miRNAs: Small Genes with Big Potential in Metazoan Phylogenetics

James E. Tarver,*[,1,2,3] Erik A. Sperling,[4] Audrey Nailor,[1,5] Alysha M. Heimberg,[1,6] Jeffrey M. Robinson,[1] Benjamin L. King,[7] Davide Pisani,[2,3,8] Philip C.J. Donoghue,[2] and Kevin J. Peterson*[,1]

[1]Department of Biological Sciences, Dartmouth College, Hanover, NH

[2]School of Earth Sciences, University of Bristol, Bristol, United Kingdom

[3]Department of Biology, The National University of Ireland, Maynooth, Kildare, Ireland

[4]Department of Earth and Planetary Sciences, Harvard University

[5]Norris Cotton Cancer Centre, Dartmouth Medical School, Hanover, NH

[6]EMBL Australia, Australian Regenerative Medicine Institute, Monash University, Clayton, Australia

[7]Mount Desert Island Biological Laboratory, Salisbury Cove, ME

[8]School of Biological Sciences, University of Bristol, Bristol, United Kingdom

***Corresponding author:** E-mail: james.tarver@nuim.ie; kevin.j.peterson@dartmouth.edu.

Associate editor: Claudia Russo

Abstract

microRNAs (miRNAs) are a key component of gene regulatory networks and have been implicated in the regulation of virtually every biological process found in multicellular eukaryotes. What makes them interesting from a phylogenetic perspective is the high conservation of primary sequence between taxa, their accrual in metazoan genomes through evolutionary time, and the rarity of secondary loss in most metazoan taxa. Despite these properties, the use of miRNAs as phylogenetic markers has not yet been discussed within a clear conceptual framework. Here we highlight five properties of miRNAs that underlie their utility in phylogenetics: 1) The processes of miRNA biogenesis enable the identification of novel miRNAs without prior knowledge of sequence; 2) The continuous addition of miRNA families to metazoan genomes through evolutionary time; 3) The low level of secondary gene loss in most metazoan taxa; 4) The low substitution rate in the mature miRNA sequence; and 5) The small probability of convergent evolution of two miRNAs. Phylogenetic analyses using both Bayesian and parsimony methods on a eumetazoan miRNA data set highlight the potential of miRNAs to become an invaluable new tool, especially when used as an additional line of evidence, to resolve previously intractable nodes within the tree of life.

Key words: **microRNA, phylogenetics, rare genomic character.**

Resolving the tree of life has been a focal ambition of the Life Sciences since the conception of evolutionary theory. This phylogenetic footprint has been sought in multifarious comparative data, from embryology to anatomy, from physiology to the fossil record. However, it was the advent of molecular phylogenetics that provided a more objective means of inferring the evolutionary history of living organisms. The rise of rapid sequencing technologies has allowed the generation of phylogenomic data sets comprising 100s to 10,000s of aligned genes (Hejnol et al. 2009; Hallström and Janke 2010). However, it is clear that currently our ability to generate phylogenomic data is advancing at a faster rate than the computational and methodological tools needed to analyze them. Thus, the largest phylogenomic data sets must be analyzed with computationally efficient but suboptimal models, diminishing or even eliminating any advantage afforded by the additional data (Philippe, Brinkmann, Lavrov, et al. 2011). Hence there has been significant interest in and utilization of large-scale mutations (above the level of single nucleotide substitutions) as phylogenetic markers (Rokas and Holland 2000; Telford and Copley 2011). Classical Rare Genomic Changes (RGCs) include retroposon integrations (SINES and LINES), insertion–deletion events, signature sequences, mtDNA genetic code variants, nuclear DNA genetic code variants, gene rearrangements in mitochondrial and chloroplast genomes, gene order, gene duplications, and chromosomal rearrangements. Here we seek to establish the case for microRNAs (miRNAs) as a new class of RGCs that has already been influential in addressing difficult phylogenetic debates (table 1).

microRNAs—The Nature of the Beast

The mature miRNA (see Bartel 2009 and Berezikov 2011 for reviews) is a small, ~22 nucleotide noncoding gene which negatively regulates the translation of protein-coding gene(s), usually by binding with imperfect complementarity to sites in the 3′ untranslated regions (UTRs) of the messenger RNAs (mRNAs), and thereby subjecting the transcript to cleavage or to blockage of its translation. miRNAs are generally transcribed from either intergenic regions or from introns

© The Author 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

Perspective

Table 1. Research Groups and Associated Publications, Which Have Used microRNAs in Phylogenetics.

