Metagenomic sequencing suggests arthropod antiviral RNA interference is highly derived, and identifies novel viral small RNAs in a mollusc and a brown alga

Fergal M. Waldron[1*], Graham N. Stone[1], Darren J. Obbard[1,2]

[1] Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, UK

[2] Centre for immunity Infection and Evolution, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, UK

[*] correspondence: fergal.waldron@ed.ac.uk

Abstract

RNA interference (RNAi)-related pathways target viruses and transposable element (TE) transcripts in plants, fungi, and ecdysozoans (nematodes and arthropods), giving protection against infection and transmission. In each case, this produces abundant TE and virus-derived 20-30nt small RNAs, which provide a characteristic signature of RNAi-mediated defence. The broad phylogenetic distribution of the Argonaute and Dicer-family genes that mediate these pathways suggests that defensive RNAi is ancient and probably shared by most metazoan (animal) phyla. Indeed, while vertebrates had been thought an exception, it has recently been suggested that mammals may also possess a functional antiviral RNAi pathway, albeit involving few small RNAs that are challenging to detect. Here we use a metagenomic approach to test for the presence of antiviral RNAi in five divergent metazoan phyla (Porifera, Cnidaria, Echinodermata, Mollusca, and Annelida), and in a brown alga. We use metagenomic RNA sequencing to identify around 80 virus-like contigs in these lineages, and small RNA sequencing to identify small RNAs derived from those viruses. Contrary to our expectations, we were unable to identify canonical (i.e. *Drosophila*-like) viral small RNAs in any of these organisms, despite the presence of abundant micro-RNAs and putative piwi-interacting piRNAs. Instead, we identified an apparently novel class of virus-derived small RNAs in the mollusc, which have a piRNA-like length distribution but lack key signatures of piRNA biogenesis, and a novel class of 21U virus-derived small RNAs in the brown alga. We also identified primary piRNAs derived from putatively endogenous copies of DNA viruses in the cnidarian and the echinoderm, and an endogenous RNA virus in the mollusc. This suggests either that the majority of metazoan phyla lack antiviral RNAi completely, such that antiviral RNAi has evolved independently in ecdysozoans, vertebrates, and molluscs, or that the antiviral RNAi response in most extant phyla—and the ancestral state of Metazoa—more closely resembles that of mammals than arthropods or nematodes.

Introduction

RNA interference-related (RNAi) pathways provide an important line of defence against parasitic nucleic acids in plants, fungi, and most metazoa (Ding *et al.*, 2004; Buchon & Vaury, 2006; Cerutti & Casas-Mollano, 2006; Segers *et al.*, 2007; Obbard *et al.*, 2009). In plants and fungi, which lack a distinct germline, Dicer and Argonaute-dependent RNAi responses suppress the expression and replication of viruses and transposable elements (TEs) through a combination of target cleavage and/or heterochromatin induction (Agius *et al.*, 2012; Chang *et al.*, 2012). This gives rise to a characteristic signature of short interfering RNAs (siRNAs) derived from both TEs and viruses (Dang *et al.*, 2011; Axtell, 2013; Nicolás & Ruiz-Vázquez, 2013; Szittya & Burgyan, 2013; Borges & Martienssen, 2015). In contrast, the best-studied metazoan (animal) lineages display two distinct signatures of defensive RNAi. First, reminiscent of plants and fungi, arthropods and nematodes exhibit a highly active Dicer-dependent antiviral pathway that is characterised by copious virus-derived siRNAs (viRNAs) peaking sharply between 20nt (e.g. Lepidoptera) and 22nt (e.g. Hymenoptera). These are cleaved from double-stranded viral RNA by Dicer, and loaded into an Argonaute-containing complex that targets virus genomes and transcripts via sequence complementarity (Barnard *et al.*, 2012; Sarkies & Miska, 2013). Second, and in contrast to plants and fungi, metazoa also possess a Piwi-dependent (piRNA) pathway that provides a defence against TEs in germline (*Drosophila* and mammals) and/or somatic cells (e.g. Porifera, Cnidaria; Grimson et al., 2008). This pathway is usually characterised by a broad peak of 26-30nt small RNAs bound by Piwi-family Argonaute proteins, and in most metazoa is thought to target TE transcripts for cleavage and genomic copies for heterochromatin induction (Czech & Hannon, 2016). It comprises both 5'U primary piRNAs cleaved by homologs of *Drosophila* Zucchini from long 'piRNA cluster' transcripts (see Yamanaka *et al.*, 2014), and secondary piRNAs generated by 'Ping-Pong' amplification.

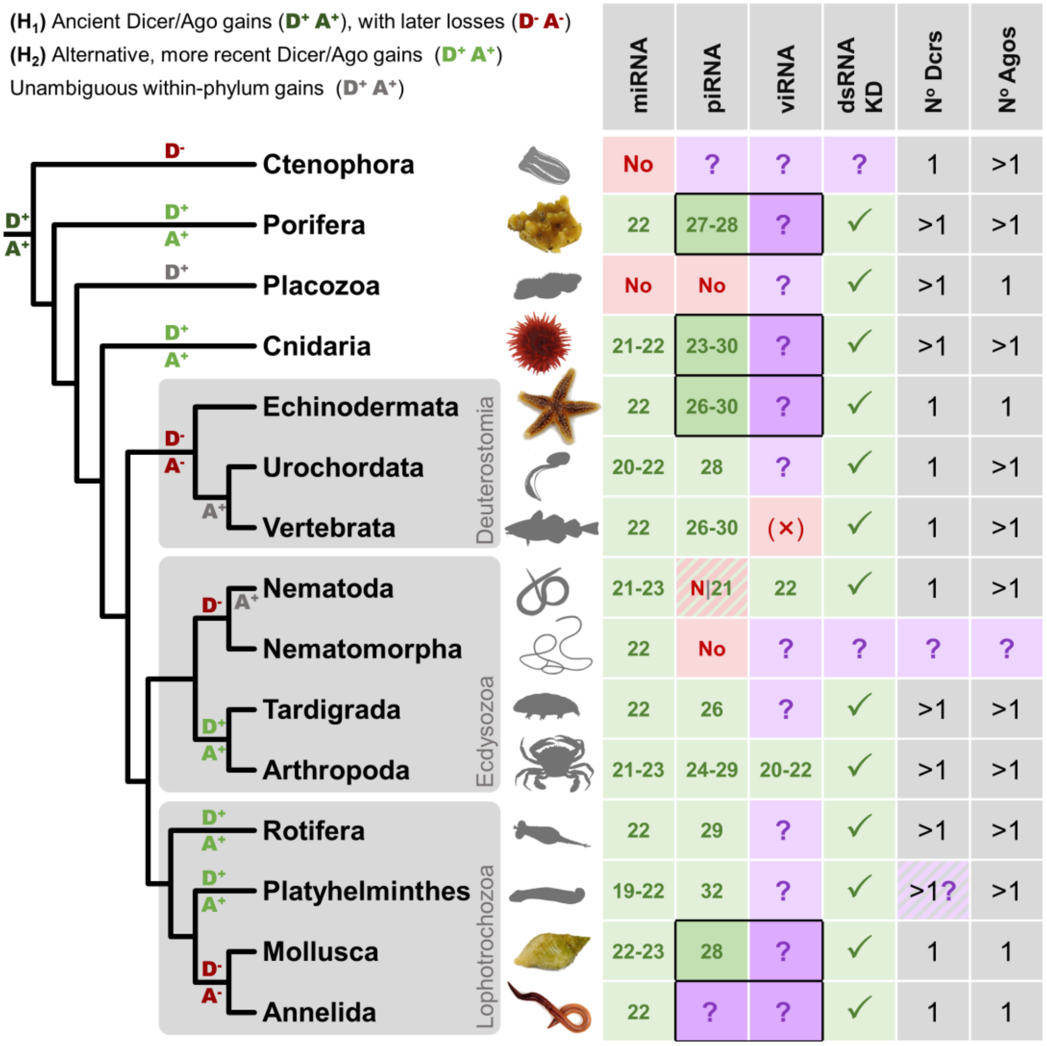
The presence of abundant viRNAs in infected plants, fungi, nematodes, and arthropods suggests that Dicer-dependent antiviral RNAi is an ancient and conserved defence (Cerutti & Casas-Mollano, 2006; Obbard *et al.*, 2009). However, antiviral RNAi has been lost in some lineages such as *Plasmodium* (Baum *et al.*, 2009), some trypanosomes (Lye *et al.*, 2010), and some *Saccharomyces* (Drinnenberg *et al.*, 2009), and/or extensively modified in others. For example, antiviral RNAi was long thought to be absent from vertebrates (Backes et al., 2014; Bogerd et al., 2014), at least in part because their viRNAs cannot easily be detect by high-throughput sequencing of the total small-RNA pool (Umbach & Cullen, 2009; Parameswaran *et al.*, 2010; Perez *et al.*, 2010; Girardi *et al.*, 2013; Backes *et al.*, 2014; Bogerd *et al.*, 2014). Recently, it has been suggested that vertebrates also possess a functional virus-targeting RNAi pathway, which can have an antiviral role in tissues lacking an interferon response (Li *et al.*, 2013; Maillard *et al.*, 2013, 2016; Benitez *et al.*, 2015) and/or in the absence of viral suppressor of RNAi (Maillard *et al.*, 2013; Li *et al.*, 2016)—although this remains contentious (see tenOever, 2017). Nevertheless, phylogenetically comprehensive experimental studies of antiviral RNAi in metazoa are not available, with studies instead focussing on arthropods such as insects (reviewed in Bronkhorst & van Rij, 2014; Gammon & Mello, 2015), crustaceans (Labreuche & Warr, 2013, and reviewed in Liu et al., 2009), chelicerates (Schnettler et al., 2014)), and on nematodes (Ashe et al., 2013; Coffman et al., 2017; Gammon et al., 2017), and vertebrates (Parameswaran *et al.*, 2010; Girardi *et al.*, 2013; Li *et al.*, 2013, 2016, Maillard *et al.*, 2013, 2016; Seo *et al.*, 2013; Backes *et al.*, 2014; Bogerd *et al.*, 2014). In particular, there have been few attempts to identify viRNAs in early-branching metazoan lineages such as Porifera or Cnidaria, in divergent Deuterostome lineages such as Echinodermata or Urochordata, or in Lophotrochozoa (including the large phyla Annelida and Mollusca). See Figure 1 for the known distribution of RNAi-pathways across the metazoa.

Broadly consistent with a wide distribution of antiviral RNAi, Argonaute and Dicer genes are detectable in most metazoan genomes (Figure 1; de Jong *et al.*, 2009; Mukherjee *et al.*, 2013; Tabach *et al.*, 2013; Casas-Mollano *et al.*, 2016). However, while Dicer and Argonaute genes would be necessary for a canonical antiviral RNAi response, their presence in most metazoa is insufficient to demonstrate one, for two reasons. First, these genes also have non-defensive roles such as transcription regulation through miRNAs (see Rajasethupathy *et al.*, 2012; Castel & Martienssen, 2013)—and a single gene can fulfil multiple roles. For example, whereas in *Drosophila* there is a distinction between the Dcr2-Ago2 antiviral pathway and the Ago1-Dcr1 mediated miRNA pathway (e.g. Lee *et al.*, 2004), in *C. elegans* a single Dicer is required for the biogenesis of both miRNAs and viRNAs (Figure 1; Grishok *et al.*, 2001; Tabara *et al.*, 2002; Ashe *et al.*, 2013). Second, RNAi pathways are labile over evolutionary timescales, with regular gene duplication, loss, and change of function (e.g. Lewis et al., 2015; Sarkies et al., 2015). For example, the Piwi-family Argonaute genes that mediate anti-TE defence in metazoa were ancestrally present in Eukaryotes, but were lost independently in plants, fungi, brown algae, and most nematodes (Cerutti & Casas-Mollano, 2006; Mukherjee et al., 2013; Swarts et al., 2014; Sarkies et al., 2015). In contrast, non-Piwi Argonautes were lost in many Alveolates, Excavates and Amoebozoa (Burroughs et al., 2014; Swarts et al., 2014) while Piwis were retained in these lineages. At the same time, new RNAi mechanisms have arisen, such as the 22G RNAs of nematodes (Yigit *et al.*, 2006; Pak & Fire, 2007; Sarkies *et al.*, 2015) and the recent gain of an antiviral role for Piwi in *Aedes* mosquitoes (Morazzani et al., 2012; Vodovar et al., 2012). Taken together, the potential for multiple functions, and for gains and losses of function, make it challenging to confidently predict the phylogenetic distribution of antiviral RNAi from the distribution of the required genes (see Casas-Mollano *et al.*, 2016).

Thus, although antiviral RNAi is predicted to be shared by most extant eukaryotes (see tenOever, 2016; Koonin, 2017), in the absence of experimental studies, its distribution across metazoan phyla remains largely unknown (Figure 1). This contrasts sharply with our knowledge of other RNAi-related pathways, such as the micro-RNA (miRNA) mediated control of gene expression, which is conserved across plants, brown algae, fungi, and almost all metazoans (Moran *et al.*, 2017), and the presence of TE-derived piRNAs in most metazoans: Porifera (Grimson *et al.*, 2008; Funayama *et al.*, 2010), Cnidaria (Grimson *et al.*, 2008; Juliano *et al.*, 2013), Ctenophora (Alié *et al.*, 2011), Vertebrata (Aravin *et al.*, 2006; Houwing *et al.*, 2007), Arthropoda (Brennecke *et al.*, 2007; Cai *et al.*, 2012; Miesen *et al.*, 2015; Swarts *et al.*, 2017), some Nematoda (Sijen & Plasterk, 2003; but see Sarkies *et al.*, 2015), Platyhelminthes (Zhou *et al.*, 2015), but not Placozoa (Grimson et al., 2008). In eukaryotes that lack direct experimental evidence for viRNAs, the presence of an inducible RNAi response to experimentally applied long double-stranded RNA might indicate a potential for antiviral RNAi (Figure 1). This been reported for Excavates (Ishikawa *et al.*, 2008), Heterkonts (Takahashi *et al.*, 2007) Amoebozoa (Kaur & Lohia, 2004), trypanosomes (Ngo *et al.*, 1998), and amongst metazoa in Porifera (Rivera *et al.*, 2011), Cnidaria (Wittig et al., 2011), Placozoa (Jakob et al., 2004), Arthropoda (Yu *et al.*, 2013), Nematoda (Fire *et al.*, 1998), and several lineages of Lophotrochozoa including planarian flatworms (Sánchez Alvarado & Newmark, 1999), bivalve molluscs (Fabioux et al., 2009), rotifers (Snell *et al.*, 2011) and annelids (Yoshida-Noro & Tochinai, 2010).

Thus, although circumstantial evidence suggests a near-universal potential for antiviral RNAi in metazoa, we still lack experimental evidence of exogenous viral processing. Here we seek to examine the phylogenetic distribution of viRNAs, and thus elucidate the phylogenetic distribution of a canonical (i.e. *Drosophila*-like or plant-like) antiviral RNAi response, through metagenomic sequencing. We combine rRNA-depleted RNA sequencing with small-RNA sequencing to detect both viruses and viRNAs in pooled samples of six deeply divergent lineages. First, we include two early branching metazoan phyla: a sponge (*Halichondria panicea*: Porifera, Demospongiae) and a sea anemone (*Actinia* *equina*: Cnidaria, Anthozoa) that branch basally to the divergence between deuterostomes and protostomes (Figure 1). Second, a starfish (*Asterias rubens*: Echinodermata, Asteroidea) that branches basally to vertebrates within the Deuterostomia. Third, two divergent species of Lophotrochozoa, the clade which forms the sister group to Ecdysozoa within the protostomes: a dog whelk (*Nucella lapillus*: Mollusca, Gastropoda) and earthworms (Annelida, Oligochaeta). Finally, to explore the deep history of antiviral RNAi within the eukaryotes, we included the brown alga *Fucus serratus* (Phaeophyceae, Heterokonta), which represents a fourth origin of multicellularity separate from plants, fungi, and metazoa.

