	*Margaritifera margaritifera*

	Trigoniida	Trigoniidae	*Neotrigonia margaritacea*
Anomalodesmata	Laternulidae	Pandorida	*Laternula elliptica*
	Lyonsiidae	Pandorida	*Lyonsia floridana*
Imparidentia	Venerida	Acticidae	*Arctica islandica*
	Venerida	Cyrenidae	*Corbicula fluminea*
	Venerida	Mactridae	*Mactra chinensis*
	Venerida	Veneridae	*Paratapes textilis*
	Venerida	Veneridae	*Ruditapes philippinarum* (F and M)
	Venerida	Veneridae	*Ruditapes decussatus*
	Venerida	Glossidae	*Glossus humanus*
	Myida	Myidae	*Mya arenaria*
	Sphaeriida	Sphaeriidae	*Sphaerium nucleus*
	Adapendonta	Pharidae	*Sinonovacula constricta*
	Galeommatida	Galeommatidae	*Galeomma turtoni*
Outgroups	Dentaliida	Dentalidae	*Graptacme eborea*
	Octopoda	Octopodidae	*Octopus bimaculoides*
	Chitonida	Acanthochitonidae	*Acanthochitona crinita*
	Lepetellida	Haliotidae	*Haliotis tuberculata*

Table 1. List of species included in the phylogenetic analysis divided by higher classification taxa, orders and families according to Carter and colleagues (2011) and WoRMS database (WoRMS Editorial Board 2022).

The three maximum-likelihood (ML) trees and the single Bayesian tree inferred from the mt-OXPHOS dataset were never significantly different and did not show any alternative resolution of major clades (Fig. 2a, Supplementary figure S2 and table S2). Protobranchia were basal, exception made for *Aequiyoldia eightsii* (Nuculanida), which clusters within Amarsipobranchia. Autobranchia were fully supported by all four trees. Then, the tree was divided into Amarsipobranchia and Palaeoheterodonta, both fully supported. The Amarsipobranchia were divided into Heterodonta and a clade comprised by *A. eightsii* and Pteriomorphia. Within this clade a polytomy between *A. eightsii,* Mytilida (*Perna viridis*, *Bathymodiolus azoricus*, *Mytilus edulis*) and the other pteriomorphians was recovered. Heterodonta were split into Imparidentia and Anomalodesmata, both fully supported.

The ML and Bayesian trees inferred from the nu-OXPHOS dataset were never significantly different and did not show any alternative resolution of major clades (Fig. 2b, Supplementary figure S3 and table S2). Protobranchia were basal, but monophyletic in the MrBayes tree only (Fig. 2b); according to the other trees this group was not monophyletic or not robustly supported (Supplementary figure S3). As for the mt-OXPHOS dataset, Autobranchia were split into Palaeoheterodonta and monophyletic Amarsipobranchia. Amarsipobranchia were divided into Pteriomorphia and Heterodonta, and the latter clade was split into Anomalodesmata and Imparidentia; all these clades were fully supported. Within Pteriomorphia, Mytilida are the sister group of remaining OTUs.

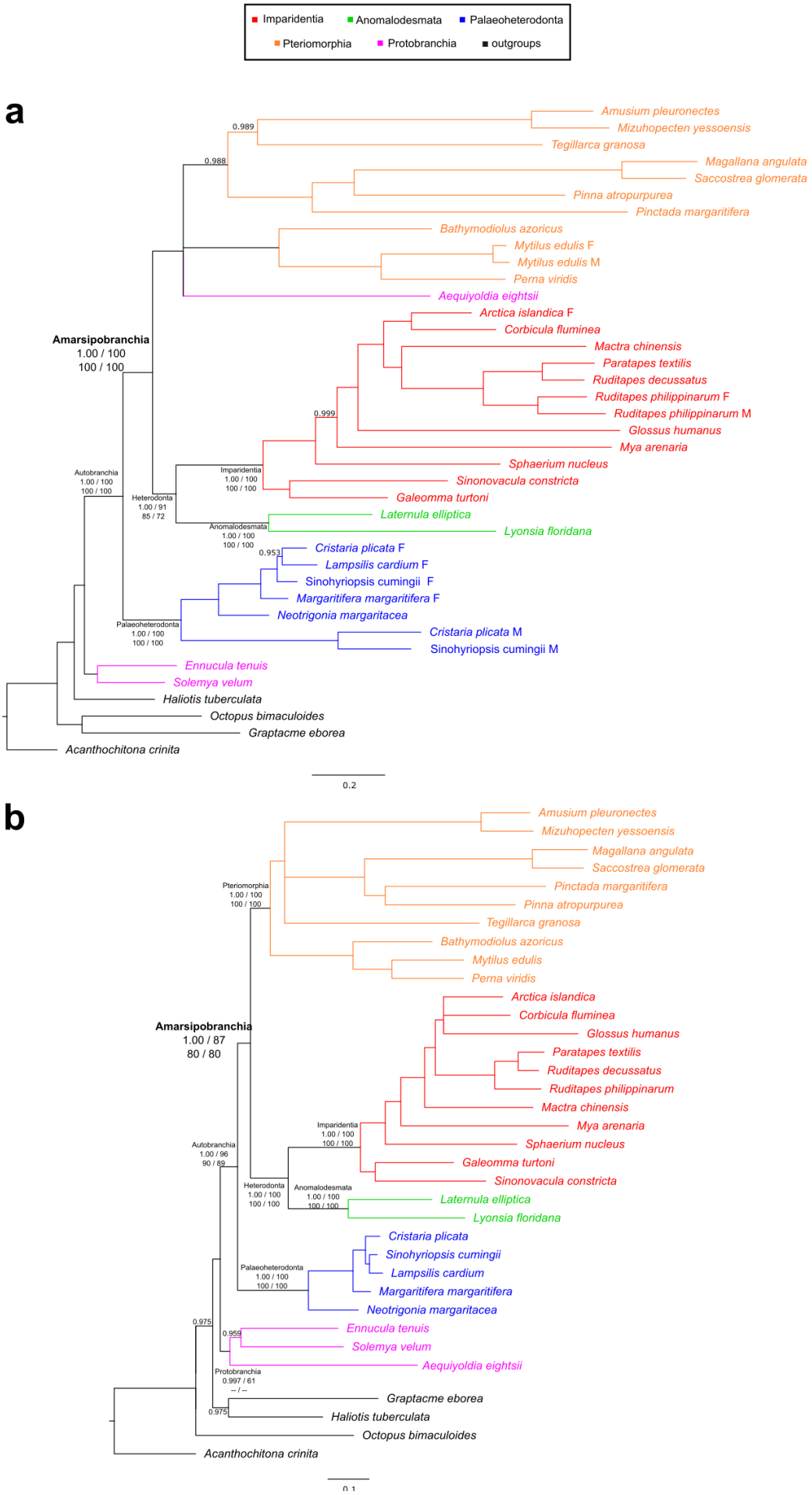

Imparidentia
Anomalodesmata
Palaeoheterodonta
Pteriomorphia
Protobranchia
outgroups
a
Amusium pleuronectes
Mizuhopecten yessoensis
Tegillarca granosa
Magallana angulata
Saccostrea glomerata
Pinna atropurpurea
Pinctada margaritifera
Bathymodiolus azoricus
Mytilus edulis F
Mytilus edulis M
Perna viridis
Aequiyoldia eightsii
Arctica islandica F
Corbicula fluminea
Mactra chinensis
Paratapes textilis
Ruditapes decussatus
Ruditapes philippinarum F
Ruditapes philippinarum M
Glossus humanus
Mya arenaria
Sphaerium nucleus
Sinonovacula constricta
Galeomma turtoni
Laternula elliptica
Lyonsia floridana
Cristaria plicata F
Lampsilis cardium F
Sinohyriopsis cumingii F
Margaritifera margaritifera F
Neotrigonia margaritacea
Cristaria plicata M
Sinohyriopsis cumingii M
Ennucula tenuis
Solemya velum
Haliotis tuberculata
Octopus bimaculoides
Graptacme eborea
Acanthochitona crinita
Amarsipobranchia
1.00 / 100
100 / 100
Autobranchia
1.00 / 100
100 / 100
Heterodonta
1.00 / 91
85 / 72
Imparidentia
1.00 / 100
100 / 100
Anomalodesmata
1.00 / 100
100 / 100
Palaeoheterodonta
1.00 / 100
100 / 100
0.989
0.988
0.999
0.953
0.2
b
Amusium pleuronectes
Mizuhopecten yessoensis
Magallana angulata
Saccostrea glomerata
Pinctada margaritifera
Pinna atropurpurea
Tegillarca granosa
Bathymodiolus azoricus
Mytilus edulis
Perna viridis
Arctica islandica
Corbicula fluminea
Glossus humanus
Paratapes textilis
Ruditapes decussatus
Ruditapes philippinarum
Mactra chinensis
Mya arenaria
Sphaerium nucleus
Galeomma turtoni
Sinonovacula constricta
Laternula elliptica
Lyonsia floridana
Cristaria plicata
Sinohyriopsis cumingii
Lampsilis cardium
Margaritifera margaritifera
Neotrigonia margaritacea
Ennucula tenuis
Solemya velum
Aequiyoldia eightsii
Graptacme eborea
Haliotis tuberculata
Octopus bimaculoides
Acanthochitona crinita
Amarsipobranchia
1.00 / 87
80 / 80
Pteriomorphia
1.00 / 100
100 / 100
Autobranchia
1.00 / 96
90 / 89
Imparidentia
1.00 / 100
100 / 100
Heterodonta
1.00 / 100
100 / 100
Anomalodesmata
1.00 / 100
100 / 100
Palaeoheterodonta
1.00 / 100
100 / 100
Protobranchia
0.997 / 61
-- / --
0.975
0.959
0.975
0.1

Figure 2. Bayesian trees inferred from the two OXPHOS datasets. (a) The mt-OXPHOS tree inferred through MrBayes. (b) The nu-OXPHOS tree inferred through MrBayes. Notably, both trees support the Amarsipobranchia hypothesis.The posterior proability on each node is reported when lower than 1.00; nodes with posterior probability lower than 0.95 were collapsed. Major nodes are annotated and support values of each of the four trees inferred for the present work are shown, as follows: MrBayes poterior probability, partitioned and mixture-model IQ-TREE UFBoot values, and RAxML bootstrap value. A double dash instead of the support means that the clade is not monophyletic in that tree. Red, Imparidentia; green, Anomalodesmata; blue, Palaeoheterodonta; orange, Pteriomorphia, purple, Protobranchia; outgroups are shown in black.

The ML and Bayesian trees inferred from the sncRNAs dataset were never significantly different and did not show any alternative resolution of the main clades (Fig. 3a, Supplementary figure S4 and table S2). Overall, several phylogenetic relationships were not resolved and some species were placed in unexpected major clades. After the separation of *Ennucula tenuis*, there was a polytomy with 6 branches: Heteroconchia; Mytilida + Ostreida, exception made for *Pinna atropurpurea*; Pectinida; *A. eightsii* + *P. atropurpurea*; *Tegillarca granosa*; *Solemya velum* (Fig. 3a). Heteroconchia were divided into Palaeoheterodonta and Heterodonta. Heterodonta were split into Anomalodesmata and Imparidentia, even if within the latter clade the palaeoheterodont *Margaritifera margaritifera* was recovered, which does belong to freshwater mussels.

The ML and Bayesian trees inferred from the glycolysis dataset were never significantly different and did not show any alternative resolution of major clades (Fig. 3b, Supplementary figure S5 and table S2). A long branch led to the Bivalvia node, which further separated Pteriomorphia from other bivalves, leading to the paraphyly of Autobranchia. Namely, Protobranchia and Heteroconchia clustered into a monophyletic group that was supported by all four trees. Heteroconchia were split into Palaeoheterodonta and Heterodonta. The latter clade was divided in Anomalodesmata and Imparidentia; all these clades were fully supported. Within major clades all relationships were resolved and supported and the pteriomorphian and imparidentian species clustered in the expected orders.

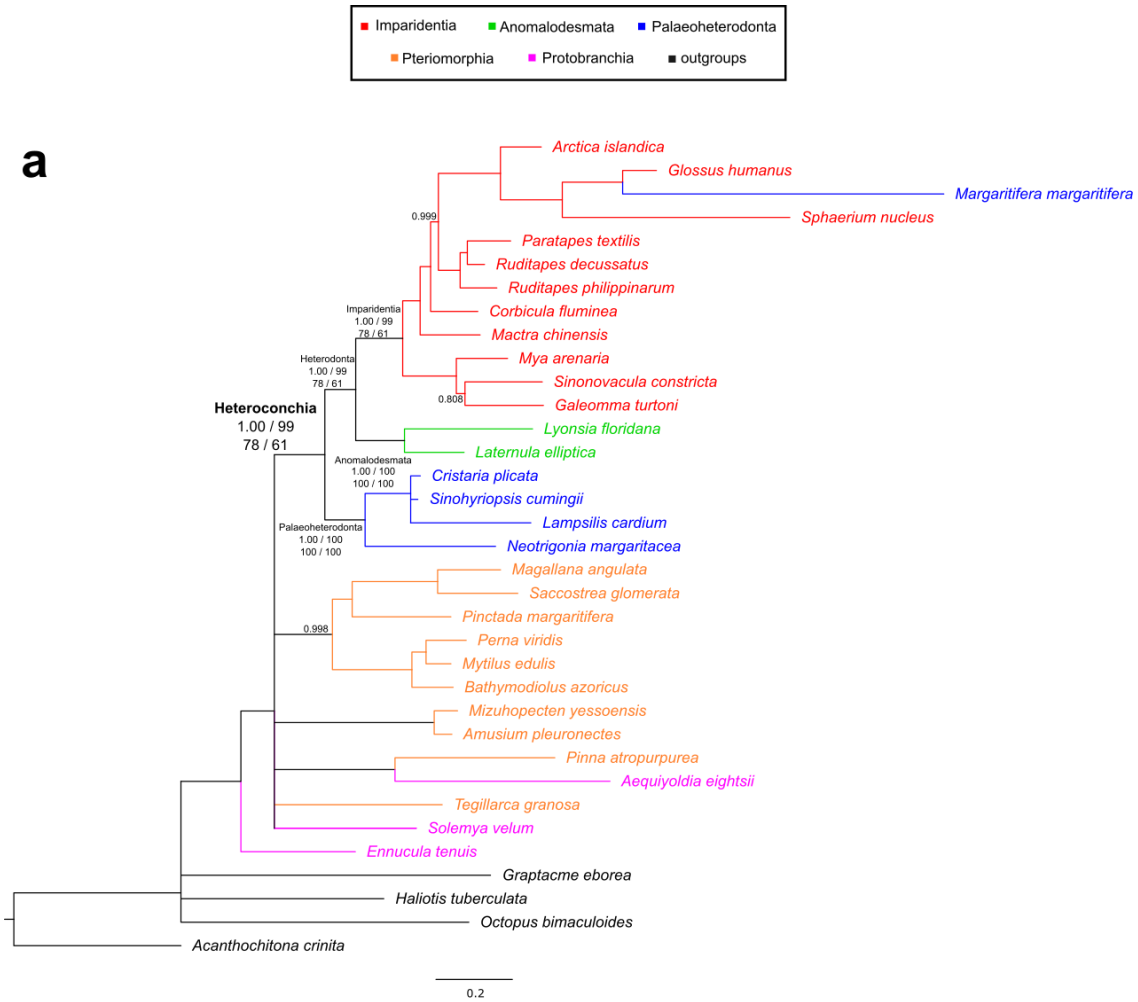

Imparidentia
Anomalodesmata
Palaeoheterodonta
Pteriomorphia
Protobranchia
outgroups
a
Arctica islandica
Glossus humanus
Margaritifera margaritifera
Sphaerium nucleus
0.999
Paratapes textilis
Ruditapes decussatus
Ruditapes philippinarum
Imparidentia
1.00 / 99
78 / 61
Corbicula fluminea
Mactra chinensis
Heterodonta
1.00 / 99
78 / 61
Mya arenaria
Sinonovacula constricta
0.808
Heteroconchia
1.00 / 99
78 / 61
Galeomma turtoni
Lyonsia floridana
Laternula elliptica
Anomalodesmata
1.00 / 100
100 / 100
Cristaria plicata
Sinohyriopsis cumingii
Palaeoheterodonta
1.00 / 100
100 / 100
Lampsilis cardium
Neotrigonia margaritacea
Magallana angulata
Saccostrea glomerata
Pinctada margaritifera
0.998
Perna viridis
Mytilus edulis
Bathymodiolus azoricus
Mizuhopecten yessoensis
Amusium pleuronectes
Pinna atropurpurea
Aequiyoldia eightsii
Tegillarca granosa
Solemya velum
Ennucula tenuis
Graptacme eborea
Haliotis tuberculata
Octopus bimaculoides
Acanthochitona crinita
0.2
b
Corbicula fluminea
Glossus humanus
Arctica islandica
Paratapes textilis
0.961
Ruditapes decussatus
Ruditapes philippinarum
Mya arenaria
0.996
Mactra chinensis
Imparidentia
1.00 / 100
100 / 100
Sphaerium nucleus
Heterodonta
1.00 / 100
100 / 100
Sinonovacula constricta
Galeomma turtoni
Heteroconchia
1.00 / 100
100 / 100
Laternula elliptica
Anomalodesmata
1.00 / 100
100 / 100
Lyonsia floridana
Cristaria plicata
Sinohyriopsis cumingii
Protobranchia +
Heteroconchia
0.964 / 87
75 / 75
Lampsilis cardium
Margaritifera margaritifera
Palaeoheterodonta
1.00 / 100
100 / 100
Neotrigonia margaritacea
Ennucula tenuis
Aequiyoldia eightsii
Solemya velum
Protobranchia
1.00 / 99
99/ 98
Amusium pleuronectes
Mizuhopecten yessoensis
Bathymodiolus azoricus
Mytilus edulis
Perna viridis
Pteriomorphia
1.00 / 100
100/ 100
Magallana angulata
Saccostrea glomerata
Pinna atropurpurea
Pinctada margaritifera
Tegillarca granosa
Haliotis tuberculata
Octopus bimaculoides
Graptacme eborea
Acanthochitona crinita
0.2

Figure 3. Bayesian trees inferred from the sncRNAs and glycolysis datasets. (a) The sncRNAs tree inferred through MrBayes. (b) The glycolysis tree inferred through MrBayes. Notably, both trees support the Heteroconchia hypothesis. The posterior proability on each node is reported when lower than 1.00; nodes with posterior probability lower than 0.95 were collapsed. Major nodes are annotated and support values of each of the four trees inferred for the present work are shown, as follows: MrBayes poterior probability, partitioned and mixture-model IQ-TREE UFBoot values, and RAxML bootstrap value. A double dash instead of the support means that the clade is not monophyletic in that tree. Red, Imparidentia; green, Anomalodesmata; blue, Palaeoheterodonta; orange, Pteriomorphia, purple, Protobranchia; outgroups are shown in black.

Concluding, notwithstanding some issues with the major clade of Protobranchia which blurred the comparison and the substantial overlapping of all phylogenetic trees, the Amarsipobranchia hypothesis was supported in both OXPHOS datasets, while the Heteroconchia hypothesis was supported in the glycolysis and sncRNAs datasets. Henceforth, I will use mt-topology to refer to the Amarsipobranchia hypothesis and nuc-topology for the Heteroconchia hypothesis.

Phylogenetic signal and its distribution across markers and complexes

Markers belonging to the same dataset may support a different phylogenetic signal. Gene concordance factor (gCF), site concordance factor (sCF; Minh et al. 2020) and ultrafast bootstrap approximation (Hoang et al. 2018) (UFBoot) were calculated for the Heteroconchia and Amarsipobranchia (which represent alternative resolutions of a node). The mt-OXPHOS dataset showed high support for Amarsipobranchia according to each value (UFBoot = 100; gCF = 30.8; sCF = 48.6), and low support for the Heteroconchia (UFBoot = 0; gCF = 0; sCF = 25.5). Despite a non-zero gCF suggests more markers concordant with the nuc-topology than with the mt-topology, the nu-OXPHOS dataset similarly favors mt-topology (UFBoot = 87; gCF = 3.57; sCF = 37.2) against nuc-topology (UFBoot = 12; gCF = 5.36; sCF = 32.5). Regarding the sncRNAs and glycolysis datasets, markers are more concordant with Heteroconchia, since the UFBoot, gCF and sCF calculated for this topology are considerably higher (Supplementary table S3).

For the two OXPHOS datasets I clustered the markers according to the OXPHOS complexes; the sCF for each complex was computed; moreover, it was compared to the sitewise log-likelihood score (SLS) calculated for both topologies. The difference between the mt-topology sitewise log-likelihood score and the nuc-topology sitewise log-likelihood score (ΔSLS) can tell which topology is favored by each site: sites with $\Delta SLS > 0$ supports the mt-topology; sites with $\Delta SLS < 0$ support the nuc-topology. Moreover, by summing all the ΔSLS within a complex I obtained a complexwise log-likelihood score (ΔCLS; Table 2; Castoe et al., 2009; Shen et al., 2017). Since the summed ΔCLS highly depends on the number of sites within each complex, I divided the

ΔCLS for the number of sites of each complex (average ΔCLS). For the mitochondrial markers that belong to CI I made a distinction between those *nadh* genes that in Palaeoheterodonta are on the plus strand (CI-ps) from those *nadh* genes located on the minus strand (CI-ms), since I was willing to test if the mt-topology is mostly supported in the genes that are in different strands in Palaeoheterodonta and Amarsipobranchia (i.e. *cytb, nadh1,2,6*; see Introduction).

All the mitochondrial groups (Table 2) show a positive ΔCLS; a positive average ΔCLS; more sites that strongly support the mt-topology; more sites in the alignment that agree with the mt-topology. The only exception is CI-ps, where the ΔCLS and average ΔCLS are negative, although the other statistics follow the pattern of the other groups.

Complexes III to V of the nu-OXPHOS dataset (Table 2) support Amarsipobranchia; sites that strongly support the mt-topology (with ΔSLS > 0.5) are more than those supporting the nuc-topology and most sites in the alignment agree with the mt-topology. Contrastingly, CI and CII do not support Amarsipobranchia. In CII there is an equal number of sites for either topology, while in the CI those with a ΔSLS > 0.5 are more. The sCF calculated on the nuc-topology is higher in CII and almost equal in CI with respect to the sCF calculated on the mt-topology.

Group	ΔCLS	Average ΔCLS	%ΔSLSs > 0.5	%ΔSLSs < −0.5	mt-sCF	nuc-sCF
nu-OXPHOS dataset						
CV	20.1	0.0079	0.90%	0.35%	40.9	30.7
CIV	6.2	0.0043	1.39%	1.04%	42.6	31.0
CIII	8.9	0.0097	1.19%	0.59%	42.1	33.6
CII	−7.9	−0.0073	0.83%	0.83%	32.9	36.3
CI	−1.0	−0.00019	0.96%	0.72%	35.4	35.0
mt-OXPHOS dataset						
CV	5.9	0.0380	3.22%	1.29%	48.1	26.6
CIV	25.3	0.0272	3.11%	0.75%	54.6	22.8

CIII (*cytb*)	6.3	0.0181	1.97%	0.28%	42.6	28.4
CI-ms	11.4	0.0110	2.14%	0.77%	49.7	24.8
CI-ps	−7.2	−0.0090	1.62%	0.85%	38.1	32.4

Table 2. The phylogenetic signal of nu and mt-OXPHOS markers grouped by complexes. The CI mt-markers are split into two groups: CI-ms is comprised by nadh1,2,6 and CI-ps is comprised by nadh3,4,4l,5. For each group it was calculated: ΔCLS; average ΔCLS; percentage of sites with %ΔSLSs > 0.5; percentage of sites with %ΔSLSs < −0.5; sCF for the mt-topology; sCF for the nuc-topology.

Overall, in all complexes ΔCLS, average ΔCLS and sCF variate together; statistics related to the strongly supporting sites do not always follow the same pattern, since CI shows a negative ΔCLS but a higher number of sites with ΔSLS > 0.5.

To test whether the mt-topology phylogenetic signal is mostly retained in the nu-OXPHOS subunits that interact with the mitochondrial subunits, I calculated the sCF referred to each marker and I split the markers into two groups: those that are in direct contact with the mitochondrial counterparts and those that are not. The sCF values of the "contact" nu-OXPHOS markers are significantly higher than the values of "non-contact" nu-OXPHOS makers (p-value=0.006363; Fig. 4).

In the sncRNAs and glycolysis datasets the ΔSLS was calculated on the whole matrix. In both datasets the average ΔSLS is negative and there are more sites strongly supporting Heteroconchia (Supplementary table S3).

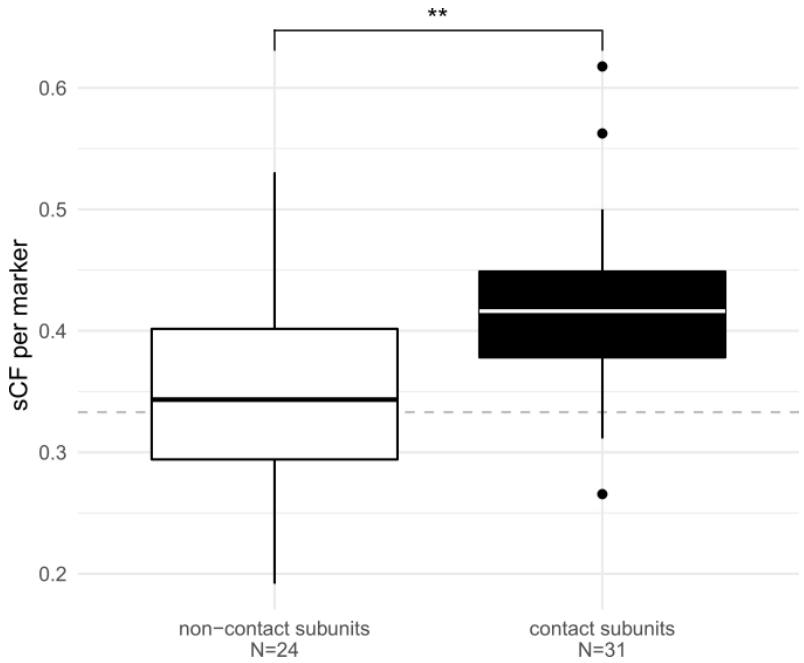


Figure 4. The phylogenetic signal in "contact" and "non-contact" nu-OXPHOS subunits. Boxplot comparing the sCF of nu-OXPHOS markers in direct contact with mitochondrial subunits and the sCF of nu-OXPHOS markers that are not in direct contact. Dashed gray line at 0.33 marks the threshold below which the branch with the highest figure of sCF between the three possible resolutions is not the one that support the mt-topology. Significance calculated through a Student's t-test (t=−2.862, p-value=0.006363, d.f.=23, 30).

Nucleotide composition and mitochondrial topology

I placed attention on the nucleotide asymmetry between the two mitochondrial strands, which can be assessed calculating the AT skew and GC skew (Reyes et al. 1998). In Bivalvia the plus strand is richer in guanines and thymines than the minus strand (Yu and Li 2011; Sun et al. 2018). Thus, I also analyzed possible dissimilarities in the guanine and thymine content (G+T content) between markers that in Pteriomorphia, Imparidentia, Anomalodesmata and Palaeoheterodonta are on the same strand (i.e., *atp6,8, cox1-3, nadh3-5*; Supplementary table S4).

For each mt-OXPHOS marker of each species I calculated the AT skew, the GT content, the frequency of codons with guanines or thymines at the first and the second position (GT-rich codons) and the GT content at the third position of four-fold degenerated codons (Fig. 5). Among Imparidentia, Anomalodesmata, Pteriomorphia and Palaeoheterodonta the markers show an AT skew < 0 and a GT content > 0.5. On average, Palaeoheterodonta show the highest values in all the statistics but the AT skew (Fig. 5a). Indeed, Palaeoheterodonta are always significantly different from the other groups, with the only exception of Anomalodesmata in GT-rich codons (Fig. 5b). On the other hand, the comparisons between Pteriomorphia and Imparidentia are never significant. Regarding the outgroups and Protobranchia values, data show a high standard deviation in most of the cases. The only exception is *A. eightsii,* whose statistics are in line with the values of Imparidentia and Pteriomorphia.

Finally, I studied if the nucleotide compositional patterns outlined in the protein coding regions were extended to the unassigned regions (URs): I downloaded the mitochondrial genomes available on NCBI of all the species that belong to Imparidentia, Anomalodesmata, Palaeoheterodonta, Pteriomorphia and Protobranchia. Then, I calculated the GT content in the URs of the genomes. The GT content of URs calculated on 92 Palaeoheterodonta entries is significantly higher than the one calculated on 77 Pteriomorphia entries, 70 Imparidentia entries and 4 Protobranchia entries. Conversely, it is not significantly higher than the one calculated on 6 Anomalodesmata entries (Supplementary figure S6). For what concerns the other comparisons, no clade is significantly different from any other.

Overall, the nucleotide composition of Palaeoheterodonta mt-OXPHOS markers is most of the times significantly different from the one of other major clades. In particular, I detected a higher GT content. This pattern is reflected in all codon positions as well as in the URs. On the other hand, statistics are overlapping between Imparidentia and Pteriomorphia.

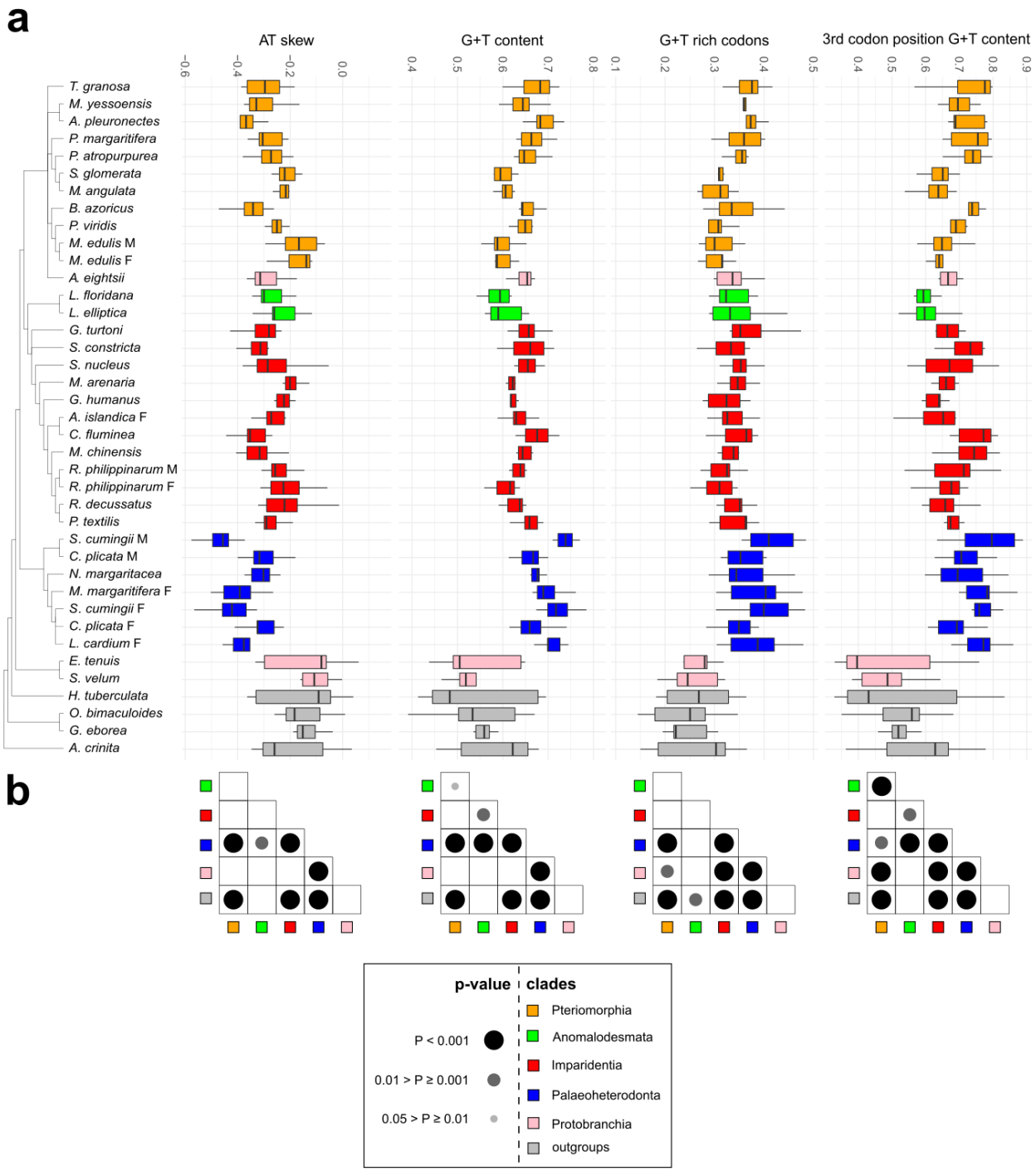


Figure 5. Nucleotide and codon composition statistics in mt-OXPHOS markers compared across OTUs. (a) OTUs are reported according to the mitochondrial consensus tree. The x-axis is divided in four boxplots with no outliers, each one reports a different statistic calculated on a set of mt-OXPHOS markers (*atp6,8, cox1-3, nadh3-5*). From left to right, plots report the AT skew, the GT content, the frequency of codons that have T or G at first and second codon position, and the GT content of the third codon position in four-fold degenerated codons, respectively. (b) Each table reports the significance of pair-wise comparisons between the values

reported in the plot right above grouped according to the six clades. The significance is calculated through the Dunn test with the Bonferroni correction. Black and grey dots inside the table mark the significant comparisons; as reported in the legend, the bigger and the darker the dot, the more significant the comparison.

2.3 Discussion

For all four datasets, the more recent nodes were resolved and highly supported. In Imparidentia and Pteriomorphia the OTUs were always placed in the expected orders and major clades (i.e. Pteriomorphia, Anomalodesmata, Imparidentia, Palaeoheterodonta and Protobranchia). Only few exceptions were detected, i.e. the position of *A. eightsii* in the sncRNAs and mt-OXPHOS trees and the position of *M. margaritifera* in the sncRNAs trees. The latter was likely a long branch attraction bias (Felsenstein 1978), since the final branches of the OTU and its sister species were the longest in the tree (Fig. 3a). Generally speaking, I regard to these misplacements as minor phylogenetic issues in the broader figure of deep evolutionary relationships among bivalves, which do not significantly blur the topology connecting major clades.

Major clades were retrieved with higher support and with better resolution from the OXPHOS datasets with respect to the glycolysis and sncRNAs dataset. Overall, OXPHOS genes are known to be more conservative, therefore these markers might be more informative in the resolutions of cladogenetic events dating to the Ordovician, approximately 470-480 million years ago (Mya; Cope 1996; Sánchez 2006; Fang and Sanchez 2012).

