bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Functional lability of RNA-dependent RNA polymerases in animals

Natalia Pinzón[1☯], Stéphanie Bertrand[2], Lucie Subirana[2], Isabelle Busseau[1], Hector Escrivá[2], Hervé Seitz[1*]

1 Institut de Génétique Humaine, UMR 9002 CNRS and université de Montpellier, 141, rue de la Cardonille, 34396 Montpellier CEDEX 5, France
2 Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins, BIOM, F-66650 Banyuls-sur-Mer, France

☯Current address: INSERM U981, Gustave Roussy Cancer Campus, Paris-Saclay University, Villejuif, France
* herve.seitz@igh.cnrs.fr

Abstract

RNA interference (RNAi) requires RNA-dependent RNA polymerases (RdRPs) in many eukaryotes, and RNAi amplification constitutes the only known function for eukaryotic RdRPs. Yet in animals, classical model organisms can elicit RNAi without possessing RdRPs, and only nematode RNAi was shown to require RdRPs. Here we show that RdRP genes are much more common in animals than previously thought, even in insects, where they had been assumed not to exist. RdRP genes were present in the ancestors of numerous clades, and they were subsequently lost at a high frequency. In order to probe the function of RdRPs in a deuterostome (the cephalochordate *Branchiostoma lanceolatum*), we performed high-throughput analyses of small RNAs from various *Branchiostoma* developmental stages. Our results show that *Branchiostoma* RdRPs do not appear to participate in RNAi: we did not detect any candidate small RNA population exhibiting classical siRNA length or sequence features. Our results show that RdRPs have been independently lost in dozens of animal clades, and even in a clade where they have been conserved (cephalochordates) their function in RNAi amplification is not preserved. Such a dramatic functional variability reveals an unexpected plasticity in RNA silencing pathways.

Author summary

RNA interference (RNAi) is a conserved gene regulation system in eukaryotes. In non-animal eukaryotes, it necessitates RNA-dependent RNA polymerases ("RdRPs"). Among animals, only nematodes appear to require RdRPs for RNAi. Yet additional animal clades have RdRPs and it is assumed that they participate in RNAi. Here, we find that RdRPs are much more common in animals than previously thought, but their genes were independently lost in many lineages. Focusing on a species with RdRP genes (a cephalochordate), we found that it does not use them for RNAi. While RNAi is the only known function for eukaryotic RdRPs, our results suggest additional roles. Eukaryotic RdRPs thus have a complex evolutionary history in animals, with frequent independent losses and apparent functional diversification.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

Small interfering RNAs (siRNAs) play a central role in the RNA interference (RNAi) response. Usually loaded on a protein of the AGO subfamily of the Argonaute family, they recognize specific target RNAs by sequence complementarity and typically trigger their degradation by the AGO protein [1]. In many eukaryotic species, normal siRNA accumulation requires an RNA-dependent RNA polymerase (RdRP). For example in plants, RdRPs are recruited to specific template RNAs and they generate long complementary RNAs [2–4]. The template RNA and the RdRP product are believed to hybridize, forming a long double-stranded RNA which is subsequently cleaved by Dicer nucleases into double-stranded siRNAs (reviewed in [5]). In fungi, RdRPs have also been implicated in RNAi and in RNA-directed heterochromatinization [6–9], but the exact nature of their products remains elusive: fungal RdRPs are frequently proposed to polymerize long RNAs which can form Dicer substrates after annealing to the RdRP template [10–12]. But the purified *Neurospora crassa*, *Thielavia terrestris* and *Myceliophthora thermophila* QDE-1 RdRPs tend to polymerize essentially short (9–21 nt) RNAs *in vitro*, suggesting that they may generate Dicer-independent small RNAs [13, 14]. In various unicellular eukaryotes, RdRPs have also been implicated in RNAi and related mechanisms (*e.g.*, see [15, 16]). It is usually believed that their products are long RNAs that anneal with the template to generate a Dicer substrate, and that model has gained experimental support in one organism, *Tetrahymena* [17].

Among eukaryotes, animals are thought to constitute an exception: most classical animal model organisms (*Drosophila* and mammals) can elicit RNAi without the involvement of an RdRP [1]. Only one animal model organism was shown to require RdRPs for RNAi: the nematode *Cænorhabditis elegans* [18, 19]. In nematodes, siRNAs made by Dicer only constitute a minor fraction of the total siRNA pool: such "primary" siRNAs recruit an RdRP on target RNAs, triggering the production of short antisense RNAs named "secondary siRNAs" [20–22]. Secondary siRNAs outnumber primary siRNAs by $\approx$ 100-fold [20] and the major class of secondary siRNAs (the so-called "22G RNAs") is loaded on proteins of the WAGO subfamily of the Argonaute family [22, 23]. WAGO proteins appear to be unable to cleave RNA targets [23]. Yet WAGO/secondary siRNA/cofactor complexes appear to be much more efficient at repressing mRNA targets than AGO/primary siRNA/cofactor complexes [24], possibly by recruiting another, unknown, nuclease. In contrast to Dicer products (which bear a 5´ monophosphate), direct RdRP products bear a 5´ triphosphate. 22G RNAs are thus triphosphorylated on their 5´ ends [20]. Another class of nematode RdRP products, the "26G RNAs", appears to bear a 5´ monophosphate, and it is not clear whether they are matured from triphosphorylated precursors, or whether they are directly produced as monophosphorylated RNAs [25–27].

The enzymatic activity of RNA-dependent RNA polymerization can be mediated by several unrelated protein families [28]. Most of these families are specific to viruses (*e.g.*, PFAM ID #PF00680, PF04196 and PF00978). Viral RdRPs are involved in genome replication and transcription in RNA viruses, and they share common structural motifs [29]. On the other hand, RdRPs involved in RNAi in plants, fungi and nematodes belong to a family named "eukaryotic RdRPs" (PFAM ID #PF05183). While viral RdRPs are conceivably frequently acquired by virus-mediated horizontal transfer, members of the eukaryotic RdRP family are thought to be inherited vertically only [30]. The eukaryotic RdRP family can be further divided into three subfamilies, named α, β and γ based on sequence similarity. Phylogenetic analyses suggest these three subfamilies derive from three ancestral RdRPs that could have coexisted in the most recent common ancestor of animals, fungi and plants [31].

Besides eukaryotic RdRPs, other types of RdRP enzymes have been proposed to exist in various animals. It has been suggested that human cells express an atypical

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RdRP, composed of the catalytic subunit of telomerase and a non-coding RNA [32]. While that complex exhibits RdRP activity *in vitro*, functional relevance of that activity is unclear, and other mammalian cells were shown to perform RNAi without RdRP activity [33]. More recently, bat species of the *Eptesicus* clade were shown to possess an RdRP of viral origin, probably acquired upon endogenization of a viral gene at least 11.8 million years ago [34].

Here we took advantage of the availability of hundreds of metazoan genomes to draw a detailed map of predicted RdRP genes in animals. We found RdRP genes in a large diversity of animal clades, even in insects, where they had escaped detection so far. Even though RdRP genes are found in diverse animal clades, they are lacking in many species, indicating that they were frequently and independently lost in many lineages. Furthermore, the presence of RdRP genes in non-nematode genomes raises the possibility that additional metazoan lineages possess an RdRP-based siRNA amplification mechanism. We sequenced small RNAs from various developmental stages in one such species with 6 candidate RdRP genes, the cephalochordate *Branchiostoma lanceolatum*, using experimental procedures that were designed to detect both 5´ mono- and tri-phosphorylated RNAs. Our analyses did not reveal any evidence of the existence of secondary siRNAs in that organism. While RNAi is the only known function for eukaryotic RdRPs, we thus propose that *Branchiostoma* RdRPs do not participate in RNAi.

Materials and methods

Bioinformatic analyses of protein sequences

Predicted animal proteome sequences were downloaded from the following databases: NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes/`), VectorBase (`https://www.vectorbase.org/download/`), FlyBase (`ftp://ftp.flybase.net/releases/FB2015_03/`), JGI (`ftp://ftp.jgi-psf.org/pub/JGI_data/`), Ensembl (`ftp://ftp.ensembl.org/pub/release-81/fasta/`), WormBase (`ftp://ftp.wormbase.org/pub/wormbase/species/`) and Uniprot (`http://www.uniprot.org/`). The predicted *Branchiostoma lanceolatum* proteome was obtained from the *B. lanceolatum* genome consortium. RdRP HMMer profiles were downloaded from PFAM v. 31.0 (`http://pfam.xfam.org/`): 19 viral RdRP family profiles (PF00602, PF00603, PF00604, PF00680, PF00946, PF00972, PF00978, PF00998, PF02123, PF03035, PF03431, PF04196, PF04197, PF05788, PF05919, PF07925, PF08467, PF12426, PF17501) and 1 eukaryotic RdRP family profile (PF05183). Candidate RdRPs were selected by **hmmsearch** with an E-value cutoff of 10^{-2}. Only those candidates with a complete RdRP domain according to NCBI's Conserved domain search tool (`https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi`) were considered (tolerating up to 20% truncation on either end of the domain). One identified candidate, in the bat *Rhinolophus sinicus*, appears to be a plant contaminant (it is most similar to plant RdRPs, and its genomic scaffold [ACC# LVEH01002863.1] only contains that gene): it was not included in Figure 1 and in Supplementary Fig. S1.

The *Branchiostoma* Hen1 candidate was identified using **HMMer** on the predicted *B. lanceolatum* proteome, with an HMMer profile built on an alignment of *Drosophila melanogaster*, *Mus musculus*, *Danio rerio*, *Nematostella vectensis* and *Arabidopsis thaliana* Hen1 sequences.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Phylogenetic tree reconstruction

Amino acid sequences of the eukaryotic RdRP domain (Pfam #PF05183) were retrieved from PFAM [35], and supplemented with the RdRP domains of the proteins identified in the 538 animal proteomes (*cf* above). Sequences were aligned using **hmmalign** [36] using the HMM profile of the PF05183 RdRP domain. Sequences for which the domain was incomplete were deteled from the alignment. Sites used to reconstruct the phylogenetic tree were selected using **trimAl** [37] on the Phylemon 2.0 webserver [38]. Bayesian inference (BI) tree was inferred using **MrBayes 3.2.6** [39], with the model recommended by **ProtTest 1.4** [40] under the Akaike information criterion (LG+Γ), at the CIPRES Science Gateway portal [41]. Two independent runs were performed, each with 4 chains and one million generations. A burn-in of 25% was used and a fifty majority-rule consensus tree was calculated for the remaining trees. The obtained tree was customized using **FigTree v.1.4.0**.

Sample collection

Mediterranean amphioxus (*Branchiostoma lanceolatum*) males and females were collected at le Racou (Argelès-sur-mer, France) and were induced to spawn as previously described [42]. Embryos were obtained after fertilization in Petri dishes filled with filtered sea water and cultivated at 19°C. Total RNA was extracted from 8, 15, 36 and 60 hours post fertilization (hpf) embryos (three independent batches for each stage, pooled before small RNA gel purification) as well as from males (6 pooled individuals) and females (4 pooled individuals) using the RNeasy mini kit (for embryonic samples) and the RNeasy midi kit (for adult samples) (Qiagen).