Surprisingly, although we find viral RNA sequences to be common and abundant, we do not find abundant viRNAs from RNA viruses in most of the sampled species, suggesting that they lack a canonical antiviral RNAi response. Specifically, we detect no viRNAs from RNA viruses infecting the earthworms, the sponge, or the sea anemone, suggesting that insect- or nematode-like antiviral RNAi is absent from these lineages. In contrast, we do detect viRNAs from RNA viruses in the dog whelk and the brown alga. In both cases these viRNAs derive from both strands of the virus. However, in the dog whelk they peak broadly at 26-30nt—as would be expected of piRNAs, but lacking the 5'U or ‘ping-pong’ signature—and in the brown alga they peak sharply at 21nt and are exclusively 5'U. This suggests the presence of distinct and previously unrecognised antiviral RNAi responses in these two lineages. Finally, we identify primary piRNA-like 26-30nt 5'U small-RNAs derived from putatively endogenous copies of viruses in the sponge, the starfish, and the dog whelk, and TE-derived piRNAs in all the metazoan lineages examined. Taken together, these findings imply that the true diversity of defensive RNAi strategies employed by eukaryotes may have been underestimated, and that antiviral RNAi is either lacking from many metazoan phyla, or more closely resembles the RNAi response reported for mammals, than that of arthropods and nematodes.



	miRNA	piRNA	viRNA	dsRNA KD	Nº Dcrs	Nº Agos
Ctenophora	No	?	?	?	1	>1
Porifera	22	27-28	?	✓	>1	>1
Placozoa	No	No	?	✓	>1	1
Cnidaria	21-22	23-30	?	✓	>1	>1
Echinodermata	22	26-30	?	✓	1	1
Urochordata	20-22	28	?	✓	1	>1
Vertebrata	22	26-30	(×)	✓	1	>1
Nematoda	21-23	N|21	22	✓	1	>1
Nematomorpha	22	No	?	?	?	?
Tardigrada	22	26	?	✓	>1	>1
Arthropoda	21-23	24-29	20-22	✓	>1	>1
Rotifera	22	29	?	✓	>1	>1
Platyhelminthes	19-22	32	?	✓	>1?	>1
Mollusca	22-23	28	?	✓	1	1
Annelida	22	?	?	✓	1	1

Figure 1: Distribution of RNA-interference pathways across the metazoa

Phylogeny of selected metazoan phyla (topology follows Giribet, 2016) with a table recording the reported range of modal lengths for miRNAs, piRNAs, and viRNAs detectable by routine sequencing (miRNA modes taken from miRbase). Entries marked 'No' have been reported to be absent, and those marked '?' are untested. Focal taxa in this study are marked in colour, and the target table entries are outlined. Although antiviral RNAi has recently been reported in mammals (Li *et al.*, 2013, 2016, Maillard *et al.*, 2013, 2016), vertebrate viRNAs are marked '(×)' because these viRNAs cannot be detected by simple bulk sequencing of wild-type hosts and viruses, as used here. Note that piRNAs are absent from some, but not all, nematodes (Sarkies *et al.*, 2015). The column 'dsRNA KD' records whether dsRNA knockdown of gene expression using long dsRNA (i.e. a Dicer substrate) has been reported, as this may suggest the presence of an RNAi pathway capable of producing viRNAs from replicating viruses. The 'Dcrs' and 'Agos' columns record the inferred number of Dicers and (non-Piwi) Argonautes ancestrally present in each phylum, although the number of Dicers in Platyhelminthes is contentious as the putative second Dicer lacks the majority of expected Dicer domains. Broadly speaking, there are two competing hypotheses for the histories of Dicers and (non-Piwi) Argonautes in metazoa, e.g. (de Jong *et al.*, 2009; Mukherjee *et al.*, 2013, 2014). The first (H_1), posits that an early duplication in Dicer and/or Argonaute (marked D^+ and A^+ in dark green on the phylogeny) gave rise to at least two very divergent homologues of each gene, followed by subsequent losses (D^- and A^- in dark red). The second (H_2), suggests that divergent homologues are the result of more recent duplications (D^+ and A^+ in pale green), and where homologs have high divergence it is as a result of rapid evolution. Note that these hypotheses are independent for Argonautes and Dicers, and one may be ancient but the other recent. For Dicers, at least, the 'ancient' duplication is arguably better supported (Mukherjee et al., 2013), although it remains extremely difficult to determine orthology between the duplicates. In addition, Dicers and Argonautes have unambiguously diversified within some phyla (important examples marked A^+ and D^+ in grey)—as seen for the large nematode-specific WAGO clade of Argonautes (reviewed in Buck & Blaxter, 2013), and the multiple Argonautes in vertebrates.

Results

New virus-like sequences identified by metagenomic sequencing

Using the Illumina platform, we generated strand-specific 150 nt paired-end sequence reads from ribosome-depleted RNA, extracted from metagenomic pools of each of six different species: the breadcrumb sponge (*Halichondria panacea*, Porifera); the beadlet sea anemone (*Actinia equina*, Cnidaria); the common starfish (*Asterias rubens*, Echinodermata); the dog whelk (*Nucella lapillus*, Mollusca); mixed earthworm species (*Amynthas* and *Lumbricus* spp., Annelida), and a brown alga (the 'serrated wrack', *Fucus serratus*, Fucales, Phaeophyceae, Heterokonta). See S1_Table for collection data. Gut contents were excluded by dissection, and contaminating nematodes excluded by a PCR screen prior to pooling (Materials and Methods; S1_Table). Reads were assembled separately for each species using Trinity v2.2.0 (Grabherr et al., 2011; Haas et al., 2013), resulting in between 104,000 contigs for the sponge and 235,000 contigs for the earthworms. Unannotated contigs are provided in supporting file S1_Data. To identify viruses, we used Diamond v0.7.11.60 (Buchfink et al., 2014) to search with translated open reading frames (ORFs) against all virus proteins from the NCBI nr database, all predicted proteins from Repbase (Bao et al., 2015), and all proteins from the NCBI RefSeq_protein database (see Materials and Methods). After excluding some low-quality matches to large DNA viruses and matches to phage, this identified nearly 900 potentially virus-like contigs (S2_Data). These matches were examined and manually curated to generate 85 high-confidence virus-like contigs between 0.5 and 12kbp (mean 3.7Kbp), which are the focus of this study. We have provided provisional names for these viruses following the model of Shi *et al.*, (2016) and sequences have been submitted to GenBank under accession numbers MF189971-MF190055.

The majority of these virus-like contigs were related to positive sense RNA viruses (+ssRNA), including *ca.* 20 contigs from the Picornavirales, 10 Weivirus contigs, and around 5 contigs each from Hepeviruses, Nodaviruses, Sobemoviruses, and Tombusviruses. We also identified 18 putative dsRNA virus contigs (Narnaviruses, Partitiviruses and a Picobirnavirus) and 11 negative sense RNA virus (-ssRNA) contigs (5 bunya-like virus contigs, 3 chuvirus-like contigs, and two contigs each from Rhabdoviridae and Orthomyxoviridae). Our curated viruses included five DNA virus-like contigs, all of which were related to the single-stranded DNA Parvoviridae. Sequences very similar to our Caledonia Starfish parvo-like viruses 1, 2 and 3 are detectable in the publicly-available transcriptomes of *Asterias* starfish species (Figure S1; Hennebert *et al.*, 2015). Although some of the virus-like contigs are likely to be near-complete genomes, including several +ssRNA viruses represented by single contigs of >9kbp, many are partial genomes representing only the polymerase, which tends to be highly conserved (Holmes, 2009). We identified virus-like contigs from all of the sampled taxa, although numbers varied substantially, with only three in the earthworm pool and around 40 in the sponge. This may represent differences in host species biology, but more likely reflects the different range of tissues sampled, and/or differences in sampling effort (S1_Text). A detailed description of each virus is provided in S2_Table.

After initially assigning viruses to potential taxonomic groups based on BLASTp hits, we applied a maximum likelihood approach to protein sequences to infer the phylogenetic relationships of each virus. Many derived from large poorly-studied clades recently identified by metagenomic sequencing (C.-X. Li *et al.*, 2015; Shi *et al.*, 2016), and most are related to viruses from other invertebrates. For example, five of the sponge picornavirales were broadly spread across the 'Aquatic picorna-like viruses' clade of Shi *et al.*, (2016), with closest known relatives that infect marine Lophotrochozoa and Crustacea. Associated with the breadcrumb sponge we identified sequences related to the recently described "Weivirus" clade known from marine molluscs (Shi *et al.*, 2016), and from beadlet anemone we identified sequences related to Chuviruses of arthropods (C.-X. Li *et al.*, 2015; Shi *et al.*, 2016). Some of the virus-like sequences are closely-related to well-studied viruses, for example Millport beadlet anemone dicistro-like virus 1 and Caledonia beadlet anemone dicistro-like virus 2 are both very closely related to Drosophila C virus (Jousset et al., 1972; Brun & Plus, 1980) and Cricket Paralysis virus (Reinganum et al., 1970). Others are notable because they lack very close relatives, or because they fall closest to lineages not previously known to infect invertebrates. These include the Caledonia dog whelk rhabdo-like virus 2 sequence, which is represented by a nucleoprotein and falls between the Rabies/Lyssaviruses, and Barns Ness dog whelk orthomyxo-like virus 1—for which the PB2 polymerase subunit falls between Infectious Salmon Anaemia virus and the Influenza/Thogoto virus clade (Figure 2; the PA polymerase subunit shows similarity to the Thogoto viruses, but not other Orthomyxoviruses.) All phylogenetic trees are presented with support values and GenBank sequence identifiers in S1_Figure, and the alignments used for phylogenetic inference and newick-format trees with support values are provided in S3_Data and S4_Data respectively.

Evidence supporting the viruses as bone fide infectious agents of the target hosts

In addition to avoiding gut content and/or nematode contamination, we sought to provide four lines of corroborating evidence that these virus-like sequences represent infections of the targeted hosts. First, we estimated the representation of potential hosts in each pool by mapping RNA-seq forward reads to the Cytochrome Oxidase 1 (COI, a highly expressed eukaryotic gene) contigs that could be identified in our assemblies. COI reads that could not be matched to the target host species amounted to less than 0.2% of the target's own COI reads in every case, arguing against substantial contamination with non-target taxa such as parasites or commensals. Contamination was higher in the brown alga, perhaps reflecting the challenge of recovering RNA from this taxon (S1_Text). In this case we identified around 10 contaminating taxa, amounting to 5% of the COI reads (including taxa that we might expect to live as ectocommensals on seaweeds, such as a bryozoan with 3.6% and a tunicate with 1.2%). We also identified some cross-contamination and/or adapter-switching between libraries that shared an Illumina lane (e.g. Kircher et al., 2012; Ballenghien et al., 2017), with a mean of < 0.2% of COI reads deriving from the other libraries in the lane. Nevertheless, an average of 99.78% of the mapped COI reads in each invertebrate library derived from the targeted species (93% in the brown alga), suggesting that any viruses of contaminating species would need to be at a very high titre to be detected and erroneously attributed to the target host (read counts are provided in S3_Table).

Second, we remapped reads to the 85 focal virus contigs to measure the number of virus-derived reads relative to host COI. We reasoned that sequence reads from genuine infections are likely to appear in a single host species and to have high representation, whereas viruses present only as surface or sea-water contaminants would be present at low titre and seen in association with the multiple hosts that were collected together. We only identified one virus present at an appreciable titre in more than one host pool, suggesting that the virus-like sequences do not in general represent biological or experimental contaminants, and that the majority of viruses infected only one of the sampled host species. The exception was a 1.3 kbp partiti-like virus contig (Caledonia partiti-like virus 1), which displayed substantial numbers of reads in both the anemone and the sponge—perhaps indicative of closely related viruses infecting these highly divergent taxa. Four viruses were present at a very high level (>1% of COI in at least one library), including Caledonia beadlet anemone dicistro-like virus 2, Millport beadlet anemone dicistro-like virus, Lothian earthworm picorna-like virus 1, and in the brown alga, Barns Ness serrated wrack bunya/phlebo-like virus 1. In total, 18 of the 85 virus contigs were present at >0.1% of host COI in at least one library, and all but 8 were present at >0.01% of COI (S3_Table).

Third, we recorded which strand each read derived from, as actively replicating DNA viruses and -ssRNA viruses generate substantial numbers of positive sense mRNAs. Note that, although +ssRNA viruses also produce complementary (negative sense) RNA during replication, the positive to negative strand ratio is usually very high (e.g. 50:1 to 1000:1 in Drosophila C Virus), making the negative strand hard to detect by metagenomic sequencing. As expected, all of the -ssRNA viruses in our sample (Orthomyxoviridae, Rhabdoviridae, Bunyaviridae/Arenaviridae-like, Chuvirus-like) displayed substantial numbers of reads from both strands, consistent with active replication. We also detected negative-sense reads for many of the +ssRNA viruses, but not at a substantially higher rate than seen for host mRNAs such as COI (S3_Table). These data provide strong evidence that all of the negative sense RNA viruses we detected comprise active infections, and are consistent with replication by the other viruses. Surprisingly, only one of the five DNA viruses (Millport starfish parvo-like virus 1) showed the strong positive sense bias expected of mRNAs, whereas the other four displayed a large negative sense bias. This suggests that these parvovirus-like sequences derive from expressed Endogenous Viral Elements (‘EVEs’; Katzourakis & Gifford, 2010) rather than active viral infections.

Fourth, we selected 53 of the putative virus contigs for further verification by PCR (Materials and Methods; S2_Table). For most of these, we confirmed that the template was detectable by RT-PCR but not by (RT-negative) PCR, confirming that the viruses were not present in DNA form, i.e. were not EVEs (Materials and Methods; S2_Table). The exceptions were Caledonia dog whelk rhabdo-like virus 2 and (as expected) the DNA parvovirus-like contigs, which did appear in RT-negative PCR. to We then estimated virus prevalence in the wild, using RT-PCR to survey all of our samples in pools of between 7 and 30 individuals. The majority of viruses had an estimated prevalence in the range 0.79-100% (S4_Table), with some virus-like sequences present in all sub-pools of the species. These ‘ubiquitous’ sequences included Caledonia dog whelk rhabdo-like virus 2, Caledonia starfish parvo-like virus 2, Caledonia starfish parvo-like virus 3, Caledonia beadlet anemone parvo-like virus 1, and 13 of the sponge viruses. This suggests that these sequences are common or that they are ‘fixed’ in the population, which could be consistent with integration into the host genome (i.e. an EVE). However, given the sampling scheme, a sponge virus at >36% prevalence has a 95% chance of being indistinguishable from ubiquitous. In addition, with the exception of Caledonia dog whelk rhabdo-like virus 2, none of the RNA viruses could be amplified from a DNA template. Taken together, the use of tissue dissection in RNA preparation, the distribution of viruses across sequencing pools, the host distribution of related viruses, the abundance and strand specificity of virus reads, the absence of DNA copies (for all but one of the RNA viruses), and the variable prevalence of the putative viruses in wild populations, support the majority of these sequences as *bone fide* viral infections of the sampled species.

Only the dog whelk displays small RNAs derived from RNA viruses

Based on our knowledge of antiviral RNA interference in plants, fungi, arthropods, and nematodes we expected viral infections to be associated with large numbers of Dicer-generated viRNAs, with a narrow size distribution peaking between 20nt (as seen in Lepidoptera; Zografidis *et al.*, 2015) and 22nt (as seen in chelicerates, hymenopterans, and nematodes; Félix *et al.*, 2011; Chejanovsky *et al.*, 2014; Schnettler *et al.*, 2014). We therefore sequenced small RNA libraries from each pool, generating an average of around 60 million small RNA reads per library. These included untreated RNA, RNA treated with 5' polyphosphatase (to remove 5’ triphosphates, thus facilitating the detection of viRNAs generated by synthesis), and oxidised RNA (to increase the representation of small RNAs bearing a 3' 2-O-methyl group, such as piRNAs and viRNAs). To ensure that we did not exclude viRNAs that had been edited (e.g. by ADAR; see Samuel, 2012), or that contained untemplated bases (e.g. 3' adenylation or uridylation; Ameres et al., 2010), our mapping approach permitted at least two high base-quality mismatches within a 21nt sRNA. We also confirmed that remapping with local alignment, which permits any number of contiguous mismatches at either end of the read, did not substantially alter our results.

We successfully recovered abundant miRNAs in all of the metazoan samples, with between 20% (sponge) and 80% (starfish) of 20-23nt RNAs from untreated libraries mapping to known miRbase miRNAs (Kozomara & Griffiths-Jones, 2014). Consistent with the absence of a 3' 2-O-methyl group, these miRNA-like reads had much lower representation in the oxidised libraries, there comprising only 0.4% (earthworms) to 14% (dog whelk) of 20-23nt RNAs. We also identified characteristic peaks of small RNAs derived from ribosomal RNA at 12nt and 18nt in the sponge, at 12nt and 16nt in the sea anemone, and in oxidised libraries from all organisms. The only exception to this overall pattern was for the sea anemone, in which oxidation had no effect on the number of miRNAs, although did strongly affect the overall size distribution of rRNA-derived sRNAs. This suggests the presence of a 3' 2-O-methyl group in sponge miRNAs (S2_Figure).

