Draft genome assemblies and predicted microRNA complements of the intertidal lophotrochozoans *Patella vulgata* (Mollusca, Patellogastropoda) and *Spirobranchus (Pomatoceros) lamarcki* (Annelida, Serpulida)

Nathan J Kenny [1,2*], Erica K O Namigai [2*], Ferdinand Marlétaz [2], Jerome H.L. Hui [1**], Sebastian M Shimeld [2**]

[1] Simon F.S. Li Marine Science Laboratory of School of Life Sciences and Center for Soybean Research of the State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong

[2] Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

* Contributed equally

** Communicating authors:
Jerome H.L. Hui. jeromehui@cuhk.edu.hk
Sebastian M Shimeld, sebastian.shimeld@zoo.ox.ac.uk

Abstract

MicroRNAs (miRNA) are small non-coding RNAs that act post-transcriptionally to regulate gene expression levels. Some studies have indicated microRNAs may have low homoplasy, and as a consequence the phylogenetic distribution of microRNA families has been used to study animal evolutionary relationships. Limited levels of lineage sampling, however, may distort such analyses. Lophotrochozoa is an under-sampled taxon that includes molluscs, annelids and nemerteans, among other phyla. Here, we present two novel draft genomes, those of the limpet *Patella vulgata* and polychaete *Spirobranchus (Pomatoceros) lamarcki.* Surveying these genomes for known microRNAs identifies numerous potential orthologues, including a number that have been considered to be confined to other lineages. RT-PCR demonstrates that some of these (*miR-1285, miR-1287, miR-1957, miR-1983* and *miR-3533*), previously thought to be found only in vertebrates, are expressed. This study provides genomic resources for two lophotrochozoans and reveals patterns of microRNA evolution that could be hidden by more restricted sampling.

Key Words: spiralian, lophotrochozoan, annelid, mollusc, genome, microRNA, phylogeny

1. Introduction

Originally discovered in the nematode *Caenorhabditis elegans* in 1993 (Lee et al 1993), microRNAs (miRNAs) did not attract much attention until the discovery of the first conserved miRNAs in animals in 2000 (Pasquinelli et al 2000). miRNAs have important and widespread roles in many aspects of the biology of animals, plants and even some viruses, with some playing evolutionarily-ancient roles (Bartel 2004, Axtell and Bartel 2005, Plaisance-Bonstaff and Renne 2011). For example, it is likely that the majority of mammalian mRNAs are regulated by miRNAs (Friedman et al 2009), while miR-1 appears to have an ancient role in muscle development (Kloosterman and Plasterk 2006). As a consequence, over the last decade the study of miRNAs has become a rapidly moving field in a range of contexts, most commonly in study of the post-transcriptional regulation of gene expression (see Bartel, 2009, Hui et al., 2013a,b) but also in the field of phylogenetic reconstruction (Tarver et al 2013; Kenny et al 2015).

One reason that miRNAs have been utilized in phylogenetic reconstruction was an initially reported low rate of homoplasy (Tarver et al 2013). Initial investigation suggested that once miRNAs were incorporated into genomes, they would seldom be lost (Sempere et al 2006). Further studies suggested that they might be used as slow-evolving genomic characters, such that mapping their gain across a cladogram would allow the derivation of evolutionary relationships (for examples, see Wheeler et al 2009, Tarver et al 2013). This approach has been used to shed light on several recalcitrant cases in animal phylogeny (e.g., Rota-Stabelli et al 2010, Campbell et al 2011, Philippe et al 2011). More recently, however, it has been noted that heterogeneous rates of gain and loss of miRNA loci, as well as their secondary loss, may be more common than previously suspected, especially in some lineages, and also that sampling error has affected some historic analyses (Fromm et al 2013, Thomson et al 2014, Quah et al 2015). Attempts have been made to correct some of these problems via the re-analysis of previously published datasets (Field et al 2014). While work remains to be done in this regard, if these problems can be addressed miRNA remain potentially useful for the reconstruction of phylogeny, and their flanking sequences have also been shown to contain useful phylogenetic signal at the intra-ordinal and -familial levels (Kenny et al 2015).

Uneven genome sampling across animal phylogeny, however, remains a limitation. As shown in Figure 1A, three major clades make up the bilaterians (Halanych et al 1995, Aguinaldo et al 1997). Deuterostomia, and particularly Chordata, are relatively well sampled genomically, and Ecdysozoa are also well represented (Kenny et al 2013). However the third assemblage, variously named Lophotrochozoa or Spiralia, is relatively poorly represented in both genomic and miRNA databases. This assemblage includes a number of phyla including Mollusca, Annelida, Brachiopoda, Nemertea and Platyhelminthes, with some authors using Lophotrochozoa and Spiralia synonymously, while

others reserve Lophotrochozoa for a subgroup of these phyla (for example see Struck et al 2014). The uneven distribution of characters, such as stereotypical spiral cleavage, the occurrence of a trochophore larval stage and the presence of a lophophore feeding organ, make it difficult to decide which synapomorphy best represents this clade, which was originally based on molecular phylogeny (Halanych et al 1995). We use Lophotrochozoa as inclusive of all these taxa, though whichever nomenclature is adopted their poor sampling has a range of ramifications for phylogeny reconstruction using miRNA data, and in particular the inference of gains of 'novel' miRNA sequences in single clades.

Here, we present the draft genomic sequences of two marine lophotrochozoans- the gastropod mollusc *Patella vulgata* and the serpulid annelid *Spirobranchus (Pomatoceros) lamarcki.* These species are members of diverse and ecologically vital phyla, and to our knowledge are only the fourth mollusc and third annelid genome resources to be published (after Takeuchi et al 2012, Zhang et al 2012, Simakov et al 2013). These genomes will therefore be useful for a range of investigations.

The common European limpet *P. vulgata* (Fig 1B) is a univalve gastropod and typical true limpet of the family Patellidae. It is distributed throughout Europe, as far north as the Arctic Circle and as far south as Portugal. It is found attached to firm substrates from the high shore to the edge of the sublittoral zone, although it predominates in areas of wave action. The order Patellogastropoda , to which *P. vulgata* belongs, can be found worldwide, and is well described with members widely used as models in studies of ecology, development and evolution (Lindberg, 2008; Nakano & Sasaki, 2011). A mantle-derived transcriptome also exists for this species (Werner et al 2013).

S. lamarcki (Fig 1C) is a tube-building serpulid worm which is widespread in intertidal and sublittoral zones around the United Kingdom and northern Europe. They attach to firm substrates and are noted for their detrimental effect on shipping, earning them the common name 'keelworm' (Hamer et al., 2001). *S. lamarcki* is also a useful model for embryological work, as it provides a readily-accessible source of embryonic and larval material (McDougall et al., 2006) and both a *S. lamarcki* EST dataset (Takahashi et al 2009) and embryonic transcriptome (Kenny & Shimeld, 2012) are available. *S. lamarcki* is a member of the newly redefined Sedentaria class, as is the only other available polychaete genome *Capitella teleta*, however it is relatively phylogenetically distant from both *C. teleta* and *Helobdella robusta*, the other published annelid genome (Struck et al 2011, Simakov et al 2013). We note that *S. lamarcki* has recently been the subject of taxonomic revision and the name is synonymous with *Pomatoceros lamarcki* and *Pomatoceros lamarckii* (which are widely used in the literature) (ten Hove 2015).

Our draft genome assemblies recover 578,961,269 and 964,274,156 bp of sequence for *P. vulgata* and *S. lamarcki* respectively. Using the known catalogue of metazoan miRNAs as the queries

for BLAST searches for initial assignation of identity, several unexpected miRNA candidate loci (*miR-1285, miR-1287, miR-1957, miR-1983* and *miR-3533*) were found to be present in these lophotrochozoans. This study provides new genomic resources for an undersampled clade, and suggests that broader sampling will be useful for revealing the evolutionary history of miRNAs.