Sequencing analyses

The BL09945 locus was PCR-amplified from adult female DNA, cloned in the pGEM-T easy vector (cat. #A1360; Promega, Madison, WI, USA) and sequenced by MWG Eurofins Genomics (Ebersberg, Germany).

For Small RNA-Seq, 18–30 nt RNAs were gel-purified from total RNA (using between 92 and 228 μg total RNA per sample). One quarter of the small RNA preparation was kept untreated before library preparation (for "Libraries #1"). One quarter was incubated for 10 min at room temperature in 100 μL of freshly-prepared 60 mM sodium borate (pH=8.6), 25 mM sodium periodate, then the reaction was quenched with 10 μL glycerol (for "Libraries #2"). One quarter was treated with 1.25 U Terminator exonuclease (Epicentre, Madison, WI, USA) in 25 μL 1X Terminator reaction buffer A for 1h at 30°C, then the reaction was quenched with 1.25 μL 500 mM EDTA (pH=8.0) and ethanol-precipitated. RNA was then treated with 5 U Antarctic phosphatase (New England Biolabs, Ipswich, MA, USA) in 20 μL 1X Antarctic phosphatase buffer for 30 min at 37°C, the enzyme was heat-inactivated, then RNA was precipitated, then phosphorylated by 15 U T4 PNK (New England Biolabs) with 50 nmol ATP in 50 μL 1X T4 PNK buffer for 30 min at 37°C, then the enzyme was heat-inactivated (for "Libraries #3"). One quarter was treated successively with Terminator exonuclease, Antarctic phosphatase, T4 PNK then boric acid and sodium periodate, with the same protocols (for "Libraries #4"). Small RNA-Seq libraries were then generated using the TruSeq Small RNA library preparation kit (Illumina, San Diego, CA, USA), following the manufacturer's instructions.

Libraries were sequenced by the MGX sequencing facility (CNRS, Montpellier, France). Read sequences were aligned on the *B. lanceolatum* genome assembly [43] using **bowtie2**. A database of abundant non-coding RNAs was assembled by a search for orthologs for human and murine rRNAs, tRNAs, snRNAs, snoRNAs and scaRNAs;

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

deep-sequencing libraries were also mapped on that database using **bowtie2**, and matching reads were flagged as "abundant ncRNA fragments". For pre-miRNA annotation, every *B. lanceolatum* locus with a Blast E-value $\leqslant 10^{-6}$ to any of the annotated *B. floridae* or *B. belcheri* pre-miRNA hairpins in miRBase v.22 was selected. Reads matching these loci were identified using **bowtie2**. For the measurement of miRNA abundance during development, hairpins were further screened for their **RNAfold**-predicted secondary structure and their read coverage: Supplementary Table S1 only lists unbranched hairpins with at least 25 bp in their stem, with a predicted $\Delta G_{folding} \leqslant -15\ kcal.mol^{-1}$, generating mostly 21- to 23-mer RNAs, and with at least 20 ppm read coverage on any nucleotide of the hairpin.

RNA-Seq data was taken in [43] for embryonic and juvenile samples. Adult sample libraries were prepared and sequenced by "Grand plateau technique régional de génotypage" (SupAgro-INRA, Montpellier). mRNA abundance data was extracted using **vast-tools** [44].

Extragenomic contig assembly and annotation

Small RNA reads that fail to map on the *B. lanceolatum* genome or transcriptome according to **bowtie2** were collected and assembled using **velvet** [45], with k values ranging from 9 to 19 for better sensitivity [46].

Contigs at least 50 bp in length were then compared to the NCBI non-redundant nucleotide collection (as of October 31, 2018) by **megablast** on the NCBI server with default parameters. Contigs with a detected similarity to known sequences in the collection were annotated with phylogenetic information (see Table 2) using the NCBI "Taxonomy" database.

Code availability

Source code, detailed instructions, and intermediary data files are accessible on GitHub (`https://github.com/HKeyHKey/Pinzon_et_al_2018`) as well as on `https://www.igh.cnrs.fr/en/research/departments/genetics-development/systemic-impact-of-small-regulatory-rnas/165-computer-programs`.

Data availability

Deep-sequencing data has been deposited at NCBI's Short Read Archive under BioProject accession #PRJNA419760 (for Small RNA-Seq) and BioSample accession #SAMN09381006 and SAMN09381007 (for adult RNA-Seq). Sequences of the re-sequenced *B. lanceolatum* BL09945 locus have been deposited at GenBank under accession #MH261373 and #MH261374.

Results

A sporadic phylogenetic distribution of RdRP genes

Previous analyses showed that a few animal genomes contain candidate RdRP genes [28, 31, 34, 47]. Rapid development of sequencing methods recently made many animal genomes available, allowing a more complete coverage of the phylogenetic tree. A systematic search for RdRP candidates (including every known viral or eukaryotic RdRP family) in 538 predicted metazoan proteomes confirms that animal species possessing RdRPs are unevenly scattered in the phylogenetic tree, but they are much more abundant than previously thought: we identified 98 metazoan species with convincing eukaryotic RdRP genes (see Figure 1A). Most RdRPs identified in animal

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

predicted proteomes belong to the eukaryotic RdRP family, but 3 species (the Enoplea *Trichinella murrelli*, the Crustacea *Daphnia magna* and the Mesozoa *Intoshia linei*) possess RdRP genes belonging to various viral RdRP families (in green, dark blue and light blue on Figure 1A), which were probably acquired by horizontal transfer from viruses. Most sequenced nematode species appear to possess RdRP genes. But in addition, many other animal species are equipped with eukaryotic RdRP genes, even among insects (the Diptera *Clunio marinus* and *Rhagoletis zephyria*), where RdRPs were believed to be absent [47, 48].

Fig 1. Phylogenetic distribution of RdRP genes in metazoans. A. Proteome sequences from 538 metazoans were screened for potential RdRPs. For each clade indicated on the right edge, n is the number of species analyzed in the clade, and piecharts indicate the proportion of species possessing RdRP genes (with each RdRP family represented by one piechart, according to the color code given at the top left). B. An HMMer search identifies 6 candidate RdRPs in the predicted *Branchiostoma lanceolatum* proteome. Only 2 candidates have a complete RdRP domain (represented by a red bar with round ends; note that apparent domain truncations may be due to defective proteome prediction). A white star indicates that every catalytic amino acid is present. Candidate BL02069 also possesses an additional known domain, AAA_12 (in yellow).

Our observation of eukaryotic family RdRPs in numerous animal clades therefore prompted us to revisit the evolutionary history of animal RdRPs: eukaryotic RdRPs were probably present in the last ancestors for many animal clades (including insects, mollusks, deuterostomes) and they were subsequently lost independently in most insects, mollusks and deuterostomes. It has been recently shown that the last ancestor of arthropods possessed an RdRP, which was subsequently lost in some lineages [47]: that result appears to be generalizable to a large diversity of animal clades. The apparent absence of RdRPs in some species may be due to genome incompleteness, or to defective proteome prediction. Excluding species with low numbers of long predicted proteins ($\geqslant$ 500 or 1,000 amino acids) indeed eliminates a few dubious proteomes, but the resulting distribution of RdRPs in the phylogenetic tree is only marginally affected, and still suggests multiple recent RdRP losses in diverse lineages (see Supplementary Fig. S1).

Alternatively to multiple gene losses, such a sporadic phylogenetic distribution could be due to frequent horizontal transfer of RdRP genes in animals. In order to assess these two possibilities, it is important to better understand the evolution of metazoan RdRPs in the context of the whole eukaryotic RdRP family. We therefore used sequences found in all eukaryotic groups for phylogenetic tree reconstruction. The supports for deep branching are low and do not allow us to propose a complete evolutionary history scenario of the whole eukaryotic RdRP family (see Figure 2A). However, metazoan sequences are forming three different groups, which were named RdRP α, β and γ according to the pre-existing nomenclature [31], and their position in relation to non-metazoan eukaryotic sequences does not support an origin through horizontal gene transfer. The only data that would support horizontal gene transfer pertains to the metazoan sequences of the RdRP β group (see Figure 2C). Indeed, sequences of stramenopiles and a fungus belonging to parasitic species are embedded in this clade. For the RdRP α and γ groups, the phylogeny strongly suggests that they derive from at least two genes already present in the common ancestor of cnidarians and bilaterians and that the scarcity of RdRP presence in metazoans would be the result of many secondary gene losses. Even the *Strigamia maritima* RdRP was probably not acquired by a recent horizontal transfer from a fungus, as has been proposed [47]: when assessed against a large number of eukaryotic RdRPs, the *S. maritima* sequence clearly clusters within metazoan γ RdRP sequences. In summary, we conclude that RdRPs

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

were present in the last ancestors of many animal clades, and they were recently lost independently in diverse lineages.

Fig 2. Eukaryotic RdRP phylogeny supports the vertical transfer scenario. Bayesian phylogenetic tree of the eukaryotic RdRP family. α, β and γ clades of eukaryotic RdRPs have been defined by [31]. Sectors highlighted in grey are detailed in panels B, C and D for clarity. Scale bar: 0.4 amino acid substitution per position. Posterior probability values are indicated for each node in panels B–D.

Experimental search for RdRP products in *Branchiostoma*

In an attempt to probe the functional conservation of RdRP-mediated RNAi amplification among metazoans, we decided to search for secondary siRNAs in an organism where RdRP candidates could be found, while being distantly related to *C. elegans*. We reasoned that endogenous RNAi may act as a gene regulator during development or as an anti-pathogen response. Thus siRNAs are more likely to be detected if several developmental stages are probed, and if the analyzed specimens are gathered in a natural ecosystem, where they are naturally challenged by pathogens. From these considerations it appears that the most appropriate organism is a cephalochordate species, *Branchiostoma lanceolatum* [49]. In good agreement with the known scarcity of gene loss in that lineage [50], cephalochordates also constitute the only bilaterian clade for which both RdRP α and γ sequences can be found, thus increasing the chances of observing RNAi amplification despite the diversification of eukaryotic RdRPs into three groups. According to our HMMer-based search, the *B. lanceolatum* genome encodes 6 candidate RdRPs, three of which containing an intact active site DbDGD (with b representing a bulky amino acid; [51]) (see Figure 1B). The current *B. lanceolatum* genome assembly contains a direct 1,657 bp repeat in one of the 6 RdRP genes, named BL09945. This long duplication appears to be an assembly artifact: we cloned and re-sequenced that locus and identified two alleles (with a synonymous mutation on the 505th codon; deposited at GenBank under accession numbers MH261373 and MH261374), and none of them contained the repeat. In subsequent analyses, we thus used a corrected version of that locus, where the 1,657 bp duplication is removed.

In most metazoan species, siRNAs (as well as miRNAs) bear a 5´ monophosphate and a 3´ hydroxyl [52, 53]. The only known exceptions are "22G" secondary siRNAs in nematodes (they bear a 5´ triphosphate; [20]), which may be primary polymerization products by an RdRP; Ago2-loaded siRNAs and miRNA in *Drosophila*, which are 3´-methylated on their 2´ oxygen after loading on Ago2 and unwinding [54, 55]; and a subset of "26G" secondary siRNAs in nematodes (those which are loaded on the ERGO-1 Argonaute protein), which also bear a 2´-*O*-methyl on their 3´ end [56–58].