Despite our identification of more than 40 RNA virus-like contigs associated with the sponge, 17 in the sea anemone, and three in the earthworms, we were unable to detect a signature of abundant viRNAs in any of these three organisms. On average, less than 0.002% of 17-35nt RNAs from these organisms mapped to the RNA virus contigs, and those that did map were enriched for shorter lengths (17-19nt), lacked a clearly defined size distribution, and were less common in the oxidised than non-oxidised libraries (S2_Figure; S3_Table)—features consistent with non-specific degradation products, rather than viRNAs. (Note that the starfish sample lacked detectable RNA viruses, precluding the identification of RNA-virus viRNAs).

The only metazoan sample to display a clear viRNA signature was the dog whelk (*Nucella lapillus*), with 0.14% of oxidised small RNAs derived from four of the seven RNA virus-like contigs. These included both contigs of Barns Ness dog whelk orthomyxo-like virus 1, Caledonia dog whelk rhabdo-like virus 1, and Caledonia dog whelk rhabdo-like virus 2. A Narnavirus-like contig and a very low titre Bunyavirus-like contigs were not major sources of viRNAs. However, these small RNAs did not show the expected size or strand signature of canonical Dicer-generated viRNAs (Figure 3; S3_Figure). Instead, viRNA lengths formed a broad distribution from 26 to 30nt (peaking at 28nt), more consistent with piRNAs seen in the *Drosophila* and mammalian germlines. These small RNAs were derived almost entirely from the negative-sense (i.e. genomic) strand of Barns Ness dog whelk orthomyxo-like virus 1 (Figure 3 A and B) and Caledonia dog whelk rhabdo-like virus 2 (Figure 3 D), but from both stands of Caledonia dog whelk rhabdo-like virus 2 (Figure 3 C and E). Although this size distribution is more consistent with the piRNA pathway, only those from Caledonia dog whelk rhabdo-like virus 2 (a suspected EVE, see above) displayed the strong 5'U bias expected of primary piRNAs (Figure 3D), and none showed any evidence of ping-pong amplification. In all three cases, the putative dog whelk viRNAs were derived from the whole length of the viral genome—albeit with strong hotspots in Caledonia dog whelk rhabdo-like virus 2. Relative to miRNAs, these RNA-virus derived viRNAs were much more strongly represented in the oxidised library than the untreated library, with the miRNA:viRNA ratio increasing 300-fold—consistent with the presence of a 3' 2-O-methyl group (S2_Figure, S3_Figure).

The sea anemone and starfish display 5'U 26-30nt RNAs from DNA virus-like contigs

DNA viruses are a source of Dicer-mediated viRNAs in arthropods and in plants, and antiviral RNAi pathways are important for antiviral immunity to DNA viruses in both groups (reviewed in Rajeswaren & Pooggin, 2012; Bronkhorst et al., 2013). Although our RNA sequencing strategy was intended to detect RNA viruses, we also identified four novel parvo/densovirus-like contigs (Parvoviridae; single-stranded DNA) in the starfish, and one in the sea anemone. These sequences were a substantial source of small RNAs in both organisms, particularly the starfish—contributing 0.3% of small RNAs in the untreated libraries and 3.4% of small RNAs in the oxidised library. In four of the five cases, the these small RNAs were almost exclusively negative sense, were 26 to 30nt in length (peaking at 28nt), and were very strongly biased toward U in the 5' position—resembling primary piRNAs (Figure 4). However, the high prevalence and/or negative strand RNAseq bias of these four source contigs is consistent with expressed genomic integrations (EVEs) rather than active viral infections. In the other case, Caledonia starfish parvo-like virus 1, both positive and negative sense reads were detectable, the negative sense reads again displayed a strong 5' U bias, but the positive sense reads displayed a postion-10 ‘A’ ping-pong signature (Figure 4 B), as expected of piRNAs. Relative to miRNAs, these putative piRNAs were much more strongly represented in the oxidised library than the untreated library, consistent with the presence of a 3' 2-O-methyl group (S2_Figure; S3_Figure).

All of the sampled metazoa display somatic TE-derived piRNAs

Transposable elements and TE-derived transcripts represent a major source of piRNAs in the germlines of *Drosophila* (Brennecke et al., 2007), *C. elegans* (Das et al., 2008; Lee et al., 2012), mice (Deng & Lin, 2002; Kuramochi-Miyagawa, 2004), and zebrafish (Houwing et al., 2007). Piwi-interacting RNAs are also detectable in Cnidaria and Porifera, although their tissue specificity is unclear (Grimson et al., 2008). In addition, TE transcripts in *Drosophila* are also processed by Dicer to generate 21nt endo-siRNAs. We therefore selected a total of 146 long high-confidence TE contigs from our assemblies to analyse TE-derived small RNAs (these were chosen based on length and similarity, and to best illustrate small RNA properties; contigs are provided in S5_Data). We identified large numbers of TE-derived putative piRNAs in the somatic tissues of all the sampled organisms (Figure 5). In total, between 0.17% (starfish) and 1.7% (dog whelk) of untreated small RNA reads mapped to the 146 high-confidence TE contigs (S3_Data; S3_Figure; S4_Figure). In every case except the anemone, the putative piRNAs were more highly represented in the oxidised library than in untreated or polyphosphatase-treated libraries (1.4-6% of oxidised reads), suggesting that they are 3' 2-O-methylated and result from cleavage rather than synthesis. Despite very large numbers of piRNAs for some TE contigs, we did not observe endo-siRNA -like small RNAs similar those observed in *Drosophila* (e.g. Czech *et al.*, 2008).

We observed putative piRNAs derived from one or both strands of the TEs (Figure 5). Where they derived predominantly from a single strand they were generally strongly 5'U-biased (consistent with primary piRNAs). Where they derived from both strands, those from the second strand presented evidence of ‘ping pong’ amplification (i.e. no 5' U bias, and a strong ‘A’ bias at position ten; Figure 5; S4_Figure). However, the piRNA size distribution varied substantially among organisms and TEs. In the sponge, the length of the 5' U-biased piRNAs either peaked at 23-24nt, or presented a broader bimodal distribution peaking at 23-24nt and 27-29nt. Where piRNAs derived from both strands, the strand with a ping-pong signature showed a shorter length distribution (22-23nt). In a few cases the putative sponge piRNAs from both strands showed a strong 5'U bias with no evidence of ping-pong amplification. In the sea anemone we consistently identified a strong peak of 5'-U biased sRNAs peaking at 28-29nt on one strand, but a generally bimodal distribution from the second ‘ping-pong’ strand (if piRNAs were present), peaking at around 23nt and 28nt. Again, both strands occasionally displayed a 5'-U bias and no evidence of ping-pong amplification. The patterns were again similar in the starfish and the earthworms, except that size distributions were unimodal, peaking at 29-30nt in the 5'-U biased strand and 25-26nt (starfish) and 26-27nt (earthworms) in the ‘ping-pong’ strand.

As with viRNAs, the only exception to this general pattern was seen in the dog whelk. In addition to TE-like contigs that displayed a classical piRNA-like signature (28nt 5'U RNAs from one strand; 26-

28nt ‘ping-pong’ RNAs from the opposite strand), a small number of TE-like contigs in the dog whelk had an sRNA signature that resembled that of the dog whelk viruses Barns Ness dog whelk orthomyxo-like virus 1 and Caledonia dog whelk rhabdo-like virus 1. In these TE-like contigs, the sRNAs were derived from one or both strands, peaked broadly at 26-30nt, and lacked any bias in base composition or evidence of ‘ping-pong’ (Figure 5 E and F). This indicates that some TEs are processed as if they were viruses, perhaps suggesting retrovirus-like horizontal transmission (e.g. Gypsy, S4_Figure D). A minority of TE-like contigs displayed an intermediate pattern, with a weak 5'U-bias from one strand, and a broad peak that lacked a pong-pong signature from the other strand. Such an intermediate pattern could result either from a single TE targeted by two different mechanisms, or from cross-mapping of sRNAs derived from different copies of the same TE inserted in different locations/contexts. As before, our permissive mapping approach and re-mapping using local alignments reduces the possibility that a large category of sRNAs escaped detection.

Virus and TE-derived 21nt 5'U RNAs are present in a brown alga

Virus-derived small RNAs have been well characterised in plants, fungi, and some metazoa, but other major eukaryotic lineages such as Heterokonts, Alveolates, Excavates and Amoebozoa have received less attention. In principle, a metagenomic approach could also be applied to these lineages, but the difficulty of collecting large numbers of individuals of a single lineage makes this challenging for single-celled organisms. Here we have taken advantage of multicellularity in the brown algae (Phaeophyceae, Heterokonta) to test for the presence of viRNAs using the serrated wrack, *Fucus serratus*. Based on a single pooled sample of tissue from 100 individuals, we identified large numbers of small RNAs with a tight distribution between 22 and 23nt, peaking sharply at 21nt. Almost all of the 21nt sRNAs were 5' U (S2_Figure), as has been seen for sRNAs in diatoms (Bacillariophyceae, Heterokonta; Rogato et al., 2014). Although miRNAs have been described for two other brown algae, *Ectocarpus siliculosus* (Cock et al., 2010; Tarver et al., 2015) and *Saccharina japonica* (Cock et al., 2017), we were unable to identify homologues of known miRbase miRNAs among these reads. This may reflect a lack of sensitivity, as the miRNA complements of the studied brown algae are highly divergent (Cock *et al.*, 2017), and miRNAs of *Fucus serratus* may be sufficiently divergent again to be undetectable based on sequence similarity. In contrast, 1.8% of small RNAs corresponded to the selected TE contigs. These were derived from both strands, but as expected given the absence of Piwi, displayed no evidence of ‘ping-pong’ amplification—with sRNAs from both strands showing a 5' U bias. We also detected viRNAs corresponding to a -ssRNA bunya-like virus (Barns Ness serrated wrack bunya/phelobo-like virus 1; Figure 3 E; S3_Figure). Numbers were relatively small, comprising 0.01% of all small RNA reads, but these were derived from both strands along the full length of the contig, peaked sharply at 21nt, and were almost exclusively 5'U. We did not detect a viRNA signature from a further two -ssRNA or from four dsRNA virus-like contigs, although their titre was very low compared to Barns Ness serrated wrack bunya/phelobo-like virus 1.

The phylogenetic distribution and expression of RNAi-pathway genes

We sought to examine whether the phylogenetic distribution and expression of RNAi pathway genes in our samples was consistent with the small RNAs we observed. As expected, based on the presence of abundant miRNAs and/or an antiviral pathway, and given what is known for their close relatives (Grimson et al., 2008; Shoguchi et al., 2013; Gao et al., 2014; Coruh et al., 2015; Huang et al., 2015; Bollmann et al., 2016; Liew et al., 2016; Rosani et al., 2016), we identified two deeply divergent Dicer transcripts in the sea anemone, and a single Dicer transcript in each of the other metazoan species. The single Dicers seen in the starfish, dog whelk, and earthworms were more similar to Dicer-1 from the *Drosophila* miRNA pathway than to arthropod Dicer-2-like genes that mediate antiviral RNAi.

Similarly consistent with an antiviral RNAi and/or a miRNA pathway, and with what is known for their close relatives (Buck & Blaxter, 2013; Hu et al., 2013; Moran et al., 2013; Schnettler et al., 2014; Singh et al., 2015; Liew et al., 2016; Rosani et al., 2016; Takeuchi et al., 2016), we identified two deeply divergent (non-Piwi) Argonaute transcripts in the sponge and in the anemone (S6_Table), and single Argonaute transcripts in the dog whelk and in the starfish. We identified three distinct Argonaute transcripts in the mixed-earthworm species pool, although these may represent the species present. The dog whelk, starfish, and earthworm Argonautes were all more closely related to arthropod Ago-1 (which binds miRNAs but rarely viRNAs) and to vertebrate Argonautes, than to insect Ago2-like genes that mediate antiviral RNAi. It is likely that these genes mediate the miRNA pathway in these organisms, although it is possible that they may also mediate the production of novel viRNAs seen in the dog whelk. We also identified a single Dicer and Argonaute in the *Fucus*, which is consistent with what has been seen in other brown algae (Cock *et al.*, 2010, 2017; Tarver *et al.*, 2015), and with the presence of both miRNAs and viRNAs.

In metazoa, the piRNA pathway suppresses transposable element transcripts, and activity and is mediated by homologs of the *Drosophila* nuclease ‘Zucchini’ and the piwi-family Argonaute proteins Ago3 and Piwi/Aub. In mammals, fish, *C. elegans* and *Drosophila*, this pathway is primarily active in the germline and its associated somatic tissues (Deng & Lin, 2002; Kuramochi-Miyagawa, 2004; Brennecke *et al.*, 2007; Houwing *et al.*, 2007; Das *et al.*, 2008; Lee *et al.*, 2012), whereas in sponges and cnidarians—which lack a segregated germline—Piwi homologs are ubiquitously expressed (Denker *et al.*, 2008; Funayama *et al.*, 2010). Consistent with our finding of TE-derived piRNAs displaying a canonical ‘ping-pong’ signature, we identified single Zucchini, Ago3 and Piwi homologs in four of the five metazoans surveyed (S6_Table). The exception was the sea anemone, in which we could only identify a single Piwi (more similar to *Drosophila* Piwi/Aub than to Ago-3). Surprisingly, although we did not identify canonical piRNAs in the brown alga, we did identify a possible Piwi-like transcript. However, its relatively low expression and apparent similarity to Piwi genes from the Lophotrochozoa suggest it most likely derives from the contaminating bryozoan identified by COI reads (above). Finally, consistent with the altered small RNA profile associated with oxidation, we were able to identify a single homolog of the RNA methyl transferase Hen-1 in each of the metazoan species, although we were unable to detect a homolog in the brown alga. These sequences have been submitted to GenBank under accession numbers MF288049-288076.

Figure 2: Phylogenetic relationships of virus-like contigs from the dog whelk

Mid-point rooted maximum likelihood phylogenetic trees for each of the virus-like contigs associated with viRNAs in the dog whelk (*Nucella lapillus*). New virus-like contigs described here are marked in red, sequences marked 'TSA' are derived from public transcriptome assemblies of the species named, and the scale is given in amino acid substitutions per site. Panels are: (A) rhabdoviruses related to lyssaviruses, inferred using the protein sequence of the nucleoprotein (the only open reading frame available from this contig, which is likely an EVE); (B) orthomyxoviruses related to influenza and thogoto viruses, inferred using the protein sequence of PB1; (C) rhabdoviruses and chuviruses, inferred from the RNA polymerase. Support values and accession identifiers are presented in S1_Figure and S4_Data, and alignments in S3_Data. Given the high level of divergence, alignments and inferred trees should be treated as tentative.