2. Materials and Methods

2.1 P. vulgata *DNA extraction and genome sequencing*

Adult *P. vulgata* were collected from Tinside, Plymouth, UK. Gonads were dissected from a single male and left in a petri dish in filtered seawater to allow sperm to disassociate from somatic tissue. Large fragments of somatic tissue were removed from the petri dish, and the liquid including sperm transferred to a 15 mL tube. This was then spun at 4000 RPM at 4°C for 5 minutes. The supernatant was then removed, and the pellet washed thrice in 3 times its volume of 1x PHB (0.1 M EDTA, 50M Tris, 2.5% SDS, in distilled water), and spun at 4000 RPM at 4°C to pellet following each wash step. 1 mL PHB containing 3 µL of 5 M NaCl and 60 µL of 10 mgμL^{-1} Proteinase K was then added to the pellet, which was gently pipetted. This was then left overnight at 50°C. After digestion, the solution was phenol/chloroform extracted, a process which was repeated three times (at which point no identifiable protein layer was observed). The DNA pellet was then ethanol precipitated. The washed pellet was then left to air dry at room temperature, and resuspended in 100 µL milliQ filtered water. DNA concentration was determined using a Nanodrop 1000 spectrophotometer.

A sample of genomic DNA was prepared for sequencing by the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (Oxford, UK) with nominal fragment size libraries of 200 bp and 500 bp. Genomic DNA from the same sample has been stored for future sequencing. A single lane of Illumina HiSeq 2000 was generated, with 100 bp paired end read length. Initial assessments of the quality of the genomic data were performed using FastQC (Andrews 2010). The NCBI SRA has been used as the long term repository for raw read data, which are available under the accession number SRP055157.

2.2 S. lamarcki *DNA extraction and genome sequencing*

S. lamarcki adults were collected from the coast of Tinside, Plymouth, UK and maintained in an aquarium at 12°C. Genomic DNA (gDNA) was extracted from the sperm of a single adult worm. Sperm were homogenized using an RNase-free polypropylene pellet pestle, washed three times with 2x PHB buffer, and digested overnight in 0.015M NaCl and Proteinase K (0.6 µgml^{-1}) at 50°C. gDNA

was extracted by three phenol-chloroform extractions followed by one chloroform extraction with a 15-30 min rotation. Samples were extracted twice with chloroform and incubated in 0.1 volumes of 5M Sodium acetate and 2.5 volumes of 100% ethanol at -20°C overnight. gDNA was washed twice with 70% ethanol, air-dried at room temperature, and resuspended in 100 µl of distilled water. Sequencing was performed by the Wellcome Trust Centre for Human Genetics (Oxford, UK) using a single lane on the Illumina HiSeq 2000 with 100 bp paired-end reads, multiplexing two libraries (nominally 201bp and 500bp fragment library sizes, including read length). Some gDNA was retained for future sequencing, Quality was assessed using FastQC (Andrews et al 2010). Raw reads have been uploaded to the NCBI SRA, and are available under the accession number SRP055158.

2.3 Genome Assembly and Coverage

Several assembly programs (Velvet, ABySS, SOAPdenovo) were trialled at a variety of k-mer lengths and the 'best' assembly determined empirically from contig sizes and genome coverage as estimated from total number of base pairs in long (> 1kb) contigs. Genome sizes were estimated for both species using *k*-mer spectrum approach by counting all occurrence of 21-mers in sequenced data with Jellyfish (Figure 2), where *k* coverage is translated into actual sequencing coverage using the equation $C_k=C\times(L-k+1)/L$, where C is real coverage, L is read length, k is k-mer size and C_k is k-mer coverage. Read cleaning and error correction was attempted for *P. vulgata*, but was found to decrease quality of final assembly as assayed by the metrics used above. Error correction was performed with Quake (Kelley et al., 2010) using a 19-mer for *S. lamarcki*. For *P. vulgata* ABySS 1.3.4 (Simpson et al 2009) at a k-mer length of 57 (abyss-pe driver script and all default settings) was used for further analyses, while for *S. lamarcki* SOAPdenovo2 (Luo et al., 2012) at a k-mer length of 51 was used. Final metrics relating to genome assemblies presented here can be seen in Table 1. Genome assemblies themselves are available from the Oxford Research Archive under DOI: 10.5287/bodleian:xp68kh25x.

To assess the quality of genome assemblies, we first assayed the recovered fraction of sets of conserved eukaryotic orthologues in those assemblies using CEGMA (Parra et al 2007) and BUSCO Version 1.1b (Simão et al 2015). We then determined the mapping rates of RNA-seq data using the STAR splice-aware aligner to map previously published transcriptomic data (Method: Dobin et al. 2013. Transcriptomes: Werner et al 2013, Kenny et al 2012, available under accession SRA055301).

2.4 miRNA searches and identification

Genomes were compared to all known metazoan miRNA sequences, as downloaded from miRbase (Griffiths-Jones et al 2008) on the 6th of February 2013 using BLASTN with the following settings: -word_size 11 -reward 5 -penalty -4 -gapopen 8 -gapextend 6. Putative miRNA sequences obtained by blast were checked to confirm that both arms of the putative miRNA were present, for general homogeneity of 5' (seed) sequence, for robust hairpin structures and for a lack of similarity to known protein, tRNA or rRNA sequences. These criteria are similar to those found in Tarver et al (2012) and Quah et al (2015), although as small RNA libraries were not sequenced as a part of our investigation, criteria related to processing and overhang listed in Tarver et al (2012) were not included in our process. We also performed BLASTN comparison of identified contigs to the NCBI nr database to exclude the possibility that contamination with human DNA (or DNA from other species represented in this database) underlay the identification of candidate miR loci.

2.5 Transcriptional validation of predicted miRNAs

RNA was extracted from *P. vulgata* (head, mantle and foot samples) and *S. lamarcki* (whole adult) samples using a miRVana kit (Life Technologies) and residual gDNA removed using a Qiagen RNeasy kit with on-column DNase treatment according to the manufacturer's protocol. No correction to the protocol was made for small RNA size, so some small RNAs may have been lost at this step. cDNA was generated from RNA using a Takara Primescript Reverse Transcriptase (RT) kit, with negative RT controls made using equal quantities of RNA but no Primescript 5X solution. Primers designed using Primer3Plus were used to perform PCR at the following settings: 94°C 3 minutes, 35x (94°C 30 sec, 55°C 30 sec, 72°C 30 sec), 72°C 7 minutes. Samples were run in 2% w/v agarose gels alongside 1kb+ (Invitrogen) ladder. Primers used to perform PCR, alongside details of band sizes and miRNA identity can be seen in Table 2. Gel bands were excised and DNA extracted using a Qiagen Qiaquick gel extraction kit. Bands were cloned into pMD18T vector, transformed into DL-1 *E. coli* and plasmid DNA miniprepped with a Qiagen Qiaprep spin miniprep kit after blue/white selection. Sequencing was performed by Techdragon (Hong Kong) using M13-47 primer on an Applied Biosystems 3730xl DNA Analyzer.