The mt-OXPHOS trees were mostly coherent with the previous mitochondrial phylogenetic analyses, exception made for the monophyly of the Heterodonta with Anomalodesmata inferred from our analysis (Doucet-Beaupré et al. 2010; Plazzi and Passamonti 2010; Plazzi et al. 2011; Stöger and Schrödl 2013; Plazzi et al. 2016; Sun et al. 2018; Piccinini et al. 2021). Since all the nu-OXPHOS trees supported the Amarsipobranchia clade (Fig. 1b), our data confirmed that the mt and nu-OXPHOS markers share the same phylogenetic signal, which is different from that inferred from transcriptome-wide analyses or other nuclear markers (Giribet and Wheeler 2002; Kocot et al. 2011; Smith et al. 2011; Bieler et al. 2014; González et al. 2015).

Among the interacting sites of coevolving proteins there are epistatic interactions, which lead the sites of both proteins to evolve at the same rate (Avila-Herrera and Pollard 2015). Bivalvia OXPHOS subunits show a positive ERC, which is the most solid clue of protein coevolution (Wolfe and Clark 2015; Yan et al. 2019). Our data enforce the hypothesis of mito-nuclear coevolution in bivalves, depicting a clear relationship between the phylogenetic signal of interacting subunits. Moreover, they provide an overview on how the phylogenetic signal of OXPHOS subunits may be biased under this type of interaction. The CF and ΔSSL analyses suggested that OXPHOS markers did not equally support the Amarsipobranchia, yet the two dataset were largely coherent with each other: CIII-V markers from both OXPHOS datasets largely supported the mt-topology; contrastingly, CI did not show a clear pattern, and the nuclear-only Complex II favours the nuc-topology. Moreover, the nu-OXPHOS subunits in contact with the mitochondrial counterparts were significantly more concordant with the

mt-topology than the subunits that are not directly in contact. Accordingly, previous analyses reported that the CII is the only complex that shows uncorrelated rates of evolution compared to the other subunits (Piccinini et al. 2021).

Finally, the support and concordance statistics (BP, PP, UFBoot, gCF and sCF) calculated for the mt-OXPHOS dataset on the Amarsipobranchia node were always equal to or higher than those calculated for the nu-OXPHOS dataset. Thus, the mt-topology in the first dataset was more consistent: more sites and markers agreed with this topology and the signal was less susceptible to resampling.

The mito-nuclear coevolution is expected to be mainly driven by slightly deleterious mitochondrial mutations that are compensated by the nuclear genome (Osada and Akashi 2012; Sloan et al. 2018). Even if previous data did not show signal of nuclear compensation (Piccinini et al. 2021), it is tempting to conclude that the mitochondrial genome acquired the mutations leading to the mt-topology at first, and then the phylogenetic signal has been traced by interacting sites in the nuclear markers through nuclear compensation.

Pteriomorphia and Imparidentia share some unique mitochondrial features: their gene order is highly rearranged, but all genes are on the same strand. Contrastingly, Palaeoheterodonta show a highly conserved gene order, with a set of genes on the minus strand (*nadh1,2,6* and *cytb*; see Introduction for further details). According to the nucleotide composition analyses, the Palaeoheterodonta mt-OXPHOS markers were significantly GT-richer in each codon position as well as in URs, while Pteriomorphia and Imparidentia did not show any significant difference (Fig. 5, Supplementary figure S6). Thus, this pattern was accounted for either synonymous and non-synonymous substitutions and it was extended also to URs. Mito-nuclear coevolution largely explains why the nu and mt-OXPHOS markers support a common topology.

Mito-nuclear discordance is a quite common phenomenon and a multitude of processes can cause it (Funk and Omland 2003). The introgression of mitochondrial lines from a phylogenetically distant population is widely used to explain mito-nuclear discordance (Funk and Omland 2003; Toews and Brelsford 2012). In some cases it has been hypothesized that a set of nuclear genes might cointrogress to avoid mito-nuclear incompatibilities (Sloan et al. 2017). The mito-nuclear cointrogression would explain very well our data, since the phylogenetic artifact is mostly supported by the nuc-OXPHOS markers that directly interact with the mt-OXPHOS markers, whereas almost all mt-OXPHOS markers support the mt-topology. In this case, the use of a single mitochondrial strand and other features would be apomorphies arisen along a single branch and subsequently acquired by the other branch through introgression. Having said that, other evidences of cointrogression are limited and only restricted to populations within the same genus (Beck et al. 2015; Sloan et al. 2017; Morales et al. 2018). In our case the discordance mainly resides in the resolution of deep nodes, which originated around 480 Mya (Cope 1996; Fang 2006; Sánchez 2008), between clades that already evolved quite different life

habits (Fang 2006). Under this scenario, the mito-nuclear cointrogression might not be the most likely hypothesis.

Another source of mito-nuclear discordance can be found in how markers are located on the two mitochondrial strands (Hassanin et al. 2005). In mollusks, whose mitochondrial genome is highly rearranged, the nucleotide bias is also reflected in amino acid bias (Sun et al. 2018). My results showed that the signal supporting the mt-topology (the Amarsipobranchia clade) is not only retained in the set of markers that in Palaeoheterodonta are on the minus strand. Instead, the mt-OXPHOS CIV-V markers on the plus strand clearly favor the Amarsipobranchia hypothesis (Table 2). The higher GT content in Palaeoheterodonta is consistent throughout different parts of the mitochondrial genome, from coding to unassigned regions. Therefore, the nucleotide substitutions that have led to this pattern are likely to be produced by a process that act on the whole genome. Possible candidates might be the mitochondrial transcription and replication, which are indeed notable source of deamination (Saccone et al. 1999; Lawless et al. 2020). Moreover, mitochondrial replication constitutes the main source of mitochondrial point mutations, at least in humans (Zheng et al. 2006).

The position of all genes on the same strand is probably linked to the fact that the two clades do not show any significant difference in GT content. It is tempting to hypothesize that transcription involves the coding strand only for these mtDNAs and, thus, the aforementioned deamination effect may be less pronounced. Indeed, even smithRNAs, which were recently described in the imparidentian *R. philippinarum* (see Chapter 1), were annotated on the same coding strand (Pozzi et al. 2017), thus corroborating the idea that only one strand is transcribed in these clades.

The use of a single strand seems also linked to the mitochondrial architecture: among most of the metazoan taxon that share this feature it has been detected a higher mitochondrial rearrangement rate (Gissi et al. 2008; Plazzi et al. 2016; Malkócs et al. 2022). Likewise, Pteriomorphia and Imparidentia show highly rearranged mitochondrial genomes (Ren et al. 2010). An additional clue is the behavior of *A. eigthsii*: the protobranch species cluster with Pteriomorphia and shows similar nucleotide composition features (Fig. 2a, Fig. 5a). Indeed, although no mitochondrial genome has been annotated from the order Nuculanida, it is possible that *A. eigthsii* mitogenome harbours all the genes on the heavy strand, since all its mitochondrial genes show AT skew < 0 and GC skew > 0 (Supplementary table S5).

If the hypothesis holds true, it is reasonable to consider the different transcriptional patterns among Bivalvia as the most likely source that has led the mitochondrial genome to support a different phylogenetic signal, namely a biased one. Since I demonstrated that also non-synonymous mutations have shaped the GT content pattern, the modifications of the amino acid sequences could have altered the epistatic interactions between

nuclear and mitochondrial OXPHOS subunits, leading the OXPHOS markers to support the same phylogenetic artifact.

2.4 Conclusion

The results obtained from the phylogenetic analysis of Piccinini and colleagues (2021) has been confirmed by our work, since markers of both OXPHOS datasets support the same biased topology, regardless of the phylogenetic pipeline used. Moreover, I depicted how the coevolution process affected the phylogenetic signal in different set of OXPHOS markers, concluding that the artifactual topology is mainly supported by the OXPHOS subunits that interact more directly.

Considering that the phylogenetic signal is more stable and stronger in the mt-OXPHOS markers, I suggest that the biased topology arose for these markers at first, then it has been acquired also by the nu-OXPHOS markers through the coevolution of interacting subunits. This model agrees with the pattern of evolution hypothesized for the mito-nuclear coevolution. That is, the mito-nuclear coevolution is mainly driven by slightly deleterious mitochondrial mutations that are compensated by the nuclear genome (Osada and Akashi 2012; Sloan et al. 2018).

Our data suggest a relationship between the mt-topology supporting Amarsipobranchia and the gene rearrangements in the Bivalvia mitochondrial genome. The clades that harbour all the mitochondrial genes on a single strand and show a similar nucleotide composition (Pteriomorphia, Heterodonta, and possibly *A. eightsii*) are grouped together in a monophyletic clade. On the other side, Palaeoheterodonta show a peculiar nucleotide composition, which is not only due to the genes located on the minus strand. Indeed, genes such *cox1-3*, *atp6,8, nadh3-5*, even if they are located on the plus strand, show a higher GT content compared to the Amarsipobranchia ones. Overall, the difference in GT content between OTUs may be a source of possible phylogenetic artifacts. Further analyses will be focused on understanding how the nuclear subunits compensated differently during the evolution of Palaeoheterodonta, Pteriomorphia and Heterodonta.

Finally, according to the data, the reliability of the Amarsipobranchia clade should be reconsidered. At the state of the art, although many mitochondrial phylogenies confirmed the Amarsipobranchia clade (Stöger and Schrödl 2013; Plazzi et al. 2016; Sun et al. 2018; Piccinini et al. 2021), no phylogeny supports Amaripobranchia when based on nuclear markers (exception made for the nu-OXPHOS markers; Fig. 2b, Fig. S2; Piccinini et al. 2021). On the other side, the Heteroconchia clade has been retrieved by genome-wide, transcriptomic, and morphological analyses (Kocot et al. 2011; Smith et al. 2011; Bieler et al. 2014; González et al. 2015). If the evolutionary scenario depicted in our discussion is correct, then the taxon Amarsipobranchia cannot be supported anymore and has to be considered a phylogenetic artifact: the Heteroconchia clade should be regarded as a more reliable hypothesis instead.

2.5 Materials and Methods

The datasets

My phylogenetic analyses were performed on four datasets: mt and nu-OXPHOS genes, glycolytic pathway genes, and genes related to the biogenesis of sncRNAs. All markers were retrieved from the transcriptomes used by Piccinini and colleagues (2021): the transcriptomes of 35 molluscan species were assembled (Table 1). When available, the mt-OXPHOS markers from both sexes were retrieved for those species that show mitochondrial Doubly Uniparental Inheritance (DUI; Breton et al. 2007; Zouros and Rodakis 2019; Passamonti and Plazzi 2020).

Information about the assembly of transcriptomes is detailed in the aforementioned paper (Piccinini et al. 2021). Briefly, the annotation of transcripts was performed using BLASTx (Camacho et al. 2009) against a user-defined database and HMMER (Mistry et al. 2013) against the Pfam database 30.0 (El-Gebali et al. 2019); the user-defined database contains sequences of all genes of that dataset available for Bivalvia on NCBI.

Clam homologs for the first three datasets were extracted following the gene lists available in the Kyoto Encyclopedia of Genes and Genomes (KEGG: Kanehisa 2000; Kanehisa 2019; Kanehisa et al. 2021; Kanehisa et al. 2022), which provides a curated database of enzymes involved in specific biochemical pathways: namely, the Oxidative phosphorylation pathway (KEGG entry: map00190) and the Glycolysis / Gluconeogenesis pathway (KEGG entry: map00010). Regarding genes for the fourth dataset, i.e. genes related to the biogenesis of sncRNAs, I identified a set of genes shared across Metazoa (Ha and Kim 2014; Lewis et al. 2016). Entries available on NCBI and UniProt (Bateman et al. 2021) were included in the database (Supplementary table S6). Annotation was performed using BLASTp (Camacho et al. 2009).

Paralogs were recurrent among the markers associated to glycolysis. Therefore, I devised a method to conservatively distinguish paralogs from orthologs. I inferred the ML tree from each single marker putting orthologs together, which was obtained using IQ-TREE1.7 (Nguyen et al. 2015) with mixture model as model of evolution, 1,000 UFBoot (Hoang et al. 2018) replicates, and constraining the Bivalvia clade. Through the analysis of topologies, more than one group of clear monophyletic orthologs were detected in some cases, namely in the markers with KEGG ID K00002, K00128, K00129, K00149, K00627, K00844, K01596, K01623, K01689, K01785, K01895, K03103, K08074, and K13953 (Supplementary table S7). In these cases, groups of orthologs were split and considered as single markers. Aiming to ensure that the phylogenetic signal supported by the glycolysis matrix after this scrutiny was coherent, I retained two different datasets associated to glycolysis genes: a larger dataset with all markers obtained in this way (total-glyco) and a dataset with markers that showed no evidence of paralogs (partial-glyco). All subsequent analyses were carried out independently for both datasets; since differences in results were negligible, I am confident that I identified

paralogs correctly, thus in the results I mean the total-glyco dataset only when referring to the “glycolysis” dataset.

Phylogenetic reconstruction

We performed the phylogenetic analysis using amino acid sequences, since I were more interested in deep relationships and nucleic sequences are bound to saturate along long branches. First, I aligned sets of homologous markers with PSI-Coffee (Chang et al. 2012). Then, to remove the uninformative or misleading sites for the analysis, I used and combined the results of five different masking algorithms (Plazzi et al. 2016): BMGE (Criscuolo and Gribaldo 2010), Aliscore (Kück et al. 2010), Gblocks (Castresana 2000), ZORRO (Wu et al. 2012) and Noisy (Dress et al. 2008). This step was performed by masking_package 1.1, downloaded from GitHub and available at https://github.com/mozoo/masking_package. To include the indels in the phylogenetic reconstruction I ran GapCoder (Young and Healy 2003) on every alignment.

To assign the best-fitting evolutionary model to each marker of the matrix I used PartitionFinderProtein (Lanfear et al. 2012). All markers belonging to the same dataset were concatenated together. For each dataset I obtained four trees. (i) One tree was obtained through IQ-TREE 1.7 with the dataset partitioned according to the PartitionFinder results. (ii) One tree was obtained through IQ-TREE 1.7 with the mixture model as model of evolution (Nguyen et al. 2015). (iii) One tree was obtained through RAxML version 8.2.11 (Stamatakis 2014) with the dataset partitioned according to the PartitionFinder results, using the CAT model instead of the Gamma model (Stamatakis 2006). 1,000 bootstrap replicates were executed for each run, to test the robustness of the nodes, and the UFBoot approximation was chosen for IQ-TREE. (iv) The fourth tree is based on the Bayesian inference, obtained through MrBayes (Ronquist and Huelsenbeck 2003) with the dataset partitioned according to the PartitionFinder results. Number of generations was set to 10,000,000; the convergence between runs were manually checked to set the burn-in value. To set this value, I looked at the standard deviation of average split frequency over generations; moreover, I took the Potential Scale Reduction Factor (PSRF; Gelman and Rubin 1992) into consideration. In each analysis the monophyly of Bivalvia was constrained and in the Bayesian analysis the outgroup was set to be the polyplacophoran *Acanthochitona crinita* (Table 1).

Analyses on topologies and markers

At the end of phylogenetic analysis, four trees were obtained for each dataset through four different pipelines, as described above. To test whether the trees obtained from the same dataset are significantly different or not,

I performed the Shimodaira-Hasegawa test (SH-test; Shimodaira and Hasegawa 1999), exploiting the RAxML option "-f H".

The support of each site for the Amarsipobranchia hypothesis ("mt-topology") and the Heteroconchia hypothesis ("nuc-topology") was calculated through the ΔSLS. Sites with ΔSLS > 0.5 or ΔSLS < −0.5 were retained as sites with strong support for either hypothesis (Castoe et al. 2009; Shen et al. 2017). To calculate the sitewise log-likelihood I exploited the RAxML option "-f g" providing the RAxML ML tree when the sitewise log-likelihood was calculated on the mt-topology. A tree with the nuc-topology was obtained by running the phylogenetic analysis with the same settings, but constraining the Heteroconchia clade (as suggested by Shen and colleagues; 2017), and the resulting ML tree was used to calculate the nuc-topology sitewise log-likelihood.

The sCF and the gCF were calculated through IQ-TREE 1.7 (again with 1,000 UF-bootstrap replications) with the option "--cf-verbose" to study phylogenetic signal between and within partitions (Minh et al. 2020). Each dataset was partitioned into single markers in order to calculate the sCF per marker and the gCF. Then, the matrices were partitioned according to complexes to obtain sCF per complex. The nu-OXPHOS subunits in direct contact with the mitochondrial counterparts were defined according to the list of Piccinini and colleagues (2021).

Custom-tailored python and R (R Core Team 2021) scripts were used to analyze and plot the nucleotide and amino acid composition, using Biopython (Cock et al. 2009) and ggplot2. Since mitochondrial URs are missing from transcriptomes, their nucleotide composition was calculated for a list of NCBI indexes obtained through the alMighto database (Formaggioni et al. 2021): a single entry was selected for each species in the database belonging to Palaeoheterodonta, Imparidentia, Anomalodesmata, Pteriomorphia or Protobranchia.

For DUI species the mtDNA of both sexes was selected. The guanine and thymine content in URs was obtained through a customized version of the HERMES tool (Plazzi et al. 2021). The significance of the comparisons was calculated through the Kruskal and Wallis test (Kruskal and Wallis 1952), followed by the Dunn's test (Dinno 2017) with Bonferroni's correction.

2.6 References

Avila-Herrera A, Pollard KS. 2015. Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species. *BMC Bioinformatics* 16:268.

Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bursteinas B, et al. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49:D480–D489.

Beck EA, Thompson AC, Sharbrough J, Brud E, Llopart A. 2015. Gene flow between *Drosophila yakuba* and *Drosophila santomea* in subunit V of cytochrome *c* oxidase: A potential case of cytonuclear cointrogression. *Evolution (N Y)* 69:1973–1986.

Bieler R, Mikkelsen PM. 2006. Bivalvia - a look at the Branches. *Zool J Linn Soc* 148:223–235.

Bieler R, Mikkelsen PM, Collins TM, Glover EA, González VL, Graf DL, Harper EM, Healy J, Kawauchi GY, Sharma PP, et al. 2014. Investigating the Bivalve Tree of Life – an exemplar-based approach combining molecular and novel morphological characters. *Invertebr Syst* 28:32.

Breton S, Beaupré HD, Stewart DT, Hoeh WR, Blier PU. 2007. The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? *Trends in Genetics* 23:465–474.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Carter J, Altaba C, Anderson L, Yancei T. 2011. A Synoptical Classification of the Bivalvia (Mollusca). *Paleontological Contributions*.

Carter JG, Altaba CR, Anderson LC, Araujo R, Biakov Alexander, Bogan A, Campbell D, Campbell M, Chen J, Cope CW. 2011. A Synoptical Classification of the Bivalvia (Mollusca). *Paleontological Contributions*.

Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences* 106:8986–8991.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol Biol Evol* 17:540–552.

Chang J-M, di Tommaso P, Taly J-F, Notredame C. 2012. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. *BMC Bioinformatics* 13:S1.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.

Combosch DJ, Collins TM, Glover EA, Graf DL, Harper EM, Healy JM, Kawauchi GY, Lemer S, McIntyre E, Strong EE, et al. 2017. A family-level Tree of Life for bivalves based on a Sanger-sequencing approach. *Mol Phylogenet Evol* 107:191–208.

Cope JCW. 1996. The early evolution of the Bivalvia. In: Taylor JD, editor. Origin and Evolutionary Radiation of the Mollusca. Oxford: Oxford University Press. p. 361–370.

Cope JCW. 2002. Diversification and biogeography of bivalves during the Ordovician Period. In: Crame JA, Owen AW, editors. Palaeobiogeography and Biodiversity Change: the Ordovician and Mesozoic-Cenozoic Radiations. London: Geological Society of London. p. 25–52.

Cope JCW, Babin C. 1999. Diversification of bivalves in the Ordovician. *Geobios* 32:175–185.

Cope JCW, Kříž J. 2013. The Lower Palaeozoic palaeobiogeography of Bivalvia. In: Harper DAT, Servais T, editors. Geological Society, London, Memoirs. Vol. 38. London: Geological Society of London. p. 221–241.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10:210.

Dinno A. 2017. dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums. Available from: https://CRAN.R-project.org/package=dunn.test

Doucet-Beaupré H, Breton S, Chapman EG, Blier PU, Bogan AE, Stewart DT, Hoeh WR. 2010. Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. *BMC Evol Biol* 10:50.

Dress AW, Flamm C, Fritzsch G, Grünewald S, Kruspe M, Prohaska SJ, Stadler PF. 2008. Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology* 3:7.

Dreyer H, Steiner G, Harper EM. 2003. Molecular phylogeny of Anomalodesmata (Mollusca: Bivalvia) inferred from 18S rRNA sequences. *Zool J Linn Soc* 139:229–246.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* 47:D427–D432.

Fang Z, Sanchez TM. 2012. Treatise Online no. 43: Part N, Revised, Volume 1, Chapter 16: Origin and early evolution of the Bivalvia. *Treatise Online* 0.

Fang ZJ. 2006. An introduction to Ordovician bivalves of southern China, with a discussion of the early evolution of the Bivalvia. *Geological Journal* 41:303–328.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Syst Zool* 27:401.

Formaggioni A, Luchetti A, Plazzi F. 2021. Mitochondrial Genomic Landscape: A Portrait of the Mitochondrial Genome 40 Years after the First Complete Sequence. *Life* 11:663.

Forsythe ES, Williams AM, Sloan DB. 2021. Genome-wide signatures of plastid-nuclear coevolution point to repeated perturbations of plastid proteostasis systems across angiosperms. *Plant Cell* 33:980–997.

Funk DJ, Omland KE. 2003. Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annu Rev Ecol Evol Syst* 34:397–423.

Gelman A, Rubin DB. 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7.

Ghiselli F, Gomes-dos-Santos A, Adema CM, Lopes-Lima M, Sharbrough J, Boore JL. 2021. Molluscan mitochondrial genomes break the rules. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376:20200159.

Giribet G. 2008. Bivalvia. In: Ponder WF, Lindberg DR, editors. Phylogeny and Evolution of the Mollusca. Berkeley: University of California Press. p. 105–142.

Giribet G, Distel DL. 2003. Bivalve phylogeny and molecular data. In: Lydeard C, Lindberg D, editors. Molecular Systematics and Phylogeography of Mollusks. Washington DC: Smithsonian Institution Press. p. 45–90.

Giribet G, Wheeler W. 2002. On bivalve phylogeny: a high-level analysis of the Bivalvia (Mollusca) based on combined morphology and DNA sequence data. *Invertebrate Biology* 121:271–324.

Gissi C, Iannelli F, Pesole G. 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity (Edinb)* 101:301–320.

González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, Taylor JD, Giribet G. 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. *Proceedings of the Royal Society B: Biological Sciences* 282:20142332.

Gosling EM. 2003. Bivalve Molluscs: Biology, Ecology and Culture. Oxford: Fishing News Books

Guerra D, Plazzi F, Stewart DT, Bogan AE, Hoeh WR, Breton S. 2017. Evolution of sex-dependent mtDNA transmission in freshwater mussels (Bivalvia: Unionida). *Sci Rep* 7:1551.

Ha M, Kim VN. 2014. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* 15:509–524.

Harper EM, Dreyer H, Steiner G. 2006. Reconstructing the Anomalodesmata (Mollusca: Bivalvia): morphology and molecules. *Zool J Linn Soc* 148:395–420.

Harper EM, Hide EA, Morton B. 2000. Relationships between the extant Anomalodesmata: a cladistic test. In: Haper EM, Taylor JD, Crame JA, editors. The Evolutionary Biology of the Bivalvia. London: The Geological Society of London. p. 129–143.

Hassanin A, Léger N, Deutsch J. 2005. Evidence for Multiple Reversals of Asymmetric Mutational Constraints during the Evolution of the Mitochondrial Genome of Metazoa, and Consequences for Phylogenetic Inferences. *Syst Biol* 54:277–298.

Havird JC, Whitehill NS, Snow CD, Sloan DB. 2015. Conservative and compensatory evolution in oxidative phosphorylation complexes of angiosperms with highly divergent rates of mitochondrial genome evolution. *Evolution (N Y)* 69:3069–3081.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 35:518–522.

Kanehisa M. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28:27–30.

Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. *Protein Science* 28:1947–1951.

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 49:D545–D551.

Kanehisa M, Sato Y, Kawashima M. 2022. <scp>KEGG</scp> mapping tools for uncovering hidden features in biological data. *Protein Science* 31:47–53.

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, et al. 2011. Phylogenomics reveals deep molluscan relationships. *Nature* 477:452–456.

Kruskal WH, Wallis WA. 1952. Use of Ranks in One-Criterion Variance Analysis. *J Am Stat Assoc* 47:583.

Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, Misof B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool* 7:10.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Mol Biol Evol* 29:1695–1701.

Lawless C, Greaves L, Reeve AK, Turnbull DM, Vincent AE. 2020. The rise and rise of mitochondrial DNA mutations. *Open Biol* 10:200061.

Lemer S, Bieler R, Giribet G. 2019. Resolving the relationships of clams and cockles: dense transcriptome sampling drastically improves the bivalve tree of life. *Proceedings of the Royal Society B: Biological Sciences* 286:20182684.

Lewis SH, Salmela H, Obbard DJ. 2016. Duplication and Diversification of Dipteran Argonaute Genes, and the Evolutionary Divergence of Piwi and Aubergine. *Genome Biol Evol* 8:507–518.

Lynch M. 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol Biol Evol* 13:209–220.

Malkócs T, Viricel A, Becquet V, Evin L, Dubillot E, Pante E. 2022. Complex mitogenomic rearrangements within the Pectinidae (Mollusca: Bivalvia). *BMC Ecol Evol* 22:29.

Minh BQ, Hahn MW, Lanfear R. 2020. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Mol Biol Evol* 37:2727–2733.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 41:e121–e121.

Morales HE, Pavlova A, Amos N, Major R, Kilian A, Greening C, Sunnucks P. 2018. Concordant divergence of mitogenomes and a mitonuclear gene cluster in bird lineages inhabiting different climates. *Nat Ecol Evol* 2:1258–1267.

Morris NJ. 1980. A new Lower Ordovician bivalve family, the Thoraliidae (? Nuculoida), intepreted as actinodont deposit feeders. *Bulletin of the British Museum Natural History (Geology)* 34:265–272.

Morton B. 1996. The evolutionary history of the Bivalvia. In: Taylor J D, editor. Origin and Evolutionary Radiation of the Mollusca. Oxford: Oxford University Press. p. 337–359.

Morton B, Machado FM. 2019. Chapter One - Predatory marine bivalves: A review. *Adv Mar Biol* 84:1–98.

Myra KA. 1963. Marine molluscan genera of Western North America: an illustrated key. Stanford: Stanford University Press

Newell ND. 1965. Classification of the Bivalvia. *Am Mus Novit* 2206:1–25.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 32:268–274.

Nielsen R. 2005. Molecular Signatures of Natural Selection. *Annu Rev Genet* 39:197–218.

Osada N, Akashi H. 2012. Mitochondrial–Nuclear Interactions and Accelerated Compensatory Evolution: Evidence from the Primate Cytochrome c Oxidase Complex. *Mol Biol Evol* 29:337–346.

Passamaneck YJ, Schander C, Halanych KM. 2004. Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. *Mol Phylogenet Evol* 32:25–38.

Passamonti M, Plazzi F. 2020. Doubly Uniparental Inheritance and beyond: The contribution of the Manila clam *Ruditapes philippinarum*. *Journal of Zoological Systematics and Evolutionary Research* 58:529–540.

Piccinini G, Iannello M, Puccio G, Plazzi F, Havird JC, Ghiselli F. 2021. Mitonuclear Coevolution, but not Nuclear Compensation, Drives Evolution of OXPHOS Complexes in Bivalves. *Mol Biol Evol* 38:2597–2614.

Plazzi F, Ceregato A, Taviani M, Passamonti M. 2011. A Molecular Phylogeny of Bivalve Mollusks: Ancient Radiations and Divergences as Revealed by Mitochondrial Genes. *PLoS One* 6:e27147.

Plazzi F, Passamonti M. 2010. Towards a molecular phylogeny of Mollusks: Bivalves' early evolution as revealed by mitochondrial genes. *Mol Phylogenet Evol* 57:641–657.

Plazzi F, Puccio G, Passamonti M. 2016. Comparative Large-Scale Mitogenomics Evidences Clade-Specific Evolutionary Trends in Mitochondrial DNAs of Bivalvia. *Genome Biol Evol* 8:2544–2564.

Plazzi F, Puccio G, Passamonti M. 2017. Burrowers from the Past: Mitochondrial Signatures of Ordovician Bivalve Infaunalization. *Genome Biol Evol* 9:956–967.

Plazzi F, Puccio G, Passamonti M. 2021. HERMES: An improved method to test mitochondrial genome molecular synapomorphies among clades. *Mitochondrion* 58:285–295.

Plazzi F, Ribani A, Passamonti M. 2013. The complete mitochondrial genome of Solemya velum (Mollusca: Bivalvia) and its relationships with Conchifera. *BMC Genomics* 14:409.

Pozzi A, Plazzi F, Milani L, Ghiselli F, Passamonti M. 2017. SmithRNAs: Could Mitochondria "Bend" Nuclear Regulation? *Mol Biol Evol* 34:1960–1973.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ren J, Liu X, Jiang F, Guo X, Liu B. 2010. Unusual conservation of mitochondrial gene order in Crassostreaoysters: evidence for recent speciation in Asia. *BMC Evol Biol* 10:394.

Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* 15:957–966.

Rockenbach K, Havird JC, Monroe JG, Triant DA, Taylor DR, Sloan DB. 2016. Positive Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. *Genetics* 204:1507–1522.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238:195–209.

von Salvini-Plawen L, Steiner G. 1996. Synapomorphies and plesiomorphies in higher classification of Mollusca. In: Taylor JD, editor. Origin and Evolutionary Radiation of the Mollusca. Oxford: Oxford University Press. p. 29–51.

Sánchez TM. 2006. Taxonomic position and phylogenetic relationships of the bivalve Goniophorina Isberg and related genera from the Early Ordovician of northwestern Argentina. *Ameghiniana* 43:113–122.

Sánchez TM. 2008. The early bivalve radiation in the Ordovician Gondwanan basins of Argentina. *Alcheringa: An Australasian Journal of Palaeontology* 32:223–246.

Sánchez TM, Babin C. 2003. Distribution paléogéographique des mollusques bivalves durant l'Ordovicien. *Geodiversitas* 25:243–259.

Sharma PP, González VL, Kawauchi GY, Andrade SCS, Guzmán A, Collins TM, Glover EA, Harper EM, Healy JM, Mikkelsen PM, et al. 2012. Phylogenetic analysis of four nuclear protein-encoding genes largely corroborates the traditional classification of Bivalvia (Mollusca). *Mol Phylogenet Evol* 65:64–74.

Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol* 1:0126.

Shimodaira H, Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* 16:1114–1116.

Sloan DB, Havird JC, Sharbrough J. 2017. The on-again, off-again relationship between mitochondrial genomes and species boundaries. *Mol Ecol* 26:2212–2236.

Sloan DB, Warren JM, Williams AM, Wu Z, Abdel-Ghany SE, Chicco AJ, Havird JC. 2018. Cytonuclear integration and co-evolution. *Nat Rev Genet* 19:635–648.

van der Sluis EO, Bauerschmitt H, Becker T, Mielke T, Frauenfeld J, Berninghausen O, Neupert W, Herrmann JM, Beckmann R. 2015. Parallel Structural Evolution of Mitochondrial Ribosomes and OXPHOS Complexes. *Genome Biol Evol* 7:1235–1251.

Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* 480:364–367.

Stamatakis A. 2006. Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: Proceedings 20th IEEE International Parallel & Distributed Processing Symposium. IEEE. p. 8 pp.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Starobogatov YI. 1992. Morphological basis for phylogeny and classification of Bivalvia. *Ruthenica* 2:1–25.

Stasek CR. 1963. Synopsis and discussion of the association of ctenidia and labial palps in the bivalve Mollusca. *Veliger* 6:91–97.

Steiner G, Hammer S. 2000. Molecular phylogeny of the Bivalvia inferred from 18S rDNA sequences with particular reference to the Pteriomorphia. *Geological Society, London, Special Publications* 177:11–29.

Stöger I, Schrödl M. 2013. Mitogenomics does not resolve deep molluscan relationships (yet?). *Mol Phylogenet Evol* 69:376–392.

Sun S, Li Q, Kong L, Yu H. 2018. Multiple reversals of strand asymmetry in molluscs mitochondrial genomes, and consequences for phylogenetic inferences. *Mol Phylogenet Evol* 118:222–231.

Taylor JD, Williams ST, Glover EA. 2007. Evolutionary relationships of the bivalve family Thyasiridae (Mollusca: Bivalvia), monophyly and superfamily status. *Journal of the Marine Biological Association of the United Kingdom* 87:565–574.

Taylor JD, Williams ST, Glover EA, Dyal P. 2007. A molecular phylogeny of heterodont bivalves (Mollusca: Bivalvia: Heterodonta): new analyses of 18S and 28S rRNA genes. *Zool Scr* 36:587–606.

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Mol Ecol* 21:3907–3930.

Tsubaki R, Kameda Y, Kato M. 2011. Pattern and process of diversification in an ecologically diverse epifaunal bivalve group Pterioidea (Pteriomorphia, Bivalvia). *Mol Phylogenet Evol* 58:97–104.

Waller TR. 1990. The evolution of ligament systems in the Bivalvia. In: Morton B, editor. The Bivalvia. Hong Kong: Hong Kong University Press. p. 49–71.

Waller TR. 1998. Origin of the molluscan class Bivalvia and a phylogeny of major groups. In: Johnston P A, Haggart J W, editors. Bivalves: An Eon of Evolution. Calgary: University of Calgary Press. p. 1–45.

Weng M-L, Ruhlman TA, Jansen RK. 2016. Plastid–Nuclear Interaction and Accelerated Coevolution in Plastid Ribosomal Genes in Geraniaceae. *Genome Biol Evol* 8:1824–1838.

Wolfe NW, Clark NL. 2015. ERC analysis: web-based inference of gene function via evolutionary rate covariation. *Bioinformatics* [Internet]:btv454. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv454

WoRMS Editorial Board. 2022. World Register of Marine Species.

Wu M, Chatterji S, Eisen JA. 2012. Accounting For Alignment Uncertainty in Phylogenomics. *PLoS One* 7:e30288.

Yan Z, Ye G, Werren JH. 2019. Evolutionary Rate Correlation between Mitochondrial-Encoded and Mitochondria-Associated Nuclear-Encoded Proteins in Insects. *Mol Biol Evol* 36:1022–1036.

Yonge M. 1939. The protobranchiate mollusca; a functional interpretation of their structure and evolution. *Philos Trans R Soc Lond B Biol Sci* 230:79–147.

Young ND, Healy J. 2003. GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics* 4:6.

Yu H, Li Q. 2011. Mutation and Selection on the Wobble Nucleotide in tRNA Anticodons in Marine Bivalve Mitochondrial Genomes. *PLoS One* 6:e16147.

Zheng W, Khrapko K, Coller HA, Thilly WG, Copeland WC. 2006. Origins of human mitochondrial point mutations as DNA polymerase γ-mediated errors. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 599:11–20.