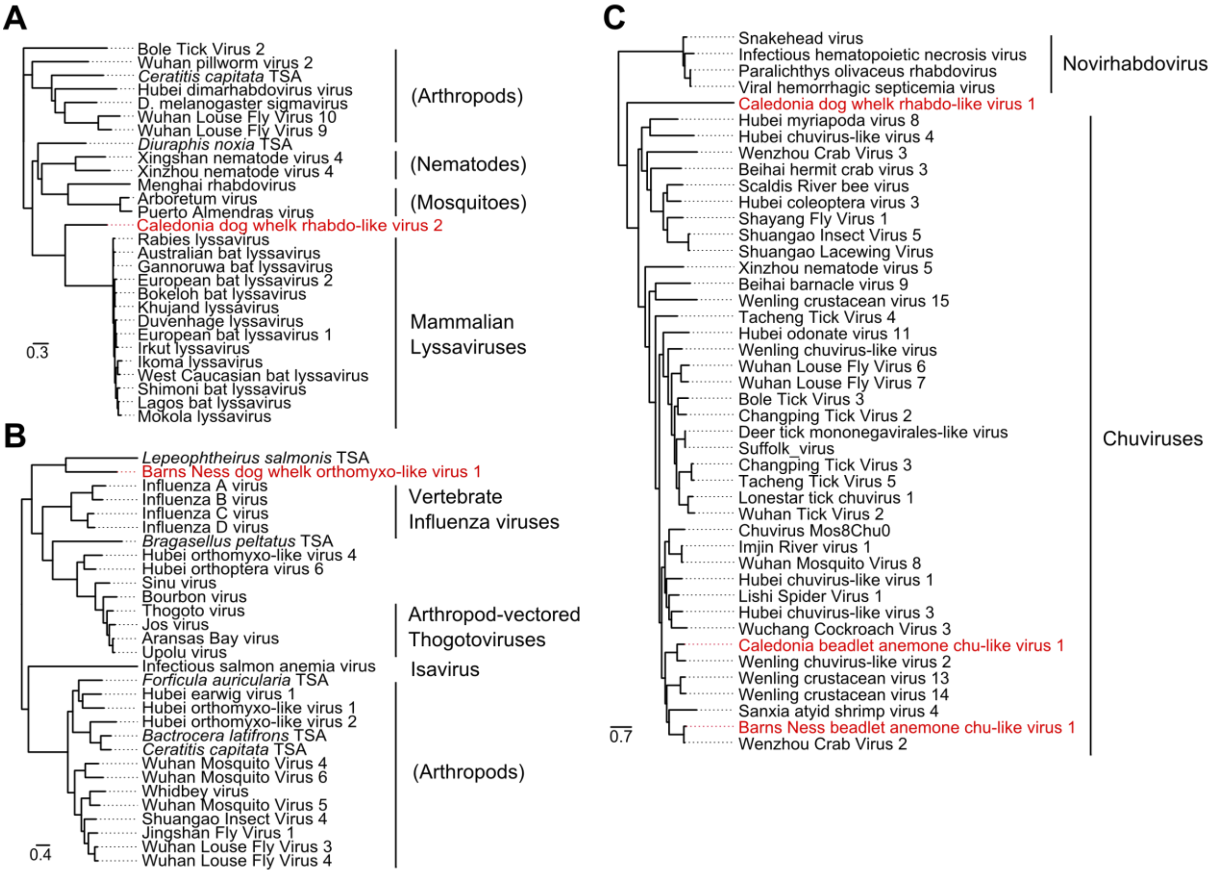


Figure 3: small RNAs from RNA virus-like contigs

Panels to the left show the distribution of 20-30nt small RNAs along the length of the virus-like contig, and panels to the right show the size distribution small RNA reads coloured by the 5' base (U red, G yellow, C blue, A green). Read counts above the x-axis represent reads mapping to the positive sense (coding) sequence, and counts below the x-axis represent reads mapping to the complementary sequence. For the dog whelk (A-D), only reads from the oxidised library are shown. Other dog whelk libraries display similar distributions and the small-RNA 'hotspot' pattern along the contig is highly repeatable (S3_Figure). Small RNAs from the two segments of the orthomyxovirus (A and B) show strong strand bias to the negative strand and no 5' base composition bias. Those from the first rhabdo-like virus (C) display little strand bias and no base composition bias, and those from the second rhabdo virus-like contig, which is a probable EVE (D), derive only from the negative strand and display a very strong 5' U bias. There were insufficient reads from the positive strand of this virus to detect a ping-pong signature. Small RNAs from the four dog whelk contigs all display 28nt peaks. Small RNAs from the *Fucus* bunya/phlebo-like virus identified in the brown alga (E) derive from both strands, and show a strong 5' U bias with a peak size of 21nt. The data required to plot the size distributions is provided in S5_Table

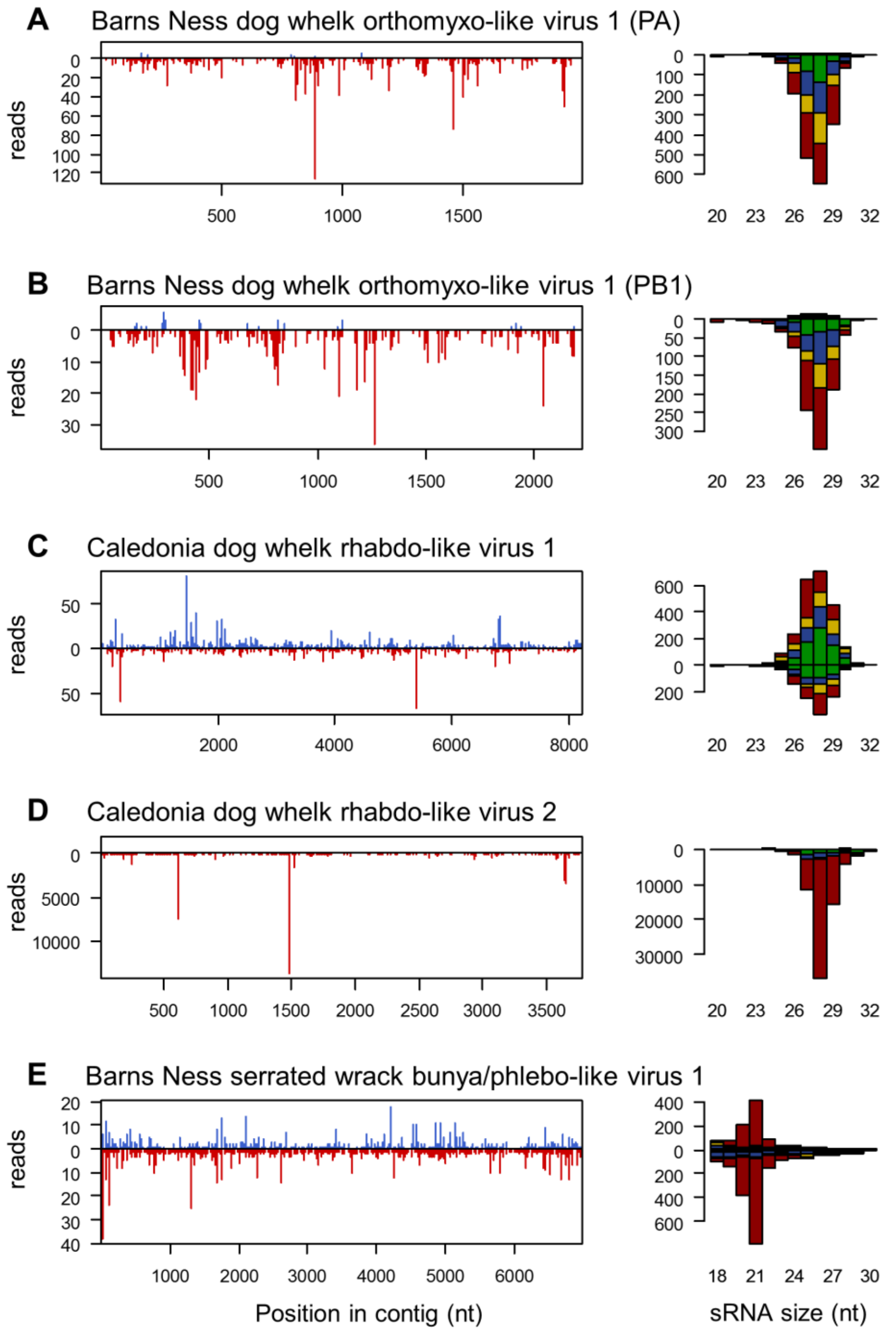


Figure 4: **small RNAs from DNA parvo/densovirus-like contigs**

Panels to the left show the distribution of 20-30nt small RNAs along the length of the parvo/densovirus-like contigs, and panels to the right show the size distribution small RNA reads coloured by the 5' base (U red, G yellow, C blue, A green). Read counts above the x-axis represent reads mapping to the positive sense (coding) sequence, and counts below the x-axis represent reads mapping to the complementary sequence. Only reads from the oxidised library are shown, but other libraries display similar distributions, and the small-RNA 'hotspot' pattern is highly repeatable (S4_Figure). For all but one of the parvo/denso-like virus contigs, the small RNAs derived exclusively from the negative sense strand and showed a strong 5'U bias, consistent with piRNAs derived from endogenous copies (see main text). For one contig (B: Millport starfish parvo-like virus 1) reads derived predominantly from the positive strand and did not display a 5' U bias. Although the number of unique small RNA sequences from this virus was small, the positive-sense small RNAs showed a slight bias to A at position 10, consistent with ping-pong (S4_Figure). The data required to plot these size distributions is provided in S5_Table

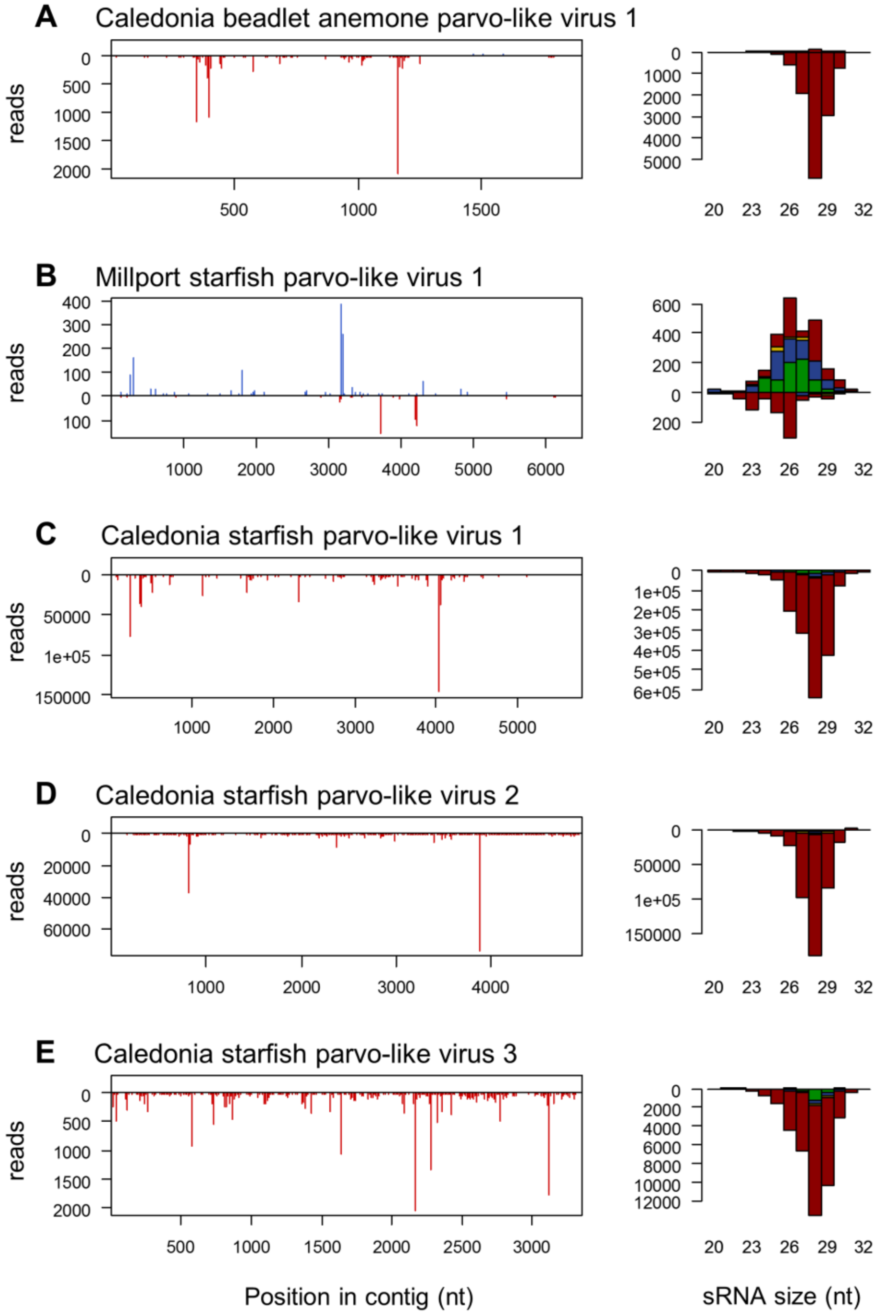
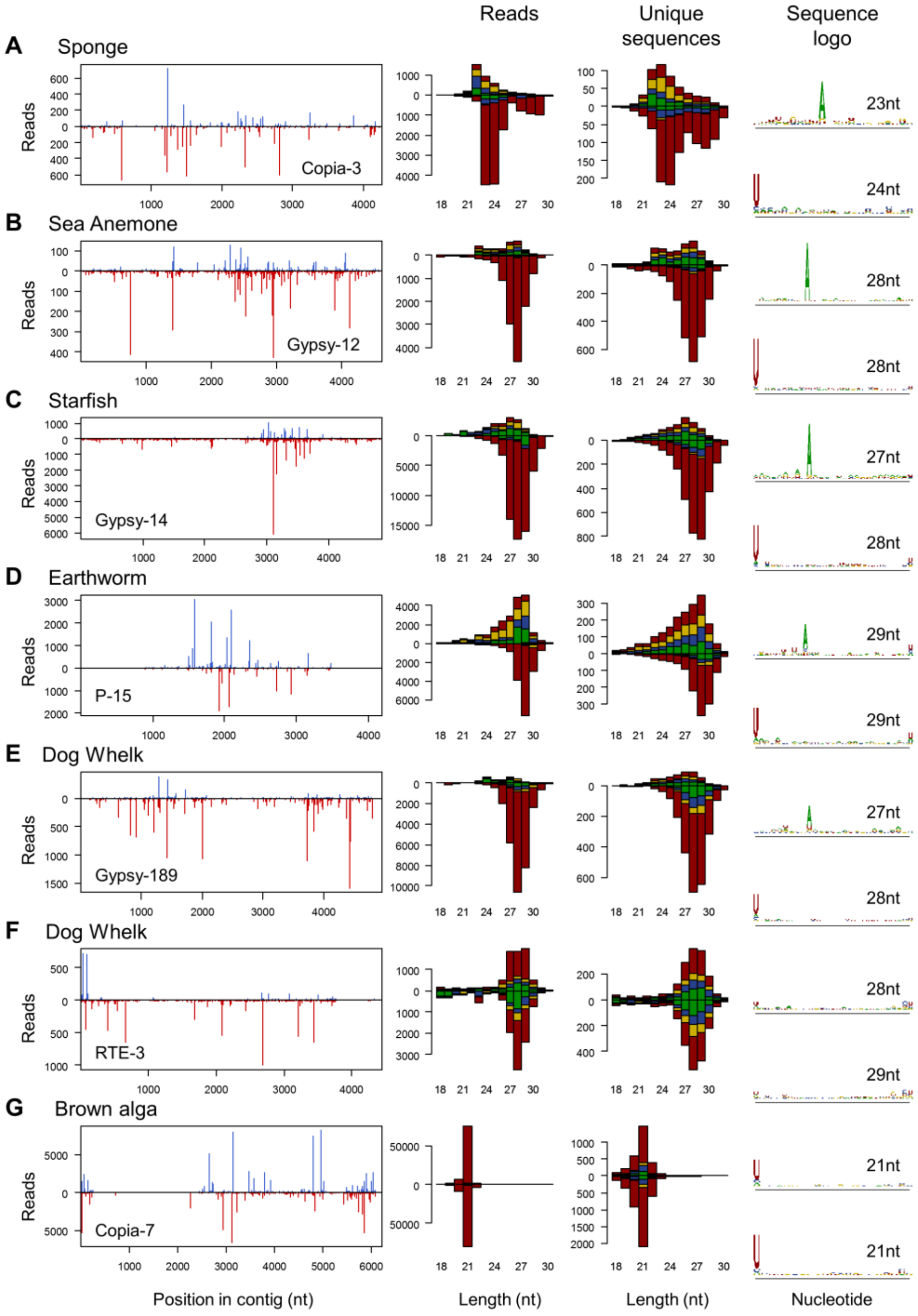


Figure 5: small RNAs from TE-like contigs

The four columns show (left to right): the distribution of 20-30nt small RNAs along the length of a TE-like contig; the size distribution of small RNA reads (U red, G yellow, C blue, A green); the size distribution for unique sequences; and the sequence 'logo' of unique sequences for the dominant sequence length. Read counts above the x-axis represent reads mapping to the positive sense (coding) sequence, and counts below the x-axis represent reads mapping to the complementary sequence. For the sequence logos, the upper and lower plots show positive and negative sense reads respectively, and the y-axis of each measures relative information content in bits. Where available, reads from the oxidised library are shown (A-F), but other libraries display similar distributions (S4_Figure). These examples were chosen to best illustrate the presence of the 'ping pong' signature, but other examples are shown in S4_Figure. Note that the size distribution of TE-derived small RNAs varies substantially among species, and that the dog whelk (E and F) displays at least two distinct patterns, one (F) reminiscent of that seen for some RNA virus contigs (Figure 3 C). The data required to plot these figures is provided in S5_Table



Discussion

Evidence for antiviral RNAi against -ssRNA viruses in the dog whelk and brown alga

We identified abundant viRNAs from RNA viruses in two of the six species we tested: the dog whelk (*Nucella lapillus*), and a brown alga (*Fucus serratus*). These derived from three RNA virus-like contigs in the dog whelk (two rhabdo-like viruses and an orthomyxo-like virus) and one in the brown alga (a bunya/phlebo-like virus). The viRNAs displayed some, but not all, of the expected properties of a canonical antiviral RNAi response. First, in the dog whelk, the broad length distribution around 28nt and the strong strand-bias were not consistent with Dicer processing, which is expected to generate sRNAs from both strands simultaneously, and to result in a characteristic sequence length determined by the distance between the PAZ and RNaseIII domains (MacRae, 2006). Second, the strong 5'U bias seen for both positive and negative sense viRNAs in the brown alga has not previously been reported for virus-derived sRNAs, although it is consistent with all other small RNAs of brown algae (Tarver *et al.*, 2015; Cock *et al.*, 2017). However, the viRNAs did display a distinct size distribution, they derived from the full length of the viral sequence, and in the dog whelk they were over-represented after oxidation, implying the presence of a 3' 2-O-methyl group (Figure 3, S3_Figure). These viRNA signatures therefore suggest an active response in both the dog whelk and the brown alga, and hence the presence of an antiviral RNAi pathway in these species—although there is also substantial divergence from canonical arthropod antiviral RNAi, and possibly from the ancestral state in metazoa (below).