In order to detect small RNAs with any number of 5´ phosphates, bearing either an unmodified or a methylated 3´ end, we prepared multiple Small RNA-Seq libraries (see Figure 3A). Total RNA was extracted from various embryonic stages: gastrula (8 hours post-fertilization, hpf), early neurula (15 hpf), premouth neurula (36 hpf) and larvae (60 hpf), as well as from adult male and female specimens collected from their natural ecosystem. Small (18 to 30 nt long) RNAs were gel-purified, then Small RNA-Seq libraries were prepared using either the standard Small RNA-Seq protocol (which detects 5´ monophosphorylated small RNAs, whether they bear a 3´ methylation or not; "Library #1"); or by oxidizing small RNAs with $NaIO_4$ in the presence of H_3BO_3 prior to library preparation (such treatment renders unmodified 3´ RNAs non-ligatable, hence undetectable by deep-sequencing; [59]; "Library #2"); or by treating small RNAs with the Terminator exonuclease (which degrades 5´ monophosphorylated RNAs) then with

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

phosphatase then T4 PNK (to convert 5´ polyphosphorylated RNAs and 5´ hydroxyl RNAs into monophosphorylated RNAs, suitable for Small RNA-Seq library preparation; "Library #3"); or by a combination of both treatments (to detect only small RNAs bearing a 5´ polyphosphate or a 5´ hydroxyl, and a 3´ modification; "Library #4"). If the same experiments were performed in classical animal model organisms, such as *Drosophila*, nematodes and vertebrates (where miRNAs are essentially 5´ monophosphorylated and 3´-unmodified, and piRNAs are 5´ monophosphorylated and 3´-methylated), miRNAs would be expected to be detected in Libraries #1 and piRNAs, in Libraries #1 and 2. Nematode "22G" siRNAs would be detected in Libraries #3.

Fig 3. Detection of *B. lanceolatum* small RNAs. Four libraries were prepared for each biological sample, to detect small RNAs bearing either a single 5´ phosphate (Libraries #1 and 2) or any other number of phosphates (including zero; Libraries #3 and 4), and either a 2´-OH and 3´-OH 3´ end (Libraries #1 and 3), or a protected (*e.g.*, 2´-*O*-methylated) 3´ end (Libraries #2 and 4). hpf: hours post fertilization. B. Size distribution of genome-matching adult male small RNAs, excluding reads that match abundant non-coding RNAs (rRNAs, tRNAs, snRNAs, snoRNAs or scaRNAs). Read numbers are normalized by the total number of genome-matching reads (including <18 nt and >30 nt reads) that do not match abundant non-coding RNAs, and expressed as *parts per million* (ppm). C. Size distribution of adult male small RNAs matching pre-miRNA hairpins in the sense (blue) or antisense (red) orientation.

In the course of library preparation, it appeared that Libraries #4 contained very little ligated material, suggesting that small RNAs with a 3´ modification as well as $n \geqslant 0$ (with $n \neq 1$) phosphates on their 5´ end, are very rare in *Branchiostoma* regardless of developmental stage. This observation was confirmed by the annotation of the sequenced reads: most reads in Libraries #4 did not map on the *B. lanceolatum* genome, probably resulting from contaminating nucleic acids (see Supplementary Fig. S2).

In Libraries #1 in each developmental stage, most *Branchiostoma* small RNA reads fall in the 18–30 nt range as expected. Other libraries tend to be heavily contaminated with shorter or longer reads, and 18–30 nt reads only constitute a small fraction of the sequenced RNAs (see Figure 3B for adult male libraries; see Supplementary File S1 section 1 for other developmental stages). miRNA loci have been annotated in two other cephalochordate species, *B. floridae* and *B. belcheri* (156 pre-miRNA hairpins for *B. floridae* and 118 for *B. belcheri* in miRBase v. 22). We identified the *B. lanceolatum* orthologous loci for annotated pre-miRNA hairpins from *B. floridae* or *B. belcheri*. Mapping our libraries on that database allowed us to identify candidate *B. lanceolatum* miRNAs. These RNAs are essentially detected in our Libraries #1, implying that, like in most other metazoans, *B. lanceolatum* miRNAs are mostly 22 nt long, they bear a 5´ monophosphate and no 3´ methylation (see Figure 3C for adult male libraries; see Supplementary File S1 section 2 for other developmental stages). Among the *B. lanceolatum* loci homologous to known *B. floridae* or *B. belcheri* pre-miRNA loci, 56 exhibit the classical secondary structure and small RNA coverage pattern of pre-miRNAs (*i.e.*, a stable unbranched hairpin generating mostly 21–23 nt long RNAs from its arms). These 56 loci, the sequences of the miRNAs they produce, and their expression profile during development, are shown in Supplementary Table S1.

No evidence of RdRP-based siRNA amplification in *Branchiostoma*.

In an attempt to detect siRNAs, we excluded every sense pre-miRNA-matching read and searched for distinctive siRNA features in the remaining small RNA populations.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Whether RdRPs generate long antisense RNAs which anneal to sense RNAs to form a substrate for Dicer, or whether they polymerize directly short single-stranded RNAs which are loaded on an Argonaute protein, the involvement of RdRPs in RNAi should result in the accumulation of antisense small RNAs for specific target genes. These small RNAs should exhibit characteristic features:

- a narrow size distribution (imposed either by the geometry of the Dicer protein, or by the processivity of the RdRP [24, 60]; the length of Argonaute-loaded RNAs can also be further refined by exonucleolytic trimming of 3´ ends protruding from Argonaute [22, 61–65]);
- and possibly a sequence bias on their 5´ end; it is remarkable that the known classes of RdRP products in metazoans (nematode 22G and 26G RNAs) both display a strong bias for a guanidine at their 5´ end. RNA polymerases in general tend to initiate polymerization on a purine nucleotide [66–72] and it can be expected that primary RdRP products bear either a 5´A or a 5´G. Of note: loading on an Argonaute may also impose a constraint on the identity of the 5´ nucleotide, because of a sequence preference of either the Argonaute protein or its loading machinery [73–78].

The analysis of transcriptome-matching, non-pre-miRNA-matching small RNAs does not indicate that such small RNAs exist in *Branchiostoma* (see Figure 4 for adult males, and Supplementary File S1, section 3, for the complete data set). In early embryos, 5´ monophosphorylated small RNAs exhibit the typical size distribution and sequence biases of piRNA-rich samples: a heterogeneous class of 23 to 30 nt long RNAs. Most of them tend to bear a 5´ uridine, but 23 to 26 nt long RNAs in the sense orientation to annotated transcripts tend to have an adenosine at position 10 (especially when the matched transcript exhibits a long ORF; see Supplementary File S1, section 4). Vertebrate and *Drosophila* piRNAs display very similar size profiles and sequence biases [79–85]. These 23–30 nt long RNAs may thus constitute the *Branchiostoma* piRNAs, but surprisingly, they do not appear to bear a 2´-*O*-methylation on their 3´ end (see Discussion). Note that piRNAs appear to be mostly restricted to the germ line and gonadal somatic cells in other model organisms. But they are so abundant in piRNA-expressing cells, and so abundantly maternally deposited in fertilized eggs, that they can still be readily detected in embryonic or adult whole-body small RNA samples [25, 86–90]. It is thus not surprising to observe piRNA candidates in our *Branchiostoma* whole-body Small RNA-Seq libraries.

Fig 4. Size distribution and sequence logos for transcriptome-matching small RNAs in adult males. See Supplementary File S1, section 3, for the other developmental stages. A: Library #1, B: Library #2, C: Library #3, D: Library #4. Numbers of reads are expressed as *parts per million* (ppm) after normalization to the total number of genome-matching reads that do not match abundant non-coding RNAs. For each orientation (sense or antisense-transcriptome-matching reads), a logo analysis was performed on each size class (18 to 30 nt long RNAs).

In summary, transcriptome-matching small RNAs in our *Branchiostoma* libraries contain miRNA and piRNA candidates, but they do not contain any obvious class of presumptive secondary siRNAs that would exhibit a precise size distribution, and possibly a 5´ nucleotide bias. If *Branchiostoma* RdRPs generated secondary siRNAs by polymerizing mature short antisense RNAs (similarly to nematode 22G RNAs according to the prevalent model), then such hypothetical siRNAs should be detected in libraries #3. If *Branchiostoma* RdRPs generated long antisense RNAs, that would anneal to sense RNAs to produce a Dicer substrate (similarly to fungus and plant

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RdRP-derived siRNAs according to the prevalent model), then secondary siRNAs should be detected in libraries #1. As we did not observe candidate siRNA populations in either libraries #1 or 3, our data seem to rule out the existence of secondary siRNAs in *Branchiostoma*, regardless of the mechanistical involvement of RdRPs in their production.

One could imagine that transcriptome-matching siRNAs were missed in our analysis, because of issues with the *Branchiostoma* transcriptome assembly. It is also conceivable that siRNAs exist in *Branchiostoma*, but they do not match its genome or transcriptome (they could match pathogen genomes, for example if they contribute to an anti-viral immunity). We therefore analyzed other potential siRNA types: **(i)** genome-matching reads that do not match abundant non-coding RNAs (rRNAs, tRNAs, snRNAs, snoRNAs or scaRNAs); **(ii)** reads that match transcripts exhibiting long ($\geqslant$ 100 codons, initiating on one of the three 5´-most AUG codons) open reading frames; **(iii)** reads that do not match the *Branchiostoma* genome, nor its transcriptome (potential siRNAs derived from pathogens). Once again, none of these analyses revealed any siRNA population in *Branchiostoma* (see detailed results in Supplementary File S1, sections 1, 4 and 5). This is in striking contrast to *Cænorhabditis elegans*, where antisense transcriptome-matching siRNAs (mostly 22 nt long, starting with a G) are easily detectable (see Supplementary File S1, section 6, for our analysis of publicly available *C. elegans* data; [22]).

Branchiostoma RdRP activity is not clearly detected

Our failure to detect siRNA candidates may simply be due to the fact that they are poorly abundant in the analyzed developmental stages. In order to enrich for small RNA populations derived from RdRP activity, and exclude all the other types of small RNAs, we considered small RNAs mapping on exon-exon junctions in the antisense orientation. The antisense sequence of the splicing donor (GU) and acceptor (AG) sites does not constitute a donor/acceptor pair itself, implying that any RNA antisense to a spliced RNA must have originated from the action of an RdRP on the spliced RNA — it cannot derive from the splicing of an RNA transcribed in the antisense orientation.

We therefore selected all the 18–30 nt RNA reads that map on exon-exon junctions in the annotated transcriptome, and fail to map on the genome. Such reads map almost exclusively in the sense orientation (see Table 1). When focusing on the developmental stage where some transcripts exhibit the highest observed numbers of antisense exon-exon junction reads (15 hpf embryos, for the transcripts of genes BL05604 and BL00515), it appears that these antisense junction reads are highly homogeneous in sequence (sharing the same 5´ and 3´ ends), they do not map perfectly on the spliced transcript (with 1 mismatch in each), and their total abundance remains very small (less than 10 raw reads per transcript in a given developmental stage) (see Supplementary Fig. S3). RdRP genes themselves appear to be developmentally regulated, with candidate RdRPs harboring intact active sites showing expression peaks at 8 and 18 hpf (see Supplementary Fig. S4).