3. Results

3.1 Statistics on assemblies

FastQC assessment of read quality ascertained raw read data to be excellent, with median PHRED scores above 30 through to the 100th base for both read directions in all libraries in both species. Despite this, QUAKE-treated *S. lamarcki* read genome assemblies were empirically found to be better than those constructed from raw reads alone, and were thus used for further analysis. The genomic datasets presented here comprise 578,961,269 and 964,274,156 bp of sequence for *P. vulgata* and *S. lamarcki* respectively. Genome sizes were independently estimated from raw read data using the *k*-mer spectrum approach, which assesses coverage peak based on *k*-mer count (Figure 2). We found an approximate genome size of 1.46 Gbp (25.7x base pair coverage) in *P. vulgata* using the estimate_genome_size.pl script (Ryan 2013), and an estimate of 0.95 Gbp (22x *k*-mer coverage, 28.2x base pair coverage) using the *k*-mer spectrum approach, with a single peak of *k*-mer coverage. Conversely, we observed a double peak (12x and 24x) in the *k*-mer distribution of *S. lamarcki* which is the hallmark of high heterozygosity content (Kajitani et al 2014), and when summed is consistent with a genome size of 1.25 Gb. We therefore recover approximately 50% and 80% of the estimated genomes of these species. This shortfall is likely explained by the misassembly of the repetitive fraction in these genomes, and the moderate coverage and short fragment sizes libraries used for assembly will make proper recovery of these portions of the genome challenging.

Our genome size estimates are consistent with previous experimental measures for *P. vulgata* and *S. lamarcki*, which indicate respective sizes of 1 Gbp and 1.2-1.5 Gbp according to the Animal Genome Size Database (www.genomesize.com/). The C-value (in pg) of the ten species of Patellogastropoda as extracted from this database varies between 0.43 (*Lottia gigantea*) and 0.94 (*Acmaea mitra*). The *P. vulgata* genome is therefore larger than that of all other patellogastropods, but we note that our wild-collected individual is likely highly heterozygous, which may artificially inflate our estimate. Annelid genome sizes vary widely, from the exceptionally small (*Dinophilus gyrociliatus*, 0.06 pg (Soldi et al 1994)) to the very large (*Spirosperma ferox,* 7.64 pg; Gregory and Hebert (2002)), and *S. lamarcki* is therefore within the normal range for annelid genome size.

The *P. vulgata* GC percentage, 35.18%, is similar to that found in a previous transcriptomic study (Werner et al 2013: 33.56 %), as well as resembling that seen in other molluscs. However, at 27.97%, the GC content of the *S. lamarcki* genomic assembly is exceptionally low. It is markedly lower than that seen in previous EST (42.42 %, Takahashi et al 2009) and transcriptome (43.33 %, Kenny and Shimeld 2012) analyses. It is also lower than that recorded in previously published annelid datasets (*Capitella teleta* 40%, *Helobdella robusta* 33%; Simakov et al 2013). Whether this

reflects the biology of these organisms or is the result of bias in our genomic sequencing and assembly remains to be confirmed.

Contiguity is generally low, although a sizeable percentage of the assemblies are contained in contigs longer than 1kb in length (87.5% of *P. vulgata* and 75.3% of *S. lamarcki*). At this size, contigs can easily be assayed for protein domain content, allowing identification of genes as well as some information about intronic and non-coding regions. The N50 results (*P. vulgata:* 3,160, *S. lamarcki:* 1,939) are also of sufficient length to ensure many regions of the genome have adequate contiguity for establishing whole gene sequence.

3.2 Assessment of Assembly Quality

Our genome assemblies contain 68/248 (27.42%) complete, 144/248 (58.06%) partial (*P. vulgata)* and 6/248 (2.42%) complete, 31/248 (12.50%) partial recovery (*S. lamarcki*) of the core eukaryotic gene mapping dataset as assessed by CEGMA (Parra et al 2007). To confirm these statistics, we utilized BUSCO (Simao et al 2015), which returned similar results to CEGMA. Of the BUSCO set of 843 metazoan orthologues, *P. vulgata* possessed 177 complete, 7 duplicated, 165 fragmented and 501 missing genes, for a total recovery of 41% of the dataset. *S. lamarcki'*s assembly contained 12 complete, 0 duplicated, 27 fragmented and 804 missing genes (5% recovery).

The poor recovery in *S. lamarcki* is probably a consequence of the low contiguity of the genome assembly. To test whether these genes were present but unrecovered by CEGMA or BUSCO due to low contig length, we ran TBLASTN against the genome assemblies using the 843 BUSCO orthologs as queries, with an *E* value cut-off of 1e-6. Of these, 798 (94.7%) of these had at least one hit in *P. vulgata,* and 709 (84.1%) in *S. lamarcki*. This is an overestimation of their recovery, as there will be shared domains in some of these proteins capable of generating a hit above threshold, but indicates the assemblies possess more of the coding fraction of the genomes than raw CEGMA or BUSCO output suggests. There was also only a 10.6% difference between *S. lamarcki* and *P. vulgata* using this approach, compared to 25% to 45.56% for CEGMA and BUSCO, indicating that the greater CEGMA/BUSCO recovery from *P. vulgata* may be a consequence of its increased average contig length. Finally, comparison of published transcriptome data to the assemblies using the STAR splice-aware aligner found higher levels of recovery, with 77.07% (*P. vulgata*) and 33.66% (*S. lamarcki*) of RNA-seq reads mapped. We conclude that while contiguity is limited (especially for *S. lamarcki*), we recover the majority of *P. vulgata* coding sequence and a considerable proportion *S. lamarcki* coding sequence in the assemblies. Further, low contiguity is less likely to affect the detection of miRNA loci, which are very short in length.

3.3 In silico identification of miRNA genes

Using the complete miRNA dataset contained in miRbase we recovered 45 (*P. vulgata*) and 54 (*S. lamarcki)* putative miRNAs (Supplementary File A). This is less than the number of miRNAs catalogued for *Lottia gigantea* (59) and *Capitella teleta* (129) (mirBase v21), but higher than the number reported in other lophotrochozoan species such as *Platynereis dumerilii* (34; Christodoulou et al 2010). Nineteen of these putative miRNAs were found in both *P. vulgata* and *S. lamarcki*, increasing confidence that they are bona fide miRNAs.

Of the canonical miRNA families, we recovered a large number of well-conserved examples (as described in Wheeler et al 2009) from both genomes. Eumetazoan miRNA families such as *miR-100*, and bilaterian families such as *miR-7, miR-8* and miR-*9* (among many others) were found in both genomes. We were also able to find miRNAs only previously described in molluscs or annelids (Wheeler et al 2009, Tarver et al 2013) in *P. vulgata* and *S. lamarcki* respectively, such as *miR-1985, miR-1986* and *miR-1988* in *P. vulgata* and *miR-1996* and *miR-2000* in *S. lamarcki*. These findings reinforce the categorisation of these particular miRNAs as gastropod and annelid synapomorphies respectively (e.g. Tarver et al 2013).

Surprisingly, we also noted the presence of a range of miRNA families that had been previously described as confined to lineages other than Lophotrochozoa. For example, *miR-1285, miR-1287, miR-1957, miR-1983* and *miR-3533* (undescribed outside Vertebrata), *miR-3350* (described only in silkworm; Cai et al (2010)) and *miR-494* and *miR-767* (only reported in eutherians; Tarver et al (2013)), were found in either *S. lamarcki, P. vulgata,* or both.

To ascertain whether these genes are present more broadly across the Lophotrochozoa, we first checked them against published lophotrochozoan miRNA sequences that have not been accessioned into miRbase (Xue et al 2008, Christodoulou et al 2010 and Bao et al 2014), but none of these mIRNAs have previously been noted. To check whether this was through oversight rather than true absence, we searched the published genomes of the molluscs *P. fucata, C.gigas* and *L. gigantea* and the annelids *H. robusta and C. teleta* for these sequences. miRNAs *miR-494, miR-767, miR-1287* and *miR-3350*, were not identified in the other lophotrochozoan species examined, a finding that supports reports of heterogeneous rates of gain and loss (Fromm et al 2013, Thomson et al 2014). However candidate *miR-1957, mir-1983* and *mir-3533* sequences were found in all lophotrochozoan genomes examined, suggesting that these species possess an as-yet uncharacterised diversity of miRNA genes (Supplementary File B). Some miRNAs, such as the putative *L. gigantea miR-1957*, are 100% identical to their *P. vulgata* counterpart. These three miRNAs could therefore have been

present in the common ancestor of molluscs and annelids, but be as yet unidentified in living descendant by small RNA sequencing approaches.