Zouros E, Rodakis GC. 2019. Doubly Uniparental Inheritance of mtDNA: An Unappreciated Defiance of a General Rule. *Adv Anat Embryol Cell Biol* 231:25–49.

3. The Evolution and Characterization of the RNA Interference Pathways in Lophotrochozoa

This chapter was written in collaboration with Gianmarco Cavalli, Mayuko Hamada, Tatsuya Sakamoto, Federico Plazzi and Marco Passamonti, and has been published in the journal *Genome, Biology and Evolution*. Supplementary materials are available via the link to the original article (https://doi.org/10.1093/gbe/evae098).

3.1 Introduction

Argonaute proteins are cytoplasmic proteins that play a key role in most of the RNA interference (RNAi) pathways. They interact with a small non-coding RNA (sncRNA), forming an RNA-induced silencing complex (RISC). This ribonucleoprotein complex binds and silences target transcripts, using the complementary sncRNA as a probe (Iwakawa and Tomari 2022). Argonaute proteins can be found throughout most eukaryotic clades and share a common structure, featuring four domains: N-terminal (N), PIWI-Argonaute-Zwille (PAZ), Middle (MID) and P element-induced wimpy testis (PIWI) (Kuhn and Joshua-Tor 2013). The PIWI domain resembles the RNAse H domain's structure, but only some Argonautes have been reported to cleave the target mRNA (Song et al. 2004). All the other Argonaute proteins repress the target trough proteins that interact with the RISC complex (Huntzinger and Izaurralde 2011; Wu et al. 2020).

Among the Argonaute superfamily, four different families have been characterized: *Trypanosoma*-AGO family, WAGO family, AGO-like family, and PIWI-like family (Swarts et al. 2014). *Trypanosoma*-AGO proteins have been identified only in the euglenozoan order Trypanosomatida (Garcia Silva et al. 2010). Conversely, WAGOs have been characterized in nematodes only, whereas AGO-like and PIWI-like proteins seem to be found in all animal phyla (Höck and Meister 2008; Swarts et al. 2014).

Three classes of metazoan interfering sncRNAs can be identified: micro-RNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs) (Iwakawa and Tomari 2022). Each class is matured by different pathways, is loaded by a different Argonaute protein and plays different cellular functions. Precursors of miRNAs (i.e., pri-miRNAs) are encoded by specific genes, transcribed by the RNA polymerase II (Y. Lee et al. 2004; Bartel 2018). Pri-miRNAs feature a hairpin secondary structure, with single-stranded ends at their 3' and 5' (Bartel 2018). These precursors undergo several maturation steps, starting immediately inside the nucleus, where they are targeted by the Microprocessor complex, which cleaves the overhanging nucleotides of pri-miRNAs, leaving a stem-loop structure with a 2bp offset, named pre-miRNA (Lee et al. 2003). Then, pre-miRNAs are exported in the cytoplasm, where they become substrate for DICER1, an endonuclease featuring one PAZ domain and two RNase III domains (Bernstein et al. 2001; Y.S. Lee et al. 2004). Through its catalytic activity, DICER1 removes the hairpin's loop, leaving a ~22bp ds-miRNA (Bernstein et al. 2001; Bartel 2018). Eventually, an AGO-like Argonaute binds to the ds-miRNA in the cytoplasm and disposes of one of the two strands, consequently resulting in a RISC complex. The miRNA-guided RISCs mostly target mRNAs by binding their 3'-UTR, interfering with their stability (Bartel 2018).

Unlike miRNAs, siRNAs may apparently be obtained from roughly every RNA capable of assuming a dsRNA structure (Shabalina and Koonin 2008). Therefore, siRNAs can originate from transcripts of transposons, repeated elements, or pseudogenes (Czech et al. 2008; Tam et al. 2008; Malone and Hannon 2009). siRNAs

undergo a maturation pathway that is very similar to that of miRNAs, leading to the hypothesis that they evolved from a common ancestral RNAi system (Shabalina and Koonin 2008; Moran et al. 2017). Once the dsRNA precursors reach the cytoplasm, they are processed by DICER2, a paralog of DICER1. DICER2 cleaves a ~21bp dsRNA that is loaded into an AGO-like Argonaute, called AGO2 in fruit flies (Y.S. Lee et al. 2004; Matranga and Zamore 2007; Czech et al. 2008). The resulting siRISC complex maintains one of the two strands, again using it as a probe (Matranga and Zamore 2007).

In insects, the siRNA-mediated RNAi activity is not restricted to endogenous dsRNAs, but DICER2 is also able to target dsRNAs of viral origin, producing exogenous siRNAs that are pivotal for the innate immune response (Schuster et al. 2019). *Caenorhabditis elegans* expresses a single DICER paralog, which processes endogenous and viral dsRNAs, but also primary miRNA structures (Welker et al. 2011). Nematodes are even capable of producing secondary siRNAs thanks to the RNA-dependent RNA polymerases (Matranga and Zamore 2007), which are absent in mammals and insects. In mammals, viral dsRNAs are targeted by RIG-I-Like receptors (Loo and Gale 2011), which induce an antiviral response through the activation of type I interferons (Isaacs et al. 1963; Schuster et al. 2019). The ability to process long dsRNAs seems to be related to the DICER helicase domain, which is functional in *C. elegans* DICER and insects' DICER2 (Welker et al. 2011; Sinha et al. 2018; Aderounmu et al. 2023). Contrastingly, the helicase function is not required for DICER proteins that are mainly involved in miRNAs maturation (i.e., insects' DICER1 and mammals' DICER; Jiang et al. 2005; Aderounmu et al. 2023).

Finally, piRNAs are 24-35 nt-long RNAs that originate from longer single strand precursors, consisting of either active transposons or transcripts of genomic piRNA clusters (Hirakata and Siomi 2016). Once these ssRNA precursors are exported through the nuclear pores, they undergo a rather complex maturation pathway, which takes place mostly in the perinuclear nuage (Weick and Miska 2014; Hirakata and Siomi 2016). Although piRNAs appear to be Metazoa-restricted (Grimson et al. 2008), biogenesis pathways vary significantly between clades (Weick and Miska 2014). Mature piRNAs interact with PIWI-like Argonautes, resulting in a piRISC that operates as a defence system against transposons, by means of its nuclease activity (Malone and Hannon 2009). Moreover, piRISCs have been linked to specific epigenetic modifications of chromatin (i.e. H3K9me3) (Le Thomas et al. 2013). A peculiar piRNA amplification pathway called "ping-pong cycle" has been characterized in fruit flies, and later discovered in mice as well (Brennecke et al. 2007; Weick and Miska 2014). This amplification pattern is performed by two different PIWI-like Argonautes, namely one that binds sense-piRNAs (AGO3 in fruit flies) and one that loads antisense piRNAs (Aubergine in fruit flies). As soon as one of these two proteins cleaves its target, it yields a secondary piRNA that can be loaded on the other PIWI-like protein, generating an amplification loop (Weick and Miska 2014; Hirakata and Siomi 2016). Another type of piRNA amplification is called "phasing". PIWI loads a single strand RNA (pre-

piRNA) from the 5' end, directing the endonucleolytic cleavage of the ribonuclease Zucchini at the 3' end. This process is repeated along the pre-pre-piRNA leading to phased matured piRNAs (Mohn et al. 2015; Ozata et al. 2019).

miRNAs, piRNAs and endo-siRNAs were likely to be in the last metazoan common ancestor, since they have been described in Porifera, Cnidaria and most of the metazoan phyla (Grimson et al. 2008; Wheeler et al. 2009; Moran et al. 2013; Praher et al. 2017; Calcino et al. 2018; Fridrich et al. 2020). In some clades piRNA and endo-siRNA pathways have been secondarily lost (Wynant et al. 2017; Fontenla et al. 2021), while the miRNA pathway is likely to be ubiquitous in animals (Fromm et al. 2022).

RNAi mechanisms are far less studied in Lophotrochozoa than in Deuterostomia and Ecdysozoa. This superclade includes around 14 phyla, which is notably higher than Ecdysoza and Deuterostomia, which comprises respectively eight and three phyla (Brusca et al. 2016). Moreover, Lophotrochozoa show an astonishing variability in body plans. For medical and nutritional reasons, most of the data are restricted to Mollusca and Platyhelminthes. Parasitic Platyhelminthes (i.e., Neodermata) lack the PIWI pathway, but the miRNA and endo-siRNA pathways have been reported (Fontenla et al. 2017; Fontenla et al. 2021). The PIWI pathway has been confirmed in Mollusca, where a clear signature of ping-pong amplification is visible (Jehn et al. 2018) and many miRNAs have been annotated in mollusks (Fromm et al. 2022). On the other hand, proteins related to the endo-siRNA pathway are absent in Bivalvia (Rosani et al. 2016). Outside those phyla, few data have been published. The endo-siRNA pathway seems absent in the Annelida *Capitella teleta* (Khanal et al. 2022) and the annotation of miRNA families is restricted to one Syndermata, one Brachiopoda and two Annelida species (Fromm et al. 2022).

The phylogenetic relationships within Lophotrochozoa have been strongly debated and they are not fully resolved yet. Morphological analyses agree to include Mollusca, Annelida, Brachiopoda, Phoronida, Nemertea, Bryozoa, Entoprocta, and Cycliophora in the Trochozoa clade (f.i., Kocot 2016). Some molecular analyses proved to be concordant with the Trochozoa clade (Struck et al. 2014; Laumer et al. 2015; Kocot et al. 2017; Laumer et al. 2019), but some others did not (Kocot et al. 2017; Marlétaz et al. 2019). Even the relationships among Trochozoa are far from being resolved, namely the monophyly of Polyzoa (Bryozoa+Entoprocta+Cycliophora) has regained credit recently (Khalturin et al. 2022; but see (Nesnidal et al. 2013; Kocot 2016; Bleidorn 2019; Laumer et al. 2019). Outside Trochozoa, Rouphozoa (Platyhelminthes+Gastrotricha) and Chaetognathifera (Syndermata+Micrognathozoa+Chaetognatha) are supported by most of the phylogenetic analyses (Struck et al. 2014; Laumer et al. 2015; Kocot et al. 2017; Laumer et al. 2019)

In recent years, the increase in -omics data has made it possible to compare and study the evolution of protein families along Lophotrochozoa. In this study, I exploited various -omics resources from nine lophotrochozoan phyla to annotate and characterize the diversification of the Argonaute and DICER proteins. I also analysed sncRNA libraries to annotate the three sncRNA types and confirm the presence or absence of a particular sncRNA type in some phyla. According to my results, along the Lophotrochozoa evolution the endo-siRNA pathway has been progressively lost, starting with DICER2 in Trochozoa, followed by the loss of the fruit fly AGO2-like proteins in Phoronida, Brachiopoda, Annelida, and Mollusca. This pattern is confirmed by the distribution of DICER2 and AGO2-like proteins in the analysed organisms. In contrast, the piRNA and miRNA pathways appeared to be conserved in almost all Lophotrochozoa.

3.2 Results

The Argonaute and DICER phylogeny

I annotated Argonaute proteins of 43 lophotrochozoan species by analysing 19 proteomes, 16 genomes and 8 transcriptomes. Argonaute sequences of *Homo sapiens* (Chordata), *Drosophila melanogaster* (Arthropoda), *Caenorhabditis elegans* (Nematoda) and *Nematostella vectensis* (Cnidaria) were retrieved from SwissProt and included in the phylogenetic analysis as references (Supplementary table S1). Moreover, I included metazoan species whose RNAi pathways have already been studied (see Introduction), testing whether our Argonaute and DICER annotation matches results from the literature (Table 1).

Phylum	Species	PIWI1	PIWI2	miAGO	siAGO	DICER1	DICER2	Data	Comp.
Mollusca	*Haliotis rufescens*	☑	☑	☑	☒	☑	☒	NCBI Ref.	99,4%
	Mizuhopecten yessoensis	☑	☑	☑	☒	☑	☒	NCBI Ref.	98,6%
	Gigantopelta aegis	☑	☑	☑	☒	☑	☒	NCBI Ref.	98,5%
	Pecten maximus	☑	☑	☑	☒	☑	☒	NCBI Ref.	98,5%
	Pomacea canaliculata	☑	☑	☑	☒	☑	☒	NCBI Ref.	98,2%
	Crassostrea virginica	☑	☑	☑	☒	☑	☒	NCBI Ref.	98,1%
	Aplysia californica	☑	☑	☑	☒	☑	☒	NCBI Ref.	97,8%
	Ostrea edulis	☑	☑	☑	☒	☑	☒	NCBI Ref.	96,8%
	Lottia gigantea	☑	☑	☑	☒	☑	☒	NCBI Ref.	96,5%
	Mercenaria mercenaria	☑	☑	☑	☒	☑	☒	NCBI Ref.	95,4%
	Octopus bimaculoides	☑	☑	☑	☒	☑	☒	NCBI Ref.	94,6%
	Patella vulgata	☑	☑	☑	☒	☑	☒	NCBI Ref.	90,5%
	Biomphalaria glabrata	☑	☑	☑	☒	☑	☒	NCBI Ref.	88,9%
	Saccostrea glomerata	☑	☑	☑	☒	☑	☒	NCBI Ref.	88,9%
	Acanthopleura granulata	☑	☑	☑	☒	☑	☒	Assemb.	69,8%
	Euprymna scolopes	☑	☑	☑	☒	☑	☒	Assemb.	55,5%
Annelida	*Alitta virens*	☑	☑	☑	☒	☑	☒	Assemb.	69,8%
	Piscicola geometra	☑	☑	☑	☒	☑	☒	Assemb.	43,7%
	Helobdella robusta	☑	☑	☑	☒	☑	☒	NCBI Ref.	90,2%
Brachiopoda	*Lingula anatina*	☑	☑	☑	☒	☑	☒	NCBI Ref.	98,6%
	Hemithiris psittacea	☑	☑	☑	☒	☑	☒	Transc.	90,7%
	Glottidia pyramidata	☒	☒	☒	☒	☑	☒	Transc.	75,1%
Phoronida	*Phoronis australis*	☑	☑	☑	☒	☑	☒	Assemb.	68,3%
	Phoronis vancouverensis	☑	☑	☒	☒	☑	☒	Transc.	84,9%
Nemertea	*Notospermus geniculatus*	☑	☑	☑	☑	☑	☒	Assemb.	68,4%
	Lineus longissimus	☑	☑	☑	☒	☑	☒	Assemb.	69,9%
	Tubulanus polymorphus	☑	☒	☒	☑	☑	☒	Transc.	71,8%
	Malacobdella grossa	☑	☑	☒	☑	☑	☒	Transc.	91,7%
	Paranemertes peregrina	☒	☒	☒	☑	☑	☒	Transc.	80,5%
Bryozoa	*Cryptosula pallasiana*	☑	☑	☑	☑	☑	☒	Assemb.	83,5%
	Cristatella mucedo	☒	☑	☒	☑	☑	☒	Assemb.	62,1%
	Membranipora membranacea	☑	☑	☑	☒	☑	☒	Assemb.	54,5%

	Bugulina stolonifera	☑	☒	☑	☒	☑	☒	Assemb.	54,0%
	Bugula neritina	☑	☒	☑	☒	☑	☒	Assemb.	50,7%
Entoprocta ●	*Pedicellina cernua*	☑	☑	☑	☑	☑	☒	Transc.	90,4%
Syndermata ●	*Adineta ricciae*	☑	☑	☑	☑	☑	☑	Assemb.	79,6%
	Pomphorhynchus laevis	☑	☑	☒	☒	☑	☑	Assemb.	43,2%
	Brachionus asplanchnoidis	☑	☑	☑	☑	☑	☒	Assemb.	85,4%
	Echinorhynchus gadi	☑	☑	☑	☑	☑	☑	Transc.	58,6%
Platyhelminthes ●	*Echinococcus granulosus*	☒	☒	☑	☑	☑	☑	NCBI Ref.	69,2%
	Schistosoma haematobium	☒	☒	☑	☑	☑	☑	NCBI Ref.	77,1%
	Opisthorchis viverrini	☒	☒	☑	☑	☑	☑	NCBI Ref.	64,7%
	Taenia pisiformis	☒	☒	☑	☑	☑	☑	Assemb.	36,3%
	Schmidtea mediterranea	☑	☑	☑	☑	☑	☑	Assemb.	46,5%
others	*Gallus gallus*	☑	☒	☑	☒	☑	☒	NCBI Ref.	95,3%
	Danio rerio	☑	☑	☑	☒	☑	☒	NCBI Ref.	99,6%
	Asterias rubens	☑	☑	☑	☒	☑	☒	NCBI Ref.	98,7%
	Anopheles aquasalis	☑	☑	☑	☑	☑	☑	NCBI Ref.	99,0%
	Strongyloides ratti	☒	☒	☑	☑	☑	☒	NCBI Ref.	69,2%
	Acropora muricata	☑	☑	☑	☑	☑	☑	Assemb.	60,4%

Table 1. Presence and absence of Argonaute and DICER proteins. Colored dots refer to colors assigned to the different animal phyla in Figure 1 and 2. For the four Argonaute proteins and the two DICER proteins, a green check marks the species where the protein has been annotated, a red cross marks species where the protein is absent. The "Data" column reports the type of data, proteome annotated following the NCBI Eukaryotic Genome Annotation Pipeline ("NCBI Ref."), a genome assembly ("Assemb."), or a transcriptome ("Transc."). The last column reports the BUSCO completeness score.

All the annotated Argonaute proteins were aligned, and the Maximum-Likelihood (ML) tree was inferred. The PIWI and AGO proteins of *Trypanosoma brucei* were obtained from UniProt and used as outgroups (Supplementary table S1). The phylogenetic tree of the Argonaute superfamily supported the known main families (Fig. 1; see supplementary figure S1 for the uncollapsed tree with support values for each node). The WAGO family, which is restricted to nematodes, included WAGOs and CSR-1 sequences from *C. elegans*. The PIWI-like family was characterized by AGO3, PIWI and AUB of *D. melanogaster*, HILI and HIWI of *H. sapiens* and PRG-1 of *C. elegans;* the AGO-like family comprised AGO1,2 of *D. melanogaster*, AGO1-4 of *H. sapiens* and ALG1,2 of *C. elegans.* Every family was widely supported by UFBoot and the SH-alrt test (Fig. 1).

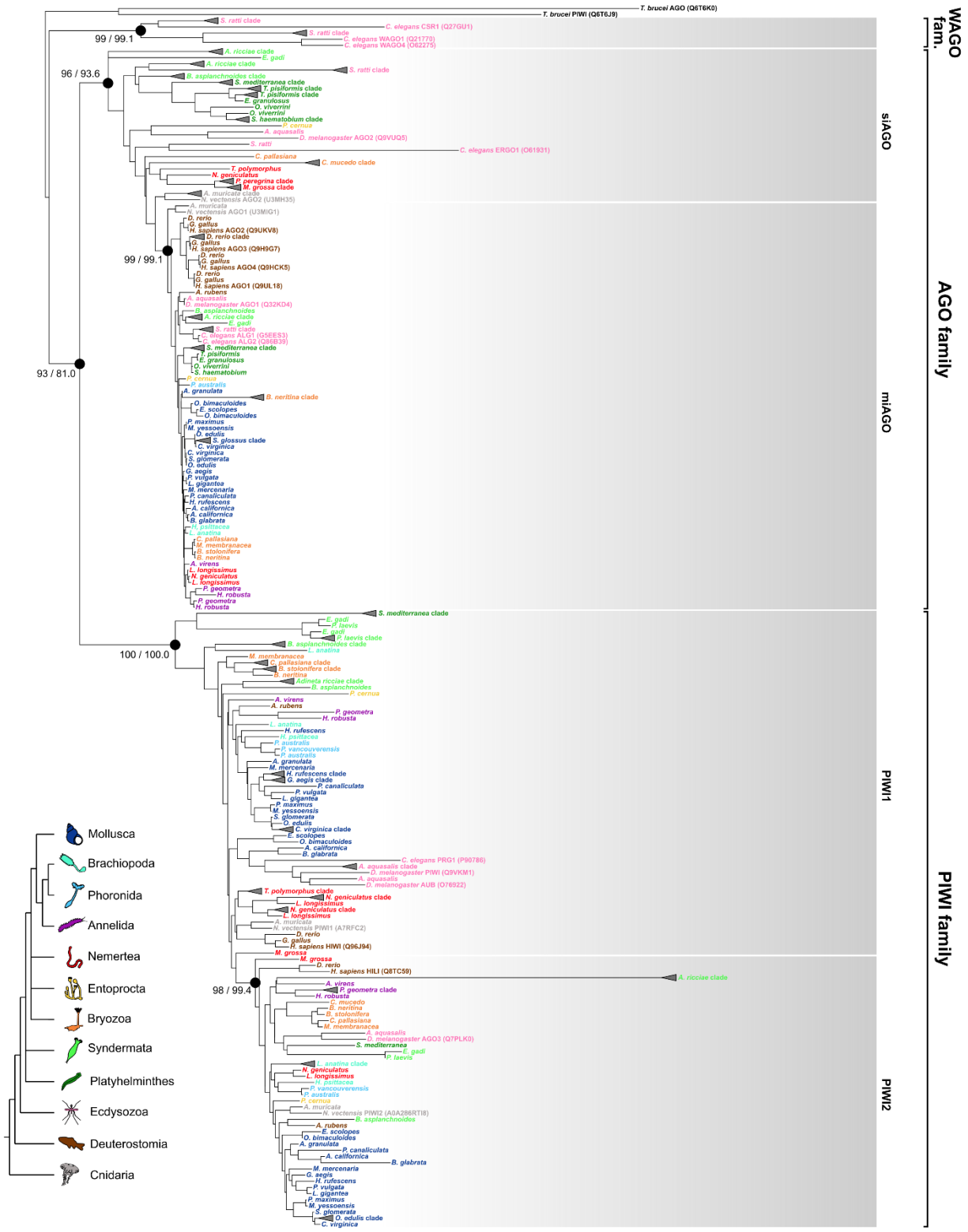

WAGO fam.
AGO family
PIWI family
siAGO
miAGO
PIWI1
PIWI2
T. brucei AGO (Q6T6K0)
T. brucei PIWI (Q6T6J9)
99 / 99.1
96 / 93.6
99 / 99.1
93 / 81.0
100 / 100.0
98 / 99.4
Mollusca
Brachiopoda
Phoronida
Annelida
Nemertea
Entoprocta
Bryozoa
Syndermata
Platyhelminthes
Ecdysozoa
Deuterostomia
Cnidaria

Figure 1. ML tree of the lophotrochozoan Argonaute proteins. For the six marked nodes the label shows UFBoot/SH-alrt value. Support values of the remaining nodes are shown in the supplementary figure S1. Clades formed by paralogs of the same species were collapsed and represented with a triangle. For reference proteins the Uniprot accession code is reported in brackets. Species are colored according to the phylum. The color legend on the bottom left reconstructs the main phylogenetic relationships according to the latest Lophotrochozoa phylogenetic analyses (Kocot et al. 2017; Bleidorn 2019; Marlétaz et al. 2019)

Within each family it was possible to identify the different Argonaute proteins. Argonaute proteins related to miRNAs were characterized by proteins like *H. sapiens* AGO1-4, *D. melanogaster* and *N. vectensis* AGO1 and *C. elegans* ALG1,2 and resulted in a monophyletic clade (UFBoot = 99 and SH-alrt = 99.1; Fig. 1). Almost all lophotrochozoan species featured a protein clustering within this clade, with very few exceptions, and at least one organism from each phylum was recovered in this clade (Fig. 1; Table 1). I will refer to this clade as the "miAGO clade". The remaining proteins within the AGO-like family were recovered as paraphyletic with respect to the miAGO clade. This group included *D. melanogaster* AGO2, *N. vectensis* AGO2 and *C. elegans* ERGO, all proteins that target endo-siRNAs (but with some exceptions: see Fridrich et al. 2020). Actually, endo-siRNA proteins are often inferred as paraphyletic with respect to the miAGO clade (Swarts et al. 2014; Praher et al. 2017; Wynant et al. 2017), I will call this group the "siAGO group". Few lophotrochozoan phyla are included in this group, since I annotated at least one siAGO protein for all the Platyhelminthes and Entoprocta species, while three other phyla featured at least one species within the group, namely Nemertea, Bryozoa and Syndermata. I did not retrieve any siAGO protein from the remaining clades, namely Mollusca, Annelida, Brachiopoda, Phoronida (Fig. 1; Table 1).

PIWI-like proteins showed a pattern similar to that of the AGO-like family. PIWI2 proteins were clustered in a monophyletic clade, characterized by already annotated PIWI2 proteins, like *H. sapiens* HILI, *N. vectensis* PIWI2 and *D. melanogaster* AGO3; all the remaining PIWI-like proteins are paraphyletic with respect to PIWI2 proteins. This grade included already annotated PIWI1 proteins like *H. sapiens* HIWI, *N. vectensis* PIWI1, *C. elegans* PRG1 and *D. melanogaster* AUB and PIWI. Both PIWI-like protein groups included at least one protein from each lophotrochozoan phylum (Table 1). PIWI-like proteins were almost absent in Platyhelminthes, apart from *Schmitdea mediterranea* (Fig. 1).

Regarding the six non-lophotrochozoan species included in the analysis, the annotation of Argonaute proteins matched the expectations: miAGO proteins were retrieved for all six species; siAGO proteins were absent in *D. rerio, G. gallus* and *A. rubens* (Deuterostomia); in all six species I annotated PIWI1 and PIWI2 proteins, exception made for *S. ratti* (i.e. that lacked of both PIWI proteins) and, unexpectedly, *G. gallus* that lacked of PIWI2 (Fig. 1; Table 1).

The phylogenetic analysis highlighted the presence of miAGO proteins in each phylum, but only some of them featured siAGO proteins. In Arthropoda, the precursor structures of siRNAs and miRNAs are processed by two distinct DICER paralogs: DICER2 and DICER1, respectively (Shabalina and Koonin 2008). DICER2 has been found in other phyla, such as Cnidaria and Platyhelminthes (Mukherjee et al. 2013), while clades lacking siAGO proteins lack DICER2 as well. I annotated DICER proteins and inferred the phylogeny to understand whether Lophotrochozoa follow the same pattern.

I annotated DICER proteins querying the lophotrochozoan sequences against an annotated metazoan DICER set (Mukherjee et al. 2013) and looking for the ribonucleases 3 domain. The phylogenetic analysis included the metazoan DICER set of Mukherjee and colleagues (2013) as references, including the *Zea mays* DICER proteins as outgroups (Fig. 2; see supplementary figure S2 for the uncollapsed tree with support values for each node). The resulting ML tree clustered the DICER proteins into two distinct groups. Recall the position of the reference sequences, DICER1 and DICER2 sequences are accordingly split into the two groups, apart from *Litopenaeus vannamei* DICER2, which is basal to all the other proteins. The monophyly of the DICER1 group is supported by the SH-alrt test (86.4) and the UFBoot (96). In contrast, the low support values of the DICER2 node (UFBoot= 58 and SH-alrt = 64.3) undermine the hypothesis of a unique common origin of DICER2 proteins. Nonetheless, the presence of DICER2 was restricted to a few lophotrochozoan phyla: all Platyhelminthes and three Syndermata species showed a DICER2 protein. On the other hand, DICER1 was annotated for every phylum: thus, Mollusca, Annelida, Phoronida, Brachiopoda, Nemertea, Bryozoa and Entoprocta reported DICER1 proteins, but no DICER2 proteins (Fig. 2; Table 1). In line with expectations, DICER2 was found in *A. muricata, A. gambiae,* while it was lacking in *D. rerio, G. gallus, A. rubens* and *S. ratti.* In contrast, DICER1 was annotated in all of them.

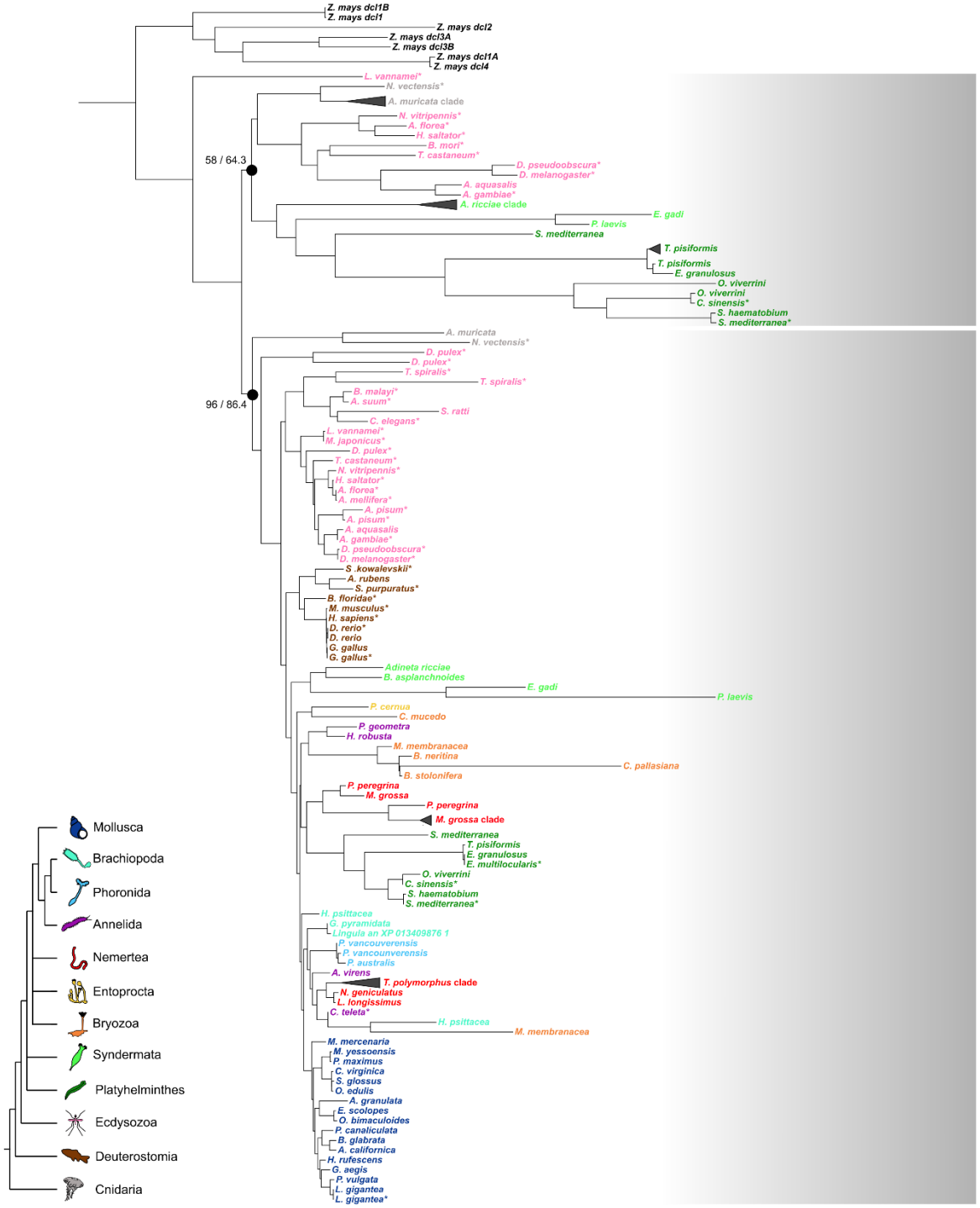


Figure 2. ML tree of the lophotrochozoan DICER proteins. For two marked nodes it is reported respectively the UFBoot and the SH-alrt value. Support values of the remaining nodes are shown in the supplementary

figure S2. Clades formed by paralogs of the same species were collapsed and represented with a triangle. References species retrieved from the analysis of Mukherjee and colleagues (2013) are marked with an asterisk. Species are colored according to the phylum. The color legend on the bottom left reconstructs the main phylogenetic relationships according to the latest Lophotrochozoa phylogenetic analyses (Kocot et al. 2017; Bleidorn 2019; Marlétaz et al. 2019).

Overall, the DICER family phylogenetic analysis confirmed that the absence of siAGO proteins coincides with the absence of DICER2, and *vice versa*, exception made for Nemertea, Bryozoa and Entoprocta, where I annotated siAGO proteins for most species; however, none of these species featured DICER2. DICER1 and miAGO proteins showed lower evolutionary rates than their paralogous counterparts: the root-to-tip distances of miAGO branches were significantly lower than the root-to-tip distances of siAGO, PIWI1 and PIWI2 branches, and the root-to-tip distances of DICER1 branches were significantly lower than their DICER2 counterparts (Supplementary figure S3). I also estimated the ratio of non-synonymous to synonymous substitution rates (ω) along the AGO family tree and the DICER family tree: I confirmed that, during the evolution of the DICER and AGO families, purifying selection on DICER1 (LRT=34.34, p-value = 4.62 x 10^{-9}; Supplementary figure S4b) and miAGO (LRT=204,64, p-value = 0; Supplementary figure Sa) has intensified compared to the rest of the family tree (Wertheim et al. 2015).

In both phylogenetic analyses the presence of a protein in some phyla was not confirmed by all the species. To understand whether it might be related to the quality of the data, I evaluated the completeness of proteomes, genomes, and transcriptomes: Argonaute and DICER proteins were missing more commonly in transcriptomes than in proteomes or genome assemblies, regardless of completeness (Table 1). All proteomes showed a similar presence/absence pattern and high completeness values, while the completeness of the genomes varied significantly between species. In general, in more complete genome assemblies I was also able to annotate more proteins (Table 1). Overall, phyla that are generally more represented in protein databases (i.e., Mollusca, Annelida, Platyhelminthes) showed a more constant presence/absence pattern between species than under-represented phyla (i.e., Nemertea, Bryozoa, Syndermata). The annotation of genome assemblies with BRAKER has heavily relied on protein databases (see Materials and methods), thus it may be possible that this method produced better annotations for clades like Mollusca, Annelida or Platyhelminthes.

Domain characterization in Argonaute and DICER proteins

The annotation of Argonaute and DICER proteins mostly relied on the annotation of peculiar domains (see Materials and Methods). However, other domains characterize the two protein families. Thus, I built domain

profiles of Argonaute and DICER domains from multiple sequences alignments. Profiles were aligned against the annotated proteins to evaluate the domain composition of lophotrochozoan Argonaute and DICER proteins.

Besides the PIWI and PAZ domains, Argonaute proteins are also characterized by N-terminal (N) and the MID domain. I failed to annotate the two domains in most of PIWI proteins, in particular the MID domain was not annotated in any Syndermata and Bryozoa PIWI1 protein, but also in each lophotrochozoan phylum I reported at least one species without the domain (Supplementary table S2). Additionally, each Syndermata PIWI2 protein lacked not only the MID domain but also the N domain. Similarly, Bryozoa PIWI2 proteins did not contain the MID domain either (Supplementary tables S2,3). On the other hand, in almost all miAGO proteins both domains were annotated (*Hemithiris psittacea* being the only exception), while in siAGO proteins almost all Platyhelminthes lacked the MID domain. After localizing the domain position in the Argonaute alignment, I examined whether certain species lacked the domains entirely. This was determined by assessing whether their sequence in that portion of the alignment was either entirely absent (i.e., with most sites being gaps) or significantly degenerated (with most sites containing amino acids, but the sequence being too degenerated for accurate domain annotation). In most cases, the sequences were found to be degenerated (Supplementary table S4).