It is formally possible that the few antisense exon-exon junction reads that we detected derive from an RNA polymerized by an RdRP. But their scarcity, as well as their extreme sequence homogeneity, suggests that they rather come from other sources (*e.g.*, DNA-dependent RNA polymerization, either from a *Branchiostoma* genomic locus or from a non-*Branchiostoma* contaminant) and map fortuitously on the BL05604 or BL00515 spliced transcript sequences. We note that *C. elegans* secondary siRNAs are highly diverse in sequence, and even low-throughput sequencing identifies antisense reads mapping on distinct exon-exon junctions [20]. We thus tend to attribute our observation of rare antisense exon-exon junction small RNAs to rare contaminants or sequencing errors, rather than to genuine RNA-dependent RNA polymerization in

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1. Genes with highest read coverage on exon-exon junctions in the sense or antisense orientation.

Whole transcriptome					
Sense (ppm)			Antisense (ppm)		
94.2			1.33		
Top genes with sense-matching junction reads			Top genes with antisense-matching junction reads		
Gene name	Sense reads (ppm)	Antisense reads (ppm)	Gene name	Sense reads (ppm)	Antisense reads (ppm)
BL01573	4.11	0	BL05604	0	0.348
BL21967	2.89	0	BL00515	0.0332	0.149
BL11086	2.29	0	BL16381	0	0.0995
BL01498	2.27	0	BL06097	0.0663	0.0663
BL23908	1.41	0	BL13214	0	0.0497
BL06958	1.38	0	BL06086	0.0497	0.0497
BL23502	1.18	0	BL05692	0	0.0332
BL20132	0.895	0	BL11851	0.0332	0.0332
BL10273	0.779	0	BL27707	0.0166	0.0332
BL06284	0.763	0	BL01135	0.0166	0.0332

Genes were sorted according to the total number of small RNA reads mapping on their exon-exon junctions in our pooled 24 Small RNA-Seq libraries. Numbers of mapped reads were normalized by the total number of genome-matching reads that do not match abundant non-coding RNAs, and expressed as *parts per million* (ppm). Top part: statistics for the whole transcriptome. Bottom part: only the top 10 genes in terms of sense-matching junction reads (left half of the table) or antisense-matching junction reads (right half of the table) are shown.

Branchiostoma.

Candidate *Branchiostoma* pathogens do not appear to be targeted by RNAi

In various other organisms, RNAi participates in the defence against pathogens (reviewed in [91]). Pathogen-specific siRNAs may exist in *Branchiostoma*, and they may have been too poorly abundant to be detected in our analyses of extragenomic, extratranscriptomic reads (see Supplementary File S1, section 5). We thus decided to interrogate specifically the populations of small RNAs mapping on *Branchiostoma* pathogen genomes. Several pathogenic bacteria (*Staphylococcus aureus*, *Vibrio alginolyticus* and *Vibrio anguillarum*; [92, 93]) have been described in various *Branchiostoma* species. We asked whether RNAi could target those pathogens *in vivo*. Focusing on the small RNA reads that do not map on the *Branchiostoma* genome or transcriptome, we observed large numbers of small RNAs deriving from these three bacterial genomes, indicating that the analyzed *Branchiostoma* specimens were in contact with those pathogens (after excluding reads that map simultaneously on 2 or 3 of these bacterial genomes, we detected 1,457,122 *S. aureus*-specific reads, 113,398 *V. alginolyticus*-specific reads and 103,153 *V. anguillarum*-specific reads in the pooled 24 Small RNA-Seq libraries; for reference: there are 125,550,314 *Branchiostoma* genome-matching reads in the pooled libraries). Small RNAs mapping on these pathogenic bacterial genomes do not display any obvious size distribution or sequence bias, thus suggesting that they constitute degradation products from longer bacterial RNAs rather than siRNAs (see Supplementary File S1, sections 7–9).

Our analyzed *Branchiostoma* specimens may also have been challenged by yet-unknown pathogens. Pooling every read that does not map on the *Branchiostoma* genome or transcriptome, across all 24 Small RNA-Seq libraries, offers the opportunity to reconstruct genomic contigs for the most abundant non-*Branchiostoma* sequences. In

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

total, we collected 23,557,012 such extragenomic, extratranscriptomic reads. 42,946 contigs at least 50 bp long could be assembled from these reads using **velvet** [45]. Of these, 4,804 contigs could be annotated by homology search (see Table 2): 291 appear to match the *Branchiostoma* genome, and the reads supporting these contigs had probably failed to map properly on the genome because of sequencing errors or sequence polymorphism.

Table 2. Annotation of reconstructed genomic contigs assembled from extragenomic, extratranscriptomic Small RNA-Seq reads.

Contig origin	Number of contigs	Contig length (bp) (mean ± st. dev.)
Total	42,946	60.1 (±10.7)
Prokaryotes	3,552	65.9 (±18.3)
Fungi	724	66.9 (±18.9)
Viridiplantae	67	63.3 (±10.3)
non-*Branchiostoma* Metazoans	131	62.8 (±12.3)
Branchiostoma	291	64.4 (±13.8)
Prokaryotic viruses	35	81.6 (±28.5)
Eukaryotic viruses	4	63.8 (±5.91)
Ambiguous, poorly described and others	38,142	59.3 (±8.92)

After assembling extragenomic, extratranscriptomic reads from all 24 libraries, contigs longer than 50 bp were annotated by a **blast** search on the NCBI non-redundant nucleic acid database.

We screened these contigs for potential *Branchiostoma* pathogens, which could be targeted by RNAi. Detected prokaryotic, fungal or non-*Branchiostoma* metazoan sequences may derive from symbiotic or commensal species rather than actual pathogens. Our analyzed adult specimens were collected from the natural environment, where unrelated organisms are expected to contaminate the samples; and our analyzed embryos were produced from gametes collected in non-sterile sea water. Following spawning, these gametes transit through the "atrium" (an open body cavity that putatively hosts various micro-organisms): so *in vitro*-fertilized embryos are also likely to be contaminated with non-pathogenic non-*Branchiostoma* species.

But we also observed several viral contigs, including 4 contigs from eukaryotic viruses. Three of them are matched by low numbers of small RNA reads, but the last one (a contig matching the *Acanthocystis turfacea* Chlorella virus 1 genome) is covered with high read counts in various developmental stages (see Supplementary Fig. S5). That virus is known to infect endosymbiotic algae of the protist *Acanthocystis turfacea*, and some reports suggest that it may also infect mammalian hosts [94], suggesting a broad tropism. Though still disputed [95, 96], this observation could suggest that *Branchiostoma* may also be sensitive to that virus. Yet, for this potential pathogen too, detected small RNA reads fail to display any size or sequence bias: they do not appear to be siRNAs (see Supplementary File S1, section 10).

Finally, we considered the possibility that some of the 38,142 un-annotated extragenomic contigs (see Table 2) may originate from unknown pathogens. We selected the 5 contigs displaying the highest read coverage (more than 200 ppm after pooling all 24 Small RNA-Seq libraries): small RNAs mapping on these hypothetical unknown pathogens also do not exhibit particular size or sequence biases, arguing against their involvement in RNAi (see Supplementary File S1, sections 11–15).

Because unambiguous RdRP-derived small RNAs could not be detected with certainty despite our efforts, and because we did not observe any small RNA population with classical siRNA size or sequence bias, we conclude that *Branchiostoma* RdRP genes are not involved in RNAi.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Discussion

In cellular organisms, the only known function for RdRPs is the generation of siRNAs or siRNA precursors. It is thus frequently assumed [32, 47] or hypothesized [34] that animal RdRPs participate in RNAi. In particular, it has recently been proposed that arthropod RdRPs are required for RNAi amplification, and arthropod species devoid of RdRPs may rather generate siRNA precursors through bidirectional transcription [47]. While this hypothesis would provide an elegant explanation to the sporadicity of RdRP gene distribution in the phylogenetic tree, the provided evidence remains disputable: it has been proposed that a high ratio of antisense over sense RNA is diagnostic of bidirectional transcription, yet it remains to be explained why RNA-dependent RNA polymerization would produce less steady-state antisense RNA than DNA-dependent polymerization.

Branchiostoma 5´ monophosphorylated small RNAs do not appear to bear a 2´-*O*-methyl on their 3´ end: Libraries #2 contain few genome-matching sequences, and their size distribution suggests they are mostly constituted of contaminating RNA fragments rather than miRNAs, piRNAs or siRNAs. In every animal model studied so far, piRNAs were shown to bear a methylated 3´ end [25, 56–58, 85, 87, 97–99]. The enzyme responsible for piRNA methylation, Hen1 (also known as Pimet in *Drosophila*, HENN-1 in nematodes), has been identified in *Drosophila*, mouse, zebrafish and nematodes [55–58, 100–102]. In order to determine whether the absence of piRNA methylation in *Branchiostoma* could be due to an absence of the Hen1 enzyme, we searched for Hen1 orthologs in the predicted *Branchiostoma* proteome. Our HMMer search identified a candidate, BL03504. Its putative methyl-transferase domain contains every known important amino acid for Hen1 activity according to [103] (see Supplementary Fig. S6), suggesting that it is functional. Further studies will be required to investigate the biological activity of that putative enzyme, and to understand why it does not methylate *Branchiostoma* piRNAs.

Focusing on small RNA reads mapping on exon-exon junctions in the antisense orientation, we did not observe convincing evidence of RdRP activity in *Branchiostoma*. Even if RdRPs do not participate in RNAi, it could have been anticipated that Small RNA-Seq libraries could capture short degradation products of RdRP-polymerized long RNAs. This observation raises the possibility that the *Branchiostoma* RdRP genes do not express any active RdRP. At least these genes are transcribed: analysis of gene expression in long RNA-Seq data [43] shows a dynamic regulation, especially for the three genes with an intact predicted active site (see Supplementary Fig. S4).

One could hypothesize that these RdRPs do not play any biological function. Yet at least two of them, BL02069 and BL23385, possess a full-length RdRP domain with a preserved catalytic site. The conservation of these two intact genes suggests that they are functionally important. It can therefore be speculated that *Branchiostoma* RdRPs play a biological role, which is unrelated to RNAi. Such a function may involve the generation of double-stranded RNA (formed by the hybridization of template RNA with the RdRP product), but it could also involve single-stranded RdRP products. Future work will be needed to identify the biological functionality of these enzymes. We also note that the fungus *Aspergillus nidulans*, whose genome encodes two RdRPs with a conserved active site, does not require any of those for RNAi [104].

Animal RdRPs thus constitute an evolutionary enigma: not only have they been frequently lost independently in numerous animal lineages, but even in the clades where they have been conserved, their biological function seems to be variable. While RNAi is an ancient gene regulation pathway [1], involving the deeply conserved Argonaute and Dicer protein families, the role of RdRPs in RNAi appears to be accessory. Even though RdRPs are strictly required for RNAi in very diverse extant clades (ranging from nematodes to plants), it would be misleading to assume that RNAi constitutes their

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

only biological function.

Supporting information

Fig. S1 Exclusion of dubious proteomes still indicates many independent RdRP losses. (file 'Supplementary_Figure_after_proteome_selection.pdf') Among the 538 analyzed proteomes, 442 contain at least 1,000 proteins of at least 1,000 amino acids (left panel) and 383 contain at least 5,000 proteins of at least 500 amino acids (right panel). Selective analysis of these species does not fundamentally change the results shown in Figure 1A. Same conventions than in Figure 1A. Some clades analyzed in Figure 1A could not be analyzed here after proteome exclusion: they are shown in grey.