3.4 Verification of Transcription

In order to discern whether these surprising putative miRNAs were in fact transcribed, we extracted short fragment RNA from head, mantle and foot tissue of *P. vulgata* and from whole adult *S. lamarcki*. We then used RT-PCR to assay for the presence of several miRNA, using primers as described in Table 2. Not all miRNA sequences identified in our analysis were found to be transcribed in these samples. At present, we could not differentiate whether this could indicate the predicted miRNAs are artefacts, or whether they are expressed in tissues or stages we did not examine. Positive results are shown in Figure 3C. In both *P. vulgata* and *S. lamarcki,* evidence for the transcription of the *miR-1983* and *miR-3533* loci was obtained (Fig. 3C): these miRs have been previously only described in mice (Barbiaz et al 2008) and/or other vertebrates (Wang et al 2009, Vegh et al 2013). In addition, we also noted the transcription of *miR-1957* (mouse specific; Kuchenbauer et al (2008)) in *P. vulgata* and *miR-1285* and *miR-1287* (vertebrate specific; Strozzi et al (2009), Brameier (2010), Meunier et al (2013)) in *S. lamarcki.* Together this provides evidence for the transcription of several of these miRNA loci. It should however be noted that miRbase shows *miR-3533* as so far only identified in chicken and cow, and in both cases as mapping to a region of the genome that also encodes an actin gene. This co-localisation with actin genes is also found in our lophotrochozoan sequences, and may both contribute to the level of homology as shown in Figure 3A and underlie its transcription.

4. Discussion and Conclusions

The two genome assemblies presented here will be useful for a range of investigations, given the currently sparse sampling of lophotrochozoans. Our preliminary assemblies and *k*-mer based genome size estimates also provide a basis for establishing appropriate strategies to improve genome assembly in these and related species. As with some other recently-sequenced marine animals, such as the pacific oyster *Crassostrea gigas* (Zhang et al 2012) and the cephalochordate *Branchiostoma belcheri* (Huang et al 2014), these two genomes are highly polymorphic. This is likely explained by very large effective population sizes deriving from planktonic larval dispersal (Romiguier et al 2014). High levels of polymorphism represent a challenge for assembly and annotation. Nevertheless, while contiguity is low, assembly is still sufficient to carry out the miRNA analysis described here.

miRNAs represent potentially useful, slow-evolving characters for the inference of phylogeny, and have been used to suggest solutions to a variety of recalcitrant problems in metazoan phylogeny. However, given the sparse and biased nature of our sampling of genomic diversity across the tree of life, it would perhaps be better to adopt caution when making claims as to the phylogenetic inter-relationships of taxa, especially in cases where only a small number of miRNAs are used as the basis for such claims, or where only single datasets are utilized as the raw material for such inference. Here, we show that an informative survey of miRNAs is possible through moderate coverage genome sequencing and assembly, and we suggest a similar approach might be applied more broadly across animal diversity. We found many (45 *P. vulgata* and 54 *S. lamarcki*) putative miRNAs in our datasets, representing a considerable proportion of the expected complements of lophotrochozoans when compared to the well-annotated *L. gigantea* and *C. teleta* genomes. The discovery of potential mollusc- and annelid-specific miRNAs in *P. vulgata* and *S. lamarcki* respectively also reinforces the likelihood that those particular genes are synapomorphies of these clades, rather than limited to the (often single species) samples where they were first described. This demonstrates the utility of extra datasets when addressing the still under-researched field of miRNA evolution.

While *miR-1957, miR-1983* and *miR-3533* (noting the caveat for mir-3533 described above) are found in both the species examined here and other sequenced lophotrochozoan species, some of the more surprising findings are limited to only one of the two assemblies described here, as can be seen in Figure 4. When coupled with their absence from other sequenced lophotrochozoan genomes, this raises the possibility of homoplasy, notably for the putative *miR-494, miR-767, miR-1287* and *miR-3350*. However sampling of animal, and particularly lophotrochozoan, phylogeny remains too sparse to assert with full confidence whether the sequences found, particularly those of *miR-494, miR-767, miR-1287* and *miR-3350,* could have arisen by convergent evolution or are the remnants of prevalent loss across metazoan phylogeny. These competing hypotheses will be better tested when a broader range of genomes are available, drawn from the widest possible range of bilaterian species. Datasets such as the ones presented here represent key comparison points in the interim. Either way, however, the identification of unexpected miRNA sequences in these species suggests that current assumptions concerning the utility of miRNAs in reconstructing phylogeny need qualification: either more loss has occurred across metazoan phylogeny than postulated previously, or convergent evolution across wide evolutionary distances is possible. The complex evolutionary history of miRNAs is interesting in light of the important role that they play in animal biology, and understanding the true nature of their gain and loss will be vital for an understanding of how miRNAs influence and are influenced by genomic evolution.

Without detailed cataloguing of miRNA family sequences in a more diverse range of animals, some of the key assumptions made about miRNAs for use in phylogeny will remain untested. Only greater density of sampling across the metazoan tree of life will reveal further the true nature of miRNA conservation and loss. The two novel lophotrochozoan genomes presented here, as well as being useful for a wide range of investigations, are another step in this process.

5. Acknowledgements

The authors thank the Elizabeth Hannah Jenkinson Fund for grants supporting the sequencing of the genomes listed here (*P. vulgata:* grant to NJK and JHLH, *S. lamarcki:* grant to EKON and JHLH). For sequencing we thank the High-Throughput Genomics unit at the Wellcome Trust Centre for Human Genetics, Oxford. NJK was supported by a Clarendon Scholarship for work on this project. We also thank the members of our laboratories for their many helpful comments and support.

6. Statement of Competing Interests

The authors declare no conflict of interests. The funding source had no input into the decision to publish or in preparation of this manuscript.

7. References

Aguinaldo, A.M.A., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., Lake, J.A., 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387: 489-493 doi:10.1038/387489a0.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Axtell, M.J., Bartel, D.P. 2005. Antiquity of microRNAs and their targets in land plants". The Plant Cell 17: 1658–1673. doi:10.1105/tpc.105.032185

Bao, Y., Zhang, L., Dong, Y., & Lin, Z. 2014. Identification and Comparative Analysis of the Tegillarca granosa Haemocytes MicroRNA Transcriptome in Response to Cd Using a Deep Sequencing Approach. PloS one 9: e93619 doi:10.1371/journal.pone.0004034 .

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116: 281–297 doi:10.1016/S0092-8674(04)00045-5 .

Bartel, D.P., 2009. MicroRNAs: Target recognition and regulatory functions. Cell 136: 215–233 doi: 10.1016/j.cell.2009.01.002.

Cai, Y., Yu, X., Zhou, Q., Yu, C., Hu, H., Liu, J., Lin, H., Yang, J., Zhang, B., Cui, P., Hu, S., Yu, J., 2010. Novel microRNAs in silkworm (*Bombyx mori*). Functional & Integrative Genomics, 10: 405-415 doi: 10.1007/s10142-010-0162-7.

Campbell, L.I., Rota-Stabellli, O., Marchioro, T., Longhorn, S. J., Edgecombe, G.D., Telford, M.J., Philippe, H., Rebecchi, L., Peterson, K.J., Pisani, D., 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest the velvet worms are the sister group of Arthropoda. Proceedings of the National Academy of Sciences, USA 108: 15920-15924 doi: 10.1073/pnas.1105499108.