Regarding DICER domain composition, I assessed the presence of the Helicase (Hel) and PAZ domain. The Hel domain was absent in most of Lophotrochozoa DICER2 proteins, with the only exception of two Syndermata paralogs (Supplementary table S5). The PAZ domain also resulted absent from most of Lophotrochozoa DICER2 proteins, exception made for one *S. mediterranea* DICER2 paralog (Supplementary table S6). Among DICER1 proteins, Hel and PAZ domains were annotated in all Mollusca species, while in other lophotrochozoan phyla the annotation of both domains was restricted to some species (Supplementary tables S5,6). Notably, the two domains resulted completely absent in Bryozoa. Contrastingly with the Argonaute analysis, DICER domains, when not recovered, were completely missing. According to the structure of DICER, the PAZ and Hel domains are towards the N-terminal end, with the Hel being the first domain of the protein (Mukherjee et al. 2013). Accordingly, all the lophotrochozoan DICER proteins lacking either the two domains or solely the Hel domain were found to be truncated at the N-terminus. The only exceptions were some DICER2 paralogs of the syndermatan *Adineta ricciae*: while five paralogs resulted truncated, three of them displayed degeneration only (Supplementary table S4).

I also evaluated the conservation of the DECH box within the Hel domain. The amino acid composition of the DECH box (i.e, aspartic acid, glutamic acid, cysteine and histidine) resulted conserved in Syndermata, Nematoda, Phoronida and Brachiopoda. Most Annelida and Gastropoda showed an aspartic acid instead of a glutamic acid on the second position (resulting in DDCH); the DECH box resulted further mutated in Bivalvia

(i.e, ENCH in Ostreida, DHCQ in Pectinida, DDCH in *Mercenaria mercenaria*), in *Biomphalaria glabrata* (DNCH), and in Cephalopoda (ECSN). Platyhelminthes also reported a highly diverged DECH box (Supplementary figure S5).

Looking for the endo-siRNA signature in small RNA libraries.

According to the phylogenomic analysis, some lophotrochozoan phyla lack pivotal proteins related to the endo-siRNA pathway. DICER2 generally processes double stranded RNAs producing two 21 bases siRNA duplexes that overlap by 19 bases. Similarly, piRNAs produced by the ping-pong cycle go in pairs that overlap by 10 bases (Antoniewski 2014; Khanal et al. 2022). Thus, siRNAs and piRNAs have a unique signature that can be identified in sncRNA libraries by looking for overlapping pairs of reads. I retrieved eight sncRNA libraries from lophotrochozoan and non-lophotrochozoan species (namely *Danio rerio, Apostichopus japonicus, Acropora muricata, Anopheles gambiae, Drosophila melanogaster, Schmitdea mediterranea, Schistosoma japonicum, Crassostrea gigas*) and I sequenced the sncRNA pool of *Notospermus geniculatus* (Nemertea) to include a species with an incomplete endo-siRNA pathway (i.e. presence of siAGOs, but absence of DICER2) in our analysis. Using the overlapping_reads.py script (Antoniewski 2014) I calculated the number of read pairs that overlapped for the same number of bases, from 4 to 20 bases. Then, I calculated the Z-Score among the number of read pairs for each overlap group. A Z-score equal to 1 means that the number of read pairs in that overlap group is one standard deviation higher from the mean size of all the overlap groups of a given species. Taking into consideration only 21 bp reads (i.e., the expected length of endo-siRNAs), species equipped with DICER2 (i.e., *D. melanogaster, A. gambiae*, *A. muricata*, *S. mediterranea* and *S. japonicum*) reported a Z-score higher in the 19-overlap group than species without DICER2 (*D. rerio* , *Apostichopus japonicus*, *C. gigas*, *N. geniculatus*; Fig. 3). Some species also reported a sharp increase in the Z-score for the 10-overlap group for 21-bp long reads only, namely *N. geniculatus*, *A. japonicus*, *C. gigas*, and *S. mediterranea*. A 10-bases overlap would correspond to the piRNA signature, although the length of piRNAs in Arthropoda ranges from 25 to 30 bp. Overall, considering the phylogenomic analysis, all the species that exhibited a complete siRNA pathway (using DICER2 and siAGO as a proxy) showed a high Z-score in the 19-overlap group (Fig. 3).

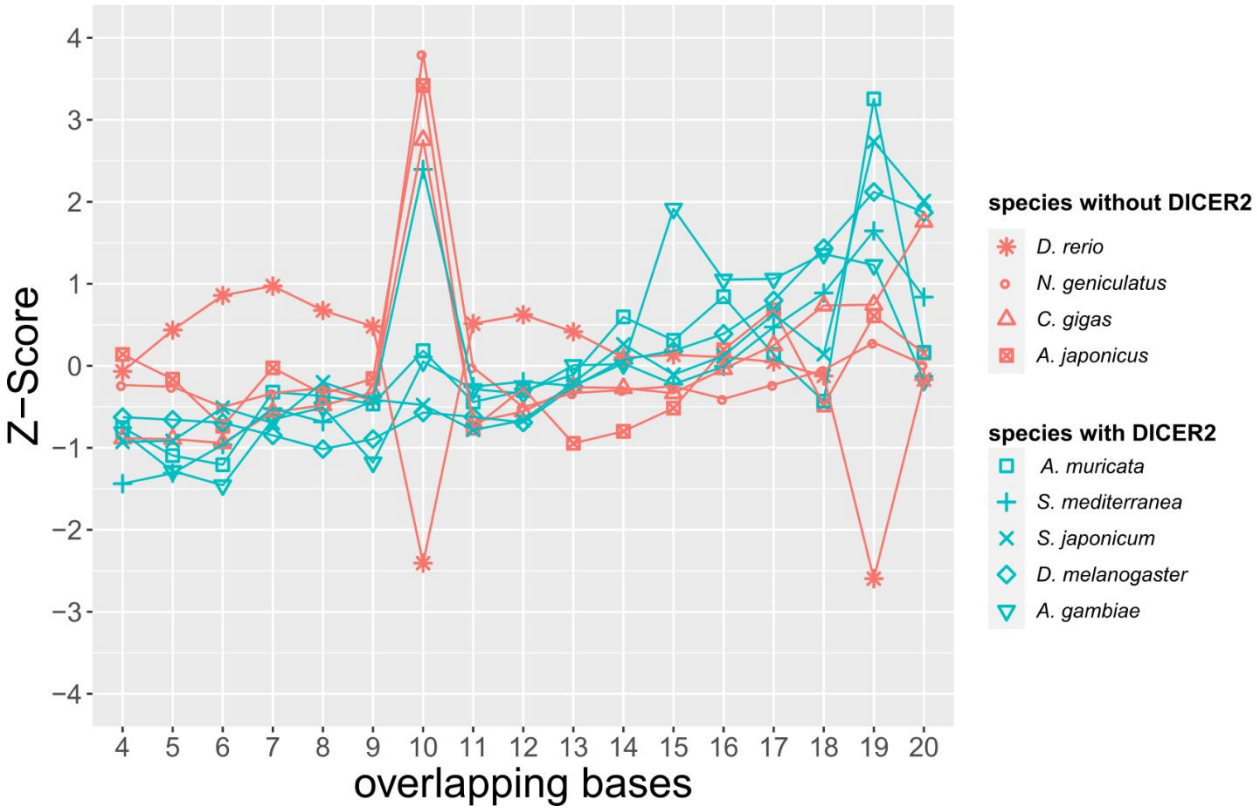


Figure 3. Evaluating the siRNA signature in sncRNA libraries. The plot reports the Z-score between the number of pairs of all possible overlaps. A Z-score greater than 1 means that pairs overlapping of that length are at least a standard deviation more numerous than the mean of all the overlaps. The two colors distinguish species equipped with DICER2 from species that lack protein.

Finally, I investigated the range of action of the three sncRNA types in three metazoan species (i.e., *C. gigas* for Lophotrochozoa, *A. gambiae* for Ecdysozoa and *D. rerio* for Deuterostomia). I annotated miRNAs, piRNAs and endo-siRNAs evaluating their expression at different sncRNA lengths (Supplementary figure S6a). In terms of reads per million (RPM), in all three species miRNAs are the most expressed and their length ranges from 20 – 25 nc. In contrast, piRNAs were annotated in the in the 24 – 30 nc length range. In most cases, I was not able to discern the endo-siRNA signal from noise, but *A. gambiae* was the species that showed the highest expression levels (Supplementary figure S6a). I also compared ovary and somatic tissue sncRNA libraries. As expected, in all three species piRNAs is the class more expressed in the ovaries. (Supplementary figure S6b). Overall, in the three species I did not report notable differences in terms of sncRNA length or differential expression in somatic/ovarian tissue among piRNA and miRNA types.

The annotation of the sRNA types was performed even for the newly sequenced small RNA libraries of *Notospermus geniculatus.* Following the guidelines of Fromm and colleagues (2015), I identified 154 *bona*

fide miRNA genes. Each miRNA gene has been annotated by blasting its preliminary structure against the MirGeneDB pre-miRNA database (Fromm et al. 2022) and checking the conservation of the seed region. According to the annotation results, 63 *N. geniculatus* miRNA genes were included in 29 miRNA families already described in other metazoan species (Supplementary table S8). The remaining 91 miRNA genes did not exhibit significant similarities with the pre-miRNAs in the database (supplementary materials). To assess the conservation of the novel miRNAs and increase the reliability of your predictions, I aligned the pre-miRNAs against the *Lineus longissimus* genome (i.e, the most closely related species to *N. geniculatus* with an assembled genome). I identified at least 28 novel miRNA genes that are shared between *N. geniculatus* and *L. longissimus*. Based on their preliminary sequences and seed regions, I clustered these 28 novel miRNA genes into 17 novel families. (Supplementary table S8, supplementary materials). Comparing the expression levels of the three different types of small RNAs, considering small RNAs with a length ranging between 20 and 24 nucleotides, miRNAs resulted the most expressed even in *N. geniculatus*. Conversely, piRNAs resulted expressed mostly in the length range 25-29 nc (Supplementary figure S7). As for *C. gigas* (Supplementary figure S6b), I was not able to discern the endo-siRNA signal from noise.

3.3 Discussion

Most lophotrochozoans have conventional miRNA and piRNA pathways

RNAi pathways play a central role in many molecular aspects, from mRNA regulation to defence mechanisms, and Argonaute proteins are the key RNAi players in all eukaryotes. All the phylogenetic analyses agree to divide the eukaryotic Argonaute superfamily into four main families, namely the *Trypanosoma*-AGO family, the WAGO family, the AGO family, and the PIWI family (Höck and Meister 2008; Garcia Silva et al. 2010; Swarts et al. 2014). Excluding the *Trypanosoma*-AGO family, all the other families are represented within animals; moreover, the PIWI and the WAGO families are restricted to animals or even nematodes, respectively (Swarts et al. 2014). It is still uncertain how the four families emerged during eukaryote evolution. For instance, several eukaryotic clades have a miRNA-like pathway, but it is not clear whether these pathways are analogous or homologous to the metazoan miRNA pathway (Moran et al. 2017). It is likely that RNAi systems diverged from an ancestral siRNA system and, considering the distribution of Eukaryota clades in the four families (Swarts et al. 2014), the divergence took place at least 1.5 billions of years ago (Strassert et al. 2021).

The inferred Argonaute phylogenetic tree confirmed and highly supported the three metazoan Argonaute families (Fig. 1). Moreover, I identified two distinct groups in the AGO and PIWI, where only one of the two groups is monophyletic. This pattern is confirmed in other Argonaute phylogenetic analyses (Swarts et al. 2014; Praher et al. 2017; Wynant et al. 2017). Recalling the deep divergence of these proteins, the signal might be saturated. Accordingly, most of the nodes at the base of the family are not strongly supported (Table 1). The same pattern has been observed in the DICER phylogeny, with DICER2 proteins being paraphyletic with respect to DICER1 proteins, which clustered in a well-supported monophyletic clade (Fig. 2) (Mukherjee et al. 2013). Concordantly, DICER2 and DICER1 are related to siAGO and miAGO proteins, respectively. Overall, within each clade the phylogenetic reconstruction is substantially in agreement with the state-of-art animal phylogeny, recalling that the signal has been inferred from single markers.

Almost all lophotrochozoans showed two distant related PIWI proteins (i.e., one AUB-like and one AGO3-like; Fig. 1; Table 1), with the only exception of Neodermata (Platyhelminthes), which lacks the whole piRNA pathway (Fontenla et al. 2021). It is likely that the ping-pong cycle, which has already been described in some mollusks (Jehn et al. 2018), has been maintained in most of lophotrochozoan. The piRNA expression of *C. gigas* is in line with that of *D. rerio* and *A. gambiae* and the differential expression analysis confirms that piRNAs are more expressed in the gonads in all three clades.

The miRNA pathway is the most ubiquitous RNAi pathway among Metazoa, and almost all species reported a miAGO and a DICER1 protein. For these proteins I even detected lower root-to-tip distances and a decrease in ω along their branches, which confirms a higher selective pressure. Therefore, the high conservation of

proteins involved in the miRNA pathway reflects the well-known conservation of miRNAs among Metazoa (Tarver et al. 2013).

Even in our case, I confirmed that the annotated miRNAs showed similar features in the three reference species; they are 20-25 nt long, they are not more expressed in the ovaries than in somatic tissue, and overall, they are by far the most expressed sncRNA class.

The conservation is also reflected in the domain composition; most of miAGO and DICER1 proteins included all the domains. However, a novel pattern has been observed in the DECH box. The DECH box is a motif present in many helicase domains and it coordinates ATP hydrolysis (Yerukhimovich et al. 2018). Although it is conserved between distant related DICER proteins (f.i., *Homo sapiens* DICER1, *D. melanogaster* DICER2, *C. elegans* DICER; Supplementary figure S5), its conservation does not imply that the Hel domain is active, since the *H. sapiens* DICER has been proved to work in an ATP-independent manner (Liu et al. 2018). The Hel domain is likely inactive also in Mollusca and Annelida, where the DECH box diverges of one or two amino acids, but it remains conserved in Phoronida, Brachiopoda and Nematoda, where it may be still active (Supplementary figure S5).

The evolution of the endo-siRNA pathway in Lophotrochozoa.

Pathways maturating endo-siRNAs are deeply diverse between metazoan clades. In the fruit fly the miRNA and the endo-siRNA pathways are separated, having a specific Argonuate and DICER protein for each pathway. In *C. elegans* a single DICER protein is responsible for the maturation of miRNAs and endo-siRNAs: endo-siRNAs are then loaded by ERGO-1 (but also other Argonaute proteins; Han et al. 2009). RNA-dependent RNA polymerases amplify the mechanism through the production of secondary siRNAs, loaded by WAGO proteins (Billi et al. 2014). In mammals, even if they lack siAGO proteins, endo-siRNAs are loaded on AGO2 (Watanabe et al. 2008), but their maturation has not been elucidated yet. Endo-siRNAs have been also reported in early diverging animals, where they are loaded by a specific siAGO (Fridrich et al. 2020).

Overall, small RNAs produced by long endogenous dsRNAs have been described in all Metazoa, but the maturation pathway evolved differently in different clades. This pathway has been certainly overlooked in Lophotrochozoa. An endo-siRNA pathway is likely to exist in most of early diverging Lophotrochozoa, since most of Platyhelminthes and Syndermata included a siAGO and DICER2 protein (Fig. 1,2). This endo-siRNA pathway looks like the one of insects and cnidarians, where two distinct DICER and AGOs interacts with two distinct small RNA types (at least for most small RNAs, see Fridrich et al. 2020). However, lophotrochozoan DICER2 proteins showed notable differences in the domain composition compared to cnidarian or insect ones.

Most of them lack the Hel and PAZ domain, which are pivotal for DICER2. The PAZ domain recognizes target dsRNAs and binds their 3' end, while the Hel domain, which seems to be inactive in all metazoan DICER1 (Aderounmu et al. 2023), allow the translocation of DICER2 along the target dsRNA, producing siRNAs processively (Kandasamy and Fukunaga 2016).

Nevertheless, I detected small RNAs with the peculiar endo-siRNA signature (i.e., 21 bp small RNA pairs with an overlap of 19 bases; Fig. 3) in the small RNA transcriptomes of *S. mediterranea* and *S. japonicum* (Platyhelminthes). Thus, it is possible that the DICER2 of early diverging Lophotrochozoa can still maturate dsRNAs without the PAZ and the Hel domain. When the Hel domain is experimentally inactivated in DICER2 of *C. elegans* or in *D. melanogaster*, the protein loses the ability of translocase along dsRNAs, but it is still able to target dsRNAs and produce endo-siRNAs (Welker et al. 2011; Sinha et al. 2018). At the same time, the inactivation of the PAZ domain leads to the production of siRNAs of altered length (Kandasamy and Fukunaga 2016). Overall, Platyhelminthes and Syndermata DICER2 might still work, and my results show that (Fig. 3), but the lack of the two domains might affect them in fidelity and efficiency.

Within Trochozoa, I did not detect DICER2, but some species belonging to Nemertea, Bryozoa and Entoprocta possess siAGO proteins. These three phyla have occasionally been placed as sister group of all the other Trochozoa (Kocot 2016; Kocot et al. 2017; Laumer et al. 2019; Khalturin et al. 2022). Thus, the endo-siRNA pathway would have been progressively lost during the evolution of Lophotrochozoa. The first step has been the loss of DICER2 in the ancestor of Trochozoa. Then, the loss of siAGO proteins in the ancestor of Mollusca, Brachiopoda, Phoronida and Annelida followed (Fig. 4). On the other hand, many other analyses do not place Entoprocta, Ectoprocta and Nemertea at the base of Trochozoa (Laumer et al. 2015; Marlétaz et al. 2019); in that scenario siAGO proteins would have been lost multiple times, depending on the phylogenetic relationships between phyla. Finally, the absence of a complete endo-siRNA pathway in Trochozoa is confirmed by the analysis of small RNA transcriptomes: I did not detect small RNAs with the endo-siRNA signature in species with an uncomplete endo-siRNA pathway (Fig. 3), namely *C. gigas* (Mollusca), *N. geniculatus* (Nemertea), but also *D. rerio* and *A. japonicus* (Deuterostomia).

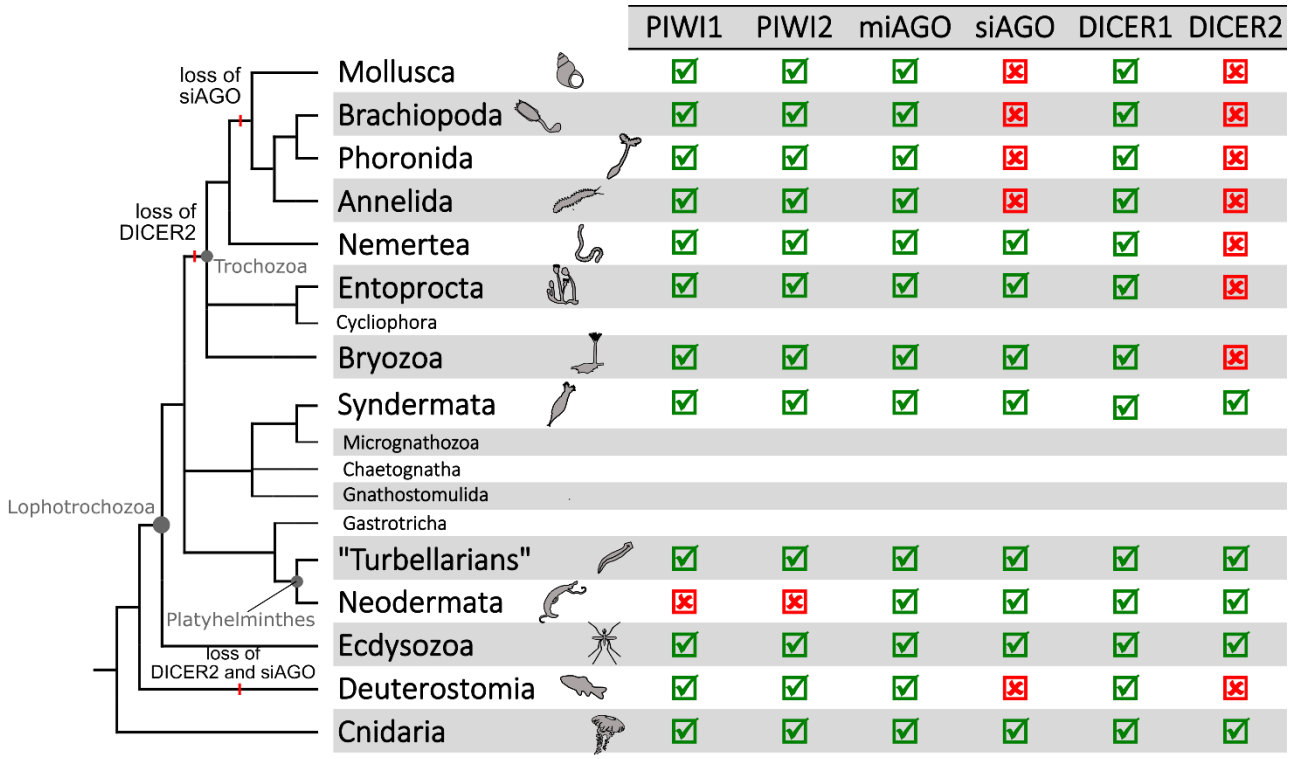


Figure 4 The loss of DICER2 and siAGO along the Metazoa evolution. The lophotrochozoan phylogenetic tree is reconstructed according to the latest Lophotrochozoa phylogenetic analyses (Kocot et al. 2017; Bleidorn 2019; Marlétaz et al. 2019). For each protein it is reported its presence, with a green check, or the absence, with a red cross, in each metazoan clade.

Unravelling the evolution of RNA interference pathways in Lophotrochozoa.

In this study I highlighted notable differences between Lophotrochozoa and other Metazoa RNAi pathways. Platyhelminthes and Syndermata have maintained an endo-siRNA machinery, but DICER2 diverges considerably from the DICER2 protein of other metazoan clades. Nevertheless, Platyhelminthes are able to produce endo-siRNAs (Fig. 3). Since the endo-siRNA pathway is highly diverse among different Metazoa clades, it is likely that also in early diverging Lophotrochozoa the endo-siRNA pathway has evolved in a unique way, not completely comparable to other metazoan pathways.

An even more unique condition was described in Nemertea, Entoprocta and Ectoprocta. These clades show an intermediate state during the loss of the endo-siRNA pathway in Lophotrochozoa. The absence of DICER2 proteins may preclude a functional endo-siRNA pathway. However, this pathway proved to be very flexible. It is possible that the siAGO protein of these clades has evolved to load small RNAs maturated from other pathways. Like in *C. elegans*, DICER1 may maturate miRNA as well as endo-siRNA. Having said that, there

are multiple scenarios where siAGO proteins may be involved, also considering the number of unconventional RNAi pathways that have been described so far (Yang and Lai 2011).

Finally, the loss of siAGO proteins and, more generally, of the endo-siRNA pathway, possibly in all Trochozoa, has been strongly supported by the joint phylogenetic analysis and analysis of sncRNA libraries. The pattern is comparable with that obtained from Deuterostomia, where the absence of the canonical (i.e., involving both DICER and siAGOs) endo-siRNA pathway has already been reported using the lack of annotated siAGO proteins as a proxy (Wynant et al. 2017). My analysis supports this hypothesis, since neither siAGO nor DICER2 homologs were retrieved in deuterostomes (Fig. 1,2). Nevertheless, in mammals, endo-siRNAs are processed by the same DICER and AGO proteins that process miRNAs (Watanabe et al. 2008; Svobodova et al. 2016). Similarly, in Trochozoa, endo-siRNAs may be matured by the miRNA pathway or other proteins related to the RNAi mechanism. As for mammals, immunoprecipitation or knockout experiments might elucidate whether Argonaute proteins, as well as other protein families, can interact with other sncRNAs types in addition to piRNAs and miRNAs. All these findings are also deeply linked to the characterization of the innate immune system. In mammals the interferon pathway has replaced RNAi mechanisms in the role of viral defense (Isaacs et al. 1963; Loo and Gale 2011; Schuster et al. 2019). An interferon defence mechanism has been described also in Mollusca (Huang et al. 2017; Qiao et al. 2021). Thus, in Lophotrochozoa an interferon system might have evolved coincidentally with the loss of the endo-siRNA pathway. Further comparative analyses might characterize the evolution of these pathways and elucidate the mechanisms that control the innate immune system in Lophotrochozoa.

3.4 Materials and Methods

Annotation of Argonaute and DICER proteins

Argonaute and DICER proteins were annotated for a wide range of omics-data. Initially, I analysed all lophotrochozoan proteomes annotated through the NCBI Eukaryotic Genome Annotation Pipeline (Thibaud-Nissen et al. 2016). To increase the sampling in underrepresented clades, I selected 17 assemblies (Supplementary table S9) and predicted gene models using the BRAKER2 automated pipeline (Stanke et al. 2008; Hoff et al. 2019; Brůna et al. 2021). To enhance the DICER and Argonaute model predictions, I enriched the Metazoa OrthoDB database provided by BRAKER2 with Argonaute and DICER sequences annotated from the lophotrochozoan NCBI proteomes. I collapsed all isoforms, retaining the longest ones, using the Perl script agat_sp_keep_longest_isoform.pl (Dainat). When few genome assemblies were available for a given phylum, I searched for Argonaute and DICER proteins in transcriptomes (Table 1). I trimmed the reads with Trimmomatic-0.39 (Bolger et al. 2014) using the following settings: ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:75. I assembled the transcriptome with Trinity v2.1.1 with default settings (Grabherr et al. 2011); I filtered out contaminants by locally aligning the transcripts against the non-redundant protein database (Sayers et al. 2022) with DIAMOND blastp (Camacho et al. 2009; Buchfink et al. 2015) and discarding all transcripts with a non-metazoan best hit.

Coding regions were predicted with Transdecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder), scanning all ORFs for homology using DIAMOND blastp and HMMER (Mistry et al. 2013). Overall, I obtained the coding sequences of 42 lophotrochozoan species and six other metazoan species. Coding sequences were translated into amino acid sequences, and then Argonaute and DICER proteins were annotated as follows: I looked for Argonaute proteins by annotating the conserved PIWI and PAZ domains in all 49 species. Domain alignments were retrieved from Pfam (PIWI accession: pfam02171; PAZ accession: pfam02170) (Mistry et al. 2021). Using HMMER, I built a profile for each multiple sequence alignment and searched the profiles against each (--e-value 10e-6). Only proteins with both domains annotated were considered Argonaute proteins and retained for phylogenetic analysis.

To annotate DICER proteins, I aligned the amino acid sequences against the bilaterian annotated set of Mukherjee and colleagues (2013) using blastp. In a second round of filtering, I retrieved the ribonuclease 3 domain alignment from Pfam (accession: pf14622.9), the only domain shared between all metazoan DICERs (Mukherjee et al. 2013); Scanning the sequences with HMMER (--e-value 10e-6), I selected only those containing the ribonucleases 3 domain. However, orthologs of the endoribonuclease DROSHA were possibly included among the annotated DICERs at this stage. Therefore, I downloaded the DROSHA orthologs from OrthoDB (reference: 9211at3208) (Zdobnov et al. 2021) and I built a custom dataset with both DICER from

Mukherjee and colleagues (2013) and DROSHA sequences. I locally aligned the set of putatively annotated DICER proteins against this dataset. I retained proteins whose best five hits were all with DICER orthologs, and discarded proteins with only DROSHA orthologs among the best five hits. No ambiguous results (i.e., proteins showing both DICER and DROSHA within the best five hits) were obtained.

The Argonaute and DICER phylogenetic trees were inferred from datasets comprising all annotated sequences from proteomes, genomes and transcriptomes, with the addition of reference sequences chosen from SwissProt (Supplementary table S1) or the bilaterian annotated set for the DICER dataset. Datasets were aligned with MAFFT v7.508 (Katoh and Standley 2013), using the options --maxiterate 1000 --localpair. Uninformative columns were masked from the alignments using Gblocks (Castresana 2000), setting -b2= (3 × number of sequences)/5 -b3=10 -b4=5 -b5=a. Additionaly, another masking tool, ClipKIT (Steenwyk et al. 2020), was used to assess the impact of the masking step on the phylogenetic analyses. Gblocks resulted in being more conservative than ClipKIT, masking most of the alignment columns (Supplementary table S10). However, the ML trees inferred from the two alignments showed no difference between each other regarding the presence of each Argonaute (i.e., miAGO, siAGO, PIWI1,2) or DICER protein in each lophotrochozoan species (Supplementary table S10, supplementary materials). Therefore, only the alignment obtained with Gblocks was used for downstream analyses.

The ML trees were inferred with IQ-Tree (Nguyen et al. 2015) using the predefined protein mixture model LG+C20+R4. To assess the robustness of the clades I calculated the ultrafast bootstrap approximation (UFBoot) with 1,000 bootstrap replicates (Hoang et al. 2018) and the SH-like approximate likelihood ratio test with 1,000 replicates (Guindon et al. 2010).

I tested whether DICER1 and miAGO proteins have experienced an intensified selection with HyPhy RELAX (Wertheim et al. 2015). I tagged all branches belonging to DICER1 and miAGO clades as foreground. All other branches belonging to DICER or AGO family clades were tagged as background. For some proteins I was not able to retrieve the respective coding sequence, namely the *Saccostrea glomerata* miAGO and DICER1, *Anopheles gambiae* DICER2, *Trobolium castaneum* and *Brugia malayi* DICER1. Those proteins were removed during the selection analysis.

The completeness of proteomes, assemblies and transcriptomes was evaluated with BUSCO v. 5.4.3 (Simão et al. 2015) using the Metazoa dataset.

The annotation of the N (accession: pfam16486) and MID (accession: pfam16487) domains for Argonaute proteins, and the Hel and PAZ domains for DICER proteins was made using HMMER, setting the e-value cut-off as the lowest e-value among outgroup sequences. The Hel and PAZ profiles were built from sequences downloaded from UniProt (Bateman et al. 2021): *D. melanogaster* DICER1 (accession: Q9VCU9), *D.*

melanogaster DICER2 (accession: A1ZAW0), *N. vectensis* DICER1 (accession: U3MHS9), *Mytilus gallus* DICER (accession: A0A140H129), *C. elegans* DICER (accession: P34529), *H. sapiens* DICER (accession: Q9UPY3).

Notospermus geniculatus sncRNA libraries sequencing and analysis of sncRNA libraries

Six specimens (three males and three females) of the nemertean *Notospermus geniculatus* were sampled in June 2018 near Ushimado (Okayama prefecture, Japan). Animals were left in seawater and 7% $MgCl_2 \cdot H_2O$ (1:1 ratio) for 15'; gonads were then dissected in 7% $MgCl_2 \cdot H_2O$ on ice and stored in RNAlater (Thermo Fisher Scientific Inc., Waltham, USA), following the manufacturer's instructions. Total RNA was extracted using a standard chloforom:TRI Reagent® (Merck KGaA, Darmstadt, Germany) protocol, following manufacturer's instructions. The TruSeq Small RNA library kit (Illumina, San Diego, USA) was used to prepare six small RNA libraries that were sequenced on an Illumina HiSeq2500 platform. Both library preparation and high-throughput sequencing were carried out at the Macrogen Inc. facility (Seoul, South Korea).

For this study, I sequenced the sncRNA pool from six samples of *N. geniculatus.* These libraries were analyzed alongside publicly available sncRNA libraries from five other species. The libraries were selected and downloaded from the Sequence Reads Archive (SRA) provided by NCBI (Supplementary table S11). Where multiple samples from the same project were available, libraries were pooled together, to obtain a single fastq file for each species. Adapters and low-quality bases were removed from reads using Cutadapt v3.9.7 (Martin 2011), with the options -e 0.2 -O 5 --quality-cutoff 6 --discard-untrimmed. Trimmed reads were mapped on the reference genome using Bowtie (Langmead et al. 2009), allowing up to 100 multiple alignments (-m 100). The distribution of overlaps between reads was estimated using the python script *overlapping_reads.py* (https://github.com/ARTbio/tools-artbio/blob/master/tools/small_rna_signatures/overlapping_reads.py; Antoniewski 2014).

For *Crassostrea gigas*, *Danio rerio*, *Anopheles gambiae* and, *Notospermus geniculatus* I annotated miRNAs, siRNAs and piRNAs. The tool miRDeep2 (Friedländer et al. 2012) was used to predict miRNAs, providing the already annotated miRNA set of that species from MiRGeneDB (Fromm et al. 2022) or miRBase (Kozomara et al. 2019), along with the annotated miRNAs of up to five closely related species. The novel predicted miRNAs were evaluated following the criteria established by Fromm and colleagues (2015) and discarding novel miRNAs with a STAR sequence coverage lower than 5 reads. Putative siRNA and piRNA pairs were predicted based on read overlaps (i.e., siRNA pairs overlap of the read length – 2, piRNA pairs with an overlap of 10 nucleotides), calculated using *overlapping_reads.py.* Pairs of siRNAs and piRNAs were

discarded when: one of the paired small RNAs had a coverage lower than 5; the logarithmic ratio of the pair exceeded 1.5; or the pair mapped on a miRNA region. Differential expression of sncRNAs between somatic tissues and ovaries was tested with edgeR (Robinson et al. 2010) with a generalized linear model and a quasi-likelihood F-test (Lund et al. 2012). I chose a stringent P-value threshold of 0.001 to consider a small RNA as significantly differentially expressed. Novel miRNA genes in *Notospermus geniculatus* were assigned to known or novel miRNA families. I locally aligned the preliminary structure of miRNA genes against the full set of MirGeneDB pre-miRNA using blastn. Novel genes were assigned to the best hit miRNA family only if they shared the same seed (Supplementary table S7). The remaining miRNA genes were clustered into novel miRNA families by blasting the pre-miRNAs against each other and comparing the seed sequence: miRNAs were clustered in the same family if they blasted against each other and the seed sequence differed by up to one nucleotide (Supplementary table S8).

3.5 References

Aderounmu AM, Aruscavage PJ, Kolaczkowski B, Bass BL. 2023. Ancestral protein reconstruction reveals evolutionary events governing variation in Dicer helicase function. *Elife* 12.

Antoniewski C. 2014. Computing siRNA and piRNA Overlap Signatures. In: Animal Endo-SiRNAs. Vol. 1173. New York, NY: Humana Press. p. 135–146.

Bartel DP. 2018. Metazoan MicroRNAs. *Cell* 173:20–51.

Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bursteinas B, et al. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49:D480–D489.

Bernstein E, Caudy AA, Hammond SM, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409:363–366.

Billi AC, Fischer SEJ, Kim JK. 2014. Endogenous RNAi pathways in C. elegans. *WormBook*:1–49.