Fig. S2 Size and quality of the Small RNA-Seq libraries. (file 'Supplementary_Figure_mapping_statistics.pdf') "No adapter" indicates that the 3´ adapter was not detected in the read. "Extragenomic" means that the adapter-trimmed read does not match the *B. lanceolatum* genome assembly. "Abundant ncRNA" means that it maps on the genome assembly, on one of the genes for known abundant non-coding RNAs (rRNAs, tRNAs, snRNAs, snoRNAs, scaRNAs). "Genome mapper, not matching abundant ncRNAs" means that it maps elsewhere in the genome assembly.

Fig. S3 Small RNA coverage in 15 hpf embryos for the two genes with highest antisense exon-exon junction read coverage. (file 'Supplementary_RdRP_template_loci_with_coverage.pdf') Exons are represented by black rectangles. Detected small RNAs mapping on these genes in the sense orientation are shown in blue, those mapping in antisense orientation are in red. For antisense reads mapping on exon-exon junctions, their precise sequence (in red) is aligned with the gene sequence (in black; splicing donor and acceptor sites are in green).

Fig. S4 Transcriptomics-based expression analysis of the 6 *Branchiostoma* RdRP genes. (file 'Supplementary_Figure_S4.pdf') For each of the six RdRP genes, mRNA abundance in various developmental stages was measured by RNA-Seq, and reported as cRPKM (corrected-for-mappability reads per kb and per million mapped reads; [105]). RdRP genes where an intact active site is predicted (see Fig. 1B) are annotated "with active site". Adult RNA-Seq data is from NCBI's BioSample accession #SAMN09381006 and SAMN09381007, other stages are from [43]. Adult male and female data were averaged. Temporal regulation of RdRP expression in embryos and juveniles was assessed by the Kruskal-Wallis test (p-values are indicated in the legend for each RdRP).

Fig. S5 Small RNA coverage of the *Acanthocystis turfacea* Chlorella virus 1 (ATCV1) genome. (file 'Supplementary_Figure_ATCV1_coverage.pdf') x axis: genomic coordinate along the ATCV1 genome. y axis: number of reads covering each bp in the viral genome. Numbers of reads are expressed as *parts per million* (ppm) after normalization to the total number of *Branchiostoma* genome-matching reads that do not match abundant non-coding RNAs.

Fig. S6 A *Branchiostoma* Hen1 candidate contains the known essential amino acids for Hen1 activity. (file 'Supplementary_Figure_Hen1.pdf') Sequences of 5 known Hen1 proteins (from *Nematostella vectensis*, *Danio rerio*, *Mus musculus*, *Arabidopsis thaliana* and *Drosophila melanogaster*) were aligned with the identified

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Branchiostoma lanceolatum Hen1 candidate (only the part of the alignment spanning amino acids 661–939 of the *Arabidopsis* protein is shown). Alignment was performed with **t-coffee** (version 11.00.8cbe486); other alignment programs (**Clustal Omega v.1.2.4**, **t-coffee v.8.93**, **Kalign v.2.03**, **MAFFT v.7.215**, but not **muscle v.3.8.31**) give the same main result: amino acids and amino acid combinations required for Hen1 catalytic activity [103] are conserved in the *Branchiostoma* candidate. Amino acids boxed in red were shown to be essential for *Arabodipsis* Hen1 activity; in orange: amino acids whose absence affects Hen1 activity without abolishing it entirely. Amino acid numbering is based on the *Arabidopsis* sequence.

File S1. Size distribution and logo analyses of various small RNA classes. (file 'Supplementary_Data.pdf') For each of the following classes, small RNA populations were analyzed as in Figure 3B, 3C and 4: reads matching the *B. lanceolatum* genome without matching abundant non-coding RNAs (section 1); reads matching *B. lanceolatum* pre-miRNA hairpins (section 2); reads matching the *B. lanceolatum* transcriptome without matching pre-miRNAs or abundant non-coding RNAs (section 3); reads matching *B. lanceolatum* mRNAs with long ORFs (section 4); reads not matching the *B. lanceolatum* genome or transcriptome (section 5); *C. elegans* small RNAs cloned with a procedure detecting 5´ mono- and polyphosphorylated RNAs [22] (section 6); reads not matching the *B. lanceolatum* genome or transcriptome, and matching the *Staphylococcus aureus* genome (section 7); reads not matching the *B. lanceolatum* genome or transcriptome, and matching the *Vibrio alginolyticus* genome (section 8); reads not matching the *B. lanceolatum* genome or transcriptome, and matching the *Vibrio anguillarum* genome (section 9); reads not matching the *B. lanceolatum* genome or transcriptome, and matching the *Acanthocystis turfacea* Chlorella virus 1 (ATCV1) genome (section 10); reads not matching the *B. lanceolatum* genome or transcriptome, and matching non-*Branchiostoma* contig #18690 (covered with 1,982.33 ppm small RNA reads across all 24 libraries) (section 11); reads not matching the *B. lanceolatum* genome or transcriptome, and matching non-*Branchiostoma* contig #7601 (covered with 1,534.35 ppm small RNA reads across all 24 libraries) (section 12); reads not matching the *B. lanceolatum* genome or transcriptome, and matching non-*Branchiostoma* contig #38312 (covered with 236.037 ppm small RNA reads across all 24 libraries) (section 13); reads not matching the *B. lanceolatum* genome or transcriptome, and matching non-*Branchiostoma* contig #3365 (covered with 223.535 ppm small RNA reads across all 24 libraries) (section 14); reads not matching the *B. lanceolatum* genome or transcriptome, and matching non-*Branchiostoma* contig #10883 (covered with 205.859 ppm small RNA reads across all 24 libraries) (section 15).

Table S1 Detection of conserved miRNAs. (file 'Supplementary_Table_1.pdf') *Branchiostoma lanceolatum* orthologs for *B. floridae* or *B. belcheri* pre-miRNA hairpins (as described in miRBase v.22) were screened for their predicted secondary structure and the abundance of the small RNAs they generate. Only those hairpins that comply with these rules are shown in this table. First column: name of orthologous pre-miRNA, and genomic coordinates in *B. lanceolatum*. Second column: sequences of the major forms of the 5´ arm and 3´ arm miRNAs, if expressed at ⩾10 ppm in at least one developmental stage (miRNAs that do not meet that criterion are flagged "low abundance"). Third column: abundance of the 5´ arm and 3´ arm miRNAs in Libraries #1 along development. Embryonic stages contain mixed sexes; adult stages are shown in blue and pink for males and females, respectively. Trimming (up to 3 nt) and templated extension of miRNA 3´ ends were considered when measuring read counts.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Acknowledgments

The authors are grateful to Dr. Darryl Conte for helpful discussions and to Julie Claycomb, Kazufumi Mochizuki and Phillip D. Zamore for critical reading of the manuscript. We thank the *B. lanceolatum* genome consortium for the assembly and annotation of the *B. lanceolatum* genome, Dr. Manuel Irimia for assistance in transcriptomics analyses, and Dr. Ferdinand Marlétaz for sharing unpublished data. This research was supported by an ATIP-Avenir grant from CNRS and Sanofi (to H.S.) and a post-doctoral fellowship from La Ligue contre le cancer (to N.P.). H.E.'s laboratory was supported by the CNRS and the ANR16-CE12-0008-01 and S.B. by the Institut Universitaire de France. We thank the MGX sequencing facility (Biocampus Montpellier, CNRS, INSERM, Univ. Montpellier, Montpellier, France). MGX acknowledges financial support from France Génomique National infrastructure, funded as part of the "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

References

1. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet. 2009;10(2):94–108.

2. Schiebel W, Haas B, Marinković S, Klanner A, Sänger HL. RNA-directed RNA polymerase from tomato leaves. II. Catalytic *in vitro* properties. J Biol Chem. 1993;268(16):11858–11867.

3. Tang G, Reinhart BJ, Bartel DP, Zamore PD. A biochemical framework for RNA silencing in plants. Genes Dev. 2003;17(1):49–63.

4. Curaba J, Chen X. Biochemical activities of Arabidopsis RNA-dependent RNA polymerase 6. J Biol Chem. 2008;283(6):3059–3066.

5. Voinnet O. Use, tolerance and avoidance of amplified RNA silencing by plants. Trends Plant Sci. 2008;13(7):317–328.

6. Cogoni C, Macino G. Gene silencing in *Neurospora crassa* requires a protein homologous to RNA-dependent RNA polymerase. Nature. 1999;399(6732):166–169.

7. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science. 2002;297(5588):1833–1837.

8. Hall IM, Shankaranarayana GD, Noma K, Ayoub N, Cohen A, Grewal SI. Establishment and maintenance of a heterochromatin domain. Science. 2002;297(5590):2232–2237.

9. Sigova A, Rhind N, Zamore PD. A single Argonaute protein mediates both transcriptional and posttranscriptional silencing in *Schizosaccharomyces pombe*. Genes Dev. 2004;18(19):2359–2367.

10. Allshire R. Molecular biology. RNAi and heterochromatin–a hushed-up affair. Science. 2002;297(5588):1818–1819.

11. Motamedi MR, Verdel A, Colmenares SU, Gerber SA, Gygi SP, Moazed D. Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. Cell. 2004;119(6):789–802.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

12. Martienssen R, Moazed D. RNAi and heterochromatin assembly. Cold Spring Harb Perspect Biol. 2015;7(8):a019323.

13. Makeyev EV, Bamford DH. Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes. Mol Cell. 2002;10(6):1417–1427.

14. Qian X, Hamid FM, El A Sahili, Darwis DA, Wong YH, Bhushan S, et al. Functional evolution in orthologous cell-encoded RNA-dependent RNA polymerases. J Biol Chem. 2016;291(17):9295–9309.

15. Kuhlmann M, Borisova BE, Kaller M, Larsson P, Stach D, Na J, et al. Silencing of retrotransposons in *Dictyostelium* by DNA methylation and RNAi. Nucleic Acids Res. 2005;33(19):6405–6417.

16. Marker S, Le A Mouël, Meyer E, Simon M. Distinct RNA-dependent RNA polymerases are required for RNAi triggered by double-stranded RNA versus truncated transgenes in *Paramecium tetraurelia*. Nucleic Acids Res. 2010;38(12):4092–4107.

17. Lee SR, Collins K. Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. Nat Struct Mol Biol. 2007;14(7):604–610.

18. Smardon A, Spoerke JM, Stacey SC, Klein ME, Mackin N, Maine EM. EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*. Curr Biol. 2000;10(4):169–178.

19. Sijen T, Fleenor J, Simmer F, Thijssen KL, Parrish S, Timmons L, et al. On the role of RNA amplification in dsRNA-triggered gene silencing. Cell. 2001;107(4):465–476.

20. Pak J, Fire A. Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. Science. 2007;315(5809):241–244.

21. Sijen T, Steiner FA, Thijssen KL, Plasterk RH. Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science. 2007;315(5809):244–247. (but note that this article's scientific integrity has been seriously questioned: `https://pubpeer.com/publications/2B00E5BEB5B75B499550D03C15EFA4`).