Christodoulou, F., Raible, F., Tomer, R., Simakov, O., Trachana, K., Klaus, S., et al. 2010. Ancient animal microRNAs and the evolution of tissue identity. Nature 463: 1084–1088. doi:10.1038/nature08744

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.,. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21 doi: 10.1093/bioinformatics/bts635.

Field, D.J., Gauthier, J.A., King, B.L., Pisani, D., Lyson, T.R., Peterson, K.J., 2014. Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. Evolution & Development 16:189-196 doi: 10.1111/ede.12081.

Friedman, R.C., Farh, K. K.-H., Burge, C.B., Bartel, D.P. 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Research 19: 92-105. doi: 10.1101/gr.082701.108.

Fromm, B., Worren, M.M., Hahn, C., Hovig, E., Bachmann, L., 2013. Substantial loss of conserved and gain of novel microRNA families in flatworms. Molecular Biology and Evolution, 30: 2619-2628 doi: 10.1093/molbev/mst155.

Gregory, T.R., Hebert, P.D.N., 2002. Genome size estimates for some oligochaete annelids. Canadian Journal of Zoology 80: 1485-1489 doi: 10.1139/z02-145.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., Enright, A. J., 2008. miRBase: tools for microRNA genomics. Nucleic Acids Research, 36: D154-D158 doi: 10.1093/nar/gkm952.

Halanych, K.M., Bacheller, J., Liva, S., Aguinaldo, A. A., Hillis, D.M. Lake, J.A., 1995. 18S rDNA evidence that the Lophophorates are Protostome Animals. Science 267: 1641–1643. doi:10.1126/science.7886451

Hamer, J. ,Walker, G., Latchford, J., 2001. Settlement of *Pomatoceros lamarkii* (Serpulidae) larvae on biofilmed surfaces and the effect of aerial drying. Journal of Experimental Marine Biology and Ecology, 260: 113 – 132 doi:10.1016/S0022-0981(01)00247-7.

Huang, S., Chen, Z., Yan, X., Yu, T., Huang, G., Yan, Q., et al 2014. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. Nature Communications 5: doi:10.1038/ncomms6896.

Hui, J.H.L., Marco, A., Hunt, S., Melling, J., Griffiths-Jones, S., Ronshaugen, M., 2013a. Structure, evolution and function of the bi-directionally transcribed iab-4 microRNA locus in insects. Nucleic Acids Research 41: 3352-3361 doi: 10.1093/nar/gks1445.

Hui J.H.L., Bendena W.G., Tobe S.S., 2013b. Future perspectives on the research of juvenile hormones and sesquiterpenoids in arthropod endocrinology and ecotoxicology. CRC Press. In Juvenile hormones and juvenoids: Modelling Biological Effects and Environmental Fate, pp. 15-30.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T,, Itoh T., 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research 24: 1384–1395 doi:10.1101/gr.170720.113.

Kelley, D.R., Schatz, M.C., Salzberg, S.L., 2010. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11, R116 doi: 10.1186/gb-2010-11-11-r116.

Kenny, N.J., Shimeld, S.M. 2012. Additive multiple k-mer transcriptome of the keelworm *Pomatoceros lamarckii* (Annelida; Serpulidae) reveals annelid trochophore transcription factor cassette. Development Genes and Evolution, 222: 325-339 doi: 10.1007/s00427-012-0416-6.

Kenny, N.J., Quah, S., Holland, P.W.H., Tobe, S.S., Hui, J.H.L., 2013. How do comparative genomics and microRNAs change our views on arthropod endocrinology and their adaptations to the environment? General and Comparative Endocrinology, 188: 16-22 doi: 10.1016/j.ygcen.2013.02.013.

Kenny, N.J., Sin, Y.W., Hayward, A., Paps, J., Chu, K.H., Hui, J.H.L., 2015 The phylogenetic utility of functional constraint on microRNA sequence evolution. Proceedings of the Royal Society B, doi:10.1098/rspb.2014.2983.

Kloosterman, W.P., Plasterk, R.H.A. 2006. The diverse functions of microRNAs in animal development and disease. Developmental Cell 11: 441-450 doi:10.1016/j.devcel.2006.09.009

Kuchenbauer, F., Morin, R.D., Argiropoulos, B., Petriv, I., Griffith, M., Heuser, M., Yung, E., Piper, J., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Hansen, C.L., Marra, M.A., Humphries, R.K. 2008. In-depth characterization of the microRNA transcriptome in a leukemia progression model. Genome Research 18: 1787-1797 doi: 10.1101/gr.077578.108.

Lee, R. C., Feinbaum, R. L., & Ambros, V. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell 75: 843-854 doi: 10.1016/0092-8674(93)90529-Y.

Lindberg, D.R. Estes, J.A., Warheit, K.I. 1998. Human influences on trophic cascades along rocky shores. Ecological Applications 8: 880-890.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T., Wang, J., 2012. SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler. Gigascience 1: 1–6 doi: 10.1186/2047-217X-1-18..

McDougall, C., Chen, W-C., Shimeld, S.M., Ferrier, D.E.K. (2006) The development of the larval nervous system, musculature and ciliary bands of *Pomatoceros lamarckii* (Annelida): heterochrony in polychaetes. Frontiers in Zoology 3: 16.

Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., Hu, H., Khaitovich, P., Kaessmann, H., 2013. Birth and expression evolution of mammalian microRNA genes. Genome Research 23: 34-45 doi: 10.1101/gr.140269.112.

Nakano, T., Sasaki, T. Recent advances in molecular phylogeny, systematics and evolution of patellagastropod limpets. J. Molluscan Studies 77: 203-217.

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics, 23: 1061-1067 doi: 10.1093/bioinformatics/btm071.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., Ruvkun, G., 2000. Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. Nature 408: 86-89 doi:10.1038/35040556.

Philippe, H., Brinkmann, H., Copley, R.R., Moroz, L.L., Nakano, H., Poustka, A.J., Wallberg, A., Peterson, K.J., and Telford, M.J., 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470: 255-258 doi: 10.1038/nature09676.

Plaisance-Bonstaff, K., Renne, R., 2011. Viral miRNAs. In Antiviral RNAi (pp. 43-66). Humana Press doi: 10.1007/978-1-61779-037-9_3.

Quah, S., Hui, J.H., Holland, P.W., 2015. A burst of miRNA innovation in the early evolution of butterflies and moths. Molecular Biology and Evolution 32: 1161-1174. doi: 10.1093/molbev/msv004

Reddy, T.B.K., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A., Kyrpides, N.C., 2014. The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification. Nucleic Acids Research gkv455v1-gkv455 doi: 10.1093/nar/gku950.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A. A.-T., Weinert, L. A., Belkhir, K., Bierne, N., Glémin, S., Galtier, N. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515: 261-263 doi.org/10.1038/nature13685

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., Telford, M., 2010. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata.

Proceedings of the Royal Society of London B Biological Sciences 278: 298-306 doi: 10.1098/rspb.2010.0590.

Ryan, J.F., 2013. estimate_genome_size.pl (version 0.03) [Computer software]. Bergen, Norway: Sars International Centre for Marine Molecular Biology. Retrieved from http://josephryan.github.com/estimate_genome_size.pl/

Sempere, L. F., Cole, C. N., McPeek, M. A., Peterson, K. J., 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 306: 575-588 doi: 10.1002/jez.b.21118.

Simakov, O., Marletaz, F., Cho, S.J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood, J., Chapman, J.A., Shapiro, H., Aerts, A., Otillar, R.P., Terry, A.Y., Boore, J.L., Grigoriev, I.V., Lindberg, D.R., Seaver, E.C., Weisblat, D.A., Putnam, N.H., Rokhsar, D.S., 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493: 526-531 doi: 10.1038/nature11696.