Bleidorn C. 2019. Recent progress in reconstructing lophotrochozoan (spiralian) phylogeny. *Org Divers Evol* 19:557–566.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* 128:1089–1103.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* 3.

Brusca R, Moore W, Shuster S. 2016. Invertebrates. Third Edition. Sinauer Associates

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60.

Calcino AD, Fernandez-Valverde SL, Taft RJ, Degnan BM. 2018. Diverse RNA interference strategies in early-branching metazoans. *BMC Evol Biol* 18:160.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol Biol Evol* 17:540–552.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, et al. 2008. An endogenous small interfering RNA pathway in Drosophila. *Nature* 453:798–802.

Dainat J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format.

Fontenla S, Rinaldi G, Smircich P, Tort JF. 2017. Conservation and diversification of small RNA pathways within flatworms. *BMC Evol Biol* 17:215.

Fontenla S, Rinaldi G, Tort JF. 2021. Lost and Found: Piwi and Argonaute Pathways in Flatworms. *Front Cell Infect Microbiol* 11.

Fridrich A, Modepalli V, Lewandowska M, Aharoni R, Moran Y. 2020. Unravelling the developmental and functional significance of an ancient Argonaute duplication. *Nat Commun* 11:6187.

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 40:37–52.

Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E, et al. 2015. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu Rev Genet* 49:213–242.

Fromm B, Høye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, Umu SU, Chabot PJ, Kang W, Aslanzadeh M, et al. 2022. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res* 50:D204–D210.

Garcia Silva MR, Tosar JP, Frugier M, Pantano S, Bonilla B, Esteban L, Serra E, Rovira C, Robello C, Cayota A. 2010. Cloning, characterization and subcellular localization of a Trypanosoma cruzi argonaute protein defining a new subfamily distinctive of trypanosomatids. *Gene* 466:26–35.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455:1193–1197.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* 59:307–321.

Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, Chu DS, Thierry-Mieg D, Thierry-Mieg J, Kim JK. 2009. 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences* 106:18674–18679.

Hirakata S, Siomi MC. 2016. piRNA biogenesis in the germline: From transcription of piRNA genomic sources to piRNA maturation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1859:82–92.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 35:518–522.

Höck J, Meister G. 2008. The Argonaute protein family. *Genome Biol* 9:210.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. In: p. 65–95.

Huang B, Zhang L, Du Y, Xu F, Li L, Zhang G. 2017. Characterization of the Mollusc RIG-I/MAVS Pathway Reveals an Archaic Antiviral Signalling Framework in Invertebrates. *Sci Rep* 7:8217.

Huntzinger E, Izaurralde E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* 12:99–110.

Isaacs A, Cox RA, Rotem Z. 1963. Foreign nucleic acids as the stimulus to make interferon. *The Lancet* 282:113–116.

Iwakawa H, Tomari Y. 2022. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. *Mol Cell* 82:30–43.

Jehn J, Gebert D, Pipilescu F, Stern S, Kiefer JST, Hewel C, Rosenkranz D. 2018. PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation. *Commun Biol* 1:137.

Jiang F, Ye X, Liu X, Fincher L, McKearin D, Liu Q. 2005. Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*. *Genes Dev* 19:1674–1679.

Kandasamy SK, Fukunaga R. 2016. Phosphate-binding pocket in Dicer-2 PAZ domain for high-fidelity siRNA production. *Proceedings of the National Academy of Sciences* 113:14031–14036.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30:772–780.

Khalturin K, Shunatova N, Shchenkov S, Sasakura Y, Kawamitsu M, Satoh N. 2022. Polyzoa is back: The effect of complete gene sets on the placement of Ectoprocta and Entoprocta. *Sci Adv* 8.

Khanal S, Zancanela BS, Peter JO, Flynt AS. 2022. The Small RNA Universe of Capitella teleta. *Front Mol Biosci* 9.

Kocot KM. 2016. On 20 years of Lophotrochozoa. *Org Divers Evol* 16:329–343.

Kocot KM, Struck TH, Merkel J, Waits DS, Todt C, Brannock PM, Weese DA, Cannon JT, Moroz LL, Lieb B, et al. 2017. Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. *Syst Biol* 66:256–282.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 47:D155–D162.

Kuhn C-D, Joshua-Tor L. 2013. Eukaryotic Argonautes come into focus. *Trends Biochem Sci* 38:263–271.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

Laumer CE, Bekkouche N, Kerbl A, Goetz F, Neves RC, Sørensen MV, Kristensen RM, Hejnol A, Dunn CW, Giribet G, et al. 2015. Spiralian Phylogeny Informs the Evolution of Microscopic Lineages. *Current Biology* 25:2000–2006.

Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Riesgo A, Andrade SCS, Sterrer W, Sørensen M V., Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B: Biological Sciences* 286:20190831.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425:415–419.

Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 23:4051–4060.

Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW. 2004. Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways. *Cell* 117:69–81.

Liu Z, Wang J, Cheng H, Ke X, Sun L, Zhang QC, Wang H-W. 2018. Cryo-EM Structure of Human Dicer and Its Complexes with a Pre-miRNA Substrate. *Cell* 173:1191-1203.e12.

Loo Y-M, Gale M. 2011. Immune Signaling by RIG-I-like Receptors. *Immunity* 34:680–692.

Lund SP, Nettleton D, McCarthy DJ, Smyth GK. 2012. Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates. *Stat Appl Genet Mol Biol* 11.

Malone CD, Hannon GJ. 2009. Small RNAs as Guardians of the Genome. *Cell* 136:656–668.

Marlétaz F, Peijnenburg KTCA, Goto T, Satoh N, Rokhsar DS. 2019. A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. *Current Biology* 29:312–318.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17:10.

Matranga C, Zamore PD. 2007. Small silencing RNAs. *Current Biology* 17:R789–R793.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res* 49:D412–D419.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 41:e121–e121.

Mohn F, Handler D, Brennecke J. 2015. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science (1979)* 348:812–817.

Moran Y, Agron M, Praher D, Technau U. 2017. The evolutionary origin of plant and animal microRNAs. *Nat Ecol Evol* 1:0027.

Moran Y, Praher D, Fredman D, Technau U. 2013. The Evolution of MicroRNA Pathway Protein Components in Cnidaria. *Mol Biol Evol* 30:2541–2552.

Mukherjee K, Campos H, Kolaczkowski B. 2013. Evolution of Animal and Plant Dicers: Early Parallel Duplications and Recurrent Adaptation of Antiviral RNA Binding in Plants. *Mol Biol Evol* 30:627–641.

Nesnidal MP, Helmkampf M, Meyer A, Witek A, Bruchhaus I, Ebersberger I, Hankeln T, Lieb B, Struck TH, Hausdorf B. 2013. New phylogenomic data support the monophyly of Lophophorata and an Ectoproct-Phoronid clade and indicate that Polyzoa and Kryptrochozoa are caused by systematic bias. *BMC Evol Biol* 13:253.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 32:268–274.

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* 20:89–108.

Praher D, Zimmermann B, Genikhovich G, Columbus-Shenkar Y, Modepalli V, Aharoni R, Moran Y, Technau U. 2017. Characterization of the piRNA pathway during development of the sea anemone *Nematostella vectensis*. *RNA Biol* 14:1727–1741.

Qiao X, Wang L, Song L. 2021. The primitive interferon-like system and its antiviral function in molluscs. *Dev Comp Immunol* 118:103997.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

Rosani U, Pallavicini A, Venier P. 2016. The miRNA biogenesis in marine bivalves. *PeerJ* 4:e1763.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 50:D20–D26.

Schuster S, Miesen P, van Rij RP. 2019. Antiviral RNAi in Insects and Mammals: Parallels and Differences. *Viruses* 11:448.

Shabalina S, Koonin E. 2008. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol* 23:578–587.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

Sinha NK, Iwasa J, Shen PS, Bass BL. 2018. Dicer uses distinct modules for recognizing dsRNA termini. *Science (1979)* 359:329–334.

Song J-J, Smith SK, Hannon GJ, Joshua-Tor L. 2004. Crystal Structure of Argonaute and Its Implications for RISC Slicer Activity. *Science (1979)* 305:1434–1437.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24:637–644.

Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A. 2020. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol* 18:e3001007.

Strassert JFH, Irisarri I, Williams TA, Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat Commun* 12:1879.

Struck TH, Wey-Fabrizius AR, Golombek A, Hering L, Weigert A, Bleidorn C, Klebow S, Iakovenko N, Hausdorf B, Petersen M, et al. 2014. Platyzoan Paraphyly Based on Phylogenomic Data Supports a Noncoelomate Ancestry of Spiralia. *Mol Biol Evol* 31:1833–1849.

Svobodova E, Kubikova J, Svoboda P. 2016. Production of small RNAs by mammalian Dicer. *Pflugers Arch* 468:1089–1102.

Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin E V, Patel DJ, van der Oost J. 2014. The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* 21:743–753.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453:534–538.

Tarver JE, Sperling EA, Nailor A, Heimberg AM, Robinson JM, King BL, Pisani D, Donoghue PCJ, Peterson KJ. 2013. miRNAs: Small Genes with Big Potential in Metazoan Phylogenetics. *Mol Biol Evol* 30:2369–2382.

Thibaud-Nissen F, DiCuccio M, Hlavina W, Kimchi A, Kitts PA, Murphy TD, Pruitt KD, Souvorov A. 2016. P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *J Anim Sci* 94:184–184.

Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev* 27:390–399.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453:539–543.

Weick E-M, Miska EA. 2014. piRNAs: from biogenesis to function. *Development* 141:3458–3471.

Welker NC, Maity TS, Ye X, Aruscavage PJ, Krauchuk AA, Liu Q, Bass BL. 2011. Dicer's Helicase Domain Discriminates dsRNA Termini to Promote an Altered Reaction Mode. *Mol Cell* 41:589–599.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol Biol Evol* 32:820–832.

Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ. 2009. The deep evolution of metazoan microRNAs. *Evol Dev* 11:50–68.

Wu J, Yang J, Cho WC, Zheng Y. 2020. Argonaute proteins: Structural features, functions and emerging roles. *J Adv Res* 24:317–324.

Wynant N, Santos D, Vanden Broeck J. 2017. The evolution of animal Argonautes: evidence for the absence of antiviral AGO Argonautes in vertebrates. *Sci Rep* 7:9230.

Yang J-S, Lai EC. 2011. Alternative miRNA Biogenesis Pathways and the Interpretation of Core miRNA Pathway Mutants. *Mol Cell* 43:892–903.

Yerukhimovich MM, Marohnic CC, Frick DN. 2018. Role of the Conserved DECH-Box Cysteine in Coupling Hepatitis C Virus Helicase-Catalyzed ATP Hydrolysis to RNA Unwinding. *Biochemistry* 57:6247–6255.

Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva E V. 2021. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 49:D389–D393.

4. Identification of Proteins Interacting with Small Mitochondrial RNAs Using *In Silico* and *In Vivo* Approaches

This chapter was written in collaboration with Federico Plazzi, Diego Carli and Marco Passamonti.

Supplementary materials are available at the end of the chapter.

4.1 Introduction

Since the emergence of the first eukaryotic cells, the mitochondrial and nuclear genomes have coevolved and competed within the same cellular environment (Gray 2012). The interactions that have emerged over the last 2 billion years have played crucial roles in several fundamental biological processes, including speciation, aging, death, and reproduction (Wolff et al. 2014; Hill 2015). Among these interactions, a new class of short RNAs (sRNAs) has been proposed. These sRNAs are transcribed from the mitochondrial genome and regulate nuclear messenger RNAs; they are known as Small MITochondrial Highly transcribed RNAs (smithRNAs). smithRNAs were first described in the Manila clam, *Ruditapes philippinarum*, and were hypothesized to be involved in the sex determination of these clams (Pozzi et al. 2017). smithRNAs were also *in silico* predicted in all the main Metazoan branches (Passamonti et al. 2020).

Many aspects about smithRNAs are still unknown. It is not clear how smithRNAs are produced and transported in the cytosol, and whether all organisms share a conserved and homologous smithRNA pathway. In animals, there are many small RNAs classes, which differ by their role in the cell and by their maturational pathway. microRNAs (miRNAs) are sRNAs around 22 nucleotides (nt) long. They maturate from RNA hairpin structures, which are transcribed by the RNA polymerases II and are then processed; first by Microprocessor, a heterotrimeric complex formed by one molecule of the endonucleases Drosha and two molecules of the partner protein DGCR8 (Bartel 2018). The Microprocessor cut the stem of the hairpin, producing a 60 nt hairpin called pre-miRNA. The pre-miRNA is further processed by another endoribonuclease, DICER, in synergy with its partner TRBP, producing the miRNA duplex (Zhang et al. 2004). The miRNA duplex is loaded by an Argonaute protein, which forms the RNA-induced silencing complex (RISC; Iwakawa and Tomari 2022). Argonaute proteins loading miRNAs belong to a sub-family of the Argonaute superfamily, the Ago family. This family includes mammals' AGO2, AGO1 of *Drosophila melanogaster* and Alg1-2 of *Caenorhabditis elegans* (Swarts et al. 2014).

A second class of sRNAs are the endogenous small-interfering RNAs (endo-siRNAs). Contrastingly to miRNAs, the biogenesis of endo-siRNAs differs among animal branches. In *D. melanogaster*, endo-siRNAs are produced from double-strand RNAs that are cleaved by DICER2, producing 21 nt RNA duplexes that are loaded on AGO2 (Czech et al. 2008). Contrastingly, in mammals, endo-siRNAs are maturated by the same DICER and Argonaute proteins that process miRNAs (Watanabe et al. 2008; Svobodova et al. 2016). In *C. elegans*, primary endo-siRNAs are produced by DICER1 and other partner proteins and loaded on the Argonaute proteins Alg3-4 and ERGO-1. RNA-dependent RNA polymerases enable the production of secondary endo-siRNAs, which are loaded by other Argonaute proteins, namely CSR-1, HRDE-1 or WAGOs (Matranga and Zamore 2007; Almeida et al. 2019).

A third class of sRNAs are Piwi-interacting RNAs (piRNAs), which are generally enriched in germline tissue and play the role of silencing mobile elements (Weick and Miska 2014). In *D. melanogaster* three PIWI proteins are encoded: PIWI, AUBERGINE(AUB) and AGO3. All three proteins show different expression patterns and load different kinds of piRNAs (Weick and Miska 2014). Moreover, AUB and AGO3 are involved in a loop that enables the secondary amplification of paired piRNAs, called the ping-pong cycle (Brennecke et al. 2007). In *Mus musculus* three PIWI proteins are present, namely MIWI, MILI and MILI2. Even in this case they are expressed in different locations at different life stages, and they take part in the ping-pong cycle (Weick and Miska 2014). In *C. elegans* we have one functional PIWI protein, PRG-1, which binds 21 nt-long sRNAs that have a 5′ bias for uridine monophosphate; these piRNAs are transcribed by the RNA polymerase II (Almeida et al. 2019).

Overall, the three main sRNAs pathway are really diverse among different animals' branches. In addition, several other pathways produce different kinds of sRNAs. These pathways, which are called "non-canonical" pathways, can be variants or sub-parts of the "canonical" ones, and they may start from different precursors. Some "non-canonical" variants of the miRNA pathways include DICER- or Microprocessor-independent pathways (Daugaard and Hansen 2017). In the latter group we find mirtrons, whose precursors are introns with a hairpin structure which is further processed by DICER and loaded on AGO2 proteins (Okamura et al. 2007). Mirtrons have been confirmed in several animal branches (Westholm and Lai 2011), and their presence has been reported even in plants (Zhu et al. 2008). tRNA-related fragments (tRFs) are another type of non-canonical sRNAs, which are produced from tRNAs. Their maturational pathway is still unclear, but it is unlikely that they are processed by the Microprocessor or DICER, since the knock down of these proteins in *H. sapiens* did not affect the tRFs signal (Kumar et al. 2014). It is more likely that some other endonucleases, in synergy with RNase P and Z, process tRFs (Kumar et al. 2016). Similarly to tRFs, small ribosomal RNA-derived fragments (rRFs) were also reported in mice and flies, and they may act as miRNAs or piRNAs (Lambert et al. 2019).

The diverse array of proteins involved in sRNA pathways allows limitless possibilities for the maturation of smithRNAs. smithRNAs precursors should be maturated by an endonuclease and, once matured, loaded on an Argonaute protein. *In-silico* data have already proposed the interaction between smithRNAs and mammals AGO2 (Pozzi and Dowling 2022). Provided the diversity of Argonaute superfamily among Metazoa, it is still not clear whether other Argonaute proteins could also interact with smithRNAs. In this paper I analysed publicly available enhanced Cross-Linking and ImmunoPrecipitation (eCLIP) and RNA Immunoprecipitation (RIP) sequencing libraries to detect putative smithRNAs-interacting proteins. According to my results, interactions with AGO2, Drosha and DGCR8 were detected. Moreover, analysing 14 Argonaute proteins from four metazoan species, also *D. melanogaster* AGO2 and *ERGO-1* in *C. elegans* reported a significant

interaction with smithRNAs. These results would suggest that smithRNAs interact with proteins belonging to the AGO-like family, rather than the PIWI or the WAGO family. Moreover, other proteins belonging to the miRNA pathway may interact with smithRNAs.

4.2 Results

Proteins involved in the miRNA pathway interact with smithRNAs

eCLIP sequencing is one of the most advanced techniques for the identifications of RNA-protein interactions. I searched for eCLIP libraries targeting proteins involved in small RNA pathway. DGCR8, DROSHA and AGO2 take part in the miRNA pathway (Bartel 2018). Therefore, I tested whether these proteins are able to interact with transcripts of mitochondrial origin. In *H. sapiens*, for each of the three proteins I analysed two experiments on different cell lines (i.e., HepG2, K562, HCT116), each experiment including two biological replicates and one or two control replicates. I performed CLIP-seq peak calling using CLIPper. The reproducibility of each peak across both replicates was evaluated using an IDR analysis, with a threshold set at 0.01. The number of reads mapping to each peak in the IP and control libraries was used to calculate fold enrichment (i.e., the number of reads in the IP library divided by the number of reads in the control library), expressed on a logarithmic scale with base 2 (log FC). I considered all peaks located on the mitochondrial genome that showed a logFC > 2 in both replicates. Peaks meeting these criteria were located in ten mitochondrial regions: six of them are located on tRNA genes and four on the 12S ribosomal RNA gene. I compared the logFC measured on those regions in all three proteins, plus three proteins that are not related to sRNA pathways and were used as controls. Moreover, I also analysed AGO2 eCLIP libraries from *Mus musculus*, measuring the logFC on mitochondrial regions homologous to the regions selected for *H. sapiens* (Fig. 1). I considered a logFC value significant when its associated P-value resulted below $1x10^{-5}$ and, according to the IDR analysis, the associated peak was reproducible in both replicates. Both DROSHA experiments reported significant logFC in tRNA-Phe and tRNA-Met genes. Moreover, significant and considerably high logFC were also in tRNA-Val, tRNA-Gly and tRNA-Pro genes for the experiment made on K562 cells. In DGCR8, I had only one significant logFC, on the tRNA-Val gene in the HepG2 experiment. Regarding AGO2 experiments, they both reported significant logFC for tRNA-Phe, tRNA-Met, tRNA-Gly and tRNA-Pro gene. Moreover, all the other regions resulted significant for at least one of the experiments, besides tRNA-Val. Most of these regions turned out to be enriched also analysing AGO2 eCLIP libraries in mouse. Contrastingly, in two of the three control proteins all logFCs were not significant, while in FXR2 three regions located on the 12S gene were significant. Therefore, I can exclude that the signal of interactions measured for DROSHA and AGO2 is due to a tendency of tRNAs and rRNAs to interact with RNA-binding proteins.

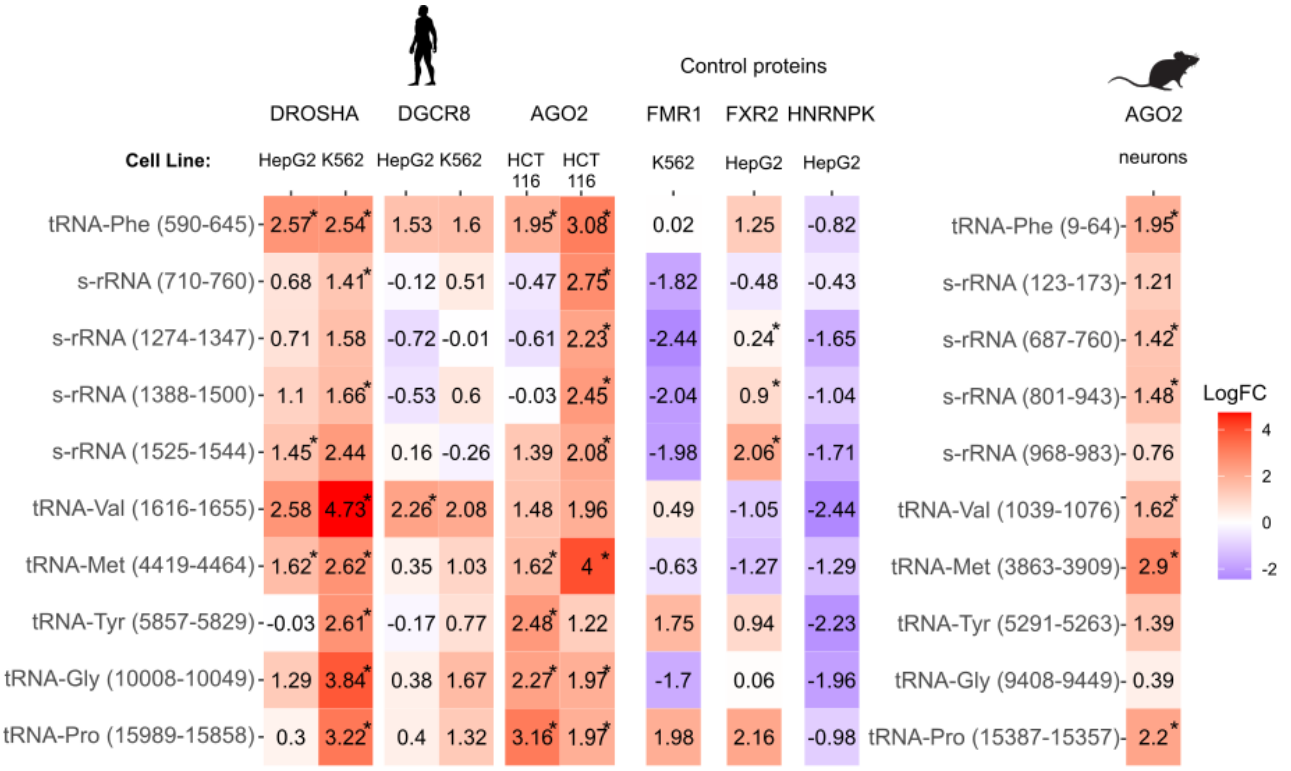


Figure 1. eCLIP-Seq analysis of sRNA related proteins. All the mt-regions that showed a logFC (log_2(n° mapping reads IP library / n° mapping reads control library)) > 2 and passed the FDR analysis in at least one of the experiments. Boxes are colored in shades of reds when logFC > 0 and in shades of blue when logFC < 0. Significant logFC values are marked with an asterisk.

Next, I tested whether the regions that interact with AGO2 and DROSHA transcribe for small RNAs that meet the abundance criteria to be considered smithRNAs (Pozzi et al. 2017; Marturano et al. 2024). To this end, I analysed miRNA libraries made available by the ENCODE project using the software SmithHunter (Marturano et al. 2024). The software detects novel smithRNAs with relatively high coverage and replicable across different libraries. I separately analysed replicates from HepG2 and K562 cells. I annotated 11 smithRNAs from HepG2 miRNA libraries and 20 smithRNAs from K562 miRNA libraries (Supplementary table S1). Among these, eight were found on both type cells. 15 smithRNAs were located on tRNAs, four on ribosomal RNAs, two on protein coding regions, and one on the d-loop. Some of these regions coincided with those detected by eCLIP analysis. In particular, smithRNA were annotated in both cell types in tRNA-Gly and tRNA-Val, while smithRNAs on tRNA-Pro and tRNA-Tyr were annotated only in one type cell and reported interactions with both proteins.

As last, I evaluated the effects of AGO2 and DGCR8 knock-out (KO) on smithRNAs transcription. I analysed miRNA seq libraries from DROSHA KO cells and AGO1 + AGO2 KO cells in *H. sapiens* (Johnson et al. 2023) and *M. musculus* (Müller et al. 2022). smithRNA transcription levels of DROSHA and AGO1+2 KO cells were compared with the transcription levels of wild-type cells. Transcription levels were expressed as mapping reads per million (RMP, that is the number of reads mapping on that position divided by the number of reads mapping in that library, multiplied by one million). I focused my attention on the ten regions that were predicted to interact with DROSHA or AGO2. In DROSHA KO cells in *H. sapiens*, the transcription level of smithRNAs was lower than in wild-type libraries. It was particularly evident in tRNA-Val, tRNA-Gly, 12S from position 1,525 to 1,544 (Fig. 2), and tRNA-Phe (Supplementary figure S1). Contrastingly, AGO2 KO libraries showed a transcription level comparable to the one of wild-type libraries. Hence, in some regions the knockout of DROSHA affected the transcription level of smithRNAs. However, this pattern was restricted to *H. sapiens*, since in *M. musculus* DROSHA KO libraries, as well as AGO2 KO libraries, showed no difference in RMP levels compared to wild-type libraries (Fig. 2; Supplementary figure S1,2). I also evaluated the effect of RNAse Z KO in *M. musculus*; this enzyme is responsible for the 3' maturation of tRNAs and mitochondrial polycistronic transcripts (Brzezniak et al., 2011; Chen et al., 2005; Siira et al., 2018). In most of the cases, smithRNAs were clearly depleted in KO libraries. However, a few tRNA and rRNAs reported a low mapping coverage in both libraries, KO and control (Supplementary figure S3).

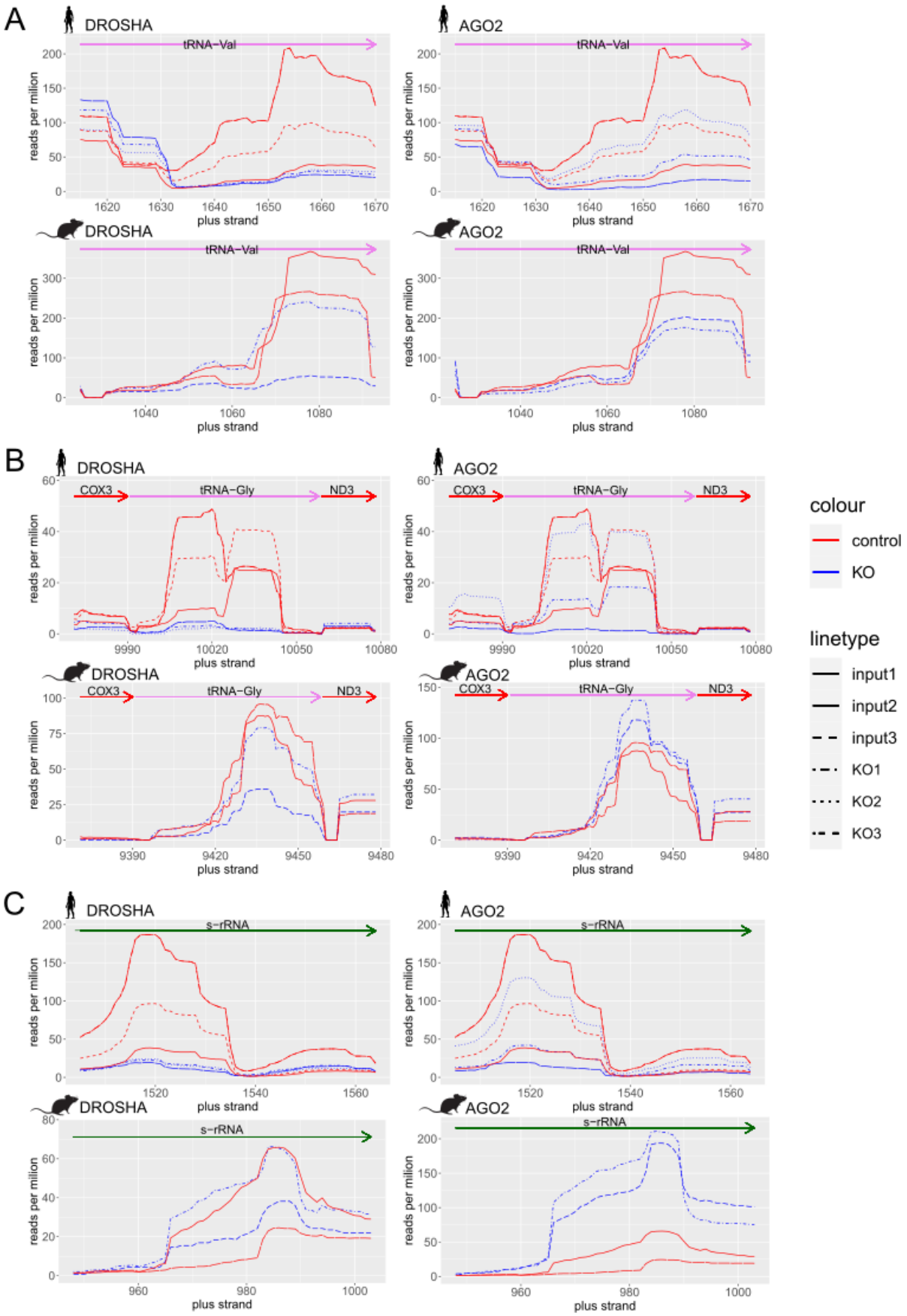

A
DROSHA
AGO2
tRNA−Val
reads per million
plus strand
B
COX3
tRNA−Gly
ND3
C
s−rRNA
colour
control
KO
linetype
input1
input2
input3
KO1
KO2
KO3

Figure 2. smithRNAs coverage in DROSHA and AGO2 KO experiments in three mitochondrial positions. Lines represent the coverage, expressed as reads per million, of three smithRNAs located on (A) tRNA-Val; (B) tRNA-Gly; (C) 12S rRNA. Colours refer to libraries from KO cells (blue) and wild-type cells (red). Line types identify replicates.

Analysing RIP-Seq data of different Argonaute proteins

The Argonaute superfamily has greatly diversified in metazoans, with different paralogs that have acquired the ability to load sRNAs with specific features (Swarts et al. 2014). Hence, it is likely that only part of the paralogs is involved in the smithRNA pathway, and it may differ in different animal branches.

eCLIP-seq is a quite new technology, which has been used for a handful of proteins in a restricted number of species. On the other hand, RIP-seq has been used for several proteins in model as well as non-model species. Therefore, I performed extensive research on SRA, looking for Bioprojects where it was performed a RIP-Seq with an Argonaute protein as IP target. It resulted in the analysis of 14 different Argonaute proteins in four metazoan species: namely, AGO1-3, Aubergine and PIWI in *D. melanogaster*, ALG-1, ERGO-1, HRDE-1, CSR-1, and PRG-1 in *C. elegans*, AGO1,2 in *Nematostella vectensis,* MIWI in *M. musculus.*

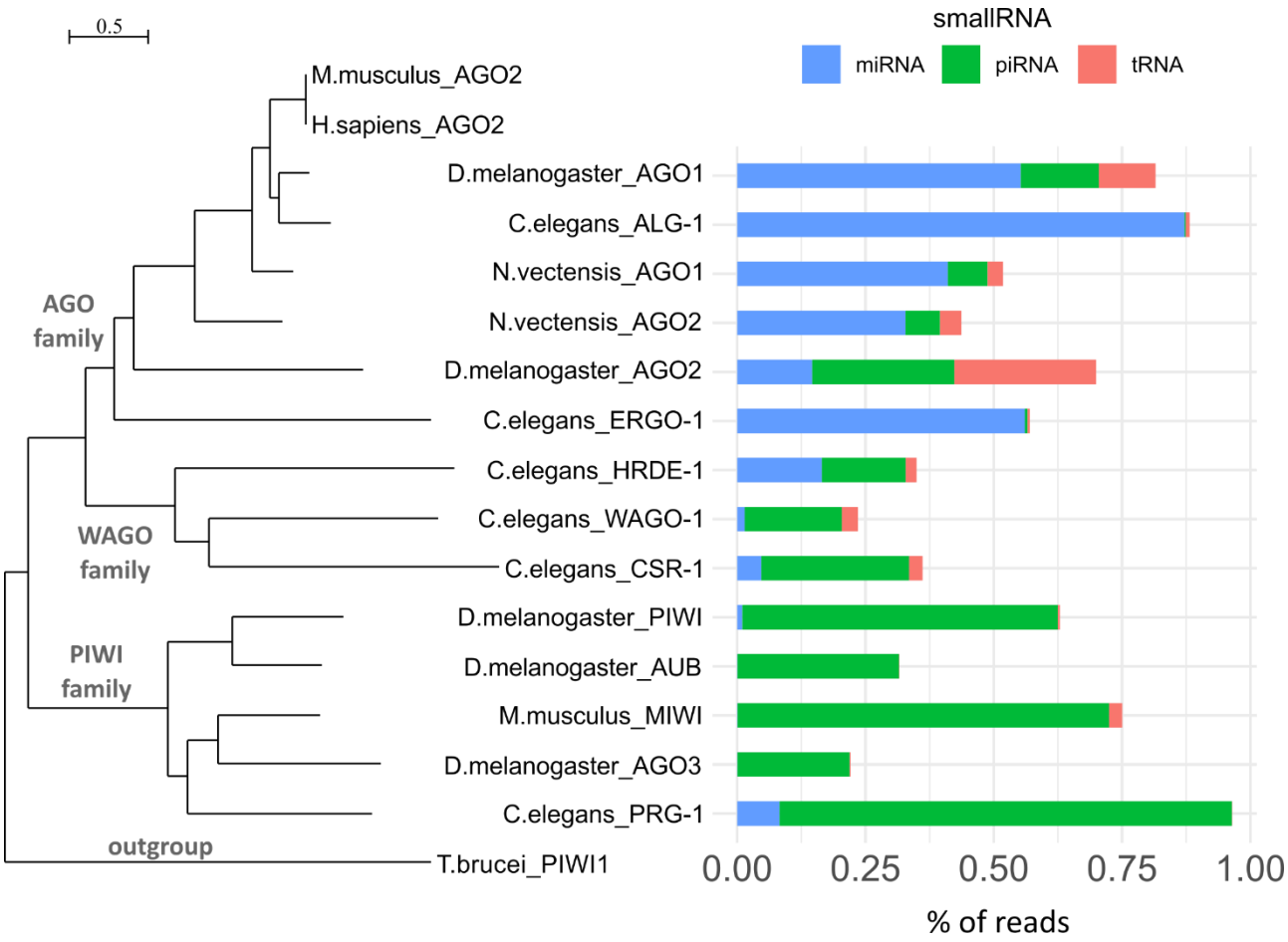


Figure 3. Small RNA composition for each RIP-seq library**.** For each RIP-seq library analysed, I measured the percentage of reads mapping on miRNAs, piRNAs and tRNAs (i.e., number of reads mapping on that class of sRNAs divided by the number of mapped reads). Phylogenetic relationships were inferred with IQ-TREE using protein sequences retrieved from Uniprot.