22. Gu W, Shirayama M, Conte DJ, Vasale J, Batista PJ, Claycomb JM, et al. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. Mol Cell. 2009;36(2):231–244.

23. Yigit E, Batista PJ, Bei Y, Pang KM, Chen CC, Tolia NH, et al. Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. Cell. 2006;127(4):747–757.

24. Aoki K, Moriguchi H, Yoshioka T, Okawa K, Tabara H. In vitro analyses of the production and activity of secondary small interfering RNAs in *C. elegans*. EMBO J. 2007;26(24):5007–5019.

25. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell. 2006;127(6):1193–1207.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

26. Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, Chu DS, et al. 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *Caenorhabditis elegans*. Proc Natl Acad Sci USA. 2009;106(44):18674–18679.

27. Vasale JJ, Gu W, Thivierge C, Batista PJ, Claycomb JM, Youngman EM, et al. Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. Proc Natl Acad Sci USA. 2010;107(8):3582–3587.

28. Wassenegger M, Krczal G. Nomenclature and functions of RNA-directed RNA polymerases. Trends Plant Sci. 2006;11(3):142–151.

29. Venkataraman S, Prasad BVLS, Selvarajan R. RNA dependent RNA polymerases: insights from structure, function and evolution. Viruses. 2018;10(2):Viruses.

30. Burroughs AM, Ando Y, Aravind L. New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing. Wiley Interdiscip Rev RNA. 2014;5(2):141–181.

31. Zong J, Yao X, Yin J, Zhang D, Ma H. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. Gene. 2009;447(1):29–39.

32. Maida Y, Yasukawa M, Furuuchi M, Lassmann T, Possemato R, Okamoto N, et al. An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. Nature. 2009;461(7261):230–235.

33. Stein P, Svoboda P, Anger M, Schultz RM. RNAi: mammalian oocytes do it without RNA-dependent RNA polymerase. RNA. 2003;9(2):187–192.

34. Horie M, , Kobayashi Y, Honda T, Fujino K, Akasaka T, et al. An RNA-dependent RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus. Sci Rep. 2016;6:25873.

35. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 1998;26(1):320–322.

36. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23(1):205–211.

37. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–1973.

38. Sánchez R, Serra F, Tárraga J, Medina I, Carbonell J, Pulido L, et al. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic Acids Res. 2011;39(Web Server issue):W470–474.

39. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–542.

40. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27(8):1164–1165.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

41. Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, et al. A RESTful API for access to phylogenetic tools via the CIPRES science gateway. Evol Bioinform Online. 2015;11:43–48.

42. Fuentes M, Benito E, Bertrand S, Paris M, Mignardot A, Godoy L, et al. Insights into spawning behavior and development of the European amphioxus (*Branchiostoma lanceolatum*). J Exp Zool B Mol Dev Evol. 2007;308(4):484–493.

43. Marlétaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature. 2018; in press (doi: 10.1038/s41586-018-0734-6).

44. Tapial J, Ha KC, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017;27(10):1759–1768.

45. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–829.

46. Massart S, Chiumenti M, De Jonghe K, Glover R, Haegeman A, Koloniuk I, et al. Virus detection by high-throughput sequencing of small RNAs: large scale performance testing of sequence analysis strategies. Phytopathology. 2018; in press (doi: 10.1094/PHYTO-02-18-0067-R).

47. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol. 2018;2(1):174–181.

48. Li H, Bowling AJ, Gandra P, Rangasamy M, Pence HE, McEwan RE, et al. Systemic RNAi in western corn rootworm, *Diabrotica virgifera virgifera*, does not involve transitive pathways. Insect Sci. 2018;25(1):45–56.

49. Bertrand S, Escrivá H. Evolutionary crossroads in developmental biology: amphioxus. Development. 2011;138(22):4819–4830.

50. Louis A, Roest Crollius H, Robinson-Rechavi M. How much does the amphioxus genome represent the ancestor of chordates? Brief Funct Genomics. 2012;11(2):89–95.

51. Iyer LM, Koonin EV, Aravind L. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. BMC Struct Biol. 2003;3:1.

52. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. 2001;15(2):188–200.

53. Hutvágner G, McLachlan J, Pasquinelli AE, Bálint E, Tuschl T, Zamore PD. A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. Science. 2001;293(5531):834–838.

54. Pélisson A, Sarot E, Payen-Groschêne G, Bucheton A. A novel repeat-associated small interfering RNA-mediated silencing pathway downregulates complementary sense gypsy transcripts in somatic cells of the *Drosophila* ovary. J Virol. 2007;81(4):1951–1960.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

55. Horwich MD, Li C, Matranga C, Vagin V, Farley G, Wang P, et al. The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. Curr Biol. 2007;17(14):1265–1272.

56. Billi AC, Alessi AF, Khivansara V, Han T, Freeberg M, Mitani S, et al. The *Caenorhabditis elegans* HEN1 ortholog, HENN-1, methylates and stabilizes select subclasses of germline small RNAs. PLoS Genet. 2012;8(4):e1002617.

57. Kamminga LM, van Wolfswinkel JC, Luteijn MJ, Kaaij LJ, Bagijn MP, Sapetschnig A, et al. Differential impact of the HEN1 homolog HENN-1 on 21U and 26G RNAs in the germline of *Caenorhabditis elegans*. PLoS Genet. 2012;8(7):e1002702.

58. Montgomery TA, Rim YS, Zhang C, Dowen RH, Phillips CM, Fischer SE, et al. PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog *henn-1*. PLoS Genet. 2012;8(4):e1002616.

59. Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, et al. Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. Science. 2008;320(5879):1077–1081.

60. Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. Single processing center models for human Dicer and bacterial RNase III. Cell. 2004;118(1):57–68.

61. Han BW, Hung JH, Weng Z, Zamore PD, Ameres SL. The 3´-to-5´ exoribonuclease Nibbler shapes the 3´ ends of microRNAs bound to *Drosophila* Argonaute1. Curr Biol. 2011;21(22):1878–1887.

62. Liu N, Abe M, Sabin LR, Hendriks GJ, Naqvi AS, Yu Z, et al. The exoribonuclease Nibbler controls 3´ end processing of microRNAs in *Drosophila*. Curr Biol. 2011;21(22):1888–1893.

63. Feltzin VL, Khaladkar M, Abe M, Parisi M, Hendriks GJ, Kim J, et al. The exonuclease Nibbler regulates age-associated traits and modulates piRNA length in *Drosophila*. Aging Cell. 2015;14(3):443–452.

64. Wang H, Ma Z, Niu K, Xiao Y, Wu X, Pan C, et al. Antagonistic roles of Nibbler and Hen1 in modulating piRNA 3´ ends in *Drosophila*. Development. 2016;143(3):530–539.

65. Hayashi R, Schnabl J, Handler D, Mohn F, Ameres SL, Brennecke J. Genetic and mechanistic diversity of piRNA 3´-end formation. Nature. 2016;539(7630):588–592.

66. Jorgensen SE, Buch LB, Nierlich DP. Nucleoside triphosphate termini from RNA synthesized *in vivo* by *Escherichia coli*. Science. 1969;164(3883):1067–1070.

67. Wu CW, Goldthwait DA. Studies of nucleotide binding to the ribonucleic acid polymerase by a fluoresence technique. Biochemistry. 1969;8(11):4450–4458.

68. Wu CW, Goldthwait DA. Studies of nucleotide binding to the ribonucleic acid polymerase by equilibrium dialysis. Biochemistry. 1969;8(11):4458–4464.

69. Miller WA, Bujarski JJ, Dreher TW, Hall TC. Minus-strand initiation by brome mosaic virus replicase within the 3´ tRNA-like structure of native and modified RNA templates. J Mol Biol. 1986;187(4):537–546.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

70. Kuzmine I, Gottlieb PA, Martin CT. Binding of the priming nucleotide in the initiation of transcription by T7 RNA polymerase. J Biol Chem. 2003;278(5):2819–2823.

71. Juven-Gershon T, Kadonaga JT. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol. 2010;339(2):225–229.

72. Hetzel J, Duttke SH, Benner C, Chory J. Nascent RNA sequencing reveals distinct features in plant transcription. Proc Natl Acad Sci USA. 2016;113(43):12316–12321.

73. Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, et al. Sorting of small RNAs into *Arabidopsis* argonaute complexes is directed by the 5´ terminal nucleotide. Cell. 2008;133(1):116–127.

74. Montgomery TA, Howell MD, Cuperus JT, Li D, Hansen JE, Alexander AL, et al. Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell. 2008;133(1):128–141.

75. Takeda A, Iwasaki S, Watanabe T, Utsumi M, Watanabe Y. The mechanism selecting the guide strand from small RNA duplexes is different among argonaute proteins. Plant Cell Physiol. 2008;49(4):493–500.

76. Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore PD. Sorting of *Drosophila* small silencing RNAs partitions microRNA* strands into the RNA interference pathway. RNA. 2010;16(1):43–56.

77. Frank F, Sonenberg N, Nagar B. Structural basis for 5´-nucleotide base-specific recognition of guide RNA by human AGO2. Nature. 2010;465(7299):818–822.

78. Seitz H, Tushir JS, Zamore PD. A 5´-uridine amplifies miRNA/miRNA* asymmetry in *Drosophila* by promoting RNA-induced silencing complex formation. Silence. 2011;2:4.

79. Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, et al. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. Genes Dev. 2006;20(16):2214–2222.

80. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, et al. Characterization of the piRNA complex from rat testes. Science. 2006;313(5785):363–367.

81. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006;442(7099):199–202.

82. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006;442(7099):203–207.

83. Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, et al. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev. 2006;20(13):1732–1743.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

84. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell. 2007;128(6):1089–1103.

85. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell. 2007;129(1):69–82.

86. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, et al. The small RNA profile during *Drosophila melanogaster* development. Dev Cell. 2003;5(2):337–350.

87. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature. 2008;455(7217):1193–1197.

88. Friedländer MR, Adamidi C, Han T, Lebedeva S, Isenbarger TA, Hirst M, et al. High-resolution profiling and discovery of planarian small RNAs. Proc Natl Acad Sci USA. 2009;106(28):11546–11551.

89. Song JL, Stoeckius M, Maaskola J, Friedländer M, Stepicheva N, Juliano C, et al. Select microRNAs are essential for early development in the sea urchin. Dev Biol. 2012;362(1):104–113.

90. Moran Y, Fredman D, Praher D, Li XZ, Wee LM, Rentzsch F, et al. Cnidarian microRNAs frequently regulate targets by cleavage. Genome Res. 2014;24(4):651–663.

91. Guo Z, Li Y, Ding SW. Small RNA-based antimicrobial immunity. Nat Rev Immunol. 2018; in press (doi: 10.1038/s41577-018-0071-x).

92. Huang G, Huang S, Yan X, Yang P, Li J, Xu W, et al. Two apextrin-like proteins mediate extracellular and intracellular bacterial recognition in amphioxus. Proc Natl Acad Sci USA. 2014;111(37):13469–13474.

93. Zou Y, Ma C, Zhang Y, Du Z, You F, Tan X, et al. Isolation and characterization of *Vibrio alginolyticus* from cultured amphioxus *Branchiostoma belcheri tsingtauense*. Biologia. 2016;71(7):757–762.

94. Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson F, Severance E, et al. Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. Proc Natl Acad Sci USA. 2014;111(45):16106–16111.