Simão, F. A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov E. M., 2015. Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs. http://busco.ezlab.org, Version 1.1b, May 2015.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J. E., Jones, S.J., & Birol, I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research 19: 1117-1123 doi: 10.1101/gr.089532.108.

Soldi, R., Ramella, L., Gambi, M.C., Sordino, P., Sella, G., 1994. Genome size in polychaetes: relationship with body length and life habit. in: Dauvin, J-C., Laubier, L., Reish D.J. (Eds), Actes de la 4ième Conférence internationale des Polychètes; Mém. Mus. Natn. Hist. Nat. 162, Paris. 129-135.

Strozzi, F., Mazza, R., Malinverni, R., & Williams, J. L. 2009. Annotation of 390 bovine miRNA genes by sequence similarity with other species. Animal Genetics 40: 125-125 doi: 10.1111/j.1365-2052.2008.01780.x.

Struck, T.H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., Bleidorn, C., 2011. Phylogenomic analyses unravel annelid evolution. Nature 470: 95–98. doi:10.1038/nature09864

Struck, T.H., Wey-Fabrizius, A.R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., Klebow, S., Iakovenko, N., Hausdorf, B., Petersen, M., Kuck, P., Herlyn, H., Hankeln, T. 2014. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. Mol. Biol. Evol. 31: 1833-1849 doi:10.1093/molbev/msu143

Takahashi, T., McDougall, C., Troscianko, J., Chen, W.C., Jayaraman-Nagarajan, A., Shimeld, S.M., Ferrier, D.E., 2009. An EST screen from the annelid *Pomatoceros lamarckii* reveals patterns of gene loss and gain in animals. BMC Evolutionary Biology 9: 240 doi: 10.1186/1471-2148-9-240.

Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E., Fujiwara, M., Shinzato, C., Hisata, K., Fujie, M., Usami, T., Nagai, K., Maeyama, K., Okamoto, K., Aoki, H., Ishikawa, T., Masaoka, T., Fujiwara, A., Endo, K., Endo, H., Nagasawa, H., Kinoshita, S., Asakawa, S., Watabe, S., Satoh, N., 2012. Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA Research 19:117-30 doi: 10.1093/dnares/dss005.

Tarver, J.E., Donoghue, P.C., Peterson, K.J., 2012. Do miRNAs have a deep evolutionary history?. Bioessays 34: 857-866 doi: 10.1002/bies.201200055.

Tarver, J.E., Sperling, E.A., Nailor, A., Heimberg, A.M., Robinson, J.M., King, B.L., Pisani, D., Donoghue, P.C.J., Peterson, K.J., 2013. miRNAs: Small Genes with Big Potential in Metazoan Phylogenetics. Molecular Biology and Evolution, 30: 2369-2382 doi: 10.1093/molbev/mst133.

ten Hove, H. 2015. Spirobranchus lamarcki (Quatrefages, 1866). In: Read, G.; Fauchald, K. (Ed.) (2015) World Polychaeta database. Accessed through: World Register of Marine Species at http://www.marinespecies.org/aphia.php?p=taxdetails&id=560033 on 2015-05-29

Thomson, R.C., Plachetzki, D.C., Mahler, D.L., Moore, B.R., 2014. A critical appraisal of the use of microRNA data in phylogenetics. Proceedings of the National Academy of Sciences USA 111: E3659-E3668 doi: 10.1073/pnas.1407207111.

Vegh, P., Foroushani, A. B., Magee, D.A., McCabe, M.S., Browne, J.A., Nalpas, N.C., Conlon, K.M., Gordon, S.V., Bradley, D.G., MacHugh, D.E., Lynn, D.J., 2013. Profiling microRNA expression in bovine alveolar macrophages using RNA-seq. Veterinary Immunology and Immunopathology 155: 238-244 doi: 10.1016/j.vetimm.2013.08.004.

Wang, Y., Brahmakshatriya, V., Zhu, H., Lupiani, B., Reddy, S.M., Yoon, B.J., Gunaratne, P.H., Kim, J.H., Chen, R., Wang, J., Zhou, H., 2009. Identification of differentially expressed miRNAs in chicken lung and trachea with avian influenza virus infection by a deep sequencing approach. BMC Genomics 10: 512 doi: 10.1186/1471-2164-10-512.

Werner, G.D., Gemmell, P., Grosser, S., Hamer, R., Shimeld, S.M., 2013. Analysis of a deep transcriptome from the mantle tissue of *Patella vulgata* Linnaeus (Mollusca: Gastropoda: Patellidae) reveals candidate biomineralising genes. Marine Biotechnology 15: 230-243 doi: 10.1007/s10126-012-9481-0.

Wheeler, B.M., Heimberg, A.M., Moy, V.N., Sperling, E.A., Holstein, T.W., Heber, S., Peterson, K.J., 2009. The deep evolution of metazoan microRNAs. Evolution and Development 11: 50-68.

Xue, X., Sun, J., Zhang, Q., Wang, Z., Huang, Y., Pan, W., 2008. Identification and characterization of novel microRNAs from *Schistosoma japonicum*. PloS one 3: e4034 doi: 10.1371/journal.pone.0004034.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490: 49–54 doi: 10.1038/nature11413.

Figure Legends

Figure 1: A representative phylogeny of bilaterally symmetrical animals, sampling at the genomic level across the Bilateria, and images of *Patella vulgata* and *Spirobranchus (Pomatoceros) lamarcki*.

A. Phylogeny is representative and based on a consensus of recent studies. Numbers cited in figure taken from GOLD (Reddy et al 2014). 'Complete' refers to the GOLD nomenclature differentiating finalised projects from those still underway. B. Oblique (left) and ventral (right) view of adult *P. vulgata.* C. Adult *S. lamarcki* (left), and calcareous tube *in situ* on a rock gathered from Plymouth seafront (right).

Figure 2: Coverage distributions of sequence read data.

21-mer distribution showing *k* mer coverage peak (22x and 12/24x in *P. vulgata* and *S. lamarcki* respectively) that are translated into actual sequencing coverages using the equation $C_k=C\times(L-k+1)/L$, where C is real coverage, L is read length, k is *k* mer size and C_k is *k* mer coverage.

Figure 3: Evidence for the Expression of "Vertebrate like" miRNA in *Patella vulgata* and *Spirobranchus (Pomatoceros) lamarcki*.

A. Example alignments of novel miRNA candidates with known orthologues. B. Hairpin structures of example miRNAs, *S. lamarcki* miR-1285 and *P. vulgata* miR-1983, as displayed by RNAfold (Gruber et al 2008) with positional entropy values displayed. C. RT-PCR results showing the expression of miRNAs in both species examined. + and - indicate analyses conducted with and without reverse transcription (RT). Note the presence of a weak band indicating amplification of small amounts of residual genomic DNA in the 5s RNA band for *P. vulgata* -RT sample (due to the high copy number of rRNA genes in the genome). We did not observe bands indicating miRNA sequence amplification in other -RT controls. Smears at the base of miR-1983 in *P. vulgata* are primer dimers, a weak band can be seen above in +RT lane.

Figure 4: Distribution of phylogenetically distant miRNA gene apparent homologues.

Phylogenetic distribution of miRNA genes of apparent homology discussed in this paper. Those shown in black have been described previously, while those shown in red, bolded text are described for the first time in this paper. Those underlined have further been shown to be transcribed in *S. lamarcki* (Annelida) or *P. vulgata* (Mollusca).

Tables

Table 1: Sequencing and assembly statistics for *Patella vulgata* and *Spirobranchus (Pomatoceros) lamarcki* genomes. Note: the word 'contig' is used as no long mate pair libraries were used for scaffolding our assemblies.