We have also considered three alternative explanations for these data. First, it is possible that the result is artefactual, and that all of the virus-like reads derive from another unknown source, such as environmental contamination. However, the large number of complementary (mRNA) sequences show these -ssRNA viruses to be active, the sequences were not identified in any of the other co-collected taxa, and the COI read counts in the dog whelk show contamination rates to be low. Contamination was higher for the brown alga, but the virus would need to be at extremely high copy number in the contaminating taxon to achieve the observed 3% of brown alga COI expression. Second, it is possible that the virus-like contigs represent expressed host loci, such as EVEs. However, sequences were not detectable by PCR in the absence of reverse transcription, in the dog whelk the low prevalence means that any putative EVE must be segregating, and prevalence differed between sampling sites—consistent with an infectious agent. Moreover, in a previous analysis of insect viruses, expressed EVEs were found to be rare relative to active viral infections: zero of 20 viruses identified by metagenomic sequencing in *Drosophila* (Webster *et al.*, 2015). Third, even if the virus-like sequences do represent real infections, it is possible that the small RNAs do not represent an active RNAi-like response. However, their distinctive size distributions, the presence of a 3' 2-O-methyl group in the dog whelk, and near 100% 5'U in the brown alga, argue strongly that these viRNAs are the result of active biogenesis, rather than degradation.

In contrast, it seems probable that the shorter rhabdo-like virus fragment from the dog whelk (Caledonia dog whelk rhabdo-like virus 2; Figure 3D) is a host-encoded EVE. First, the only open reading frame is homologous to a nucleoprotein, but we could not detect a polymerase—despite its close relationship with the Lyssaviruses (Figure 2A). Second, RNA sequencing was dominated by negative-sense reads, suggesting a lack of mRNA expression, but consistent with host-driven expression of an integrated locus. Third, the small RNAs were exclusively negative-sense and 5'U, as sometimes seen for primary piRNAs derived from EVEs. Fourth, the sequence was ubiquitous, consistent with fixation and thus genome integration. Fifth, we were able to PCR amplify a band from a DNA template. If this sequence is an EVE, this could represent an alternative antiviral RNAi mechanism, akin to the piRNA-generating EVEs seen in *Aedes* mosquitoes (Palatini *et al.*, 2017).

Evidence for RNAi against other RNA viruses

Despite the presence of more than 70 high-confidence RNA virus-like contigs, we were unable to identify an abundant or distinct population of viRNAs derived from RNA viruses in the sponge, sea anemone, or earthworms (the starfish lacked detectable RNA viruses). Whereas the -ssRNA viruses in the dog whelk produced 1-100 viRNA reads per RNAseq read (oxidised library; S5_Figure), and Barns Ness serrated wrack bunya/phelbo-like virus 1 produced *ca.* 0.1 viRNA reads per RNAseq read in the brown alga (S5_Figure), none of the other RNA viruses gave rise to ≥0.001 viRNA reads per RNAseq read. In contrast, in an equivalent analysis of *ca.* 20 RNA viruses in wild-caught *Drosophila*, all putative viruses produced viRNAs at approximately 10-1000 viRNAs per RNAseq read (Webster *et al.*, 2015). This represents a qualitative difference in the processing of RNA viruses between *Drosophila* (and other ecdysozoans, including other arthropods (Bronkhorst & van Rij, 2014; Schnettler *et al.*, 2014; Gammon & Mello, 2015) and nematodes (Félix *et al.*, 2011; Coffman *et al.*, 2017; Gammon *et al.*, 2017)) and sponges (Porifera), anemones (Cnidaria), and earthworms (Annelida). Importantly, it suggests that these lineages either do not process RNA viruses into small RNAs in the way that plants, fungi, nematodes or insects do, or that they do so at a level that is undetectable through bulk small RNA sequencing—as recently reported for mammals (Li *et al.*, 2013, 2016; Maillard *et al.*, 2016). If so, then the canonical antiviral RNAi mechanism as seen in arthropods is highly derived relative to that in other metazoan lineages. One practical consequence of this is that it will not be possible to use viRNAs to argue for active viral replication in most metazoa, or to identify viruses that lack sequence similarity to known lineages—as has been done previously for *Drosophila* (Webster *et al.*, 2015).

We believe that our sequencing strategy is likely to have detected any viRNAs that were present, as we were able to detect miRNAs, piRNAs, and small rRNAs. We would also have detected viRNAs bearing a 5' triphosphate or 3' 2-O-methyl group, as well as viRNAs that had been edited or extended by untemplated bases at the 5' or 3' end. However, the absence of evidence for abundant viRNAs is not necessarily strong evidence for their absence: it is hard to demonstrate that RNA viruses do not give rise to small RNAs in these lineages, and other explanations for an absence of viRNAs remain possible. One alternative is that all of the other RNA-virus like contigs that we identified from the sea anemone, sponge, earthworm, or dog whelk were inactive and/or encapsidated at the time of collection, and thus not subject to Dicer processing. However, this is unlikely for three reasons. First, it can be ruled out for eight of the nine highest titre dsRNA viruses in the sponge, as these all showed a strong positive-strand RNAseq bias, consistent with gene expression. Second, it is not supported by the two -ssRNA virus contigs in the earthworms, which also displayed positive sense mRNA reads (although the virus copy-number was extremely low, such that that we had little power to identify either positive sense RNAseq reads or viRNAs). Finally, although the small number of negative sense reads resulting from +ssRNA virus replication makes it hard to exclude the possibility that they were inactive, it would be surprising if all of the -ssRNA viruses and dsRNA viruses (including those in the dog whelk and brown alga) were active, but none of the +ssRNA viruses were.

A more plausible alternative is that abundant viRNAs are characteristic of a response against -ssRNA viruses in these lineages, but are not characteristic of the response against +ssRNA or dsRNA viruses. This is an appealing hypothesis, as it is also consistent with our failure to detect viRNAs from putative dsRNA narnaviruses in the dog whelk and brown alga, and to a putative +ssRNA nodavirus in the brown alga. If so, then an apparent absence of antiviral RNAi in the sponge, sea anemone and earthworms may really reflect differences in the composition of the RNA virus community, with a preponderance of -ssRNA viruses in the dog whelk and their absence from the sponge or anemone. However, even if -ssRNA viruses, but not +ssRNA viruses or dsRNA viruses, give rise to viRNAs in most metazoan lineages, then this is still in striking contrast to the antiviral RNAi response in plants, fungi, and insects (Zhang *et al.*, 2008; Szittya & Burgyan, 2013; Bronkhorst & van Rij, 2014), and again suggests that antiviral RNAi in arthropods is highly derived relative to other metazoans.

Finally, it also remains possible that the majority of sponges, sea anemones, and annelids do possess an active antiviral RNAi mechanism that generates abundant viRNAs from RNA viruses, but that the particular species we examined here have lost the ability. It is certainly the case that RNAi mechanisms are occasionally lost, as in one clade of the yeast genus *Saccharomyces* (Drinnenberg *et al.*, 2009, 2011). However, unless antiviral RNAi is lost extremely frequently in these three metazoan phyla—which is not the case in arthropods or plants—it is extremely unlikely that we would by chance select three lineages that have lost the mechanism while others retained it.

Evidence for Piwi-pathway targeting of DNA viruses in the sea anemone and starfish

We identified four parvo/denso-like virus contigs in the starfish, and one in the sea anemone. All of these sequences were (necessarily) detectable as RNAseq reads, and were associated with abundant 26-29nt piRNA-like small RNAs (Figure 4). However, RNAseq from three of the four starfish parvo/denso-like virus contigs, and the sea anemone contig, were dominated by negative sense reads. This is hard to reconcile with the normal functioning of ssDNA parvo/denso-like viruses, which replicate via a rolling circle, and may reflect host-driven transcription. For these four contigs, the small RNAs were also almost exclusively negative-sense and 5'U—as expected of primary piRNAs. In contrast, RNAseq and small RNAs reads from Millport starfish parvo-like virus 1 were almost exclusively positive (mRNA) sense, with the negative strand small RNAs showing a 5'U bias and positive strand sRNAs showing weak ‘ping-pong’ signature (S3_Figure). Together, these observations suggest that at least some of parvo/denso-like virus sequences represent expressed EVEs, but also that they are targeted by a piRNA pathway-related mechanism.

Unlike for RNA viruses, we were unable to test whether these sequences represent integrations into the host genome, as integrations are indistinguishable from viral genomic ssDNA by PCR, and both +ssDNA and -ssDNA sequences are usually encapsidated by densoviruses. However, Caledonia starfish parvo-like viruses 1, 2 and 3 are nearly identical to published starfish transcripts, and the two published sequences most similar to Caledonia beadlet anemone parvo-like virus 1 are from an anemone transcriptome and an anemone genome (S1_Figure). In addition, three of the five contigs (two in the starfish, and one in the anemone) appear to be ubiquitous in our wild sample. This ubiquitous distribution and close relationship to published sequences support the suggestion (above) that some of these sequences may be host integrations. The exceptions are Caledonia starfish parvo-like virus 1 and Millport starfish parvo-like virus 1, which both had an estimated prevalence of between 4% and 20% in the larger Millport collection. We were able to recover putatively near-complete genomes of 6.5 and 5.8 Kbp, containing the full length structural (VP1) and non-structural (NS1) genes, from Millport starfish parvo-like virus 1 and Caledonia starfish parvo-like virus 1, respectively (S2_Table).

If these sequences are EVEs, as seems very likely for four of the five, then their expression and processing into piRNAs may reflect the location of integration—for example, into or near to a piRNA generating locus (Arensburger *et al.*, 2011; Handler *et al.*, 2013). In contrast, if these sequences are not host EVEs then the high expression of negative sense transcripts and the presence of primary piRNA-like small RNAs suggests an active Piwi-pathway response targeting DNA viruses in basally-branching metazoans, which has not previously been reported. These are not mutually exclusive, and it is tempting to speculate that such integrations could provide an active defence against incoming virus infections in basal metazoans, as suggested for RNA-virus integrations in *Aedes* mosquitoes (Palatini *et al.*, 2017). If so, the low-prevalence Millport starfish parvo-like virus 1 sequence, which shared 72% sequence identity with Caledonia starfish parvo-like virus 1, but displayed positive sense transcripts, positive and negative sense piRNAs, and a 'ping-pong' signature, is a good candidate to represent an unintegrated infectious virus lineage.

Implications for the evolution of RNAi pathways

The absence of detectable viRNAs in the sponge, sea anemone, or earthworm samples, combined with the presence of 26-29nt (non-piwi) viRNAs in the mollusc and 21nt 5'U viRNAs in the brown alga, changes our perspective of the evolution of antiviral RNAi. Previously, the abundant viRNAs present in plants, fungi, nematodes and arthropods had implied that Dicer-based antiviral RNAi was ancestral to the Eukaryotes and likely to be ancestral in the Metazoa, with a recent modification (or loss: Backes et al., 2014; Bogerd et al., 2014) in the vertebrates—perhaps associated with the evolution of interferons (Benitez *et al.*, 2015). Our findings now suggest three alternative hypotheses. First, antiviral RNAi may have been absent from ancestral Metazoa, and re-evolved on at least one occasion—giving rise to the distinctively different viRNA signatures seen in nematodes, arthropods, vertebrates, and now also a mollusc. Second, the ancestral state may have been more similar to current-day mammals, which do not produce abundant easily-detected viRNAs, but may still possess an antiviral RNAi response (Li *et al.*, 2013, 2016; Maillard *et al.*, 2016). In this scenario, antiviral RNAi has been maintained as a defence—possibly since the origin of the eukaryotes—but has diversified substantially to give the distinctive viRNA signatures now seen in each lineage. Third, dsRNA, +ssRNA, -ssRNA, and DNA viruses may be targeted differently by RNAi pathways in basal metazoans, but arthropods have recently evolved a defence that gives rise to the same viRNA signature from each class. It is not possible to distinguish among these hypotheses without broader taxonomic sampling and experimental work in key lineages. For example, analyses of the Ago-bound viRNAs of Cnidaria and Porifera could help to distinguish between the first two hypotheses, and an identification of the nucleases and Argonautes and/or Piwis required for the 26-29nt mollusc viRNAs could establish whether this response is derived from a Dicer/Ago pathway or a Zucchini/Piwi like pathway. Nevertheless, it is clear that in each case the well-studied 'canonical' antiviral RNAi response of *Drosophila*, and possibly that of other Metazoa, is highly derived compared to the ancestral state.

The presence of piRNAs derived from transposable elements in the soma of all of the sampled metazoa also demonstrates a previously under-appreciated diversity of piRNA-like mechanisms. First, it argues strongly that the predominantly germline expression of the piRNA pathway in key model metazoa (vertebrates, *Drosophila*, and nematodes) is a derived state, and that "ping-pong" mediated TE-suppression in the soma is likely to be common in other metazoan phyla. Second, it suggests that the TE-derived endo-siRNAs seen in *Drosophila* are absent from most phyla, and therefore a relatively recent innovation. Third, the diversity of piRNA profiles seen among organisms—such as the bimodal length distributions of primary piRNAs in the sponge and in "ping-pong' piRNAs in the sea anemone—suggests substantial variation among metazoa in the details of piRNA biogenesis. Finally, the large numbers of primary piRNAs derived from putative endogenous copies of parvo/denso-like viruses in the starfish and sea anemone, and from the putatively endogenous rhabdo-like virus 2 in the dog whelk, suggests that the piRNA processing of endogenous virus copies may be widespread across the metazoa, perhaps even representing an additional ancient defence mechanism.

Materials and Methods

Sample collections and RNA extraction

We sampled six organisms: The breadcrumb sponge *Halichondria panacea* (Porifera: Demospongiae), the beadlet anenome *Actinia equina* (Cnidaria: Anthozoa), the common starfish *Asterias rubens* (Echinodermata: Asteroidea), the dog whelk *Nucella lapillus* (Mollusca: Gastropoda), mixed earthworm species (*Amynthas* spp. and *Lumbricus* spp.; Annelida: Oligochaeta), and the brown alga *Fucus serratus* (Heterokonta: Phaecophyceae: Fucales). Marine species were sampled from rocky shores at Barns Ness (July 2014; 56.00° N, 2.45° E), and from three sites near Millport on the island of Great Cumbrae (August 2014; 55.77° N, 4.92° E) in Scotland, UK (S1_Table, S1_Text). The terrestrial sample (mixed earthworms; *Lumbricus* spp., and *Amythas* spp.), were collected from The King's Buildings campus, Edinburgh, UK (November 2015; 55.92° N, 3.17° E). To maximise the probability of incorporating infected hosts, we included multiple individuals for sequencing (minimum: 37 sponge colonies; maximum: 164 starfish; see S1_Table for sampling details, numbers). Marine organisms were stored separately in sea water at 4°C for up to 72 hours before dissection. After dissection, the selected tissues were immediately frozen in liquid nitrogen, pooled in groups of 5-30 individuals, and ground to a fine powder for RNA extraction under liquid nitrogen (see S1_Text for details of tissue processing). Except for the brown alga *Fucus serratus*, RNA was extracted using Trizol (Life Technologies) and DNase treated (Turbo DNA-free: Life Technologies) following manufacturer's instructions. For *Fucus*, the extraction protocol was modified from Apt et al., (1995). Briefly, tissue was lysed in a CTAB extraction buffer, and RNA was repeatedly (re-)extracted using chloroform/isoamyl alchohol (24:1) and phenol-chloroform (pH 4.3), and (re-)precipitated using 100% ethanol, 12M LiCl, and 3M NaOAc (pH 5.2).