95. Kjartansdóttir KR, Friis-Nielsen J, Asplund M, Mollerup S, Mourier T, Jensen RH, et al. Traces of ATCV-1 associated with laboratory component contamination. Proc Natl Acad Sci USA. 2015;112(9):E925–E926.

96. Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson F, Severance E, et al. Reply to Kjartansdóttir et al.: Chlorovirus ATCV-1 findings not explained by contamination. Proc Natl Acad Sci USA. 2015;112(9):E927.

97. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A distinct small RNA pathway silences selfish genetic elements in the germline. Science. 2006;313(5785):320–324.

98. Kirino Y, Mourelatos Z. Mouse Piwi-interacting RNAs are 2´-O-methylated at their 3´ termini. Nat Struct Mol Biol. 2007;14(4):347–348.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

99. Fu Y, Yang Y, Zhang H, Farley G, Wang J, Quarles KA, et al. The genome of the Hi5 germ cell line from *Trichoplusia ni*, an agricultural pest and novel model for small RNA biology. Elife. 2018;7:e31628.

100. Saito K, Sakaguchi Y, Suzuki T, Suzuki T, Siomi H, Siomi MC. Pimet, the *Drosophila* homolog of HEN1, mediates 2´-O-methylation of Piwi- interacting RNAs at their 3´ ends. Genes Dev. 2007;21(13):1603–1608.

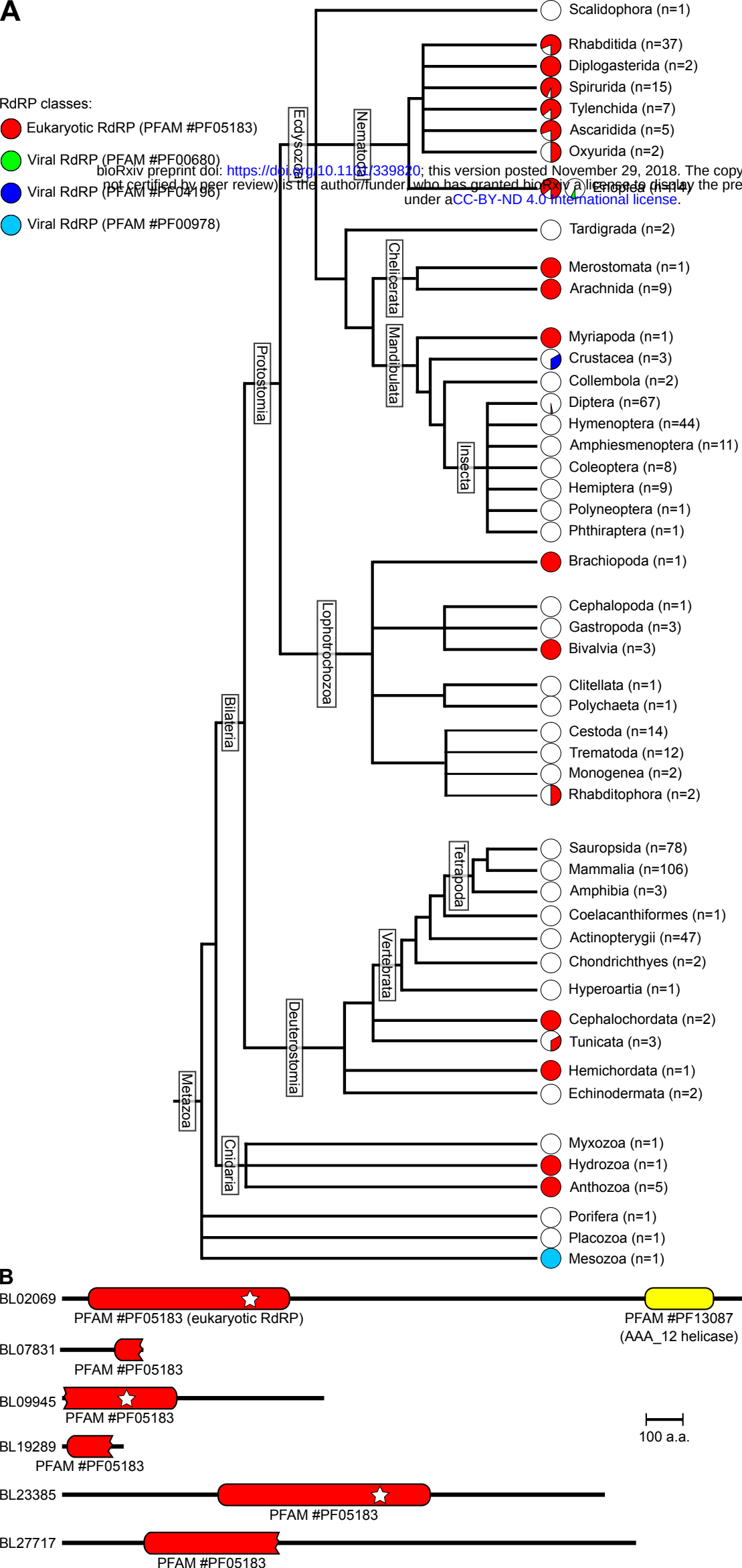
101. Kirino Y, Mourelatos Z. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. RNA. 2007;13(9):1397–1401.

102. Kamminga LM, Luteijn MJ, den Broeder MJ, Redl S, Kaaij LJ, Roovers EF, et al. Hen1 is required for oocyte development and piRNA stability in zebrafish. EMBO J. 2010;29(21):3688–3700.

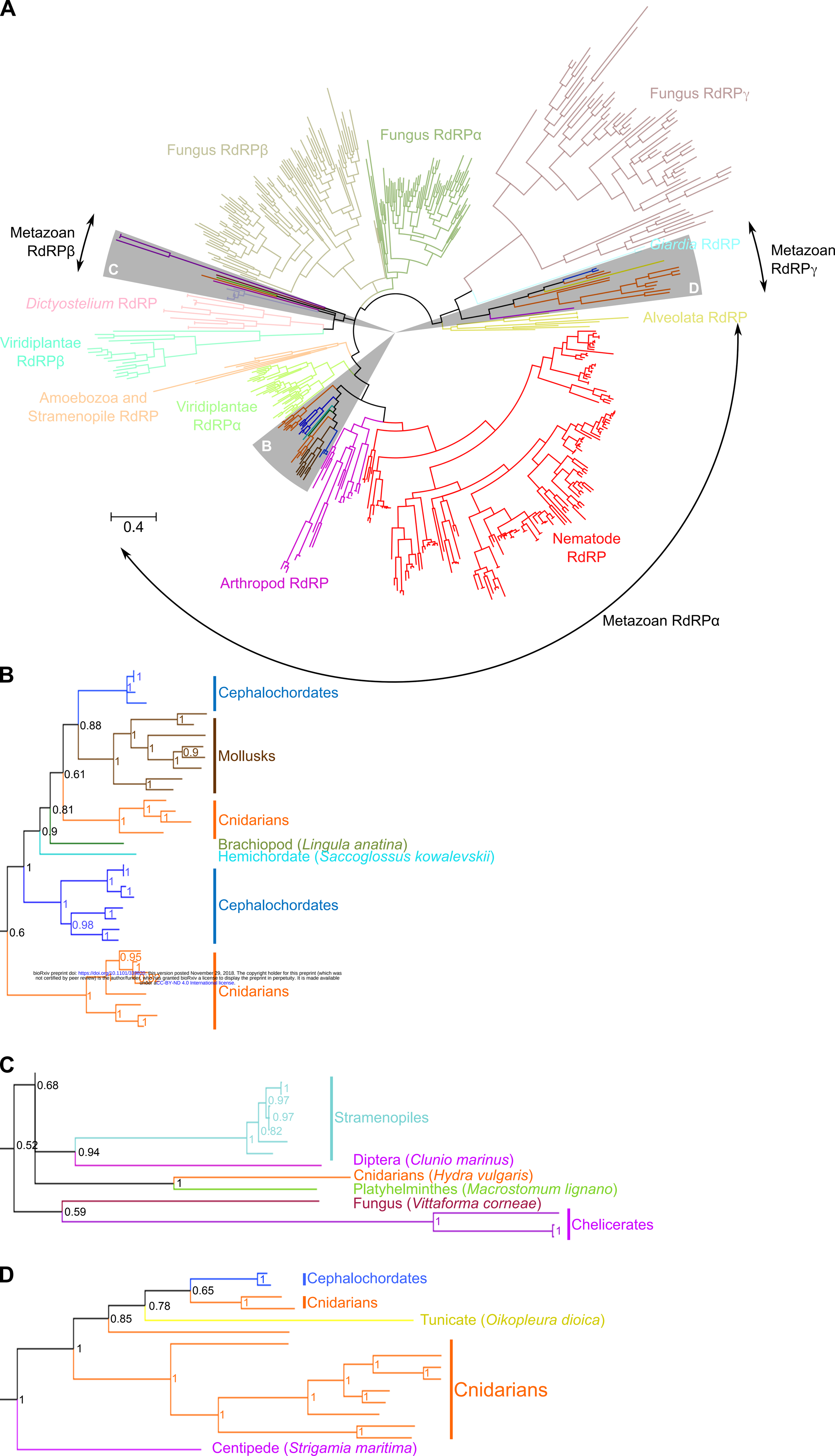
103. Huang Y, Ji L, Huang Q, Vassylyev DG, Chen X, Ma JB. Structural insights into mechanisms of the small RNA methyltransferase HEN1. Nature. 2009;461(7265):823–827.