Species/Metric (bp unless stated)	*Patella vulgata*	*Spirobranchus lamarcki*
Number of 2*100bp Paired End Fragments (200bp library)	118,329,948	119,500,405
Number of 2*100bp Paired End Fragments (500bp library)	69,107,905	67,048,915
Min contig length	300	300
Max contig length	47,432	37,132
Mean contig length	1960.27	1327.69
Std deviation of contig length	2199.21	1439.37
Median contig length	1,224	848
N50 contig length	3,160	1,939
Number of contigs	295,348	726,277
Number of contigs >=1kb	170,706	311,778
Number of contigs in N50	51,138	135,692
Number of bases in all contigs	578,961,269	964,274,156
No. of bases in contigs >=1kb	506,356,838	726,465,383
GC Content of contigs (%)	35.18	27.97
N content of contigs (%)	0.00062	0.53485
Genome size (Gbp)	1.46 (perl) /0.95 (*k* mer)	1.25
Overall Coverage	25.60x (perl) /28.2x (*k* mer)	15/30x
k mer Coverage (21mer)	22x	12/24x

Table 2: miRNAs amplified from RNA samples to confirm expression, with primer sequences and predicted band size.

miRNA:	**Primers:**	**Expected Band Size:**
P. vulgata 5s rRNA (*control)*	F: ACCACGTTGAAAACACCAGTTC R: CGGTCACCCATCCAAGTACTAA	81 bp
S. lamarcki 5s rRNA (*control)*	F: GCCATACCACGCTGAATACAC R: GCTTAACTCCCGTGATTGGA	50 bp
S. lamarcki miR-1285	F: GGATAGCACCTGTGAATAGGC R: CCAGCTATGTTGGACAGGCTA	51 bp
S. lamarcki miR-1287	F: CGAAGATTCTAGAAAAGTGGTTCGAG R: GCACCTATCACTGAATCTGTTGC	73 bp
P. vulgata miR-1957	F:GGGATGTAGCTCAGTGGTAGAG R: GAACCCGGGGCCTTTCAC	54 bp
P. vulgata miR-1983	F: AGCGCGCCGTACTTATAGACAG R: GTGAGGCTCGAACTCACAACCT	83 bp
S. lamarcki miR-1983	F: CAGCCCCAGTAGCTCAGTCG R: AAAACAATTGCCCCAGGTGA	124 bp
P. vulgata miR-3533	F: GGCATGGAATCTTCTGGTATTC R: CAAGTCTTTACGGATATCAACATCAC	78 bp
S. lamarcki miR-3533	F: CGAGACCACCTACAACAGCA R: GCGTACAAGTCCTTACGGATG	60 bp

Supplementary files

Supplementary File A: miRNAs identified by our approach and their sequence. Curated information regarding those genes chosen for transcriptional verification (.xls file).

Supplementary File B: Sequence and scaffold location of putative miRNAs in other sequenced genomes (.txt file).