Library preparation and sequencing

To avoid potential nematode contamination, an aliquot of RNA from each small (5-30 individual) pool was reverse transcribed using M-MLV reverse transcriptase (Promega) with random hexamer primers. These were screened by PCR with nematode-specific primers and conditions as described in (Floyd et al., 2005) (Forward 5'-CGCGAATRGCTCATTACAACAGC; Reverse 5'-GGCGATCAGATACCGCCC). We excluded all sample pools that tested positive for nematodes from sequencing, although they were used to infer virus prevalence (below). For each host species, RNA from the nematode-free pools were combined to give final RNA-sequencing pools in which individuals were approximately equally represented. For the sponge, sea anemone, starfish, and dog whelk this pooling was subsequently replicated, using a subset of the original small pools, resulting in sequencing pools 'A' and 'B' (S1_Table, S2_Table).

Total RNA was provided to Edinburgh Genomics (Edinburgh, UK) for paired-end sequencing using the Illumina platform. Following ribosomal RNA depletion using Ribo-Zero Gold (Illumina), TruSeq stranded total RNAseq libraries (Illumina) were prepared using standard barcodes, to be sequenced in three groups, each on a single lane. Lanes were: (i) sponge, sea anemone, starfish, and dog whelk 'A' libraries (HiSeq v4; 125nt paired-end reads; a *Drosophila suzukii* RNAseq library from an unrelated project was also included in this lane); (ii) sponge, sea anemone, starfish, and dog whelk 'B' libraries (HiSeq 4000; 150nt paired-end reads); (iii) *Fucus* and Earthworms (HiSeq 4000; 150nt paired-end reads). In total, this resulted in approximately 70M high quality read pairs (i.e. after trimming and quality control) from the sponge, 60M from the sea anemone, 70M from the starfish, 70M from the dog whelk, 130M from the earthworms, and 180M from the brown alga (S3_Table).

For small RNA sequencing, total RNA was provided to Edinburgh Genomics (Edinburgh, UK) for untreated libraries (A and B), or after treatment either with a polyphosphatase ("A: Polyphosphatase") or with sodium periodate ("B: Oxidised"). In the first case, we used a RNA 5' Polyphosphatase (Epicentre) treatment to convert 5' triphosphate groups to a single phosphate. This permits the ligation of small RNAs that result from direct synthesis rather than Dicer-mediated cleavage, such as 22G-RNA sRNAs of nematodes. In the second case, we used a sodium periodate ($NaIO_4$) treatment (S2_Text). Oxidation using $NaIO_4$ reduces the relative ligation efficiency of metazoan miRNAs that lack 3′-Ribose 2′O-methylation, relative to canonical piRNAs and viRNAs. This permits identification of 3′- 2′O-methylated sRNA populations, and is expected to enrich small RNA library for canonical piRNAs and viRNAs. TruSeq stranded total RNAseq libraries (Illumina) were prepared from treated RNA by Edinburgh Genomics, and sequenced using the Illumina platform (HiSeq v4; 50nt single-end reads), with all 'A' libraries sequenced together in a single lane, and all 'B' libraries sequenced together with *Fucus* and earthworm small RNAs, across four lanes. In total, this resulted in between 46M adaptor-trimmed small RNAs for the brown alga, and 150M for the sponge (S3_Table) Raw reads from RNAseq and small RNA sequencing are available from the NCBI Sequence Read Archive under BioProject accession PRJNA394213.

Sequence assembly and taxonomic assignment

For each organism, paired end RNAseq data were assembled *de novo* using Trinity 2.2.0 (Grabherr et al., 2011; Haas et al., 2013) as a paired end strand-specific library (--SS_lib_type RF), following automated trimming (--trimmomatic) and digital read normalisation (--normalize_reads). Where two RNAseq libraries ('A' and 'B') had been sequenced, these were combined for assembly. For the mixed earthworm assembly, which had a large number of reads, high complexity, and a high proportion of ribosomal sequences (18%), ribosomal sequences were identified by mapping to a preliminary build of rRNA derived from subsampled data, and excluded from the subsequent final assembly. To identify cytochrome oxidase 1 (COI) sequences, all COI DNA sequences from GenBank nt were used to search all contigs using BLASTn (Altschul et al., 1990), and the resulting matches examined and manually curated before read mapping. An analogous approach was taken to identify rRNA sequences, but using rRNA from related taxa for a BLASTn search.

To identify probable virus and transposable element (TE)-like contigs, all long open reading frames from each contig were identified and concatenated to provide a 'bait' sequence for similarity searches using Diamond (Buchfink et al., 2014) and BLASTp (Altschul et al., 1990). Only those contigs with an open reading frame of at least 200 codons were retained. To reduce computing time, we used a two-step search. First, a preliminary search was made using translations against a Diamond database comprising all of the virus protein sequences available in NCBI database 'nr' (mode 'blastp'; e-value 0.01; maximum of one match). Second, we used the resulting (potentially virus-like) contigs to search a Diamond database that combined all virus proteins from NCBI 'nr', with all proteins from NCBI 'RefSeq_protein' (mode 'blastp'; e-value 0.01; no maximum matches). Putatively virus-like matches from this search were retained for manual examination and curation (including assessment of coverage – see below), resulting in 85 high-confidence putative virus contigs. A similar (but single-step) approach was used to search translated sequences from Repbase (Bao et al., 2015), using an e-value of $1x10^{-10}$ to identify TE-like contigs.

Virus annotation and phylogenetic analysis

Translated open reading frames from the 85 virus-like contigs were used to search the NCBI 'RefSeq_protein' blast database using BLASTp (Altschul et al., 1990). High confidence open reading frames were manually annotated based on similarity to predicted (or known) proteins from related viruses. Where unlinked fragments could be unambiguously associated based on similarity to a related sequence or via PCR (below), they were assigned to the same virus. These contigs were provisionally named based on the collection location, host species, and virus lineage. Where available, the polymerase (or a polymerase component) from each putative virus species was selected for phylogenetic analysis. Where the polymerase was not present, sequences for phylogenetic analysis were selected to maximise the number of published virus sequences available. For the Weiviruses, bunya-like viruses, and noda-like viruses, two different proteins were used for phylogenetic inference. Published viral taxa were selected for inclusion based on high sequence similarity (identifiable by BLASTp). Translated protein sequences were aligned using T-Coffee (Notredame et al., 2000) mode 'm_coffee' (Wallace et al., 2006) combining a consensus of alignments from ClustalW (Thompson et al., 1994; Chenna et al., 2003), T-coffee (Notredame et al., 2000), POA (Lee et al., 2002), Muscle (Edgar, 2004), Mafft (Katoh & Standley, 2013), DIALIGN (Morgenstern, 2004), PCMA (Pei et al., 2003) and Probcons (Do et al., 2005). Alignments were examined by eye, and regions of ambiguous alignment at either end were removed. Phylogenetic relationships were inferred by maximum-likelihood using PhyML (version 20120412; Guindon & Gascuel, 2003) with the LG substitution model, empirical amino-acid frequencies, and a four-category gamma distribution of rates with an inferred shape parameter. Searches started from a maximum parsimony tree, and used both nearest-neighbour interchange (NNI) and sub-tree prune and re-graft (SPR) algorithms, retaining the best result. Support was assessed using the Shimodaira-Hasegawa-like nonparametric version of an approximate likelihood ratio test. All trees are presented mid-point rooted.

PCR survey for virus prevalence

To estimate virus prevalence in the five metazoan taxa, we used a PCR survey of the small sample pools (5-30 individuals) for 53 virus-like contigs. There was insufficient RNA to survey prevalence in the brown alga. Aliquots from each sample pool were reverse transcribed using M-MLV reverse transcriptase (Promega) with random hexamer primers, and 10-fold diluted cDNA screened by PCR with primers for virus-like contigs designed using Primer3 (Koressaar & Remm, 2007; Untergasser et al., 2012). To confirm that primer combinations could successfully amplify the target virus sequences, and to provide robust assays, each of four PCR assays (employing pairwise combinations of two forward and two reverse primers) were tested using combined pools of cDNA for each host, with the combination that produced the clearest amplicon band chosen as the optimal assay. We took a single successful PCR amplification to indicate the presence of virus in a pool, whereas absence was confirmed through at least 2 PCRs that produced no product. PCR primers and conditions are provided in S7_Table. Prevalence was inferred by maximum likelihood, and 2 log-likelihood intervals are reported.

RT-negative PCR survey for EVE detection

For 47 of the putative RNA virus contigs, we used PCR to verify that the sequences were not present as DNA in our sample, i.e. were not EVEs. We performed an RT-negative PCR survey of Trizol RNA extractions (which also contained DNA) using the primers and conditions provided in S7_Table. Where amplification was successful from cDNA synthesised from a DNAse-treated extraction, but not from 1:10, 1:100, or 1:0000-fold diluted RNA samples (serial dilution was necessary as excessive RNA interfered with PCR), we inferred that template DNA was absent. The remaining six (out of a total of 53 contigs for which designed PCR assays) were putative parvo/denso-like virus contigs, and were also tested as above. All six DNA virus contigs were detectable as DNA copies.

Origin of sequencing reads and small RNA properties

To identify the origin of RNA sequencing reads, quality trimmed forward-orientation RNAseq reads and adaptor-trimmed small-RNA reads between 17nt and 40nt in length (trimmed using cutadapt and retaining adaptor-trimmed reads only; Martin, 2011) were mapped to potential source sequences. To provide approximate counts of rRNA and miRNA reads, reads were mapped to ribosomal contigs from the target host taxa and to all mature miRNA stem-loops represented in miRbase (Kozomara & Griffiths-Jones, 2014), using Bowtie2 (Langmead & Salzberg, 2012) with the '--fast' sensitivity option and retaining only one mapping (option '-k 1'). To identify the number and properties of virus and TE-derived reads, the remaining unmapped reads were then mapped to the 85 curated virus-like contigs, to COI-like contigs, and to 146 selected long TE-like contigs between 2kbp and 7.5kbp from out assemblies, using the '--sensitive' option and default reporting (multiple alignments, report mapping quality). For small RNA mapping, the gap-opening and extension costs were set extremely high ('--rdg 20,20 --rfg 20,20') to exclude maps that required an indel. The resulting read mappings were counted and analysed for the distribution of read lengths, base composition, and orientation. In an attempt to identify modified or edited small RNAs, we additionally mapped the small RNA reads to the virus-like and TE-like contigs using high sensitivity local mapping options equivalent to '--very-sensitive-local' but additionally permitting a mismatch in the mapping seed region ('-N 1') and again preventing indels ('--rdg 20,20 --rfg 20,20'). This did not lead to substantially different results.

Acknowledgements

We thank Andrew Rambaut, Michael Dye, Pete Doe, and Nathan Medd for assistance with field collections; Daniel Moncrieff and the Field Studies Council Scotland for logistical support in Millport; Franklin Chow, and Susan Coelho for advice on sample preparation; Helen Gunter and Karim Gharbi from Edinburgh Genomics for advice on sequencing strategy. We thank Rob Gifford, Sam Lewis and members of the Obbard lab for discussion, and Amy Buck for discussion and for contributions to the early stages of study design.

Funding information

This work was funded by a Leverhulme Trust grant to DJO, GNS and FMW (RPG-2013-168; https://www.leverhulme.ac.uk/), and work in DJO's lab was partly supported by a Wellcome Trust strategic award to the Centre for Immunity, Infection and Evolution (WT095831; http://www.wellcome.ac.uk/).

Author contributions

Conceived and designed the experiments: DJO FMW GNS. Collected and processed field samples: FMW DJO GNS. Performed the experiments: FMW. Analysed the data: DJO FMW. Contributed reagents/materials/analysis tools: FMW DJO. Wrote the paper: FMW DJO

References

Agius, C., Eamens, A.L., Millar, A.A., Watson, J.M. & Wang, M. 2012. Antiviral Resistance in Plants. In: *Antiviral Resistance in Plants: Methods and Protocols. Methods In Molecular Biology*. (J. Watson & M.-B. Wang, eds), pp. 17–38. Springer.

Alié, A., Leclère, L., Jager, M., Dayraud, C., Chang, P., Le Guyader, H., *et al.* 2011. Somatic stem cells express Piwi and Vasa genes in an adult ctenophore: Ancient association of "germline genes" with stemness. *Dev. Biol.* **350**: 183–197.

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.

Ameres, S., Horwich, M., Hung, J.-H., Xu, J., Ghildiyal, M., Wenz, Z., *et al.* 2010. Target RNA–directed trimming and tailing of small silencing RNAs. *Science (80-.).* **328**: 1534–1539.

Apt, K.E., Clendennen, S.K., Powers, D.A. & Grossman, A.R. 1995. The gene family encoding the fucoxanthin chlorophyll proteins from the brown alga Macrocystis pyrifera. *Mol. Gen. Genet.* **246**: 455–464.

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

Arensburger, P., Hice, R.H., Wright, J.A., Craig, N.L. & Atkinson, P.W. 2011. The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics* **12**: 606.

Ashe, A., Bélicard, T., Le Pen, J., Sarkies, P., Frézal, L., Lehrbach, N.J., *et al.* 2013. A deletion polymorphism in the Caenorhabditis elegans RIG-I homolog disables viral RNA dicing and antiviral immunity. *Elife* **2**: e00994.

Axtell, M.J. 2013. Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol* **64**: 137–159.

Backes, S., Langlois, R.A., Schmid, S., Varble, A., Shim, J. V., Sachs, D., *et al.* 2014. The mammalian response to virus infection is Independent of small RNA silencing. *Cell Rep.* **8**: 114–125. The Authors.

Ballenghien, M., Faivre, N. & Galtier, N. 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biol.* **15**: 25. BMC Biology.

Bao, W., Kojima, K.K. & Kohany, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**: 11. Mobile DNA.

Barnard, A.C., Nijhof, A.M., Fick, W., Stutzer, C. & Maritz-Olivier, C. 2012. RNAi in arthropods: Insight into the machinery and applications for understanding the pathogen-vector interface. *Genes (Basel).* **3**: 702–741.

Baum, J., Papenfuss, A.T., Mair, G.R., Janse, C.J., Vlachou, D., Waters, A.P., *et al.* 2009. Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. *Nucleic Acids Res.* **37**: 3788–3798.

Benitez, A.A., Spanko, L.A., Bouhaddou, M., Sachs, D. & TenOever, B.R. 2015. Engineered mammalian RNAi can elicit antiviral protection that negates the requirement for the interferon response. *Cell Rep.* **13**: 1456–1466.

Bogerd, H.P., Skalsky, R.L., Kennedy, E.M., Furuse, Y., Whisnant, A.W., Flores, O., *et al.* 2014.

Replication of many human viruses is refractory to inhibition by endogenous cellular microRNAs. *J. Virol.* **88**: 8065–76.

Bollmann, S.R., Fang, Y., Press, C.M., Tyler, B.M. & Grünwald, N.J. 2016. Diverse Evolutionary Trajectories for Small RNA Biogenesis Genes in the Oomycete Genus Phytophthora. *Front. Plant Sci.* **7**: 1–15.

Borges, F. & Martienssen, R.A. 2015. The expanding world of small RNAs in plants. *Nat. Rev. Mol. Cell Biol.* **16**: 727–741. Nature Publishing Group.

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

Bronkhorst, A.W., Miesen, P. & van Rij, R.P. 2013. Small RNAs tackle large viruses: RNA interference-based antiviral defense against DNA viruses in insects. *Fly (Austin).* **7**: 216–223.

Bronkhorst, A.W. & van Rij, R.P. 2014. The long and short of antiviral defense: small RNA-based immunity in insects. *Curr. Opin. Virol.* **7C**: 19–28. Elsevier B.V.

Brun, P. & Plus, N. 1980. The viruses of Drosophila. In: *The genetics and biology of Drosophila* (M. Ashburner & T. R. F. Wright, eds), pp. 625–702. Academic Press, London.

Buchfink, B., Xie, C. & Huson, D.H. 2014. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**: 59–60.

Buchon, N. & Vaury, C. 2006. RNAi: a defensive RNA-silencing against viruses and transposable elements. *Heredity (Edinb).* **96**: 195–202.

Buck, A.H. & Blaxter, M. 2013. Functional diversification of Argonautes in nematodes: an expanding universe. *Biochem. Soc. Trans.* **41**: 881–6. Portland Press Ltd.

Burroughs, A.M., Ando, Y. & Aravind, L. 2014. New perspectives on the diversification of the RNA interference system: Insights from comparative genomics and small RNA sequencing. *Wiley Interdiscip. Rev. RNA* **5**: 141–181.