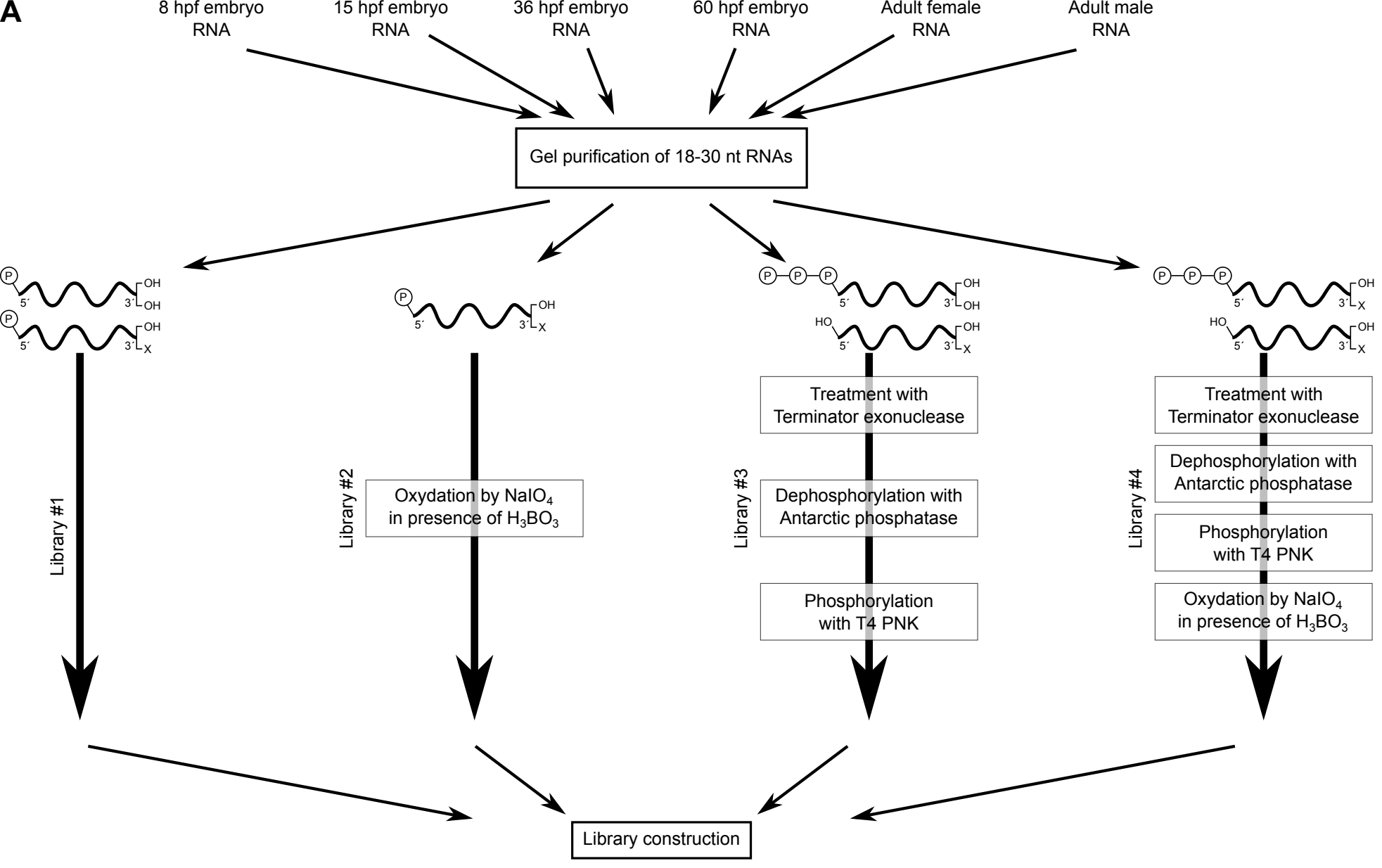
104. Hammond TM, Keller NP. RNA silencing in *Aspergillus nidulans* is independent of RNA-dependent RNA polymerases. Genetics. 2005;169(2):607–617.

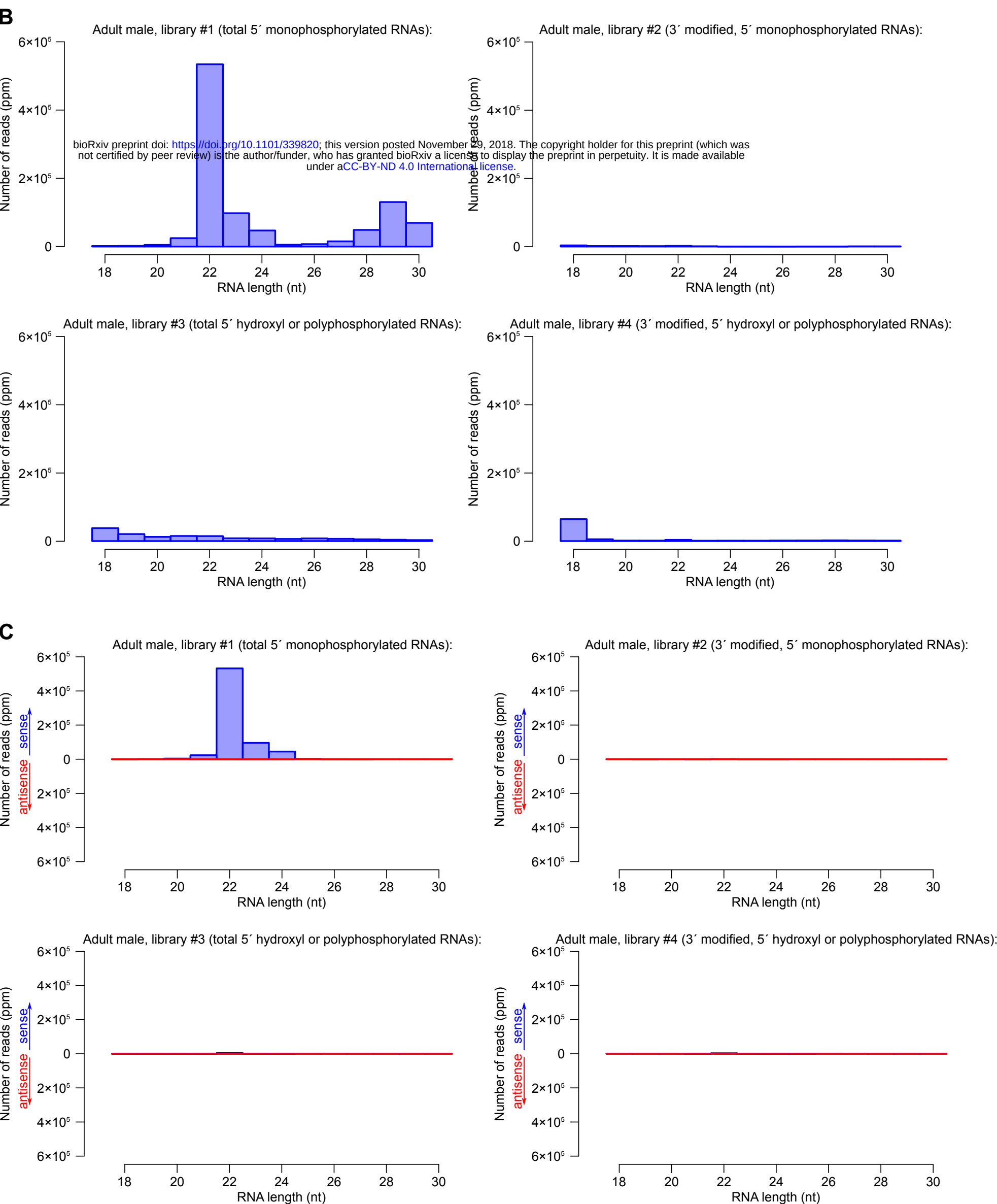
105. Labbé RM, Irimia M, Currie KW, Lin A, Zhu SJ, Brown DD, et al. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. Stem Cells. 2012;30(8):1734–1745.

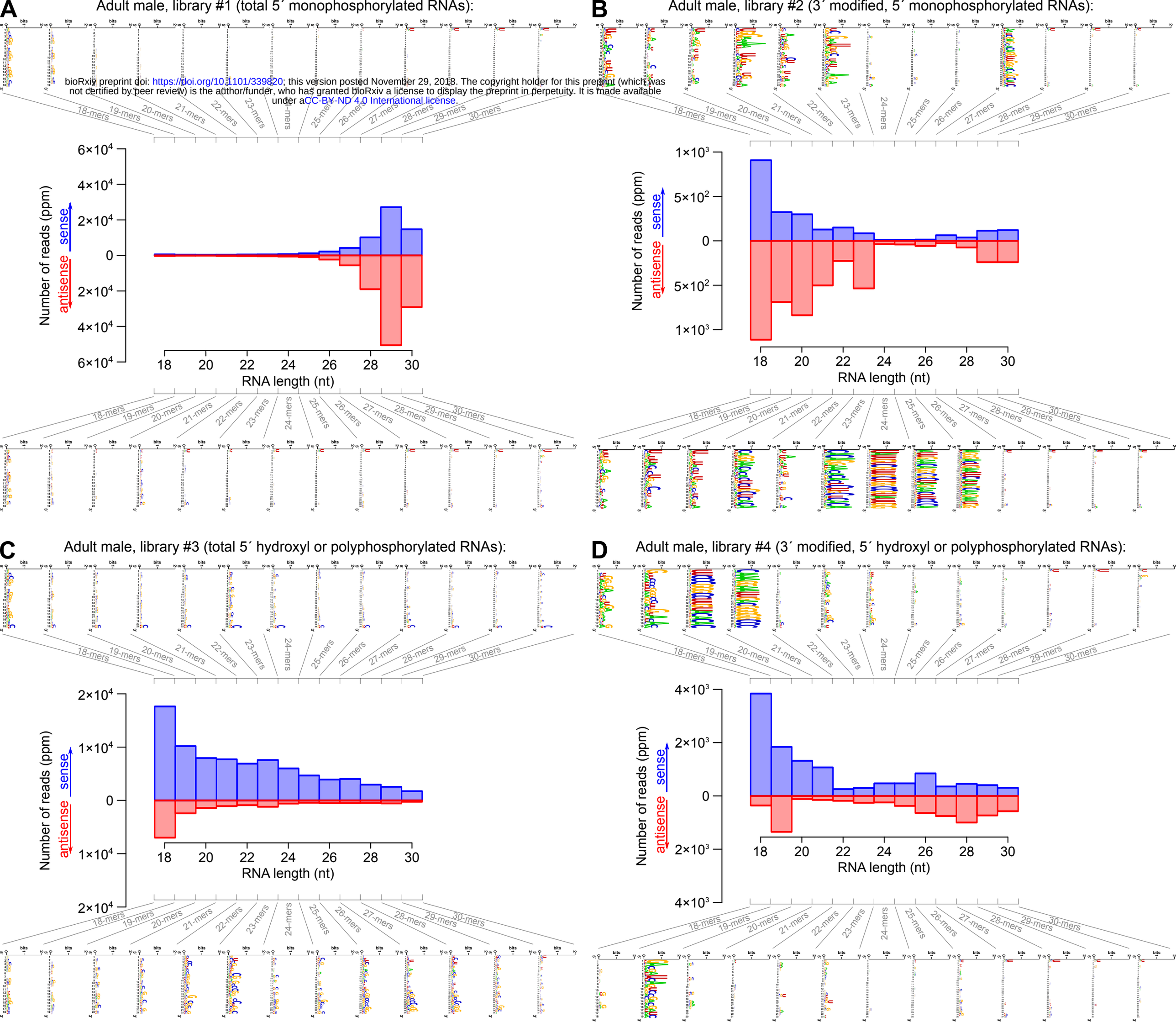


bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copy
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the pre
under aCC-BY-ND 4.0 International license.


A
Fungus RdRPγ
Fungus RdRPα
Fungus RdRPβ
Metazoan RdRPβ
Metazoan RdRPγ
Giardia RdRP
Dictyostelium RdRP
Alveolata RdRP
Viridiplantae RdRPβ
Amoebozoa and Stramenopile RdRP
Viridiplantae RdRPα
Nematode RdRP
Arthropod RdRP
Metazoan RdRPα
C
D
0.4
B
Cephalochordates
Mollusks
Cnidarians
Brachiopod (Lingula anatina)
Hemichordate (Saccoglossus kowalevskii)
Cephalochordates
Cnidarians
0.88
0.61
0.81
0.9
0.6
0.98
0.95
0.9
bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.
Stramenopiles
Diptera (Clunio marinus)
Cnidarians (Hydra vulgaris)
Platyhelminthes (Macrostomum lignano)
Fungus (Vittaforma corneae)
Chelicerates
0.68
0.52
0.94
0.59
0.97
0.97
0.82
Cephalochordates
Cnidarians
Tunicate (Oikopleura dioica)
Cnidarians
Centipede (Strigamia maritima)
0.65
0.78
0.85





bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/339820; this version posted November 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.