Figure 1

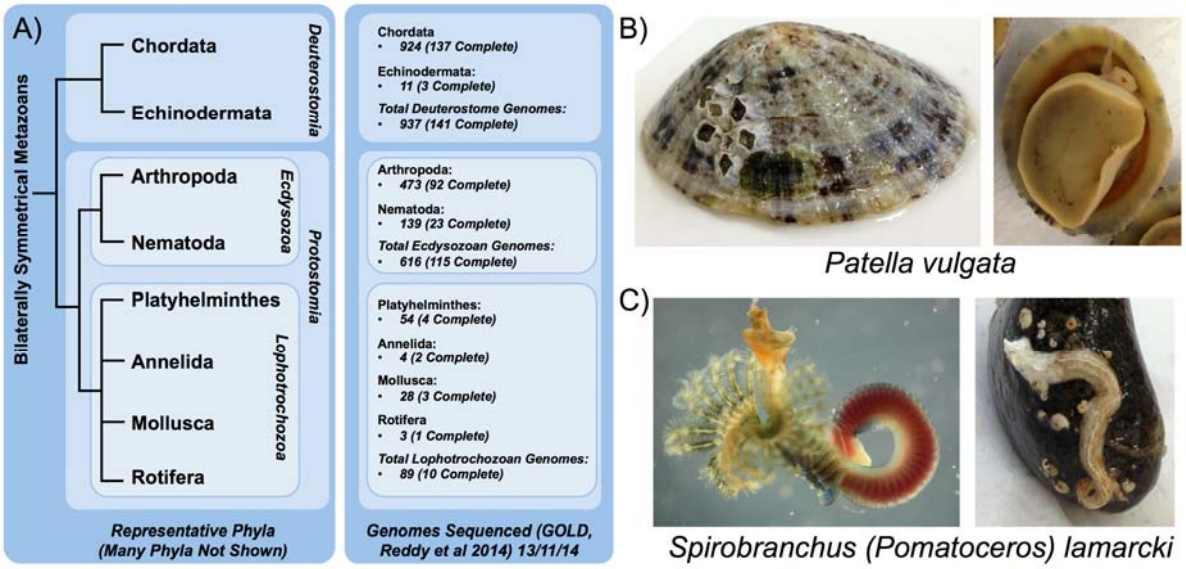


Figure 2

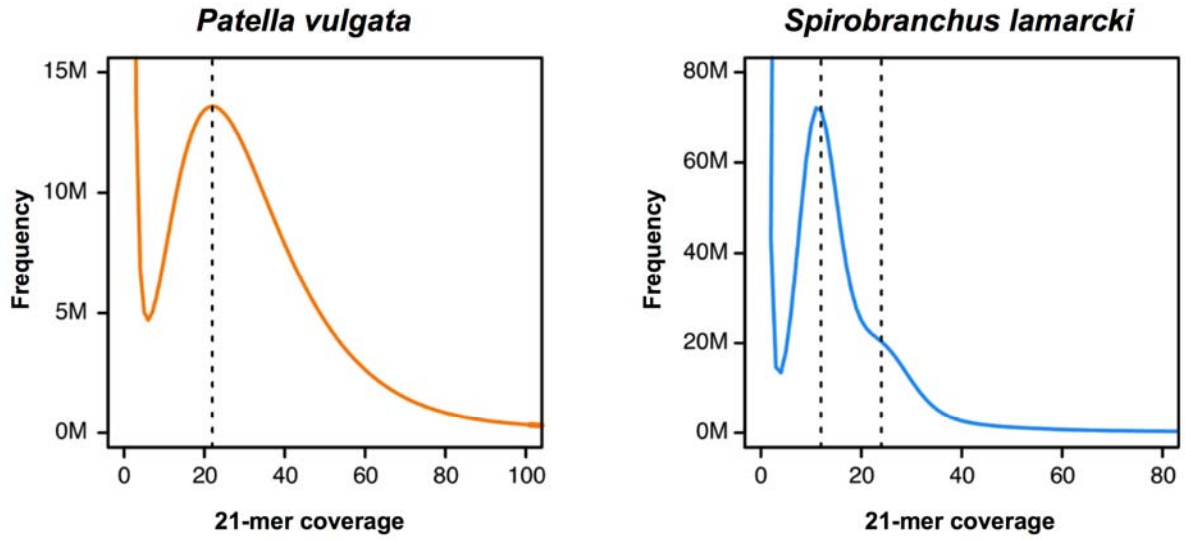


Figure 3

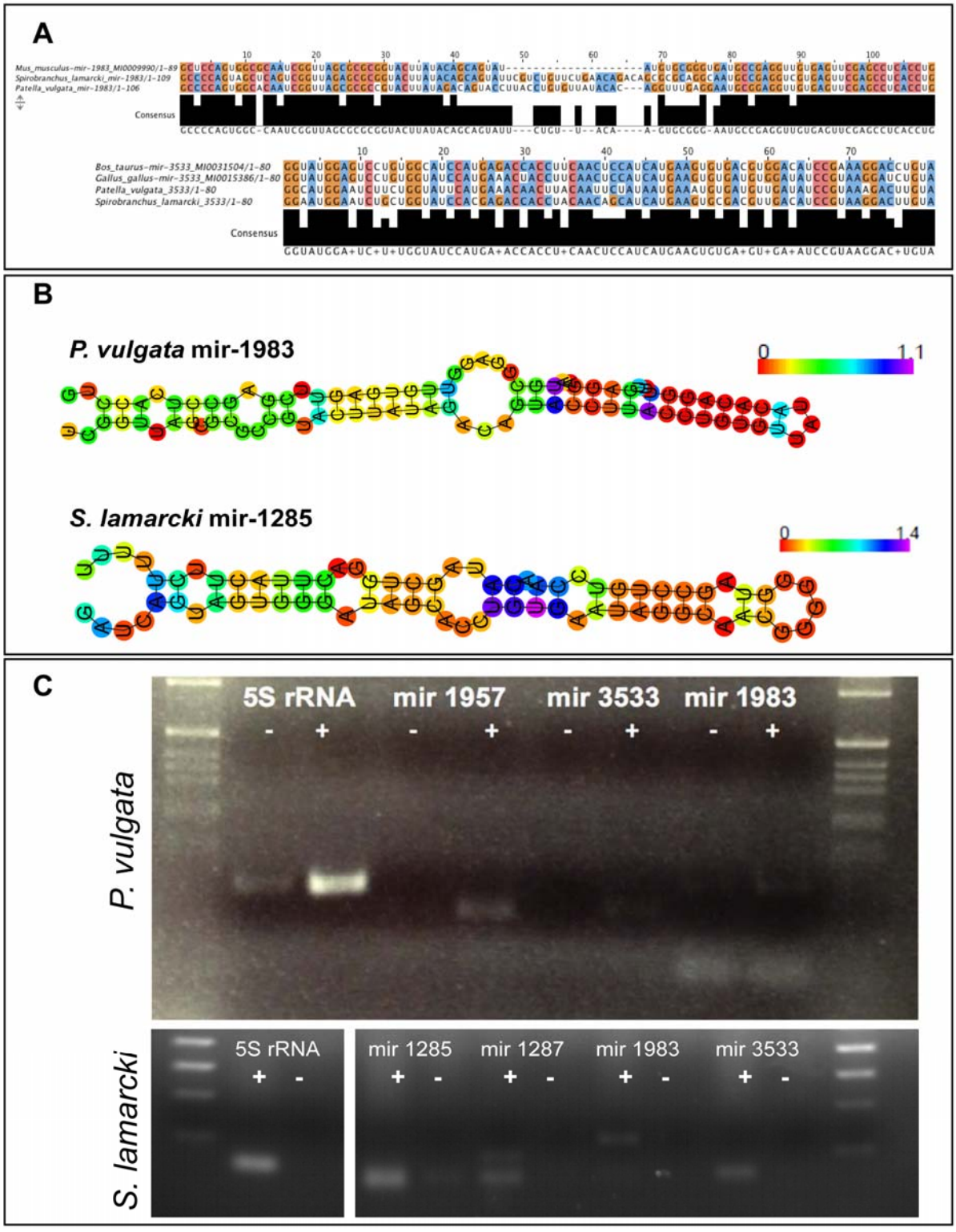
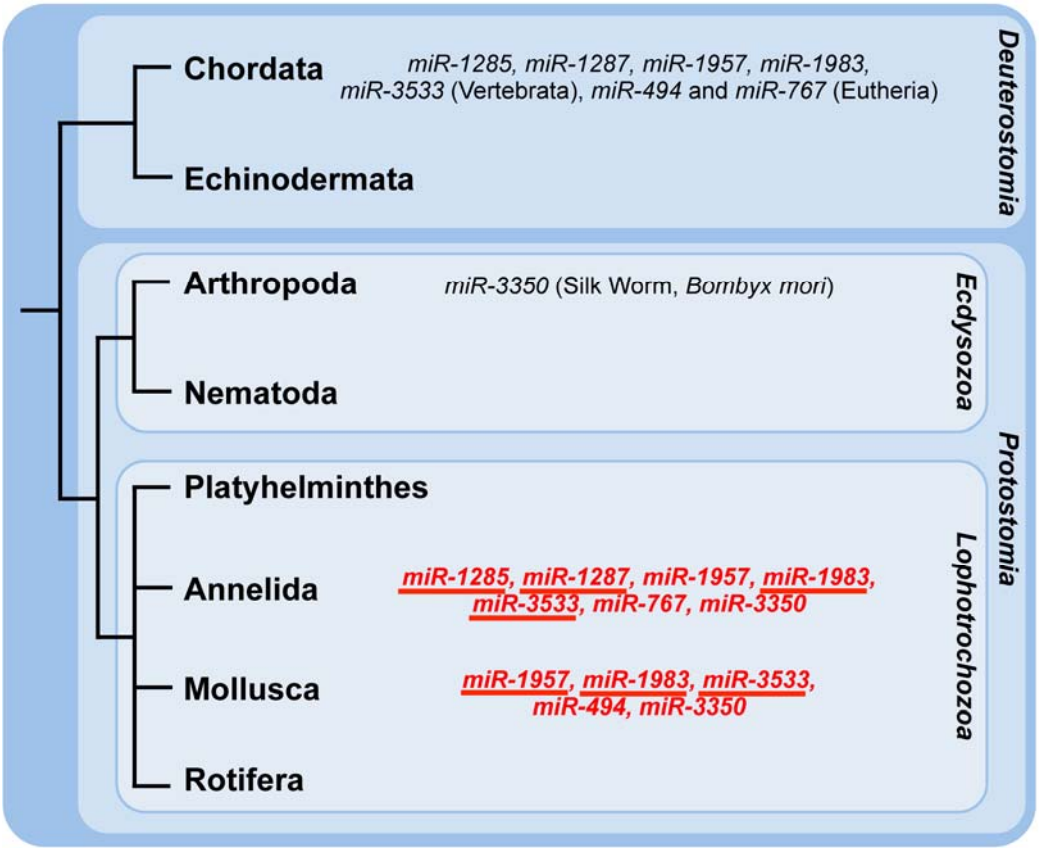

A
Mus_musculus-mir-1983_MI0009990/1-89
Spirobranchus_lamarcki_mir-1983/1-109
Patella_vulgata_mir-1983/1-106
Consensus
Bos_taurus-mir-3533_MI0031504/1-80
Gallus_gallus-mir-3533_MI0015386/1-80
Patella_vulgata_3533/1-80
Spirobranchus_lamarcki_3533/1-80
Consensus
B
P. vulgata mir-1983
0
1.1
S. lamarcki mir-1285
0
1.4
C
P. vulgata
5S rRNA
mir 1957
mir 3533
mir 1983
S. lamarcki
5S rRNA
mir 1285
mir 1287
mir 1983
mir 3533

Figure 4


Chordata
miR-1285, miR-1287, miR-1957, miR-1983, miR-3533 (Vertebrata), miR-494 and miR-767 (Eutheria)
Echinodermata
Deuterostomia
Arthropoda
miR-3350 (Silk Worm, Bombyx mori)
Nematoda
Ecdysozoa
Protostomia
Platyhelminthes
Annelida
miR-1285, miR-1287, miR-1957, miR-1983, miR-3533, miR-767, miR-3350
Mollusca
miR-1957, miR-1983, miR-3533, miR-494, miR-3350
Rotifera
Lophotrochozoa