Cai, Y., Zhou, Q., Yu, C., Wang, X., Hu, S., Yu, J., *et al.* 2012. Transposable-element associated small RNAs in Bombyx mori genome. *PLoS One* **7**.

Casas-Mollano, J.A., Zacarias, E., Ma, X., Kim, E.-J. & Cerutti, H. 2016. RNA-Mediated Silencing in Eukaryotes: Evolution of the protein synthesis machinery and its regulation. In: *Evolution of the Protein Synthesis Machinery and Its Regulation* (G. Hernanez & R. Jagus, eds), pp. 513–529. Springer.

Castel, S.E. & Martienssen, R. a. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat. Rev. Genet.* **14**: 100–12. Nature Publishing Group.

Cerutti, H. & Casas-Mollano, J.A. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.* **50**: 81–99.

Chang, S.-S., Zhang, Z. & Liu, Y. 2012. RNA Interference Pathways in Fungi: Mechanisms and Functions. *Annu. Rev. Microbiol.* **66**: 305–323.

Chejanovsky, N., Ophir, R., Schwager, M.S., Slabezki, Y., Grossman, S. & Cox-Foster, D. 2014. Characterization of viral siRNA populations in honey bee colony collapse disorder. *Virology* **454–455**: 176–183. Elsevier.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., *et al.* 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**: 3497–3500.

Cock, J.M., Liu, F., Duan, D., Bourdareau, S., Lipinska, A.P., Coelho, S.M., *et al.* 2017. Rapid Evolution of microRNA Loci in the Brown Algae. *Genome Biol. Evol.* **9**: 740–749.

Cock, J.M., Sterck, L., Rouzé, P., Scornet, D., Allen, A.E., Amoutzias, G., *et al.* 2010. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* **465**: 617–621.

Coffman, S., Lu, J., Guo, X., Zhong, J., Jiang, H., Broitman-Maduro, G., *et al.* 2017. Caenorhabditis elegans RIG-1 homolog mediates antiviral RNA interference downstream of dicer-dependent biogenesis of viral small interfering RNAs. **8**: 1–15.

Coruh, C., Cho, S.H., Shahid, S., Liu, Q., Wierzbicki, A. & Axtell, M.J. 2015. Comprehensive Annotation of Physcomitrella patens Small RNA Loci Reveals That the Heterochromatic Short Interfering RNA Pathway Is Largely Conserved in Land Plants. *Plant Cell* **27**: tpc.15.00228-.

Czech, B. & Hannon, G.J. 2016. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends Biochem. Sci.* **41**. Elsevier Ltd.

Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., *et al.* 2008. An endogenous small interfering RNA pathway in Drosophila. *Nature* **453**: 798–802.

Dang, Y., Yang, Q., Xue, Z. & Liu, Y. 2011. RNA interference in fungi: pathways, functions, and applications. *Eukaryot. Cell* **10**: 1148–55.

Das, P.P., Bagijn, M.P., Goldstein, L.D., Woolford, J.R., Lehrbach, N.J., Sapetschnig, A., *et al.* 2008. Piwi and piRNAs Act Upstream of an Endogenous siRNA Pathway to Suppress Tc3 Transposon Mobility in the Caenorhabditis elegans Germline. *Mol. Cell* **31**: 79–90.

de Jong, D., Eitel, M., Jakob, W., Osigus, H.-J., Hadrys, H., Desalle, R., *et al.* 2009. Multiple dicer genes in the early-diverging metazoa. *Mol. Biol. Evol.* **26**: 1333–40.

Deng, W. & Lin, H. 2002. < i> miwi</i>, a Murine Homolog of< i> piwi</i>, Encodes a Cytoplasmic Protein Essential for Spermatogenesis. *Dev. Cell* **2**: 819–830.

Denker, E., Manuel, M., Leclère, L., Le Guyader, H. & Rabet, N. 2008. Ordered progression of nematogenesis from stem cells through differentiation stages in the tentacle bulb of Clytia hemisphaerica (Hydrozoa, Cnidaria). *Dev. Biol.* **315**: 99–113. Elsevier Inc.

Ding, S.W., Li, H., Lu, R., Li, F. & Li, W.X. 2004. RNA silencing: A conserved antiviral immunity of plants and animals. *Virus Res.* **102**: 109–115.

Do, C.B., Mahabhashyam, M.S.P., Brudno, M. & Batzoglou, S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**: 330–340.

Drinnenberg, I.A., Fink, G.R. & Bartel, D.P. 2011. Compatibility with Killer Explains the Rise of RNAi-Deficient Fungi. *Science (80-.).* **333**: 1592–1592.

Drinnenberg, I. a, Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R., *et al.* 2009. RNAi in budding yeast. *Science* **326**: 544–550.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**: 1792–1797.

Fabioux, C., Corporeau, C., Quillien, V., Favrel, P. & Huvet, A. 2009. In vivo RNA interference in oyster -vasa silencing inhibits germ cell development. *FEBS J.* **276**: 2566–2573.

Félix, M.-A., Ashe, A., Piffaretti, J., Wu, G., Nuez, I., Bélicard, T., *et al.* 2011. Natural and Experimental Infection of Caenorhabditis Nematodes by Novel Viruses Related to Nodaviruses. Public Library of Science.

Fire, A., Xu, S., Montgomery, M., Kostas, S., Driver, S. & Mello, C. 1998. Potent and specifc genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **394**: 806–811.

Floyd, R., Rogers, A., Lambshead, P. & Smith, C. 2005. Nematode-specific PCR primers for the 18S small subunit rRNA gene. *Mol. Ecol. Notes* **5**: 611–612.

Gammon, D. & Mello, C. 2015. RNA interference-mediated antiviral defense in insects. *Curr. Opin. Insect Sci.* 111–120.

Gammon, D.B., Ishidate, T., Li, L., Gu, W., Silverman, N. & Mello, C.C. 2017. The Antiviral RNA Interference Response Provides Resistance to Lethal Arbovirus Infection and Vertical Transmission in Caenorhabditis elegans. *Curr. Biol.* **27**: 795–806. Elsevier Ltd.

Gao, Z., Wang, M., Blair, D., Zheng, Y. & Dou, Y. 2014. Phylogenetic analysis of the endoribonuclease Dicer family. *PLoS One* **9**: 1–7.

Girardi, E., Chane-Woon-Ming, B., Messmer, M., Kaukinen, P. & Pfeffer, S. 2013. Identification of RNase L-dependent, 3′ -end-modified, viral small RNAs in Sindbis virus-infected mammalian cells. *MBio* **4**: 1–10.

Giribet, G. 2016. New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. *Org. Divers. Evol.* **16**: 419–426. Organisms Diversity & Evolution.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D. a, Amit, I., *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, B.J., Chiang, H.R., King, N., *et al.* 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**: 1193–1197.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., *et al.* 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell* **106**: 23–34.

Guindon, S. & Gascuel, O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **52**: 696–704.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., *et al.* 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**: 1494–1512.

Handler, D., Meixner, K., Pizka, M., Lauss, K., Schmied, C., Gruber, F.S., *et al.* 2013. The genetic makeup of the drosophila piRNA pathway. *Mol. Cell* **50**: 762–777. Elsevier Inc.

Hennebert, E., Leroy, B., Wattiez, R. & Ladurner, P. 2015. An integrated transcriptomic and proteomic analysis of sea star epidermal secretions identifies proteins involved in defense and adhesion. *J. Proteomics* **128**: 83–91. Elsevier B.V.

Holmes, E. 2009. *The Evolution and Emergence of RNA Viruses*. Oxford University Press.

Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., *et al.* 2007. A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell* **129**: 69–82.

Hu, Y., Stenlid, J., Elfstrand, M. & Olson, A. 2013. Evolution of RNA interference proteins dicer and argonaute in Basidiomycota. *Mycologia* **105**: 1489–98.

Huang, Y., Kendall, T., Forsythe, E.S., Dorantes-Acosta, A., Li, S., Caballero-Pérez, J., *et al.* 2015. Ancient origin and recent innovations of RNA polymerase IV and V. *Mol. Biol. Evol.* **32**: 1788–1799.

Ishikawa, T., Nishikawa, H., Gao, Y., Sawa, Y., Shibata, H., Yabuta, Y., *et al.* 2008. The pathway via D-galacturonate/L-galactonate is significant for ascorbate biosynthesis in Euglena gracilis: Identification and functional characterization of aldonolactonase. *J. Biol. Chem.* **283**: 31133–31141.

Jakob, W., Sagasser, S., Dellaporta, S., Holland, P., Kuhn, K. & Schierwater, B. 2004. The Trox-2 Hox/ParaHox gene of Trichoplax (Placozoa) marks an epithelial boundary. *Dev. Genes Evol.* **214**: 170–175.

Jousset, F.-X., Plus, N., Croizier, G. & Thomas, M. 1972. Existence chez Drosophila de deux groupes de picornaviruaea de propietes serologiques et biologiques differentes. *Comptes Rendus l'Académie des Sci.* **275**: 3043–3046.

Juliano, C.E., Reich, A., Liu, N., Gotzfried, J., Zhong, M., Uman, S., *et al.* 2013. PIWI proteins and PIWI-interacting RNAs function in Hydra somatic stem cells. *Proc. Natl. Acad. Sci.* **111**: 337–342.

Katoh, K. & Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**: 772–780.

Katzourakis, A. & Gifford, R.J. 2010. Endogenous viral elements in animal genomes. *PLoS Genet.* **6**.

Kaur, G. & Lohia, A. 2004. Inhibition of gene expression with double strand RNA interference in Entamoeba histolytica. *Biochem. Biophys. Res. Commun.* **320**: 1118–1122.

Kircher, M., Sawyer, S. & Meyer, M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**: 1–8.

Koonin, E. V. 2017. Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol. Direct* **12**: 5. Biology Direct.

Koressaar, T. & Remm, M. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**: 1289–1291.

Kozomara, A. & Griffiths-Jones, S. 2014. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**: 68–73.

Kuramochi-Miyagawa, S. 2004. Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development* **131**: 839–849.

Labreuche, Y. & Warr, G.W. 2013. Insights into the antiviral functions of the RNAi machinery in penaeid shrimp. *Fish Shellfish Immunol.* **34**: 1002–1010.

Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**: 357–9.

Lee, C., Grasso, C. & Sharlow, M.F. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**: 452–464.

Lee, H.C., Gu, W., Shirayama, M., Youngman, E., Conte, D. & Mello, C.C. 2012. C. elegans piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell* **150**: 78–87. Elsevier Inc.

Lee, Y.S., Nakahara, K., Pham, J.W., Kim, K., He, Z., Sontheimer, E.J., *et al.* 2004. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117**: 69–81.

Lewis, S.H., Salmela, H., Obbard, D.J., Street, D., Buildings, K., Buildings, K., *et al.* 2015. Duplication and diversification of Dipteran Argonaute genes, and the evolutionary divergence of Piwi and Aubergine. 1–30.

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., *et al.* 2015. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife* **4**: 4–6.

Li, C.X., Shi, M., Tian, J.H., Lin, X.D., Kang, Y.J., Chen, L.J., *et al.* 2015. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife* **4**: 1–26.

Li, Y., Basavappa, M., Lu, J., Dong, S., Cronkite, D.A., Prior, J.T., *et al.* 2016. Induction and suppression of antiviral RNA interference by influenza A virus in mammalian cells. *Nat. Microbiol.* **2**: 16250.

Li, Y., Lu, J., Han, Y., Fan, X. & Ding, S.-W. 2013. RNA interference functions as an antiviral immunity mechanism in mammals. *Science* **342**: 231–4.

Liew, Y.J., Ryu, T., Aranda, M. & Ravasi, T. 2016. miRNA Repertoires of Demosponges Stylissa carteri and Xestospongia testudinaria. *PLoS One* **11**: e0149080.

Liu, H., Söderhäll, K. & Jiravanichpaisal, P. 2009. Antiviral immunity in crustaceans. *Fish Shellfish Immunol.* **27**: 79–88.

Lye, L.F., Owens, K., Shi, H., Murta, S.M.F., Vieira, A.C., Turco, S.J., *et al.* 2010. Retention and Loss of RNA interference pathways in trypanosomatid protozoans. *PLoS Pathog.* **6**.

MacRae, I.J. 2006. Structural Basis for Double-Stranded RNA Processing by Dicer. *Science (80-.).* **311**: 195–198.

Maillard, P. V, Ciaudo, C., Marchais, A., Li, Y., Jay, F., Ding, S.W., *et al.* 2013. Antiviral RNA interference in mammalian cells. *Science* **342**: 235–8.

Maillard, P. V, Veen, A.G. Van Der, Deddouche-grass, S. & Rogers, N.C. 2016. Inactivation of the type I interferon pathway reveals long double-stranded RNA-mediated RNA interference in mammalian cells. *EMBO J.* 1–14.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Miesen, P., Girardi, E. & van Rij, R.P. 2015. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. *Nucleic Acids Res.* **43**: 6545–56.

Moran, Y., Agron, M., Praher, D. & Technau, U. 2017. The evolutionary origin of plant and animal microRNAs. *Nat. Ecol. Evol.* **1**: 27. Macmillan Publishers Limited.

Moran, Y., Praher, D., Fredman, D. & Technau, U. 2013. The evolution of MicroRNA pathway protein components in Cnidaria. *Mol. Biol. Evol.* **30**: 2541–2552.

Morazzani, E.M., Wiley, M.R., Murreddu, M.G., Adelman, Z.N. & Myles, K.M. 2012. Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. *PLoS Pathog.* **8**.

Morgenstern, B. 2004. DIALIGN: Multiple DNA and protein sequence alignment at BiBiServ.

Nucleic Acids Res. **32**: 33–36.

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**: 627–41.

Mukherjee, K., Korithoski, B. & Kolaczkowski, B. 2014. Ancient origins of vertebrate-specific innate antiviral immunity. *Mol. Biol. Evol.* **31**: 140–153.

Ngo, H., Tschudi, C., Gull, K. & Ullu, E. 1998. Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. *Proc. Natl. Acad. Sci.* **95**: 14687–14692.

Nicolás, F.E. & Ruiz-Vázquez, R.M. 2013. Functional diversity of RNAi-associated sRNAs in fungi. *Int. J. Mol. Sci.* **14**: 15348–15360.

Notredame, C., Higgins, D.G. & Heringa, J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**: 205–217.

Obbard, D.J., Gordon, K.H.J., Buck, A.H. & Jiggins, F.M. 2009. The evolution of RNAi as a defence against viruses and transposable elements. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364**: 99–115.

Pak, J. & Fire, A. 2007. Distinct Populations of Primary and Secondary Effectors During RNAi in C. elegans. *Science (80-.).* **315**: 241–244.

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Tu, Z., van Rij, R., *et al.* 2017. Comparative genomics shows that viral integrations are abundant and express. *bioRxiv* 1–15. BMC Genomics.

Palmer, W.J. & Jiggins, F.M. 2015. Comparative Genomics Reveals the Origins and Diversity of Arthropod Immune Systems. *Mol. Biol. Evol.* **32**: 2111–2129.

Parameswaran, P., Sklan, E., Wilkins, C., Burgon, T., Samuel, M. a, Lu, R., *et al.* 2010. Six RNA viruses and forty-one hosts: viral small RNAs and modulation of small RNA repertoires in vertebrate and invertebrate systems. *PLoS Pathog.* **6**: e1000764.

Pei, J., Sadreyev, R. & Grishin, N. V. 2003. PCMA: Fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics* **19**: 427–428.

Perez, J.T., Varble, A., Sachidanandam, R., Zlatev, I., Manoharan, M., Garcia-Sastre, A., *et al.* 2010. Influenza A virus-generated small RNAs regulate the switch from transcription to replication. *Proc. Natl. Acad. Sci.* **107**: 11525–11530.

Rajasethupathy, P., Antonov, I., Sheridan, R., Frey, S., Sander, C., Tuschl, T., *et al.* 2012. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell* **149**: 693–707. Elsevier Inc.

Rajeswaren, R. & Pooggin, M. 2012. Role of Virus-Derived Small RNAs in Plant Antiviral Defense: Insights from DNA Viruses. In: *MicroRNAs in Plant Development and Stress Response.* (R. Sunkar, ed), pp. 261–289. Springer-Verlag; Berlin Heidelberg, Germany.

Reinganum, C., O'Loughlin, G. & Hogan, T. 1970. A nonoccluded virus of the field crickets Teleogryllus aceanicus and T. commodus (Orthoptera: Gryllidae). *J. Invertebr. Pathol.* **16**: 220–314.