RIP-Seq libraries were mapped against different small RNA libraries (i.e., miRNA, piRNA and tRNA libraries, see Materials and Methods). First, I checked where most of the reads were mapping in the IP libraries. As expected, when the IP protein belonged to the AGO-family, the small RNA pool was enriched in miRNAs. On the other hand, when the IP protein belonged to the PIWI-family, the small RNA pool was enriched in piRNAs. The WAGO-family is not linked with neither of the two small RNA classes; indeed, most of the reads in these libraries mapped on other small RNA types. I also calculated the percentage of reads that map on tRNAs. In general, any IP library resulted enriched for tRNA derived small RNAs; however, AGO-family proteins reported higher percentage values than the PIWI-like proteins, in particular AGO1 and 2 in *D. melanogaster* (Fig. 3).

To confirm the small RNA-protein interaction, I assessed whether the small RNAs identified in the IP-libraries were significantly enriched in the IP libraries compared to the control libraries. First, in the IP libraries I identified putative smithRNAs by selecting small RNAs that were mapping on the mitochondrial genome and had a coverage higher than 200, in accordance with the pipeline of Pozzi and colleagues (2017). Then, I mapped the reads from IP and control libraries against the set of putative smithRNAs and other small RNA databases (i.e., miRNAs, piRNAs, tRNAs and other small RNA classes). I identified the small RNA enriched in the IP library with a differential expression (DE) analysis. I found only three proteins out of 14 with enriched smithRNAs, namely ERGO-1 and CSR-1 in *C. elegans* and AGO2 in *D. melanogaster* (Tab. 1). ERGO-1 were found to interact with four smithRNAs: two are located on coding regions, one on the tRNA-Met, and one on the AT-region. CSR-1 interacted with one smithRNA located on the AT-region and five located on coding regions, whereas all smithRNAs enriched in AGO2-IP libraries were located on tRNAs.

Species	Protein	# sRNAs with mapped reads	# of DE sRNAs	# of DE smithRNAs	DE smithRNAs table
C. elegans	ERGO-1	3813	817	4	**Mt Region** / **LogFC** / **PValue** NADH6:112-158(+) / 1.423622 / 0.019078 COX2:9477-9509(+) / 1.257936 / 0.036384 tRNA-Met:1389-1419(+) / 1.520286 / 0.014438 AT-region:13708-13752(+) / 5.928511 / 0.001264
C. elegans	PRG-1	9472	1288	0	
C. elegans	ALG-1	3546	245	0	
C. elegans	HRDE-1	15503	5742	0	
C. elegans	CSR-1	20553	4168	6	**Mt Region** / **LogFC** / **PValue** AT-region:13315-13358(+) / 3.873227 / 1.94E-04 CYTB:4882-4926(+) / 3.785737 / 2.67E-05 CYTB:5250-5291(+) / 2.605205 / 6.00E-04 CYTB:5293-5340(+) / 3.57009 / 5.13E-05 NADH1:1794-1843(+) / 1.939445 / 1.38E-02 NADH5:12350-12390(+) / 3.392176 / 8.84E-05
C. elegans	WAGO2*	20553	3615	0	
D. melanogaster	AGO1	3810	706	0	
D. melanogaster	AGO2	6543	711	7	**Mt Region** / **LogFC** / **PValue** tRNA-Met:167-187(+) / 9.951901 / 2.23E-09

					tRNA-Ser:6190-6218(+)	2.376612	4.29E-05
					tRNA-Ser:6124-6151(+)	1.032345	3.77E-03
					tRNA-Leu2:12712-12734(-)	0.79648	2.60E-02
					tRNA-Leu1:3009-3032(+)	1.21979	4.31E-03
					tRNA-Ala:5982-6001(+)	1.995427	1.59E-04
					tRNA-Thr:9936-9968(-)	1.66145	2.40E-03
D. melanogaster	PIWI*	26699	10499	0			
D. melanogaster	Aubergine*	26699	1589	0			
D. melanogaster	AGO3*	26699	3094	0			
N. vectensis	AGO1	1255	205	0			
N. vectensis	AGO2	1255	276	0			
M. musculus	MIWI	30806	8489	0			

Table 1. For each RIP-Seq experiment, I reported the target species, the target protein, the number of sRNAs with mapped reads, the number of DE sRNAs, and the number of DE sRNAs mapping to the mt-genome. A sub-table is included for experiments showing DE mt-sRNAs. For each mt-sRNA, it provides the coordinates on the mt-genome, the fold change value expressed on a logarithmic scale with base 2, and the associated p-value. I marked proteins with only one RIP-Seq replicate available with an asterisk.

Identification of smithRNA interacting proteins in R. philippinarum

Two smithRNAs (i.e. 145t and 122nca), which were previously annotated in *R. philillipanrum* (Pozzi et al. 2017), and one nuclear miRNA (i.e., let-7) were selected for a pull-down protocol (Supplementary table S2). The precursor and mature forms of both smithRNAs and let-7 were exposed to a clam's tissue lysate and pulled down along with the interacting proteins. Through the Liquid Chromatography with tandem Mass Spectrometry (LC-MS/MS) analysis of the six samples plus one control, 6,880 proteins were reliably identified. A Principal Component Analysis (PCA) was performed on the replicate samples using all quantified proteins as variables (Fig. 4). The first principal component separated the three mature small RNAs from their precursors. Additionally, let-7 clustered with 145t, while pre-122nca clustered with pre-145t along the second principal component. I selected proteins that showed a Label-Free Quantification (LFQ) at least twofold higher in one of the conditions compared to the LFQ measured in the control samples, and analysed if there were specific GO terms enriched in the selected proteins (Supplementary table S4,5). As expected, terms related to general protein-RNA interaction, such as "RNA binding motif" and "Nucleotide-binding alpha-beta plait domain", resulted enriched in all precursor RNAs. Terms related to the spliceosome activity resulted enriched in the precursors of 145t and 122nca. In particular, several proteins that are part of the spliceosome complex (i.e., LSm2, LSm4, LSm6, LSm7, LSm8, SmD2, SmD3, snRNP-B, snRNP-E, snRNP-F; Will and Lührmann 2001) were identified in the 145t and 122nca precursor samples with higher intensity than the control sample

(Supplementary table S5). Terms related to "isomerase activity" resulted enriched in the matured let-7 and 145t samples, however the interacting isomerases were not related to a specific pathway. Finally, I checked if proteins known to be involved in small RNA pathways were enriched in some of the replicates. The only detected protein resulted TARBP2, which is required for the RISC assembly. However, this protein resulted enriched only in the precursor form of let-7.

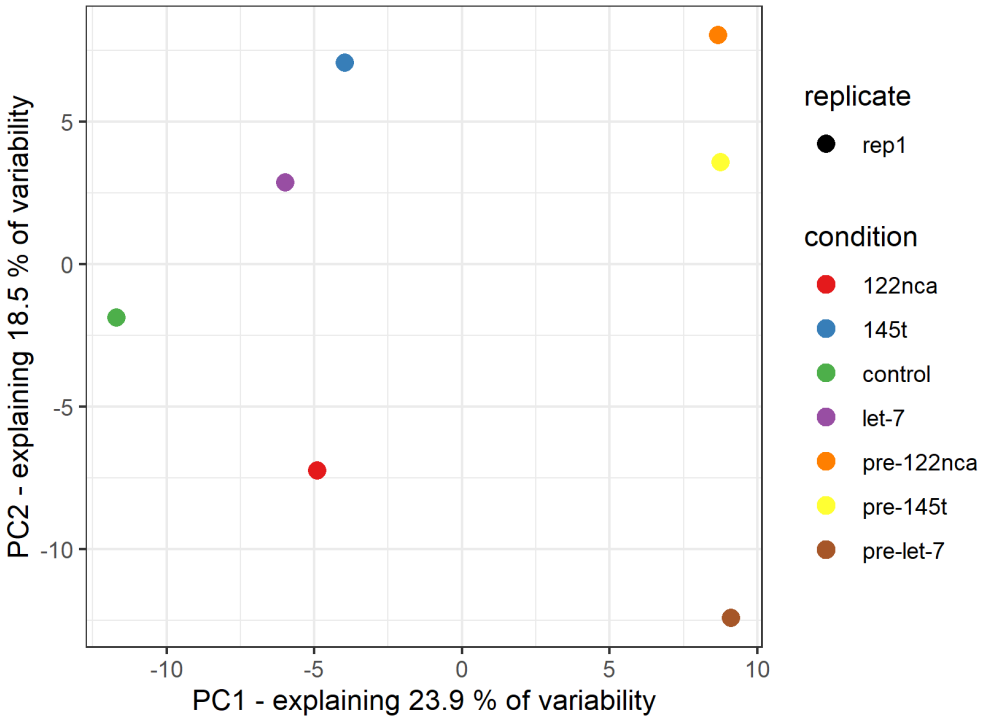


Figure 4. Protein content of each sample**.** The scatter plot summarizes the protein content in the seven samples, plotting the samples along the two principal components (PCs). Percentages of explained data variance for each PC are shown on the x and y axis.

4.3 Discussion

Possible ways of smithRNAs maturation

Nowadays, many resources are available on public repositories. These data have always been analysed to outline interactions with nuclear sRNA, while a few publications have focused on possible interaction with mitochondrial sRNAs. In the present work, using part of these resources, I identified putative smithRNA-interacting proteins, analysing animals belonging to distant branches.

The quality and quantity of data for *H. sapiens* and other mammals allowed me to compare results from eCLIP data with miRNA-Seq data from wild-type and KO cells. DROSHA appears to be a good candidate for smithRNA maturation. Several mitochondrial regions reported a significant enrichment in DROSHA eCLIP libraries (Fig. 1). Moreover, in some of those regions DROSHA KO cells reported a reduction of transcription (Fig. 2, Supplementary figure 1). The Microprocessor complex, which plays the role of cleaving pri-miRNAs in the nucleus, is formed by one monomer of DROSHA and two monomers of DGCR8 (Han et al. 2004). DGCR8 resulted to interact with only one mitochondrial RNA, located on tRNA-Val, but only in the HepG2 cell samples. In other regions, DGCR8 reported high logFC values, but those were non-significant (Fig. 1). In the Microprocessor, the two DGCR8 monomers stabilize DROSHA, which is the functional core of the complex and it can also work alone, although at lower efficiency (Nguyen et al. 2015). smithRNAs may be processed by a complete Microprocessor, but being DROSHA the main core, smithRNA precursors may appear more enriched in DROSHA eCLIP libraries than in DGCR8 ones. Another possibility is that DROSHA process smithRNA without the stabilization of DGCR8. In addition to how DROSHA process smithRNA precursors, I also wonder where they are processed. The Microprocessor is commonly located in the nucleus (Bartel 2018). However, in some tissues the alternative splicing produces a DROSHA isoform with cytosolic activity (Dai et al. 2016). On the other hand, mitochondrial RNAs can translocate to the cytosol through various mechanisms (Muneretto et al. 2024), and several mitochondrial RNAs have been found localized in the nucleus (Sriram et al. 2024). Nuclear tRFs are occasionally processed by the Microprocessor. Most tRFs are produced starting from maturated tRNAs. However, which endonucleases are involved in the processing from tRNAs to tRFs is still unknown (Kumar et al. 2016). Even smithRNAs are likely linked to the maturation of tRNAs and mitochondrial polycistronic transcripts, since the KO of RNAse Z led to the depletion of smithRNA transcription (Supplementary figure S3). smithRNAs could even be processed by other proteins, although showing clues of interaction with DROSHA. This would explain why I did not measure a depletion of smithRNAs in KO DROSHA libraries in *M. musculus.*

Finally, precursors of 145t and 122nca showed signals of interaction with proteins involved in the spliceosome complex. This complex is involved in the maturation of mirtrons and agotrons (Westhol and Lai 2011; Hansen et al. 2016). As for the Microprocessor, smithRNAs are expected to be transported to the nucleus to be

processed by the spliceosome, while evidence of cytosolic spliceosome activity remains controversial (Steitz et al. 2008).

Overall, the maturation of smithRNAs is clearly linked with the maturation of the polycistronic mitochondrial transcript and the activity of RNAse Z and P, which is the other RNAse involved in mitochondrial transcripts maturation (Fontanesi et al. 2020). However, these RNases are unlikely to be sufficient for the complete maturation of smithRNAs. Further studies on the translocation of smithRNAs are essential to determine whether DROSHA, the Microprocessor, or the spliceosome may play roles in smithRNA maturation.

Possible interactions between smithRNAs and Argonaute proteins

Argonaute proteins play a fundamental role in targeting mRNAs through the interaction of sRNAs. Besides the three main sRNA classes, Argonautes can load "non-canonical" sRNAs. tRFs can be loaded by different Argonautes: in *Homo* they were reported to interact with AGO2 (Kumar et al. 2016; Kuscu et al. 2018), but also with HIWI2 (Keam et al. 2014), which is involved in the piRNA pathway. In flies, they are more likely to be loaded by AGO2 than AGO1 (Luo et al. 2018). However, some tRFs were reported to act like piRNAs, interacting with Aub and PIWI and silencing transposons (Senti et al. 2015). Hence, I utilized RIP-Seq data on Argonaute proteins from four distant related animals, detecting mitochondrial sRNAs enriched in the IP libraries compared to the control libraries. For Argonaute proteins belonging to the PIWI family, namely Aub, PIWI and AGO3 in *D. melanogaster*, PRG-1 in *C. elegans* and MIWI in *M. musculus*, no smithRNAs were enriched in the IP-libraries. Contrastingly, in the AGO family, two proteins appear to interact with smithRNAs. Similarly to human AGO2 eCLIP libraries, in fly AGO2-IP libraries, the enriched smithRNAs were all derived from tRNAs, while for ERGO-1 only one is located on tRNA-Met. The interaction between the human AGO2 and the mitochondrial tRNA-Met has been documented in western blot analysis (Maniataki and Mourelatos 2005). Moreover, smithRNAs located on tRNA-Met have been detected across all Chordata (Pozzi and Dowling 2022). According to my results, smithRNA produced from tRNA-Met interacts with Argonuate proteins across all Bilateria. Notably, logFC values for AGO2 IP libraries over the control libraries in *Homo* and flies were the highest for the tRNA-Met regions (Fig. 1, Table 1). tRNA-Met associates with eukaryotic initiation factors 1,1A and 2 (eIF1, 1A, 2), along with the 40S rRNA, to initiate the translation on mRNAs (Maag et al. 2006; Passmore et al. 2007). eIF1A has been reported to interact with AGO2 and to facilitate the maturation of the DICER-independent miR-451 (Yi et al. 2015). Therefore, the interaction between AGO2 and tRNA-Met may be mediated by eIF1A. LC-MS/MS analysis detected a potential interaction between the initiation factor eIF4H and the precursors of let-7, 145t and 122nca, as well as the matured 122nca. Argonaute proteins can also associate with the elongation factors eEF1A (Friend et al. 2012). Consistently, eEF1A was

detected in the 122nca precursor sample by the LC-MS/MS analysis. Overall, initiation and elongation factors may mediate the interaction and loading of smithRNA precursors, which have a tRNA structure, allowing the maturation directly on the Argonaute protein, similar to miR-451.

4.4 Conclusion

In this chapter, I used both *in vitro* and *in vivo* approaches to test possible interactions between smithRNAs and a wide range of proteins. The results, although preliminar, allowed me to propose some hypothesis regarding the maturation and action of smithRNAs: precursors may be processed by DROSHA, but also by the spliceosome complex. According to the knockout analysis, only RNase Z is essential for the maturation of some smithRNAs, unlike DROSHA. Finally, among all Argonaute proteins, smithRNAs are most likely to interact with those belonging to the AGO family. The AGO-smithRNA interaction may be mediated by initiation and elongation factors, which could explain why tRNA-Met consistently interacts with AGO across different animal lineages. Overall, smithRNAs, like tRFs, mirtrons or agotrons, are likely processed by a "non-canonical" pathway, which however share some similarities with the miRNA pathway. Moreover, smithRNAs may not share a single maturation pathway; instead, depending on the location and secondary structure of their precursors, they may undergo distinct maturation steps. Further studies are needed to test these hypotheses and clarify the mode of maturation and action of smithRNAs.

4.5 Materials and Methods

eCLIP-Seq data analysis

I analysed eCLIP libraries following the pipeline made available by the ENCODE project consortium (ENCODE Project Consortium 2012; Golden et al. 2017; Whipple et al. 2020; Brannan et al. 2021; Chu et al. 2021; Supplementary table S6). Briefly, I extracted the unique molecular identifiers (UMIs) from reads with UMI-tools extract (Smith et al. 2017). I trimmed adapters and low-quality bases using Cutadapt (Martin 2011), filtering reads shorter than 18 bases, while the maximum error rate and quality cut off set at respectively at 0.1 and 10. I run the trimming step twice to remove duplicate adapters due to double ligation events. I mapped the reads against the human and mouse reference genomes (respectively hg19 and mm39) with STAR (Dobin et al. 2013), allowing up to one mismatch and aligning the reads end-to-end. Based on the mapping output, I deduplicated the reads with UMI-tools dedup (Smith et al. 2017). The cluster-finding algorithm CLIPper (Lovci et al. 2013) was used to detect genomic regions that transcribe for RNAs that interact with the target protein. eCLIP-seq libraries were normalized using the related control libraries with the perl scritp "overlap_peakfi_with_bam.pl" (https://github.com/YeoLab/merge_peaks). I determined the reproducibility of the peaks through an Irreproducible Discovery Rate (IDR) analysis, setting the IDR threshold to 0.01 (Van Nostrand et al. 2016).

Analysis on knocked-out libraries

I analysed libraries from three different experiments that sequenced the small RNA pool of samples (Supplementary table S7) where DROSHA and AGO1+2 were knocked out in *H. sapiens* (Johnson et al. 2023), where DROSHA and AGO1+2 were knocked out in *M. musculus* (Müller et al. 2022), and where RNAse Z was knocked out in *M. musculus* (Siira et al. 2018). Each library was trimmed from adapters and low-quality bases using Cutadapt, filtering reads shorter than 15 bases, while the maximum error rate and quality cut off set at respectively at 0.1 and 10. Reads were aligned with STAR, allowing up to one mismatch and aligning the reads end-to-end. Multimapping reads that were mapping on bothmitochondrial and nuclear genome were discarded from the analysis. The coverage of each library on each mt-genome position was calculated using bedtools genomecov (Quinlan and Hall 2010).

RIP-Seq data analysis

To include other Argonaute paralogs present in different metazoan species, I selected RNA Immunoprecipitation (RIP) libraries available on the Sequence Read Archive (SRA) from different Bioprojects

(Supplementary table S8). I selected libraries based on the following criteria: i) libraries constructed with a small RNA-dedicated library preparation; ii) libraries that could be compared with relative control libraries; iii) possibly, Bioprojects with at least two replicates per condition, although this was not possible for Aubergine, PIWI and AGO3 in *Drosophila melanogaster* and WAGO-1 in *Caenorhabditis elegans* (Supplementary table S8).

Each library was trimmed from adapters and low-quality bases using Cutadapt, filtering reads shorter than 15 bases, while the maximum error rate and quality cut off set at respectively at 0.1 and 10. I extracted the UMIs from reads using UMI-tools (Smith et al. 2017), when included in the library construction protocol. Reads were aligned with STAR, allowing up to one mismatch and aligning the reads end-to-end. Reads were de-duplicated, when required, with UMI-tools. Multimapping reads that were mapping on both mitochondrial and nuclear genome were discarded from the analysis. I annotated putative smithRNAs in my IP libraries following the pipeline of Pozzi and colleagues (2019). Briefly, I clustered all the reads mapping on the mitochondrial genome using cd-hit-est (Li and Godzik 2006), clustering reads with an identity score higher than 90% and a length difference cutoff of 0.5 (i.e., the shortest sequence needs to be at least half long than the length of the representative sequence). I considered the representative sequences of clusters with a coverage higher than 200 as putative smithRNAs. To test the interaction between the target Argonaute protein and the putative smithRNAs, I performed a differential expression analysis comparing RIP and control libraries, assessing whether smithRNAs were enriched in the RIP library. To perform this, I used sRNAbench (Aparicio-Puerta et al. 2019) to map reads of both libraries on the putative smithRNAs as well as the following ncRNA databases: miRbase (Kozomara et al. 2019) for miRNAs, piRBase (Wang et al. 2022) for piRNAs, GtRNAdb (Chan and Lowe 2016) for tRNAs, and the ncRNA genomic annotation made available by Ensembl (Harrison et al. 2024). Using sRNAde (Aparicio-Puerta et al. 2019), I assigned each read to a ncRNA and calculated the read count for each ncRNA, avoiding multiple assignments. I calculated the differentially expressed (DE) ncRNAs in the IP libraries compared to the control libraries using edgeR (Robinson et al. 2010; R Core Team 2021). The raw reads count table was filtered to remove low counts across libraries and normalized. Then, I calculated DE small RNAs applying a generalized linear model using a quasi-likelihood F-test.

Pull-down protocol

I identified proteins that interact with smithRNAs using a pull-down assay. Two target smithRNAs, 145t and 122nca, were selected from those characterized by Pozzi and colleagues (2018) in *R. philippinarum*. These smithRNAs were synthesized in both their mature and precursor forms (Supplementary table S3). The synthesized RNAs were biotinylated at their 3' terminus and a monophosphate group was added at their 5'

terminus. As a control, I analysed the interaction of the *R. philippinarum* miRNA let-7. The mature and precursor sequences of *R. philippinarum* let-7 were predicted using miRDeep2 (Friedländer et al., 2012), providing *R. philippianrum* small-RNA libraries available on SRA (SRR3662624-9). As for smithRNAs, the mature and precursor forms of Let-7 were synthesized, biotinylated and phosphorylated. Overall, I analysed the protein-RNA interactions of six different RNA molecules (i.e., 145t, 122nca, let7, pre_145t, pre_122nca, pre_let7).

I performed a pull-down assay for each RNA molecule following the instructions of the Pierce™ Magnetic RNA-Protein Pull-Down Kit (Thermo Scientific™, USA). Briefly, tissue lysate was obtained from the foot of *R. philippinarum* specimens using the T-PER Tissue Protein Extraction Reagent (Thermo Scientific™, USA). The biotinylated RNA was bound to streptavidin magnetic beads, and the beads were then exposed to the tissue lysate. After three washes, the interacting proteins were eluted. In addition to the six RNA samples, a control sample was included to exclude proteins that might interact nonspecifically with the magnetic beads during data analysis. For the control, no RNA was bound to the magnetic beads during the pull-down assay.

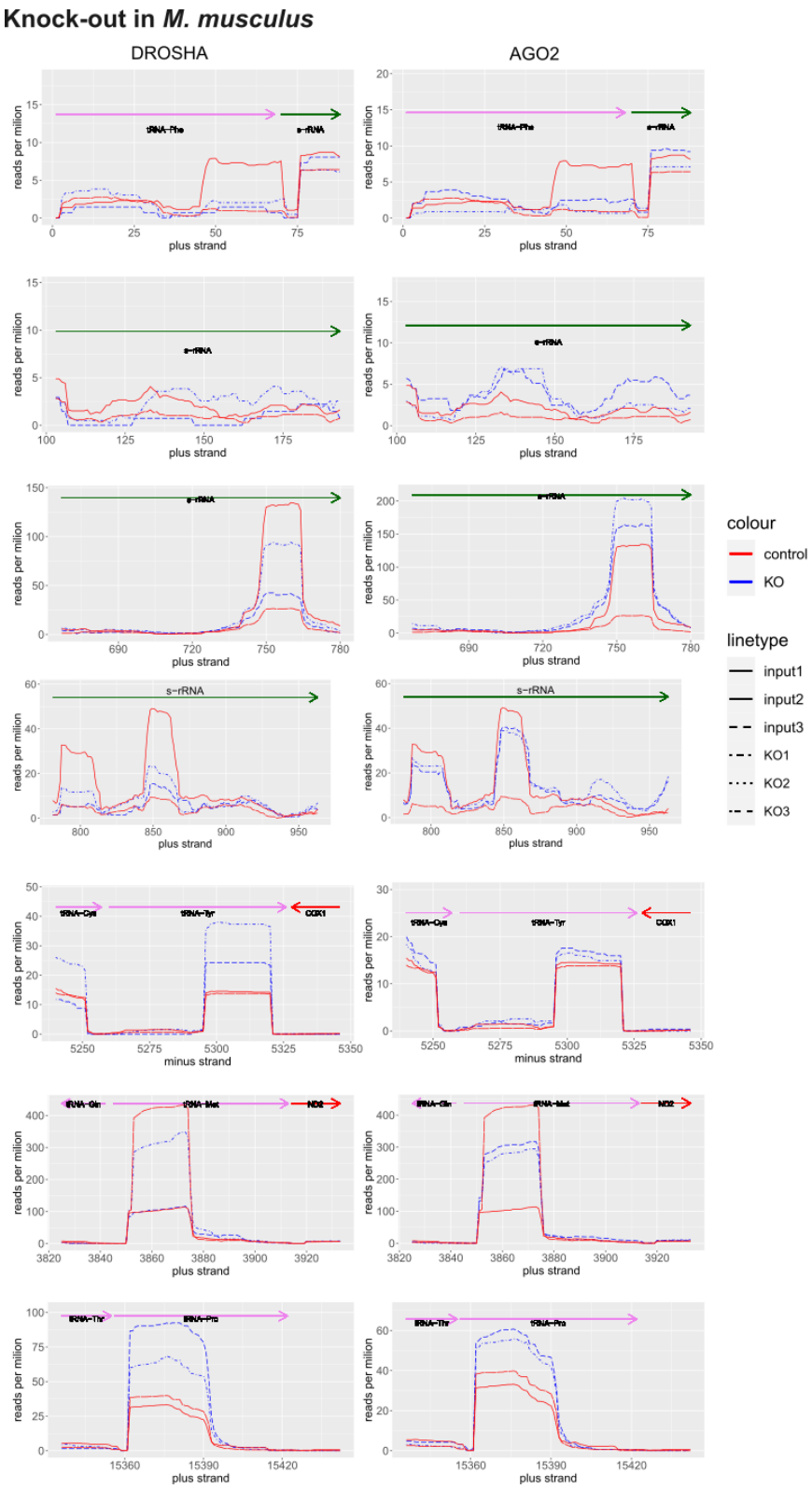
Samples were analysed at the VIB Proteomics Core (Ghent) using the LC-MS/MS technique. 10 µL of each sample was injected on an Ultimate 3000 RSLCnano system in-line connected to a Q Exactive HF mass spectrometer. The mass spectrometer was set in data-dependent mode, automatically switching between mass spectrometry (MS) and tandem MS acquisition for the 16 most abundant ion peaks for each MS spectrum. LC-MS/MS runs of all samples were searched together using the DiaNN algorithm v. 1.8.1 (Demichev et al. 2020) with mainly default search settings, including a false discovery rate set at 1% on precursor and protein level. Spectra were searched against the *Ruditapes philippinarum* refseq protein sequences in the NCBI RefSeq database (database release version of April 2024), containing 57,637 sequences, supplemented with the universal protein contaminant database (database release version of 2023_02), containing 381 sequences (Frankenfield et al. 2022).

Proteins in the *R. philippinarum* proteome were annotated by aligning them against the reviewed Swiss-Prot protein database (Bateman et al. 2021) using DIAMOND blastp (Camacho et al. 2009; Buchfink et al. 2015), retaining the best hit. Gene Ontology (GO) terms enrichment analysis was performed using A.GO.TOOL (Schölz et al. 2015). For each condition, I selected the proteins reporting a label-free quantification (LFQ) that was at least twice as high as the LFQ of that protein measured in the control as the foreground for the GO terms enrichment analysis. As background, I included the Uniprot list of all annotated proteins in the *R. philippinarum* proteome. A.GO.TOOL was run with the protein abundance correction method.

4.6 Supplementary Materials

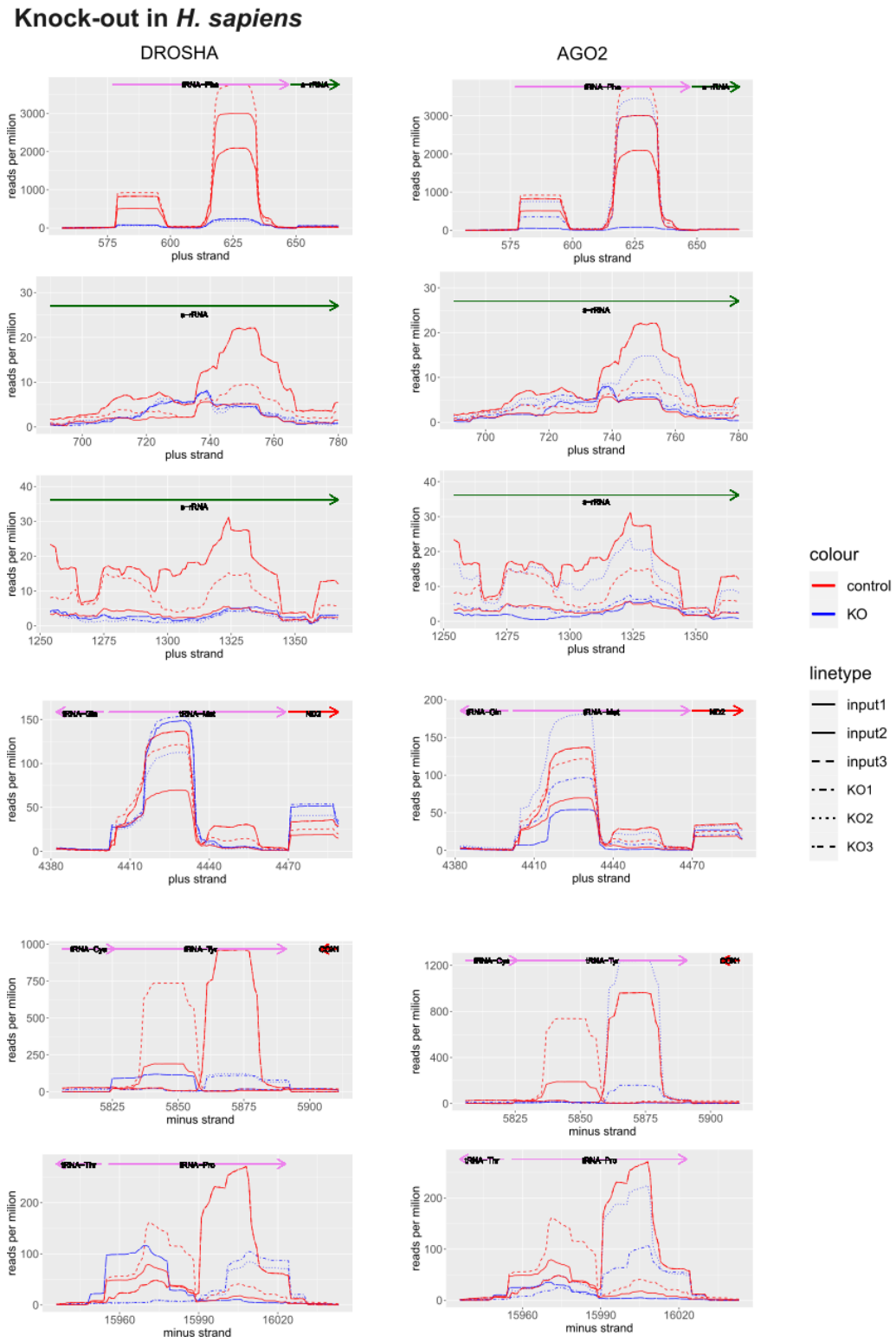
Position	HepG2 coordinates	K562 coordinates
l-rRNA (amid) +	3023-3045	3023-3045
ND2 (amid) +	X	5248-5267
tRNA-Val (5') +	1601-1625	1601-1625
tRNA-Val (3') +	1650-1673	1654-1673
tRNA-Asn (amid) -	X	5671-5692
tRNA-Asn (amid) -	X	5700-5722
tRNA-Asn (5') -	X	5706-5729
tRNA-Leu1 (5')+	X	3229-3259
tRNA-Pro (3') -	X	15952-15976
s-rRNA (amid) +	1195-1216	1195-1216
tRNA-His (3') +	X	12188-12208
tRNA-Ala (3') -	X	5583-5604
tRNA-Asn (3')	X	5653-5677
l-rRNA (amid) +	2800-2819	2800-2819
ND1 (amid) +	3424-3445	3424-3445
tRNA-Gly (3') +	10037-10061	10037-10061
tRNA-Leu1 (3')+	X	3282-3307
s-rRNA (amid) +	X	1415-1439
tRNA-Gln (5') -	4374-4400	4374-4400
tRNA-Glu (3') -	14670-14694	X
tRNA-Tyr (5') -	5870-5891	X
d-loop -	245-279	X

Supplementary table S1. The table reports the coordinates of smithRNAs detected by SmithHunter in samples from HepG2 and K562 cell types. “X” means that the smithRNA was not detected in that position for that cell type.