Rivera, A.S., Hammel, J.U., Haen, K.M., Danka, E.S., Cieniewicz, B., Winters, I.P., *et al.* 2011. RNA interference in marine and freshwater sponges: actin knockdown in Tethya wilhelma and Ephydatia muelleri by ingested dsRNA expressing bacteria. *BMC Biotechnol.* **11**: 67.

Rogato, A., Richard, H., Sarazin, A., Voss, B., Cheminant Navarro, S., Champeimont, R., *et al.* 2014. The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. *BMC Genomics* **15**: 698.

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

Samuel, C. 2012. ADARs, Viruses and Innate Immunity. *Curr. Opin. Microbiol. Immunol.* **353**: 163–195.

Sánchez Alvarado, A. & Newmark, P. 1999. Double-Stranded RNA Specifically Disrupts Gene Expression during Planarian Regeneration. *Proc. Natl. Acad. Sci. U. S. A.* **96**: 5049–5054.

Sarkies, P. & Miska, E. a. 2013. RNAi pathways in the recognition of foreign RNA: antiviral responses and host-parasite interactions in nematodes. *Biochem. Soc. Trans.* **41**: 876–880.

Sarkies, P., Selkirk, M.E., Jones, J.T., Blok, V., Boothby, T., Goldstein, B., *et al.* 2015. Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages. *PLOS Biol.* **13**: e1002061.

Schnettler, E., Tykalová, H., Watson, M., Sharma, M., Sterken, M.G., Obbard, D.J., *et al.* 2014. Induction and suppression of tick cell antiviral RNAi responses by tick-borne flaviviruses. *Nucleic Acids Res.* **42**: 1–11.

Segers, G.C., Zhang, X., Deng, F., Sun, Q. & Nuss, D.L. 2007. Evidence that RNA silencing functions as an antiviral defense mechanism in fungi. *Proc. Natl. Acad. Sci.* **104**: 12902–12906.

Seo, G.J., Kincaid, R.P., Phanaksri, T., Burke, J.M., Pare, J.M., Cox, J.E., *et al.* 2013. Reciprocal inhibition between intracellular antiviral signaling and the RNAi machinery in mammalian cells. *Cell Host Microbe* **14**.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., *et al.* 2016. Redefining the invertebrate RNA virosphere. *Nature* 1–12. Nature Publishing Group.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., *et al.* 2013. Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Curr. Biol.* **23**: 1399–408.

Sijen, T. & Plasterk, H. 2003. Transposon silencing in the Caenorhabditis elegans germ by natural RNAi. *Nature* **426**: 310–314.

Singh, R.K., Gase, K., Baldwin, I.T. & Pandey, S.P. 2015. Molecular evolution and diversification of the Argonaute family of proteins in plants. *BMC Plant Biol.* **15**: 23.

Snell, T.W., Shearer, T.L. & Smith, H.A. 2011. Exposure to dsRNA Elicits RNA Interference in Brachionus manjavacas (Rotifera). *Mar. Biotechnol.* **13**: 264–274.

Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E. V, *et al.* 2014. The evolutionary journey of Argonaute proteins. *Nat. Struct. Mol. Biol.* **21**: 743–753. Nature Publishing Group.

Swarts, D.C., Szczepaniak, M., Sheng, G., Chandradoss, S.D., Zhu, Y., Timmers, E.M., *et al.* 2017. Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. *Mol. Cell* 985–998. Elsevier Inc.

Szittya, G. & Burgyan, J. 2013. RNA interference-mediated intrinsic antiviral immunity in plants. *Curr. Top. Microbiololgy Immunol.* **371**: 153–181.

Tabach, Y., Billi, A.C., Hayes, G.D., Newman, M. a, Zuk, O., Gabel, H., *et al.* 2013. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* **493**: 694–698. Nature Publishing Group.

Tabara, H., Yigit, E., Siomi, H. & Mello, C.C. 2002. The double-stranded RNA binding protein RDE-4 interacts in vivo with RDE-1, DCR-1 and a conserved DExH-box helicase to direct RNA interference in C. elegans. *Cell* **109**: 861–871.

Takahashi, F., Yamagata, D., Ishikawa, M., Fukamatsu, Y., Ogura, Y., Kasahara, M., *et al.* 2007. AUREOCHROME, a photoreceptor required for photomorphogenesis in stramenopiles. *Proc. Natl. Acad. Sci.* **104**: 19625–19630.

Takeuchi, T., Koyanagi, R., Gyoja, F., Kanda, M., Hisata, K., Fujie, M., *et al.* 2016. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zool. Lett.* **2**: 3.

Tarver, J.E., Cormier, a., Pinzon, N., Taylor, R.S., Carre, W., Strittmatter, M., *et al.* 2015. microRNAs and the evolution of complex multicellularity: identification of a large, diverse complement of microRNAs in the brown alga Ectocarpus. *Nucleic Acids Res.* 1–15.

tenOever, B.R. 2017. Questioning antiviral RNAi in mammals. *Nat. Microbiol.* **2**: 17052.

tenOever, B.R. 2016. The Evolution of Antiviral Defense Systems. *Cell Host Microbe* **19**: 142–149. Elsevier Inc.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. 1994. ClustalW: improving the sensitivity of progressive multiple sequence aligment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl Acids Res* **22**: 4673–4680.

Umbach, J.L. & Cullen, B.R. 2009. The role of RNAi and microRNAs in animal virus replication and antiviral immunity. *Genes Dev.* **23**: 1151–64.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., *et al.* 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **40**: 1–12.

Vodovar, N., Bronkhorst, A.W., van Cleef, K.W.R., Miesen, P., Blanc, H., van Rij, R.P., *et al.* 2012. Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. *PLoS One* **7**: e30861.

Wallace, I.M., O'Sullivan, O., Higgins, D.G. & Notredame, C. 2006. M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**: 1692–1699.

Webster, C.L., Waldron, F.M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J.F., *et al.* 2015. The discovery, distribution, and evolution of viruses associated with Drosophila melanogaster. *PLOS Biol.* **13**: e1002210.

Wittig, K., Kasper, J., Seipp, S. & Leitz, T. 2011. Evidence for an instructive role of apoptosis during the metamorphosis of Hydractinia echinata (Hydrozoa). *Zoology* **114**: 11–22. Elsevier GmbH.

Yamanaka, S., Siomi, M.C. & Siomi, H. 2014. piRNA clusters and open chromatin structure. *Mob. DNA* **5**: 22.

Yigit, E., Batista, P.J., Bei, Y., Pang, K.M., Chen, C.C.G., Tolia, N.H., *et al.* 2006. Analysis of the C. elegans Argonaute Family Reveals that Distinct Argonautes Act Sequentially during RNAi. *Cell* **127**: 747–757.

Yoshida-Noro, C. & Tochinai, S. 2010. Stem cell system in asexual and sexual reproduction of Enchytraeus japonensis (Oligochaeta, Annelida). *Dev. Growth Differ.* **52**: 43–55.

Yu, N., Christiaens, O., Liu, J., Niu, J., Cappelle, K., Caccia, S., *et al.* 2013. Delivery of dsRNA for RNAi in insects: An overview and future directions. *Insect Sci.* **20**: 4–14.

Zhang, X., Segers, G.C., Sun, Q., Deng, F. & Nuss, D.L. 2008. Characterization of Hypovirus-Derived Small RNAs Generated in the Chestnut Blight Fungus by an Inducible DCL-2-Dependent Pathway. *J. Virol.* **82**: 2613–2619.

Zhou, X., Battistoni, G., El Demerdash, O., Gurtowski, J., Wunderer, J., Falciatori, I., *et al.* 2015. Dual functions of Macpiwi1 in transposon silencing and stem cell maintenance in the flatworm Macrostomum lignano. *RNA* **21**: 1885–97.

Zografidis, A., Van Nieuwerburgh, F., Kolliopoulou, A., Apostolou-Karampelis, K., Head, S.R., Deforce, D., *et al.* 2015. Viral small RNA analysis of *Bombyx mori* larval midgut during persistent and pathogenic cytoplasmic polyhedrosis virus infection. *J. Virol.* **89**: JVI.01695-15.

Supporting information

Figures

S1_Figure: Phylogenetic trees.

Maximum likelihood phylogenetic trees. Support values (approximate likelihood ratio test) and NCBI accession identifiers are provided. Viruses newly identified here are highlighted in red, and unannotated virus-like sequences from publicly-available transcriptome datasets are denoted 'TSA'. Clade names follow (C. X. Li *et al.*, 2015; Shi *et al.*, 2016). Alignments are provided in S3_Data and Newick format trees in S4_Data.

S2_Figure: Size distributions of small RNAs

Bar-plot size distributions of all small RNAs sequences. Columns correspond to species, rows to libraries. **Panel A:** All sRNAs from each library. **B**: sRNAs mapping to ribosomal sequences. Note that in most species read abundance decreases with size, indicative of degradation products, but that distinct peaks are visible in the oxidised libraries, consistent with specific short rRNAs possessing a 3' 2-O-methyl group. **C**: sRNAs mapping to known miRNA stem-loops from miRbase (Kozomara & Griffiths-Jones, 2014). The proportion of putative miRNAs decreases dramatically in all oxidised libraries except the sea anemone, suggesting that miRNAs in this species possess 3' 2-O-methyl groups. The small number of mapped miRNA reads in the brown alga is probably a result of the under-representation of close relatives in miRbase (Kozomara & Griffiths-Jones, 2014). **D**: sRNAs mapping to putative RNA virus contigs. Only the dog whelk has a large and distinctive distribution of virus-derived sRNAs, and these increase in the oxidised library, suggesting that they possess 3' 2-O-methyl groups. The small number of very short virus-derived reads in the sponge are consistent with degradation products. **E**: sRNAs mapping to DNA parvovirus-like contigs. These increase in the oxidised library, suggesting that they possess 3' 2-O-methyl groups. **F**: sRNAs mapping to selected TE-like contigs. These vary in their size range among species (21nt in the brown alga, bimodal in the sponge, peaking at 28-29nt in the other species), and increase in the oxidised library, suggesting that they possess 3'-2-O-methyl groups. Only a small proportion of TE-like contigs were used as mapping targets, and many TE-derived small RNAs remain unmapped. **G**: Unmapped sRNAs, comprising those that derived from divergent miRNAs, unrecognised viral contigs, TEs that were excluded from panel F, and all other sources. The data required to plot these figures are provided in S5_Table.

S3_Figure: Properties and repeatability of virus-derived small RNAs

Panels **A-D** are dog-whelk RNA viruses, panels **E-H** starfish DNA virus-like contigs, panel **I** is the anemone DNA virus, and panel **J** is the brown alga virus (note that only one library was made for this sample). In each panel, rows (top to bottom) represent each library: Library A, polyphosphatase-treated library A, Library B, and oxidised library B. Columns (left to right) are (i) Origin of reads from each genome position (red lines above the x-axis denote reads from the positive sense strand, blue lines below the x-axis denote reads from the negative sense strand; (ii) Bar plot of frequencies of unique sequences, bars above the x-axis denote reads from the positive sense strand, those below the x-axis denote reads from the negative sense strand, colours indicate 5' base (U red, G yellow, C blue and A green); (iii) Barplot of frequencies of reads. (iv) Sequence logo for the unique sequences of the most frequent length deriving from the positive strand (v) Sequence logo for the unique sequences of the most frequent length deriving from the negative strand. The data required to plot the size distributions are provided in S5_Table.

S4_Figure: Properties and repeatability of TE-derived small RNAs

Panels **A-R** show the small RNA properties of selected high-confidence TE-like contigs: starfish panels **A-C**, dog whelk **D-F**, sponge **G-I**, earthworms **J-L**, sea anemone **M-O**, brown alga **P-R**. Rows and columns are as in S3_Figure. The data required to plot the size distributions are provided in S5_Table.

S5_Figure: RNAseq and sRNA reads per metagenomic contig

For each metagenomic contig (pale grey) the ratio of sRNAs (20-31nt) to RNAseq reads is shown on the x-axis, and the ratio of 20-24nt sRNAs (expected viRNAs) to 25-31nt sRNAs (expected piRNAs) is shown on the y-axis. Contigs are only included if they are >0.75Kbp in length and produced at least 20 small RNAs; Contigs in dark grey have sequence similarity to known TEs, and contigs in colour correspond to the curated viruses. Based on *Drosophila*, TEs (dark grey) are expected to appear in the lower right quadrant of each plot, and viruses (colour) in the upper right (compare Figure 4 in Webster *et al.*, 2015). Only the dog whelk (panel **A**) and the brown alga (panel **D**) display sRNAs from RNA virus contigs, although DNA virus-like contigs display piRNA-like small RNAs in the sea anemone (panel **C**) and the starfish (panel **B**). No other viruses produced sufficient viRNAs to appear on these figures. All figures (except the brown alga) use data from RNAseq library B and the corresponding oxidised sRNAs (which is enriched for viRNAs over miRNAs), and sRNA counts exclude those mapping to known (miRbase) miRNAs and rRNAs.

Tables

S1_Table: Sample collection details.

Detailed descriptions of the sample collection locations, dates and numbers of individuals sampled for each target taxon, along with sample pool information, including which extraction pools were included in sequencing pools, and which were excluded due to detection of suspected nematode contamination.

S2_Table: Detailed descriptions of putative viruses and virus-like contigs.

Detailed descriptions of the candidate virus fragments identified by protein similarity search, including phylogenetic position, estimated prevalence, approximate coverage, ORF number and most similar viral proteins identified by BLASTp, detectability by RT-negative PCR, GenBank accession numbers, and additional notes.

S3_Table: Sources of RNAseq and small RNA reads.

Cytochrome oxidase coverage relative to that of the target taxon, and virus coverage for the target viruses (positive and negative strand) relative to that of COI.

S4_Table: Virus Prevalence

Estimated virus prevalence inferred by maximum likelihood (with 2 log-likelihood intervals) from an RT-PCR survey of pooled samples (methods as in Webster *et al.*, 2015).

S5_Table: Size Distribution of small RNAs

Raw counts necessary to plot Figure 3, Figure 4, Figure 5, S2_Figure, S3_Figure, S4_Figure

S6_Table: RNAi related genes identified from organisms

Counts of key RNAi related genes identified in transcriptomes of target taxa along with GenBank accession numbers for sequences.

S7_Table: PCR primers and conditions

PCR primer names and sequences, thermocycler conditions, and PCR recipes for virus prevalence, and RT-negative (EVE detection), assays.

Data

S1_Data: Raw de novo-assembled contigs

For each of the six species pools, the raw meta-transcriptomic contigs generated by Trinity are provided in compressed (gzipped) fasta format. The majority of contigs are likely to derive from the named host species and associated microbiota, although there may be a small amount of cross contamination among libraries run in the same lane. These contigs have not been curated, and are likely to include chimeric assemblies. As such, they are not suitable for submission to GenBank, and should be treated with caution.

S2_Data: Putative virus-like contigs

Raw meta-transcriptomic contigs generated by Trinity that have detectable sequence similarity (using Diamond) to virus proteins in Genbank, provided in compressed (gzipped) fasta format. Contig titles are annotated using the species name of the top match, followed by the percentage identity of that match in the sequence, and the e-value associated with that match. The contigs have not been curated, and are likely to include chimeric assemblies. As such, they are not suitable for submission to Genbank, and should be treated with caution.

S3_Data: Protein sequence alignments

Protein sequence alignments used for phylogenetic analyses are provided in compressed (gzipped) gapped fasta format, with regions of poor alignment (identified by eye) deleted. Sequence titles comprise the taxon name and NCBI accession identifier for the sequence.

S4_Data: Phylogenetic trees

Phylogenetic trees are provided in compressed (gzipped) newick format. Sequence titles comprise the taxon name and NCBI accession identifier for the original protein sequence.

S5_Data: Long high-confidence TE-like contigs

Selected meta-transcriptomic contigs generated by Trinity that have detectable sequence similarity (using Diamond) to TEs in Repbase (Bao et al., 2015), provided in compressed (gzipped) fasta format. Contig titles are annotated using the host species name and the top-match TE in Repbase.

Text

S1_Text: Sampling, tissue preparations and RNA extractions

Detailed description of the sampling, tissue preparations and RNA extractions techniques employed for each target taxon

S2_Text: RNA oxidation treatment

Protocol for sodium periodate (NaIO4) oxidation of RNA prior to library preparation, to enrich small RNA libraries for canonical piRNAs and viRNAs by reducing the relative ligation efficiency of metazoan miRNAs that lack 3′-Ribose 2′O-methylation.