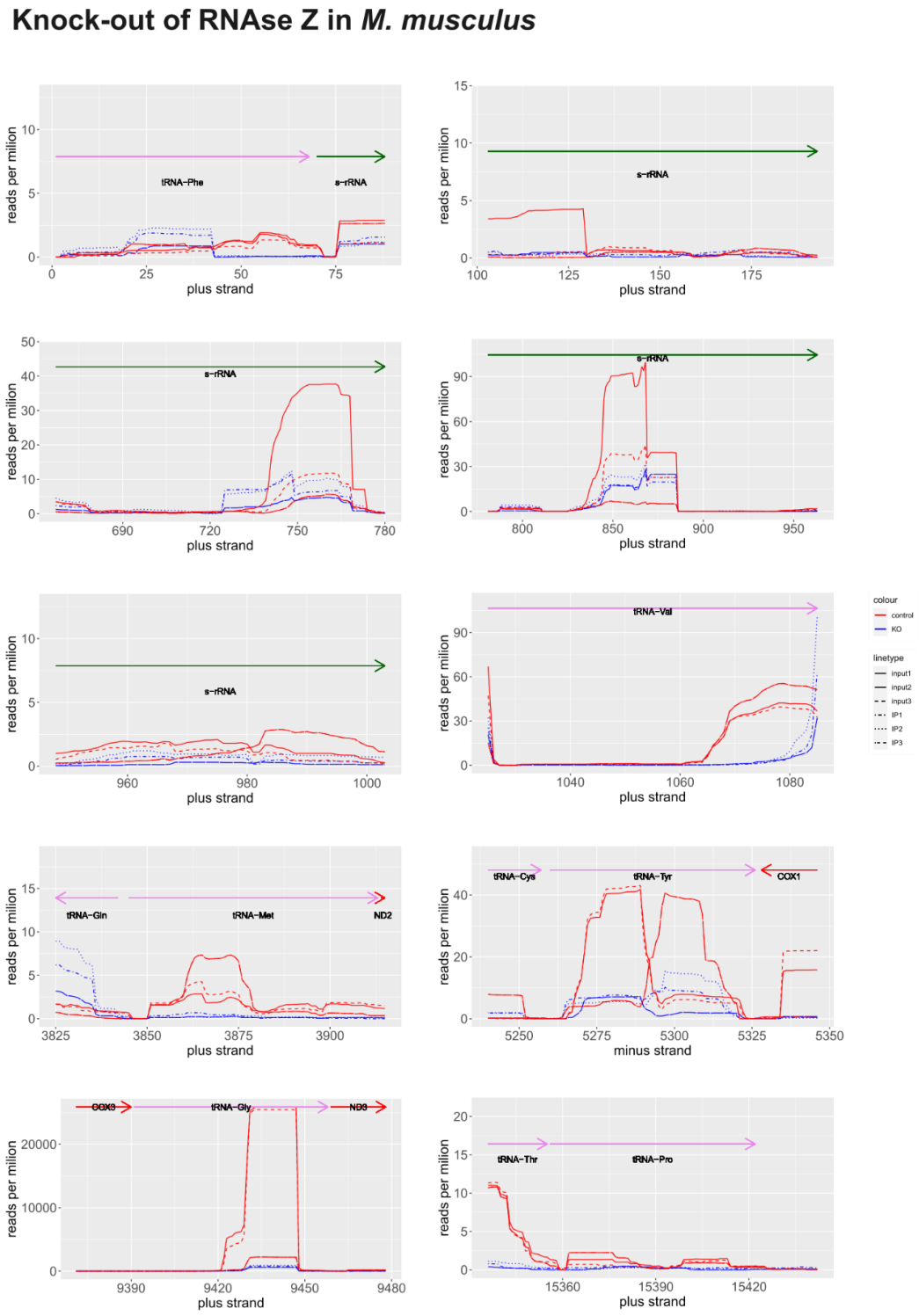


Supplementary figure S1. smithRNAs RPM in DROSHA and AGO2 KO experiments. Lines represent the coverage, expressed as reads per million (i.e., the number of reads mapping at a position multiplied by 1

million, divided by the total number of mapping reads). Colours distinguish libraries from KO cells (blue) and wild-type cells (red). Line types distinguish the replicates.



Supplementary figure S2. smithRNAs RPM in DROSHA and AGO2 KO experiments. Lines represent the coverage, expressed as reads per million (i.e., the number of reads mapping at a position multiplied by 1 million, divided by the total number of mapping reads). Colours distinguish libraries from KO cells (blue) and wild-type cells (red). Line types distinguish the replicates.



Supplementary igure S3. smithRNAs RPM in RNAse Z KO experiments. Lines represent the coverage, expressed as reads per million (i.e., the number of reads mapping at a position multiplied by 1 million, divided by the total number of mapping reads). Colours distinguish libraries from KO cells (blue) and wild-type cells (red). Line types distinguish the replicates.

Oligo Name	Sequence
smithRNA122_ncA	[Phos]GAGAAAAGCGGGGCAUGGCUAGACUUC[Btn]
pre_smithRNA_122ncA	[Phos]GAGAAAAGCGGGGCAUGGCUAGACUUCUAAUCUUUGCUAUAAGCAGUUCAACUCUGUUUUUUUCUCUA[Btn]
smithRNA_145t	[Phos]GUUGAAGUGUCAGAUAUAUGUGGUAAAUU[Btn]
pre_smithRNA_145t	[Phos]UUUGUUGAAGUGUCAGAUAUAUGUGGUAAAUUUAGAAUUUAUUUAUGGGGUUAUUCCUCUUCAAUAGUG[Btn]
let-7	[Phos]UGAGGUAGUAGGUUGUAUAGU[Btn]
pre_let-7	[Phos]UGAGGUAGUAGGUUGUAUAGUUAGAACUACACCAUACAGGAGAACUAUUCAACCUUCUAGCUUUCC[Btn]

Supplementary table S3: oligoRNA sequences tested with the pull-down protocol. Each sequences reports a phosphate group at the 5' end, and it was biotinylated at the 3' end.

term	description	p_value
122nca smithRNA		
KW-0694	RNA-binding	1.96E-06
GO:0003723	RNA binding	5.25E-06
GO:0005687	U4 snRNP	1.10E-05
KW-0507	mRNA processing	2.58E-05
145t smithRNA		
PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	8.46E-08
IPR000504	RNA recognition motif domain	2.13E-07
GO:0016018	cyclosporin A binding	2.81E-07
IPR035979	RNA-binding domain superfamily	3.84E-07
IPR012677	Nucleotide-binding alpha-beta plait domain superfamily	4.62E-07
IPR020892	Cyclophilin-type peptidyl-prolyl cis-trans isomerase, conserved site	1.07E-06
GO:0006457	protein folding	1.84E-06

term	description	p_value
pre 145t smithRNA		
GO:0005688	U6 snRNP	1.10E-08
GO:0005687	U4 snRNP	1.78E-08
GO:0070717	poly-purine tract binding	2.50E-08
IPR001163	Sm domain, eukaryotic/archaea-type	2.50E-08
IPR010920	LSM domain superfamily	3.97E-08
GO:0071005	U2-type precatalytic spliceosome	5.56E-08
IPR047575	Sm domain	9.04E-08
GO:0071013	catalytic step 2 spliceosome	9.24E-08
map03040	Spliceosome	1.04E-07
GO:0046540	U4/U6 x U5 tri-snRNP complex	1.09E-07
GO:0005684	U2-type spliceosomal complex	1.47E-07
GO:0008266	poly(U) RNA binding	1.56E-07

PF00160	Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD	2.00E-06	GO:0005685	U1 snRNP	1.56E-07
IPR002130	Cyclophilin-type peptidyl-prolyl cis-trans isomerase domain	2.00E-06	GO:0071011	precatalytic spliceosome	1.57E-07
IPR029000	Cyclophilin-like domain superfamily	2.00E-06	IPR000504	RNA recognition motif domain	1.61E-07
KW-0697	Rotamase	9.55E-06	KW-0507	mRNA processing	1.62E-07
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	1.22E-05	GO:0097525	spliceosomal snRNP complex	1.76E-07
KW-0413	Isomerase	2.67E-05	GO:0003729	mRNA binding	2.00E-07
let7 miRNA			GO:0008143	poly(A) binding	2.07E-07
GO:0016018	cyclosporin A binding	1.32E-07	GO:0034719	SMN-Sm protein complex	2.07E-07
IPR020892	Cyclophilin-type peptidyl-prolyl cis-trans isomerase, conserved site	5.03E-07	GO:0097526	spliceosomal tri-snRNP complex	2.19E-07
IPR002130	Cyclophilin-type peptidyl-prolyl cis-trans isomerase domain	9.42E-07	GO:0003730	mRNA 3'-UTR binding	2.22E-07
IPR029000	Cyclophilin-like domain superfamily	9.42E-07	GO:1990904	ribonucleoprotein complex	2.56E-07
PF00160	Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD	9.42E-07	GO:0034715	pICln-Sm protein complex	2.68E-07
KW-0697	Rotamase	4.51E-06	GO:0003727	single-stranded RNA binding	2.77E-07
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	5.79E-06	GO:0005681	spliceosomal complex	2.82E-07
map03040	Spliceosome	3.85E-05	GO:0000387	spliceosomal snRNP assembly	3.01E-07
pre 122nca smithRNA			KW-0694	RNA-binding	3.05E-07
GO:0005687	U4 snRNP	2.31E-08	IPR035979	RNA-binding domain superfamily	3.14E-07
IPR001163	Sm domain, eukaryotic/archaea-type	2.56E-08	IPR012677	Nucleotide-binding alpha-beta plait domain superfamily	3.89E-07
GO:0016018	cyclosporin A binding	3.45E-08	GO:0120115	Lsm2-8 complex	4.44E-07
IPR010920	LSM domain superfamily	3.99E-08	GO:0016070	RNA metabolic process	5.10E-07
BTA-72163	mRNA Splicing - Major Pathway	3.99E-08	KW-0747	Spliceosome	5.20E-07
GO:0071005	U2-type precatalytic spliceosome	4.36E-08	KW-0687	Ribonucleoprotein	5.23E-07

GO:0071007	U2-type catalytic step 2 spliceosome	4.93E-08	GO:0000398	mRNA splicing, via spliceosome	5.56E-07
GO:0071013	catalytic step 2 spliceosome	8.49E-08	GO:0003723	RNA binding	5.86E-07
GO:0046540	U4/U6 x U5 tri-snRNP complex	8.85E-08	GO:0090304	nucleic acid metabolic process	6.24E-07
IPR047575	Sm domain	8.85E-08	GO:0008380	RNA splicing	6.45E-07
map03040	Spliceosome	9.52E-08	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	6.60E-07
IPR000504	RNA recognition motif domain	1.02E-07	GO:0034709	methylosome	6.93E-07
GO:0071011	precatalytic spliceosome	1.38E-07	IPR002343	Paraneoplastic encephalomyelitis antigen	6.93E-07
GO:0097525	spliceosomal snRNP complex	1.55E-07	GO:0010467	gene expression	7.02E-07
GO:0005684	U2-type spliceosomal complex	1.59E-07	KW-0508	mRNA splicing	7.05E-07
GO:0005685	U1 snRNP	1.73E-07	GO:0016071	mRNA metabolic process	7.32E-07
IPR035979	RNA-binding domain superfamily	2.02E-07	PF01423	LSM domain	8.49E-07
GO:0097526	spliceosomal tri-snRNP complex	2.09E-07	GO:0006397	mRNA processing	9.23E-07
GO:0034719	SMN-Sm protein complex	2.26E-07	GO:0006396	RNA processing	9.90E-07
GO:0034715	pICln-Sm protein complex	2.32E-07	map03018	RNA degradation	1.19E-06
GO:0005681	spliceosomal complex	2.38E-07	GO:0006402	mRNA catabolic process	1.27E-06
IPR012677	Nucleotide-binding alpha-beta plait domain superfamily	2.51E-07	GO:0003676	nucleic acid binding	1.44E-06
GO:0000387	spliceosomal snRNP assembly	2.85E-07	BTA-72163	mRNA Splicing - Major Pathway	1.51E-06
KW-0687	Ribonucleoprotein	3.81E-07	GO:0005683	U7 snRNP	6.40E-06
GO:0006401	RNA catabolic process	3.85E-07	GO:0022618	protein-RNA complex assembly	7.61E-06
GO:0120115	Lsm2-8 complex	3.85E-07	GO:0009059	macromolecule biosynthetic process	1.05E-05
PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	4.34E-07	GO:0006139	nucleobase-containing compound metabolic process	2.22E-05
GO:0000398	mRNA splicing, via spliceosome	4.42E-07	GO:0010629	negative regulation of gene expression	2.39E-05
KW-0747	Spliceosome	4.61E-07	GO:0005689	U12-type spliceosomal complex	2.47E-05

GO:0008380	RNA splicing	4.90E-07	GO:0036002	pre-mRNA binding	3.39E-05
GO:0016071	mRNA metabolic process	4.97E-07	GO:0071007	U2-type catalytic step 2 spliceosome	4.54E-05
KW-0508	mRNA splicing	5.58E-07	GO:0006725	cellular aromatic compound metabolic process	4.71E-05
GO:0006396	RNA processing	5.62E-07	GO:1901363	heterocyclic compound binding	4.76E-05
GO:0034709	methylosome	6.02E-07	GO:0022613	ribonucleoprotein complex biogenesis	4.87E-05
GO:0006397	mRNA processing	6.51E-07	GO:0140513	nuclear protein-containing complex	0.0001168
GO:0005689	U12-type spliceosomal complex	7.13E-07	GO:0035770	ribonucleoprotein granule	0.0001646
GO:1990904	ribonucleoprotein complex	8.11E-07	GO:0005682	U5 snRNP	0.0001703
GO:0003676	nucleic acid binding	8.56E-07	GO:0005634	nucleus	0.0002439
KW-0507	mRNA processing	8.66E-07	GO:0005686	U2 snRNP	0.0003945
PF01423	LSM domain	8.70E-07	**pre let7 miRNA**		
GO:0005688	U6 snRNP	8.99E-07	GO:0005201	extracellular matrix structural constituent	7.32E-08
KW-0694	RNA-binding	9.12E-07	KW-0084	Basement membrane	1.56E-07
GO:0090304	nucleic acid metabolic process	9.76E-07	map04512	ECM-receptor interaction	1.57E-07
GO:0006402	mRNA catabolic process	1.04E-06	GO:0031012	extracellular matrix	1.82E-07
GO:0036002	pre-mRNA binding	1.05E-06	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	3.82E-07
GO:0010467	gene expression	1.09E-06	GO:0062023	collagen-containing extracellular matrix	3.84E-07
GO:0016070	RNA metabolic process	1.11E-06	KW-0272	Extracellular matrix	5.29E-07
GOCC:0005681	Spliceosomal complex	1.11E-06	GO:0005604	basement membrane	6.91E-07
GO:0003729	mRNA binding	1.18E-06	IPR000504	RNA recognition motif domain	1.04E-06
GO:0003723	RNA binding	1.28E-06	KW-0694	RNA-binding	1.15E-06
GO:0140513	nuclear protein-containing complex	1.35E-06	GO:0005198	structural molecule activity	1.31E-06
GO:1903312	negative regulation of mRNA metabolic process	1.37E-06	IPR035979	RNA-binding domain superfamily	1.96E-06
GO:0048255	mRNA stabilization	2.11E-06	IPR012677	Nucleotide-binding alpha-beta plait domain superfamily	2.39E-06
KW-0697	Rotamase	2.48E-06	KW-0176	Collagen	2.68E-05
GO:0003730	mRNA 3'-UTR binding	2.87E-06	GO:0005581	collagen trimer	3.05E-05

GO:0003727	single-stranded RNA binding	2.87E-06	GO:1902555	endoribonuclease complex	5.21E-05
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	3.35E-06	KW-0654	Proteoglycan	6.34E-05
GO:0000956	nuclear-transcribed mRNA catabolic process	3.77E-06	KW-0964	Secreted	8.41E-05
GO:0005682	U5 snRNP	4.21E-06			
GO:1903311	regulation of mRNA metabolic process	4.51E-06			
GO:0005683	U7 snRNP	5.76E-06			
IPR020892	Cyclophilin-type peptidyl-prolyl cis-trans isomerase, conserved site	6.79E-06			
GOCC:0005684	U2-type spliceosomal complex	7.50E-06			
GO:0005686	U2 snRNP	1.26E-05			
map03018	RNA degradation	1.34E-05			
GO:0036464	cytoplasmic ribonucleoprotein granule	1.38E-05			
GO:0070717	poly-purine tract binding	1.81E-05			
GO:0000413	protein peptidyl-prolyl isomerization	3.43E-05			
GO:0022613	ribonucleoprotein complex biogenesis	3.50E-05			
GO:0043488	regulation of mRNA stability	5.21E-05			
KW-0539	Nucleus	6.40E-05			
GO:0010629	negative regulation of gene expression	6.95E-05			
GO:0071004	U2-type prespliceosome	0.0001008			
GO:0000932	P-body	0.0001041			
KW-0963	Cytoplasm	0.0001101			
KW-1133	Transmembrane helix	0.0001868			
GO:0005634	nucleus	0.0004693			

Supplementary table S4. Enriched GO terms for the 7 samples. For each GO term it is reported its description the associated p-value.

Uniprot accession	let-7 / control LFQ	145t / control LFQ	122nca / control LFQ	pre_let-7 / control LFQ	pre_145t / control LFQ	pre_122nc / control LFQ
sp|A0A0B4KGY6|NOVA_DROME	1.046	1.112	0.96	2.647	1.338	1.506

sp|A0A0R4IBK5|R213A_DANRE	2.115	1.839	1.496	1.987	1.913	2.23
sp|A2AR02|PPIG_MOUSE	2.514	2.2	1.91	2.52	3.671	3.256
sp|A2AVA0|SVEP1_MOUSE	1.551	1.862	2.446	2.761	1.879	2.575
sp|A2AX52|CO6A4_MOUSE	1.256	1.185	1.692	1.205	3.738	2.043
sp|A4FUI2|RUXE_BOVIN	1.77	1.752	2.075	1.996	2.288	2.115
sp|A8C754|THADA_CHICK	1.224	1.338	1.25	1.443	2.053	1.475
sp|A8TX70|CO6A5_HUMAN	1.071	1.019	1.018	3.275	1.032	1.064
sp|E1BH29|ALKB5_BOVIN	0	0	0	2.09	0.991	1.466
sp|F1NV61|CASP7_CHICK	1.486	1.185	2.029	1.665	1.196	1.415
sp|F1NW29|TYDP2_CHICK	1.064	1.522	1.908	1.784	2.014	1.831
sp|F1QB54|PABPA_DANRE	0.88	1.512	1.799	1.3	2.149	1.775
sp|F1QMY1|DYT2B_DANRE	0.393	0	0	2.354	0	0
sp|F8VPK0|SKI3_MOUSE	1.863	1.636	0	1.177	2.162	2.369
sp|J3S836|VCO3_CROAD	2.735	2.413	2.544	1.348	1.35	2.86
sp|O00139|KIF2A_HUMAN	0.824	0.535	2.601	0	0.738	0.835
sp|O00338|ST1C2_HUMAN	1.141	1.201	1.033	1.454	1.158	2.188
sp|O00423|EMAL1_HUMAN	1.424	1.378	2.066	2.207	2.385	1.248
sp|O00462|MANBA_HUMAN	1.613	2.009	1.608	1.73	1.611	1.808
sp|O01761|UNC89_CAEEL	1.836	2.268	1.835	1.762	1.484	1.571
sp|O13046|WDHD1_XENLA	1.283	1.33	1.075	2.176	0.959	1.344
sp|O43390|HNRPR_HUMAN	1.104	1.25	1.512	1.618	2.383	3.914
sp|O15160|RPAC1_HUMAN	1.474	1.342	2.411	2.574	0	2.38
sp|O44437|SMD3_DROME	1.341	1.353	1.782	1.748	2.119	2.281
sp|O54701|NCKX2_RAT	1.704	2.453	0	2.392	1.582	1.902
sp|O57321|EAA1_AMBTI	1.243	1.388	1.179	1.701	2.007	1.162
sp|O60506|HNRPQ_HUMAN	1.104	1.25	1.512	1.618	2.383	3.914
sp|O62703|CTBL1_BOVIN	4.749	1.702	0	0.458	1.645	1.296
sp|O75069|TMCC2_HUMAN	2.297	2.1	2.614	2.584	1.483	0
sp|O96064|MYSP_MYTGA	1.279	1.152	1.144	2.552	1.096	1.189
sp|O97860|PPA5_RABIT	0	1.438	3.323	1.8	1.982	1.794
sp|P08183|MDR1_HUMAN	1.613	2.081	1.306	1.117	1.021	1.754
sp|P0DJG4|SMA2L_HUMAN	1.204	1.071	1.083	0	3.016	1.463
sp|P0DW91|ZTRF1_BOVIN	2.401	0	1.708	1.556	3.149	2.552
sp|P10881|LA_BOVIN	1.604	2.662	4.475	2.091	3.811	3.602
sp|P11833|TBB_PARLI	1.663	1.331	1.867	2.255	1.476	1.878
sp|P12606|ITB1A_XENLA	3.882	4.187	0	2.272	2.703	2.959
sp|P13612|ITA4_HUMAN	1.529	1.459	1.478	0.955	2.385	1.121
sp|P13667|PDIA4_HUMAN	1.336	2.056	1.219	1.284	1.375	1.767
sp|P13944|COCA1_CHICK	1.168	1.178	1.655	1.267	3.415	1.901
sp|P18172|DHGL_DROPS	1.119	1.202	1.487	1.644	2.044	1.305
sp|P21522|ROA1_SCHAM	1.857	1.742	1.075	1.544	2.091	2.102
sp|P24367|PPIB_CHICK	2.514	2.2	1.91	2.52	3.671	3.256
sp|P25228|RAB3_DROME	1.156	1.673	1.336	2.066	0.997	2.37
sp|P25782|CYSP2_HOMAM	0	0.873	1.123	1.5	1.934	2.095

sp|P26043|RADI_MOUSE	1.012	1.263	1.262	1.861	3.217	2.03
sp|P26368|U2AF2_HUMAN	1.121	1.128	0.984	1.195	1.308	3.231
sp|P26378|ELAV4_HUMAN	1.463	1.881	1.107	1.689	4.645	6.672
sp|P27393|CO4A2_ASCSU	1.24	1.137	1.05	3.19	1.051	1.074
sp|P27473|IFI44_PANTR	2.295	3.658	2.326	1.528	1.21	1.423
sp|P28740|KIF2A_MOUSE	0.824	0.535	2.601	0	0.738	0.835
sp|P29400|CO4A5_HUMAN	1.275	1.693	1.135	2.287	1.627	1.956
sp|P31943|HNRH1_HUMAN	1.219	1.656	1.921	0.902	3.236	1.438
sp|P33727|ARSB_FELCA	0.799	1.306	1.049	1.602	1.79	2.212
sp|P34611|NCL1_CAEEL	1.752	1.853	1.142	1.693	3.733	3.783
sp|P41262|GLB3_PHAPT	0.978	0.924	0.71	1.686	2.062	1.169
sp|P41366|VMO1_CHICK	1.234	1.292	1.874	2.067	1.266	2.049
sp|P41824|YBOXH_APLCA	1.241	1.383	4.643	11.751	3.911	5.422
sp|P41827|HSP74_ANOAL	1.378	2.035	1.908	2.393	2.05	1.431
sp|P48810|RB87F_DROME	2.509	1.899	1.83	2.055	2.869	3.035
sp|P49337|WNT4_CHICK	2.399	2.174	0	0	0	1.333
sp|P51907|EAA3_RAT	1.63	1.388	0.989	1.701	2.007	1.061
sp|P57789|KCNKA_HUMAN	1.658	0.57	2.137	2.102	1.507	1.085
sp|P61007|RAB8A_CANLF	1.156	1.673	1.336	2.066	1.067	2.37
sp|P61157|ARP3_BOVIN	1.662	1.728	2.709	1.202	1.286	1.085
sp|P62312|LSM6_HUMAN	1.029	1.346	1.368	2.012	5.746	6.032
sp|P62877|RBX1_HUMAN	3.936	1.083	3.393	4.331	4.343	0
sp|P63099|CANB1_BOVIN	1.646	1.973	1.698	1.438	1.146	2.226
sp|P74897|YQA3_THEAQ	1.189	1.204	1.225	2.424	1.511	1.635
sp|P79251|VATG1_BOVIN	0.837	0.963	2.242	2.082	0.918	1.041
sp|P90820|HPX2_CAEEL	0.481	1.413	1.032	1.281	2.422	1.549
sp|P97821|CATC_MOUSE	3.703	1.394	1.113	2.3	1.983	1.498
sp|P98160|PGBM_HUMAN	2.786	1.742	1.063	3.853	1.283	1.746
sp|Q00438|PTBP1_RAT	1.021	1.017	0.998	1.843	1.751	4.604
sp|Q00657|CSPG4_RAT	1.199	1.187	1.107	1.25	0.768	2.329
sp|Q01085|TIAR_HUMAN	0.853	1.002	0.946	1.013	1.765	2.727
sp|Q01459|DIAC_HUMAN	0.965	0.877	0.973	0.905	1.138	2.07
sp|Q02645|HTS_DROME	1.655	0.964	1.251	1.456	1.314	2.236
sp|Q02926|RB97D_DROME	2.509	1.899	1.83	2.055	2.869	3.035
sp|Q05793|PGBM_MOUSE	1.057	1.043	1.134	3.111	0.974	1.039
sp|Q06561|UNC52_CAEEL	1.057	0.908	1.244	3.111	0.878	0.896
sp|Q0EEE2|PTHD3_MOUSE	2.695	1.009	0.861	1.285	3.828	2.993
sp|Q0V9R3|DI3L2_XENTR	1.495	1.454	1.487	1.435	2.036	2.235
sp|Q12926|ELAV2_HUMAN	1.664	1.361	1.483	0.914	2.099	1.033
sp|Q13247|SRSF6_HUMAN	1.363	1.185	1.785	2.216	2.098	1.548
sp|Q15154|PCM1_HUMAN	0	2.002	0.978	0	1.926	1.631
sp|Q1JPH6|IF4H_BOVIN	1.68	1.803	2.694	2.348	2.621	2.902
sp|Q24498|RYR_DROME	1.76	1.479	1.436	2.845	2.87	2.232
sp|Q26486|FKBP4_SPOFR	1.073	1.385	1.719	3.919	3.761	5.148

sp|Q27874|PAT3_CAEEL	3.882	4.187	0	2.272	2.703	2.959
sp|Q28247|CO4A5_CANLF	1.275	1.203	1.135	2.972	0.993	1.155
sp|Q28F51|TADBP_XENTR	1.747	2.258	1.039	1.106	1.488	1.268
sp|Q28GD4|ELAV2_XENTR	0.932	1.086	0.898	1.065	1.698	14.132
sp|Q2HJ18|VP33B_BOVIN	1.122	1.142	1.821	2.076	0	1.51
sp|Q3MHM5|TBB4B_BOVIN	1.663	1.487	1.867	2.255	1.476	1.878
sp|Q3MHR5|SRSF2_BOVIN	4.981	3.478	3.992	4.672	4.53	5.51
sp|Q3SYV5|TSN33_BOVIN	1.853	2.016	1.577	1.906	2.018	1.628
sp|Q3SZF8|SMD2_BOVIN	1.801	1.44	1.845	1.646	2.182	2.219
sp|Q3T0Z8|RUXF_BOVIN	1.081	1.907	2.701	2.494	3.369	4.486
sp|Q3U5Q7|CMPK2_MOUSE	1.422	1.584	0.65	1.793	2.62	1.21
sp|Q3ZBK6|LSM4_BOVIN	0.741	0.886	1.143	1.005	2.756	2.967
sp|Q3ZBP3|RBMS1_BOVIN	1.981	2.015	1.508	1.713	3.604	1.954
sp|Q3ZCE0|LSM8_BOVIN	1.339	1.487	1.506	1.859	5.131	5.432
sp|Q3ZCL8|SH3L3_BOVIN	1.367	1.868	1.813	1.366	2.487	1.371
sp|Q49LS8|XKR6_TETNG	1.482	1.59	1.644	1.136	3.194	1.702
sp|Q4I8B6|AKR1_GIBZE	1.181	2.326	1.424	1.438	2.222	1.138
sp|Q4SS66|TRBP2_TETNG	0.781	0.924	1.075	3.174	1.435	1.093
sp|Q53G44|IF44L_HUMAN	2.295	3.658	2.326	1.528	1.115	1.423
sp|Q569K6|CC157_HUMAN	2.045	1.44	1.494	1.313	1.355	1.414
sp|Q58DS9|RAB5C_BOVIN	1.067	1.101	1.089	2.254	2.32	1.106
sp|Q5AYW6|DXO_EMENI	3.529	4.887	4.464	2.299	2.648	2.847
sp|Q5BKL9|CAB45_XENTR	2.036	1.707	1.551	2.197	1.603	1.073
sp|Q5BL31|ILRUN_DANRE	2.457	1.206	1.433	1.675	2.204	2.001
sp|Q5E992|PPIL1_BOVIN	2.517	3.393	0	1.96	1.654	3.047
sp|Q5G872|SCUB2_DANRE	1.551	1.189	2.446	1.331	1.207	1.561
sp|Q5R746|YTDC2_PONAB	1.595	1.102	0.951	1.511	2.66	1.09
sp|Q5R9K8|AR2BP_PONAB	2.401	2.626	0	0	0	0
sp|Q5RB68|IF2B2_PONAB	1.105	1.516	1.036	2.008	1.95	2.397
sp|Q5RC32|MDM1_PONAB	1.551	1.704	2.633	2.024	1.713	1.921
sp|Q5SXG7|VMO1_MOUSE	1.231	1.292	1.874	2.067	1.266	2.049
sp|Q5TH69|BIG3_HUMAN	2.198	2.819	2.322	2.108	2.042	2.936
sp|Q5U508|TBCE_XENLA	1.462	2.623	1.984	1.098	1.078	1.224
sp|Q5ZI72|HNRDL_CHICK	1.116	1.729	1.12	1.438	3.384	2.926
sp|Q5ZIJ9|MIB2_CHICK	2.768	1.982	2.07	3.142	1.983	2.893
sp|Q5ZJL4|CLP1_CHICK	1.748	1.479	1.021	2.172	1.185	1.5
sp|Q5ZK88|PSPC1_CHICK	1.466	2.531	1.122	1.67	1.59	1.494
sp|Q61555|FBN2_MOUSE	2.563	1.184	1.474	2.264	1.73	0
sp|Q62784|INP4A_RAT	0	3.224	1.086	0	3.319	0
sp|Q62920|PDLI5_RAT	3.543	4.423	1.616	4.946	2.655	2.029
sp|Q68EX9|CHID1_XENLA	1.463	1.123	1.163	2.851	1.985	2.735
sp|Q6GQD3|RB24A_XENLA	4.84	6.893	2.797	3.842	5.903	16.105
sp|Q6NT55|CP4FN_HUMAN	2.334	1.84	2.005	2.187	1.814	1.973
sp|Q6P4Z2|CO2A1_XENTR	1.407	1.173	1.706	2.283	1.938	2.183

sp|Q6P8M1|TATD1_MOUSE	1.024	1.277	1.343	2.172	1.602	1.938
sp|Q6P9B9|INT5_HUMAN	0.64	0.991	0	0	2.137	0
sp|Q6P9F0|CCD62_HUMAN	1.82	2.107	1.717	1.046	2.142	1.31
sp|Q6PAV2|HERC4_MOUSE	0.943	2.266	0	2.473	2.327	2.339
sp|Q6PF93|PK3C3_MOUSE	1.55	1.673	2.203	1.957	2.162	1.817
sp|Q6PHK6|PURB_DANRE	9.364	4.55	5.545	1.956	5.486	1.238
sp|Q6PIL6|KCIP4_HUMAN	1.326	0.655	0	0.566	1.26	2.2
sp|Q7LFX5|CHSTF_HUMAN	1.043	1.767	1.153	2.436	2.281	1.111
sp|Q7M456|RNOY_CRAGI	2.294	1.648	1.929	2.583	1.321	2.29
sp|Q7XBS0|DUR3_ORYSJ	1.592	1.388	1.406	1.652	2.062	1.67
sp|Q7Z2T5|TRM1L_HUMAN	0.703	0.2	1.406	0	4.908	0.829
sp|Q7Z3U7|MON2_HUMAN	1.61	2.078	2.139	1.039	0.985	1.012
sp|Q80W04|TMCC2_MOUSE	2.297	2.1	2.614	2.584	1.483	0
sp|Q868Z9|PPN_DROME	0.991	1.026	1.058	3.643	1.038	1.013
sp|Q86YT6|MIB1_HUMAN	0.754	4.185	0	1.717	2.526	2.918
sp|Q8BG18|NECA1_MOUSE	1.08	1.028	1.779	1.73	2.506	1.554
sp|Q8BJ64|CHDH_MOUSE	2.147	2.226	1.769	2.217	1.482	1.654
sp|Q8BT60|CPNE3_MOUSE	1.043	1.241	1.531	1.991	3.475	2.21
sp|Q8BTM8|FLNA_MOUSE	1.544	2.003	1.501	2.063	1.241	2.014
sp|Q8BWL5|RBMS3_MOUSE	1.981	2.015	1.508	1.713	3.604	1.954
sp|Q8CH18|CCAR1_MOUSE	0.815	1.087	3.449	0.914	0.716	0
sp|Q8CHT1|NGEF_MOUSE	1.5	1.317	1.379	2.397	1.671	1.715
sp|Q8IU26|LYS_RUDPH	1.063	1.289	1.182	1.561	2.096	2.292
sp|Q8IVL1|NAV2_HUMAN	1.752	1.525	1.811	1.323	1.505	2.002
sp|Q8JG64|PDIA3_CHICK	1.336	2.056	1.105	1.18	1.104	1.767
sp|Q8K0U4|HS12A_MOUSE	1.349	3.039	2.577	2.024	2.878	1.776
sp|Q8MJK1|CBY1_BOVIN	0	0	0	3.312	2.62	2.671
sp|Q8N3Y7|RDHE2_HUMAN	1.221	1.562	2.011	1.851	1.123	1.889
sp|Q8NFW1|COMA1_HUMAN	1.269	1.482	2.904	2.134	5.285	3.291
sp|Q8TBZ9|TEX47_HUMAN	1.857	1.954	1.645	2.05	1.444	1.568
sp|Q8VDM6|HNRL1_MOUSE	1.217	1.264	4.166	2.799	1.708	1.795
sp|Q8WW35|DYT2B_HUMAN	0.393	0	0	2.354	0	0
sp|Q91233|HSP70_ONCTS	1.378	2.035	1.908	2.393	2.05	1.431
sp|Q921F2|TADBP_MOUSE	1.519	2.258	1.039	1.046	1.488	1.268
sp|Q92614|MY18A_HUMAN	1.995	2.046	1.349	0.966	1.947	1.327
sp|Q92753|RORB_HUMAN	3.242	2.968	2.247	0	0.63	2.099
sp|Q92834|RPGR_HUMAN	0	1.19	2.505	1.49	1.076	1.288
sp|Q95KU9|NEMO_BOVIN	0.986	1.004	1.004	2.024	0.994	0.588
sp|Q96M69|LRGUK_HUMAN	1.738	1.408	1.945	2.183	1.391	1.62
sp|Q96MM6|HS12B_HUMAN	2.059	1.802	1.687	1.819	1.209	1.305
sp|Q96T60|PNKP_HUMAN	0	0	0	2.852	0	2.164
sp|Q99JR5|TINAL_MOUSE	0.978	0.91	0.807	2.759	0	0.82
sp|Q99MN1|SYK_MOUSE	0.941	0.982	0.975	1.037	1.466	3.494
sp|Q99N84|RT18B_MOUSE	2.179	2.152	3.48	2.139	1.473	1.795

sp|Q9BX84|TRPM6_HUMAN	1.573	1.45	1.281	0.793	2.078	0
sp|Q9CQQ8|LSM7_MOUSE	0	1.365	0.786	1.48	2.622	2.117
sp|Q9D0W5|PPIL1_MOUSE	2.517	3.393	0	1.96	1.654	3.047
sp|Q9D187|CIA2B_MOUSE	1.192	1.659	1.542	0	2.235	0
sp|Q9D4D4|TKTL2_MOUSE	1.479	2.027	1.719	1.647	1.133	2.051
sp|Q9DBR1|XRN2_MOUSE	1.716	1.986	3.719	1.415	1.553	1.987
sp|Q9ERH8|S28A3_MOUSE	2.023	1.196	1.257	1.089	1.397	1.05
sp|Q9H0D6|XRN2_HUMAN	1.716	1.986	3.719	1.415	1.553	1.987
sp|Q9JIL8|RAD50_RAT	8.954	12.554	8.755	0	0	0
sp|Q9N1Q0|RSMB_NOTEU	1.861	1.712	1.924	1.824	2.074	2.415
sp|Q9QXT5|EGFL7_MOUSE	0.641	3.842	2.242	2.375	2.198	2.611
sp|Q9RBP5|ISOH_RHOSX	0.629	0.648	2.003	1.069	0.588	0.951
sp|Q9SKB3|PARG1_ARATH	0.984	0.888	2.279	0.983	0.834	0.832
sp|Q9UI40|NCKX2_HUMAN	1.704	2.453	0	2.392	1.582	1.902
sp|Q9UKR8|TSN16_HUMAN	1.532	1.219	2.135	1.414	1.488	1.684
sp|Q9UMY4|SNX12_HUMAN	2.098	1.624	1.924	2.379	1.811	1.88
sp|Q9VI13|PAK_DROME	0.955	1.666	2.706	1.113	1.402	0.996
sp|Q9VJY9|LOQS_DROME	0.781	0.924	1.075	3.174	1.435	1.093
sp|Q9VPW8|PINO_DROME	1.778	2.068	1.362	1.427	1.44	1.843
sp|Q9VVE5|MSIR6_DROME	15.53	1.438	2.085	3.676	12.04	1.682
sp|Q9VXE0|RUXG_DROME	1.521	1.234	2.101	1.655	1.856	2.107
sp|Q9Y2I8|WDR37_HUMAN	1.353	1.26	0.862	0.966	1.223	2.076
sp|Q9Y333|LSM2_HUMAN	1.064	1.224	1.101	1.342	2.496	2.576
sp|Q9Y4D2|DGLA_HUMAN	1.8	1.927	2.379	2.077	1.653	2.003
sp|Q9Y4J8|DTNA_HUMAN	0.949	1.159	2.157	1.39	1.103	1.215
sp|Q9Y573|IPP_HUMAN	2.279	1.501	1.248	0.84	0	1.66
sp|Q9YHZ6|CDC45_XENLA	0	0.795	0	1.877	2.744	3.708
sp|Q9YIC0|EF1A_ORYLA	1.342	1.537	1.195	1.274	1.136	2.229

Supplementary table S5. Enriched proteins detected in the LC-MS/MS analysis in the six samples over the control sample. Each row reports the Uniprot accession number of the detected protein, and for each samples, the ratio of the LFQ detected in that sample over the LFQ detected in the control sample.

Protein	Cell Type	IP1 accession	IP2 accession	Control1 accession	Control2 accession	N° of reported peaks	N° of peaks passing IDR cutoff
DROSHA	HepG2	ENCLB583RAS	ENCLB058IYX	ENCBL855VXV	n/a	48655	3932
DROSHA	K562	ENCLB778RFV	ENCLB893EAL	ENCLB373HPU	n/a	20844	5411
DGCR8	HepG2	ENCLB026DPK	ENCB909IOJ	ENCLB745ADQ	n/a	29866	4196
DGCR8	K562	ENCLB465FHS	ENCLB991LAW	ENCLB098QVZ	n/a	5469	1077
AGO2	HCT116	SRR13067820	SRR13067821	SRR13067822	SRR13067823	117542	7138
AGO2	HCT116	SRR5027862	SRR5027863	SRR5027856	SRR5027857	20020	6059
FMR1	K562	ENCFF436TNC	ENCFF736XNI	ENCFF328KAL	n/a	31325	3194

FXR2	HepG2	ENCFF032TZM	ENCFF385GCC	ENCFF936GTY	n/a	24536	6205
HNRNPK	HepG2	ENCFF329QRR	ENCFF457EXY	ENCFF019JFZ	n/a	63998	17343
AGO2	neurons	SRR10513943	SRR10513944	SRR10513938	SRR10513939	26036	1395

Supplementary table S6. The table reports all the eCLIP experiments analysed with the associated accession codes of immunoprecipitated and input libraries (in some cases only one input library was available), the cell type, the number of peaks detected by Clipper and the number of peaks that passed the IDR analysis**.**

Species	Condition	Rep.	SRA
Homo sapiens	DROSHA KO	1	SRR21714672
Homo sapiens	DROSHA KO	2	SRR21714673
Homo sapiens	DROSHA KO	3	SRR21714674
Homo sapiens	AGO1+AGO2 KO	1	SRR21714684
Homo sapiens	AGO1+AGO2 KO	2	SRR21714685
Homo sapiens	AGO1+AGO2 KO	3	SRR21714686
Homo sapiens	WT	1	SRR21714696
Homo sapiens	WT	2	SRR21714697
Homo sapiens	WT	3	SRR21714698
Mus musculus	DROSHA KO	1	SRR6757505
Mus musculus	DROSHA KO	2	SRR6757506
Mus musculus	WT	1	SRR6757511
Mus musculus	WT	2	SRR6757512
Mus musculus	RNase Z KO	1	SRR6790393-5
Mus musculus	RNase Z KO	2	SRR6790396-9
Mus musculus	RNase Z KO	3	SRR6790400-3
Mus musculus	WT	1	SRR6790380-3
Mus musculus	WT	2	SRR6790384-7
Mus musculus	WT	3	SRR6790388-91

Supplementary table S7: for each library it is reported the species, whether a gene was knocked out or it was a wild type (WT) sample, the replicate, and the SRA code.

Species	IP protein	Replicate ID	SRA
Caenorhabditis elegans	ALG-1	1	SRR2230088
Caenorhabditis elegans	ALG-1	2	SRR2230091
Caenorhabditis elegans	ALG-1	3	SRR2230094
Caenorhabditis elegans	control (ALG-1)	1	SRR2230081
Caenorhabditis elegans	control (ALG-1)	2	SRR2230084
Caenorhabditis elegans	control (ALG-1)	3	SRR2230087
Caenorhabditis elegans	HRDE-1	1	SRR12567593
Caenorhabditis elegans	HRDE-1	2	SRR12567594

Caenorhabditis elegans	HRDE-1	3	SRR12567597
Caenorhabditis elegans	HRDE-1	4	SRR12567598
Caenorhabditis elegans	HRDE-1	5	SRR12567601
Caenorhabditis elegans	HRDE-1	6	SRR12567602
Caenorhabditis elegans	control (HRDE-1)	1	SRR12567591
Caenorhabditis elegans	control (HRDE-1)	2	SRR12567592
Caenorhabditis elegans	control (HRDE-1)	3	SRR12567595
Caenorhabditis elegans	control (HRDE-1)	4	SRR12567596
Caenorhabditis elegans	control (HRDE-1)	5	SRR12567599
Caenorhabditis elegans	control (HRDE-1)	6	SRR12567600
Caenorhabditis elegans	ERGO-1	1	SRR20334703
Caenorhabditis elegans	ERGO-1	2	SRR20334741
Caenorhabditis elegans	control (ERGO-1)	1	SRR20334704
Caenorhabditis elegans	control (ERGO-1)	2	SRR20334742
Caenorhabditis elegans	PRG-1	1	SRR20334695
Caenorhabditis elegans	PRG-1	2	SRR20334733
Caenorhabditis elegans	control (PRG-1)	1	SRR20334696
Caenorhabditis elegans	control (PRG-1)	2	SRR20334734
Caenorhabditis elegans	WAGO9	1	SRR18266347
Caenorhabditis elegans	WAGO9	2	SRR18266348
Caenorhabditis elegans	CSR1	1	SRR18266353
Caenorhabditis elegans	CSR1	2	SRR18266354
Caenorhabditis elegans	control (CSR1)	1	SRR18266351
Caenorhabditis elegans	control (CSR1)	2	SRR18266352
Drosophila melanogaster	AGO1	1	SRR13314104
Drosophila melanogaster	AGO1	2	SRR13314105
Drosophila melanogaster	AGO1	3	SRR13314106
Drosophila melanogaster	control (AGO1)	1	SRR13314092
Drosophila melanogaster	control (AGO1)	2	SRR13314093
Drosophila melanogaster	control (AGO1)	3	SRR13314094
Drosophila melanogaster	AGO2	1	SRR13314116
Drosophila melanogaster	AGO2		SRR13314117
Drosophila melanogaster	AGO2	3	SRR13314118
Drosophila melanogaster	control (AGO2)	1	SRR13314110
Drosophila melanogaster	control (AGO2)	2	SRR13314111
Drosophila melanogaster	control (AGO2)	3	SRR13314112
Drosophila melanogaster	PIWI	1	SRR2042569
Drosophila melanogaster	Aubergine	1	SRR2042570
Drosophila melanogaster	AGO3	1	SRR2042571
Drosophila melanogaster	control (PIWI,Aub,AGO3)	1	SRR2042568
Nematostella vectensis	AGO1	1	SRR10960921
Nematostella vectensis	AGO1	2	SRR10960922
Nematostella vectensis	AGO2	1	SRR10960923
Nematostella vectensis	AGO2	2	SRR10960924

Nematostella vectensis	control (AGO1,AGO2)	1	SRR10960925
Nematostella vectensis	control (AGO1,AGO2)	2	SRR10960926
Mus musculus	HIWI	1	SRR11818515
Mus musculus	HIWI	2	SRR11818516
Mus musculus	control (HIWI)	1	SRR11818517
Mus musculus	control (HIWI)	2	SRR11818518

Supplementary table 8: for each RIP-Seq library I reported the species, the target proteins and the SRA code.

4.7 References

Almeida MV, Andrade-Navarro MA, Ketting RF. 2019. Function and Evolution of Nematode RNAi Pathways. *Noncoding RNA* 5:8.

Aparicio-Puerta E, Lebrón R, Rueda A, Gómez-Martín C, Giannoukakos S, Jaspez D, Medina JM, Zubkovic A, Jurak I, Fromm B, et al. 2019. sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res* 47:W530–W535.

Bartel DP. 2018. Metazoan MicroRNAs. *Cell* 173:20–51.

Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bursteinas B, et al. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49:D480–D489.

Brannan KW, Chaim IA, Marina RJ, Yee BA, Kofman ER, Lorenz DA, Jagannatha P, Dong KD, Madrigal AA, Underwood JG, et al. 2021. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. *Nat Methods* 18:507–519.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* 128:1089–1103.

Brzezniak LK, Bijata M, Szczesny RJ, Stepien PP. 2011. Involvement of human ELAC2 gene product in 3' end processing of mitochondrial tRNAs. *RNA Biol* 8:616–626.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* 44:D184–D189.

Chen Yang, Beck A, Davenport C, Chen Yuan, Shattuck D, Tavtigian S V. 2005. Characterization of TRZ1, a yeast homolog of the human candidate prostate cancer susceptibility gene ELAC2 encoding tRNase Z. *BMC Mol Biol* 6:12.

Chu Y, Yokota S, Liu J, Kilikevicius A, Johnson KC, Corey DR. 2021. Argonaute binding within human nuclear RNA and its impact on alternative splicing. *RNA* 27:991–1003.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, et al. 2008. An endogenous small interfering RNA pathway in Drosophila. *Nature* 453:798–802.

Dai L, Chen K, Youngren B, Kulina J, Yang A, Guo Z, Li J, Yu P, Gu S. 2016. Cytoplasmic Drosha activity generated by alternative splicing. *Nucleic Acids Res*:gkw668.

Daugaard I, Hansen TB. 2017. Biogenesis and Function of Ago-Associated RNAs. *Trends in Genetics* 33:208–219.

Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. 2020. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 17:41–44.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Fontanesi F, Tigano M, Fu Y, Sfeir A, Barrientos A. 2020. Human mitochondrial transcription and translation. In: The Human Mitochondrial Genome. Elsevier. p. 35–70.

Frankenfield AM, Ni J, Ahmed M, Hao L. 2022. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *J Proteome Res* 21:2104–2113.

Friend K, Campbell ZT, Cooke A, Kroll-Conner P, Wickens MP, Kimble J. 2012. A conserved PUF–Ago–eEF1A complex attenuates translation elongation. *Nat Struct Mol Biol* 19:176–183.

Golden RJ, Chen B, Li T, Braun J, Manjunath H, Chen X, Wu J, Schmid V, Chang T-C, Kopp F, et al. 2017. An Argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature* 542:197–202.

Gray MW. 2012. Mitochondrial Evolution. *Cold Spring Harb Perspect Biol* 4:a011403–a011403.

Han J, Lee Y, Yeom K-H, Kim Y-K, Jin H, Kim VN. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* 18:3016–3027.

Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Becker A, Bennett R, Berry A, Bhai J, et al. 2024. Ensembl 2024. *Nucleic Acids Res* 52:D891–D899.

Hill GE. 2015. Mitonuclear Ecology. *Mol Biol Evol* 32:1917–1927.

Iwakawa H, Tomari Y. 2022. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. *Mol Cell* 82:30–43.

Johnson KC, Johnson ST, Liu J, Chu Y, Arana C, Han Y, Wang T, Corey DR. 2023. Consequences of depleting TNRC6, AGO, and DROSHA proteins on expression of microRNAs. *RNA* 29:1166–1184.

Keam SP, Young PE, McCorkindale AL, Dang THY, Clancy JL, Humphreys DT, Preiss T, Hutvagner G, Martin DIK, Cropley JE, et al. 2014. The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. *Nucleic Acids Res* 42:8984–8995.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 47:D155–D162.

Kumar P, Anaya J, Mudunuri SB, Dutta A. 2014. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol* 12:78.

Kumar P, Kuscu C, Dutta A. 2016. Biogenesis and Function of Transfer RNA-Related Fragments (tRFs). *Trends Biochem Sci* 41:679–689.

Kuscu C, Kumar P, Kiran M, Su Z, Malik A, Dutta A. 2018. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA* 24:1093–1105.

Lambert M, Benmoussa A, Provost P. 2019. Small Non-Coding RNAs Derived from Eukaryotic Ribosomal RNA. *Noncoding RNA* 5:16.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, et al. 2013. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* 20:1434–1442.

Luo S, He F, Luo J, Dou S, Wang Y, Guo A, Lu J. 2018. Drosophila tsRNAs preferentially suppress general translation machinery via antisense pairing and participate in cellular starvation response. *Nucleic Acids Res* 46:5250–5268.

Maag D, Algire MA, Lorsch JR. 2006. Communication between Eukaryotic Translation Initiation Factors 5 and 1A within the Ribosomal Pre-initiation Complex Plays a Role in Start Site Selection. *J Mol Biol* 356:724–737.

Maniataki E, Mourelatos Z. 2005. Human mitochondrial tRNA Met is exported to the cytoplasm and associates with the Argonaute 2 protein. *RNA* 11:849–852.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17:10.

Marturano G, Carli D, Cucini C, Carapelli A, Plazzi F, Frati F, Passamonti M, Nardi F. 2024. SmithHunter: a workflow for the identification of candidate smithRNAs and their targets. *BMC Bioinformatics* 25:286.

Matranga C, Zamore PD. 2007. Small silencing RNAs. *Current Biology* 17:R789–R793.

Müller M, Schaefer M, Fäh T, Spies D, Hermes V, Ngondo RP, Peña-Hernández R, Santoro R, Ciaudo C. 2022. Argonaute proteins regulate a specific network of genes through KLF4 in mouse embryonic stem cells. *Stem Cell Reports* 17:1070–1080.

Muneretto G, Plazzi F, Passamonti M. 2024. Mitochondrion-to-nucleus communication mediated by RNA export: a survey of potential mechanisms and players across eukaryotes. *Biol Lett* 20.

Nguyen TA, Jo MH, Choi Y-G, Park J, Kwon SC, Hohng S, Kim VN, Woo J-S. 2015. Functional Anatomy of the Human Microprocessor. *Cell* 161:1374–1387.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 13:508–514.

Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC. 2007. The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. *Cell* 130:89–100.

Passamonti M, Calderone M, Delpero M, Plazzi F. 2020. Clues of in vivo nuclear gene regulation by mitochondrial short non-coding RNAs. *Sci Rep* 10:8219.

Passmore LA, Schmeing TM, Maag D, Applefield DJ, Acker MG, Algire MA, Lorsch JR, Ramakrishnan V. 2007. The Eukaryotic Translation Initiation Factors eIF1 and eIF1A Induce an Open Conformation of the 40S Ribosome. *Mol Cell* 26:41–50.

Pozzi A, Dowling DK. 2022. New Insights into Mitochondrial–Nuclear Interactions Revealed through Analysis of Small RNAs. *Genome Biol Evol* 14.

Pozzi A, Plazzi F, Milani L, Ghiselli F, Passamonti M. 2017. SmithRNAs: Could Mitochondria “Bend” Nuclear Regulation? *Mol Biol Evol* 34:1960–1973.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

Schölz C, Lyon D, Refsgaard JC, Jensen LJ, Choudhary C, Weinert BT. 2015. Avoiding abundance bias in the functional annotation of posttranslationally modified proteins. *Nat Methods* 12:1003–1004.

Senti K-A, Jurczak D, Sachidanandam R, Brennecke J. 2015. piRNA-guided slicing of transposon transcripts enforces their transcriptional silencing via specifying the nuclear piRNA repertoire. *Genes Dev* 29:1747–1762.

Siira SJ, Rossetti G, Richman TR, Perks K, Ermer JA, Kuznetsova I, Hughes L, Shearwood AJ, Viola HM, Hool LC, et al. 2018. Concerted regulation of mitochondrial and nuclear non-coding <scp>RNA</scp> s by a dual-targeted <scp>RN</scp> ase Z. *EMBO Rep* 19.

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 27:491–499.

Sriram K, Qi Z, Yuan D, Malhi NK, Liu X, Calandrelli R, Luo Y, Tapia A, Jin S, Shi J, et al. 2024. Regulation of nuclear transcription by mitochondrial RNA in endothelial cells. *Elife* 13.

Svobodova E, Kubikova J, Svoboda P. 2016. Production of small RNAs by mammalian Dicer. *Pflugers Arch* 468:1089–1102.

Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin E V, Patel DJ, van der Oost J. 2014. The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* 21:743–753.

Wang J, Shi Y, Zhou H, Zhang P, Song T, Ying Z, Yu H, Li Y, Zhao Y, Zeng X, et al. 2022. piRBase: integrating piRNA annotation in all aspects. *Nucleic Acids Res* 50:D265–D272.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453:539–543.

Weick E-M, Miska EA. 2014. piRNAs: from biogenesis to function. *Development* 141:3458–3471.

Westholm JO, Lai EC. 2011. Mirtrons: microRNA biogenesis via splicing. *Biochimie* 93:1897–1904.

Whipple AJ, Breton-Provencher V, Jacobs HN, Chitta UK, Sur M, Sharp PA. 2020. Imprinted Maternally Expressed microRNAs Antagonize Paternally Driven Gene Programs in Neurons. *Mol Cell* 78:85-95.e8.

Will CL, Lührmann R. 2001. Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol* 13:290–301.

Wolff JN, Ladoukakis ED, Enríquez JA, Dowling DK. 2014. Mitonuclear interactions: evolutionary consequences over multiple biological scales. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369:20130443.

Yi T, Arthanari H, Akabayov B, Song H, Papadopoulos E, Qi HH, Jedrychowski M, Güttler T, Guo C, Luna RE, et al. 2015. eIF1A augments Ago2-mediated Dicer-independent miRNA biogenesis and RNA interference. *Nat Commun* 6:7194.

Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. 2004. Single Processing Center Models for Human Dicer and Bacterial RNase III. *Cell* 118:57–68.

Zhu Q-H, Spriggs A, Matthew L, Fan L, Kennedy G, Gubler F, Helliwell C. 2008. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res* 18:1456–1465.

5. Conclusion of the Characterization and Evolution of Mito-Nuclear Interactions

The phenotype of the eukaryotic cell is the result of the expression of at least two genomes, the nuclear and the mitochondrial genome (Hill 2015). Mito-nuclear interactions are never unidirectional: every biological process involved in the mito-nuclear crosstalk is, in the end, crucial for the evolution of both genomes.

For years, Amarsipobranchia has been considered one of the hypotheses coping with the phylogeny of the Bivalvia (Plazzi and Passamonti 2010). However, the monophyly of Heterodonta and Pteriomorphia is more likely a phylogenetic artifact. The distribution of genes across the two mitochondrial strands can affect various factors, such as transcription levels (Shtolz and Mishmar 2023) and rearrangement rates (Gissi et al. 2008). These factors can, in turn, influence the nucleotide composition of mitochondrial markers. Indeed, nucleotide compositional biases have repeatedly been a cause of discordance between mitochondrial and nuclear topologies (Hassanin et al. 2005; Quattrini et al. 2023). Remarkably, in bivalves, at least two clades have converged on the same mitochondrial architecture, with all mitochondrial genes located on the same strand. In the second chapter "Mito-nuclear coevolution and phylogenetic artifacts: the case of bivalve mollusks", I hypothesized that Nuculanida might also exhibit this gene distribution. Recent sequencing of the *Yoldia hyperborea* mitogenome (PP541907.1) has confirmed this hypothesis, revealing that three independent clades have converged on the same mitochondrial architecture (i.e., Pteriomorphia, Nuculanida and Heterodonta). The acquisition of this mitochondrial architecture is not unique to these Bivalvia clades; it is also present in Annelida, Brachiopoda, Platyhelminthes, Cnidaria, and Porifera, where all mitochondrial genes are located on one strand, and this feature appears to be irreversible in these phyla (Gissi et al. 2008). Overall, this particular mitochondrial architecture may be more likely to emerge in clades with a low-energy demand, whose mitogenomes are under relaxed selective pressure and with elevated evolutionary rates (Jakovlić et al. 2023). Conversely, the acquisition of this mitochondrial architecture may be the driver of a higher reorganization of the mitochondrial genome and accelerated evolutionary rates, though both scenarios may explain relaxed pressure on the mitogenome of certain clades. In general, Pteriomorphia and Heterodonta exhibit a relaxation in selective pressures on the mitochondrial genome, which may have contributed support for the Amarsipobranchia clade.

The Amarsipobranchia artifact is not limited to mitochondrial genes; nuclear OXPHOS genes, particularly those closely interacting with their mitochondrial counterparts, also support this topology. The evolutionary rates of mitochondrial and nuclear OXPHOS subunits are highly correlated, a pattern observed in mammals (Weaver et al. 2022), insects (Yan et al. 2019), and bivalves themselves (Piccinini et al. 2021). This correlation might be explained by the hypothesis that detrimental mutations accumulating in mitochondrial OXPHOS genes are compensated by adaptive mutations in nuclear OXPHOS genes (Levin et al. 2014). However,

compensatory evolution only partially accounts for the correlation between interacting subunits. Alternatively, relaxation of selective pressure can lead to significant changes in evolutionary rates, further increasing the rate correlation between subunits within the same pathway (Little et al. 2024).

RNA interference (RNAi) has long been recognized as a mechanism that protects against viruses and mobile elements, while also modulating specific cellular functions (Shabalina and Koonin 2008). However, RNAi is also an effective tool for communication between genomic players. In some host-symbiont systems, small RNAs are exchanged to control molecular functions of the respective partner (Bermúdez-Barrientos et al. 2020). It is therefore unsurprising that in one of the most ancient symbioses, the origin of the eukaryotic cell, small RNAs participate in interactions between host and symbiont. The evolution of RNA interference pathways has diverged significantly across animal lineages. Argonaute and DICER, key proteins in the RNAi system, play different roles depending on the metazoan clade (Swarts et al. 2014). Despite their importance, these two protein families have been poorly studied in lophotrochozoan phyla. My study of the evolution of these proteins reveals the loss of the endo-siRNA pathway during Lophotrochozoan evolution. Early-diverging phyla like Platyhelminthes and Syndermata retain a complete and functional endo-siRNA pathway. In contrast, most Trochozoa phyla (i.e., Mollusca, Annelida, Brachiopoda, and Phoronida) have lost this pathway entirely. Interestingly, other Trochozoa (i.e., Nemertea, Bryozoa and Entoprocta) exhibit an intermediate state: their genomes still encode the Argonaute protein, but lack the DICER responsible for processing them. The role of this Argonaute protein in organisms that can no longer process endo-siRNAs remains unclear. These findings underscore the plasticity of RNAi pathways across organisms. RNAi may evolve differently based on an organism's life strategy. For example, the loss of Piwi Argonaute proteins in parasitic flatworms has been linked to parasitism (Fontenla et al. 2021), as has the loss of certain miRNA families in parasitic Syndermata (Herlyn et al. 2024). Similarly, the loss of endo-siRNAs in Trochozoa may have been driven by specific selective pressures, with the Argonaute protein possibly co-opted for a different function in some of these lineages.

RNAi mechanisms have enabled the emergence of novel patterns, even in the context of mito-nuclear interactions. In the freshwater mussel *Potamilus streckersoni*, mitochondria may influence sex determination by encoding a male-restricted smithRNA that targets a gene involved in female development (Smith et al. 2023). Understanding the mode of action of a potential smithRNA pathway is therefore crucial to elucidating the evolution of complex traits such as sex determination. My initial focus was on the interaction between smithRNAs and Argonaute proteins. While some human smithRNAs have been shown to interact with AGO2 (Pozzi and Dowling 2022), other Argonaute proteins could also be involved. My analysis confirmed the interaction between AGO2 and human smithRNAs and identified two additional Argonaute candidates: ERGO-1 in *C. elegans* and AGO2 in *D. melanogaster*. Both proteins are known to load endo-siRNAs, yet the analysis of RIP-Seq libraries revealed their tendency to bind small RNAs derived from mitochondrial tRNAs.

Notably, in both cases, the strongest interaction was observed with the small RNA mapping to mitochondrial Met-tRNA. Since AGO2 interacts with the initiation factor eIF4H, it is possible that initiation and elongation factors mediate the loading of smithRNAs onto Argonaute proteins. My data also revealed other potential candidates involved in smithRNA maturation, such as the Microprocessor or the Spliceosome complex. However, both scenarios would require the transport of smithRNA precursors from the mitochondria to the nucleus. It is clear, though, that smithRNA biogenesis is closely linked to the maturation of the mitochondrial polycistronic transcript, as knocking out RNAse Z significantly reduced the coverage of most human smithRNAs.

In conclusion, the results of my thesis highlight the close connection between the nucleus and mitochondria, as well as the evolutionary implications of this interaction. A shift in the mitochondrial architecture of certain bivalve clades has produced a phylogenetic artifact that, due to the strict co-evolution between OXPHOS subunits, is supported even by nuclear OXPHOS genes. Similarly, mitochondrially encoded small RNAs interact with nuclear-encoded proteins to mature and target messenger RNAs involved in specific cellular functions. As is the case for any symbiotic relationship, the two genomes become interdependent, and the sum of their interactions contributes to the complexity of the eukaryotic cell.

5.1 Future Perspectives

Every result opens the door to multiple scientific questions. In this final paragraph, I will briefly discuss some questions that emerged during the analysis and discussion of the results, which I have not addressed in my thesis project.

Mito-nuclear discordances (i.e., when the phylogenetic signal inferred from mitochondrial markers differs from the signal inferred from nuclear markers) are commonly described in intrageneric or intra-family phylogenies and have been linked to phenomena such as incomplete lineage sorting, introgression, and phylogeographic patterns (Toews and Brelsford 2012). However, it is not clear whether mito-nuclear discordances at deep nodes are due to similar causes. In Squamata, the clades Serpentes and Agamidae have experienced convergent evolution in mitochondrial OXPHOS markers (Castoe et al. 2009). Due to this molecular convergence, Serpentes and Agamidae are supported as a monophyletic group by mitochondrial OXPHOS genes. Together with my colleague Oscar Wallnoefer, we annotated the mitochondrial and nuclear OXPHOS genes of 56 Squamata species. We detected a signal of convergent evolution in a subset of the nuclear OXPHOS genes, which also supported the monophyly of Agamidae + Serpentes (Wallnoefer et al., in prep.). Other cases of mito-nuclear discordances at deep nodes have been reported (Hassanin et al. 2005), but the causes behind this discordance have not always been addressed (Quattrini et al. 2023). Moreover, these studies have not considered how the nuclear counterparts co-evolve alongside these mitochondrial artifacts. In an era where -omics data are available for a wide range of organisms, we have the opportunity to understand how OXPHOS subunits co-evolve and compensate for each other to maintain aerobic respiration.

Proteins involved in the endo-siRNA pathway and their role vary among different animal branches remarkably. Once again, the loss of the endo-siRNA pathway during the evolution of Lophotrochozoa highlights the diversity of the phyla that comprise this group. The absence of this pathway is not restricted to Trochozoa, as all deuterostomes also lack a DICER and Argonaute protein dedicated to this pathway. In these cases, endo-siRNAs have been replaced by other pathways that fulfil the same functions without relying on RNAi mechanisms. One example is the interferon pathway, which senses and responds to viral RNAs in deuterostomes (Loo and Gale 2011). Contrastingly, insects depend on the endo-siRNA pathway for the antiviral defence (Schuster et al. 2019). An interferon defence mechanism has been described in Mollusca (Huang et al. 2017; Qiao et al. 2021). Therefore, it would be interesting to test whether, during the evolution of Lophotrochozoa, the endo-siRNA pathway was replaced by the interferon pathway. In this scenario, I would expect the emergence of the interferon pathway in all Trochozoa, exhibiting a defence mechanism similar to that of Deuterostomia.

During my thesis, I attempted to identify some candidates involved in smithRNA maturation. In my opinion, the most promising candidates are the eukaryotic initiation and elongation factors (eIFs and eEFs). These proteins interact with AGO2 (Friend et al. 2012), and eIF4H has been linked to the maturation of the DICER-independent miR-451 (Yi et al. 2015). I believe eIFs and eEFs may play a role not only in the maturation of smithRNAs but also in nuclear tRNA-related fragments (tRFs). Unfortunately, this link has never been tested. Knocking out eIF and eEF genes and performing small-RNA sequencing on knockout cells may elucidate whether these proteins are involved in the maturation of different small RNA types.

Overall, many aspects of the interaction between the nucleus and mitochondria remain overlooked. RNAi mechanisms seem to play a fundamental role in shaping these interactions. However, further studies are required to understand the significance of mitochondrial RNAi in different metazoan branches and the pathways that regulate these mechanisms.

5.2 References

Bermúdez-Barrientos JR, Ramírez-Sánchez O, Chow FW-N, Buck AH, Abreu-Goodger C. 2020. Disentangling sRNA-Seq data to study RNA communication between species. *Nucleic Acids Res* 48:e21–e21.

Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences* 106:8986–8991.

Fontenla S, Rinaldi G, Tort JF. 2021. Lost and Found: Piwi and Argonaute Pathways in Flatworms. *Front Cell Infect Microbiol* 11.

Friend K, Campbell ZT, Cooke A, Kroll-Conner P, Wickens MP, Kimble J. 2012. A conserved PUF–Ago–eEF1A complex attenuates translation elongation. *Nat Struct Mol Biol* 19:176–183.

Gissi C, Iannelli F, Pesole G. 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity (Edinb)* 101:301–320.

Hassanin A, Léger N, Deutsch J. 2005. Evidence for Multiple Reversals of Asymmetric Mutational Constraints during the Evolution of the Mitochondrial Genome of Metazoa, and Consequences for Phylogenetic Inferences. *Syst Biol* 54:277–298.

Herlyn H, Hebrom AA, Tosar JP, Mauer K, Schmidt H, Dezfuli B, Hankeln T, Bachman L, Sarkies P, Peterson K, et al. 2024. Substantial hierarchical reductions of genetic and morphological traits in the evolution of rotiferan parasites. *bioRxiv*.

Hill GE. 2015. Mitonuclear Ecology. *Mol Biol Evol* 32:1917–1927.

Huang B, Zhang L, Du Y, Xu F, Li L, Zhang G. 2017. Characterization of the Mollusc RIG-I/MAVS Pathway Reveals an Archaic Antiviral Signalling Framework in Invertebrates. *Sci Rep* 7:8217.

Jakovlić I, Zou H, Ye T, Zhang H, Liu X, Xiang C-Y, Wang G-T, Zhang D. 2023. Mitogenomic evolutionary rates in bilateria are influenced by parasitic lifestyle and locomotory capacity. *Nat Commun* 14:6307.

Levin L, Blumberg A, Barshad G, Mishmar D. 2014. Mito-nuclear co-evolution: the positive and negative sides of functional ancient mutations. *Front Genet* 5.

Little J, Chikina M, Clark NL. 2024. Evolutionary rate covariation is a reliable predictor of co-functional interactions but not necessarily physical interactions. *Elife* 12.

Loo Y-M, Gale M. 2011. Immune Signaling by RIG-I-like Receptors. *Immunity* 34:680–692.

Parry L, Tanner A, Vinther J. 2014. The origin of annelids. *Palaeontology* 57:1091–1103.

Piccinini G, Iannello M, Puccio G, Plazzi F, Havird JC, Ghiselli F. 2021. Mitonuclear Coevolution, but not Nuclear Compensation, Drives Evolution of OXPHOS Complexes in Bivalves. *Mol Biol Evol* 38:2597–2614.

Plazzi F, Passamonti M. 2010. Towards a molecular phylogeny of Mollusks: Bivalves' early evolution as revealed by mitochondrial genes. *Mol Phylogenet Evol* 57:641–657.

Pozzi A, Dowling DK. 2022. New Insights into Mitochondrial–Nuclear Interactions Revealed through Analysis of Small RNAs. *Genome Biol Evol* 14.

Qiao X, Wang L, Song L. 2021. The primitive interferon-like system and its antiviral function in molluscs. *Dev Comp Immunol* 118:103997.

Quattrini AM, Snyder KE, Purow-Ruderman R, Seiblitz IGL, Hoang J, Floerke N, Ramos NI, Wirshing HH, Rodriguez E, McFadden CS. 2023. Mito-nuclear discordance within Anthozoa, with notes on unique properties of their mitochondrial genomes. *Sci Rep* 13:7443.

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238:195–209.

Schuster S, Miesen P, van Rij RP. 2019. Antiviral RNAi in Insects and Mammals: Parallels and Differences. *Viruses* 11:448.

Shabalina S, Koonin E. 2008. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol* 23:578–587.

Shtolz N, Mishmar D. 2023. The metazoan landscape of mitochondrial DNA gene order and content is shaped by selection and affects mitochondrial transcription. *Commun Biol* 6:93.

Smith CH, Mejia-Trujillo R, Breton S, Pinto BJ, Kirkpatrick M, Havird JC. 2023. Mitonuclear Sex Determination? Empirical Evidence from Bivalves. *Mol Biol Evol* 40.

Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin E V, Patel DJ, van der Oost J. 2014. The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* 21:743–753.

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Mol Ecol* 21:3907–3930.

Weaver RJ, Rabinowitz S, Thueson K, Havird JC. 2022. Genomic Signatures of Mitonuclear Coevolution in Mammals. *Mol Biol Evol* 39. Available from: https://dx.doi.org/10.1093/molbev/msac233

Yan Z, Ye G, Werren JH. 2019. Evolutionary Rate Correlation between Mitochondrial-Encoded and Mitochondria-Associated Nuclear-Encoded Proteins in Insects. *Mol Biol Evol* 36:1022–1036.

Yi T, Arthanari H, Akabayov B, Song H, Papadopoulos E, Qi HH, Jedrychowski M, Güttler T, Guo C, Luna RE, et al. 2015. eIF1A augments Ago2-mediated Dicer-independent miRNA biogenesis and RNA interference. *Nat Commun* 6:7194.