Home Institution	Authors (Year)	Clade
Dartmouth College, USA	Heimberg et al. (2010)	Cyclostomes
	Lyson et al. (2012)	Tetrapods
	Peterson et al. (2013)	Hemichordates
	Sperling et al. (2009)	Annelids
	Sperling et al. (2010)	Sponges
University of Leipzig, Germany	Helm et al. (2012)	Myzostomids
The Natural History Museum, London, UK	Pisani et al. (2012)	Echinoderms
Yale University, USA	Sperling et al. (2011)	Brachiopods
Universite' de Montreal, Canada	Philippe et al. (2011)	Deuterostomes
North Carolina State University, USA	Wiegmann et al. (2011)	Dipterans
National University of Ireland Maynooth, Ireland	Campbell et al. (2011)	Arthropods
University College London, UK	Rota-Stabelli et al. (2011)	Mandibulates
Tongji University, China	Cai and Zhang (2010)	Deuterostomes

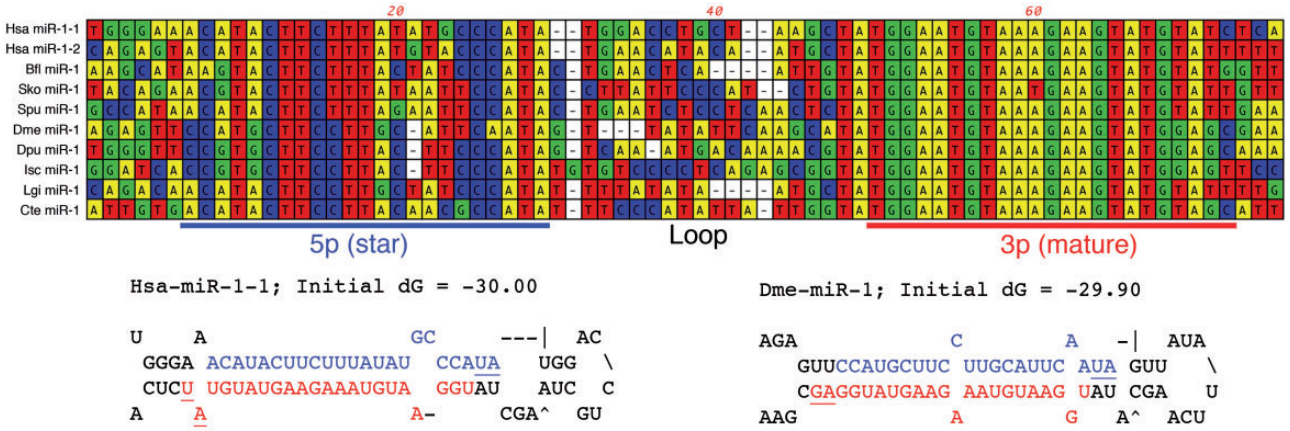

Fig. 1. Alignment and secondary structure of representative sequences of miR-1. The precursor sequences of miR-1 from *Homo sapiens* (Hsa), the lancelet *Branchiostoma floridae* (Bfl), the hemichordate *Saccoglossus kowalevskii* (Sko), the sea urchin *Strongylocentrotus purpuratus* (Spu), the fruit fly *Drosophila melanogaster* (Dme), the water flea *Daphnia pulex* (Dpu), the deer tick *Ixodes scapularis* (Isc), the limpet *Lottia gigantea* (Lgi), and the polychaete worm *Capitella teleta* (Cte), as well as the reported mature (3p) and star (5p) reads from *Drosophila melanogaster* (Dme), were aligned using the default settings of ClustalW (MacVector v.9.5.2). Note the conservation of the mature (red) and star (blue) regions of the sequence. Shown below the alignment are the secondary structures of two of these sequences, *H. sapiens* (left) and *D. melanogaster* (right).

as a long primary transcript (pri-miRNA) that then folds into a characteristic hairpin-like structure, which is recognized by the microprocessor–enzyme complex involving the proteins Drosha and Pasha (see Krol et al. 2010 and Starega-Roslan et al. 2011 for recent reviews). The microprocessor then cleaves the pri-miRNA into a ~70 nucleotide precursor miRNA (pre-miRNA) (fig. 1). This pre-miRNA is exported to the cytoplasm, where it is further processed by the RNAseIII enzyme Dicer to form a ~22 nucleotide RNA duplex with 2 nucleotide overhangs at each 3′-end. The duplex separates subsequently into two distinct strands, representing the 5′ and 3′ arms (fig. 1). While either product can be loaded into an Argonaute (AGO)-containing protein complex (Hui et al 2013), there is usually a preference for one arm, often termed the mature, which then regulates target mRNAs (reviewed by Huntzinger and Izaurralde 2011); the opposing arm is termed the star sequence.

miRNAs are named in sequential order of discovery, with identical or near-identical mature sequences in the same or different organisms given the same number (Ambros et al. 2003). miRNAs given different numbers have different primary sequences and are assumed to have evolved independently of other prior-named miRNAs, whereas those that are given the same number are either orthologues or paralogues within the same miRNA family, such as human miR-1, for which two paralogues exist (fig. 1). In some cases, misnumberings have occurred, resulting in the annotation of members of the same family with different numbers. For example, the mature sequences from the first four miRNA families named in *Drosophila* show that each of these families consists of unique mature sequences with respect to the other families and, thus, share little similarity in their primary sequences (fig. 2A). However, the two miR-2 sequences and the miR-13 sequence are evidently all members of the same miRNA gene family given that they share the same seed (nucleotide positions 2–8) and 3′ complementarity (positions 14–16) as miR-2 (Marco et al. 2012). Indeed, the seed and the 3′-complimentary motifs are the two most highly conserved

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

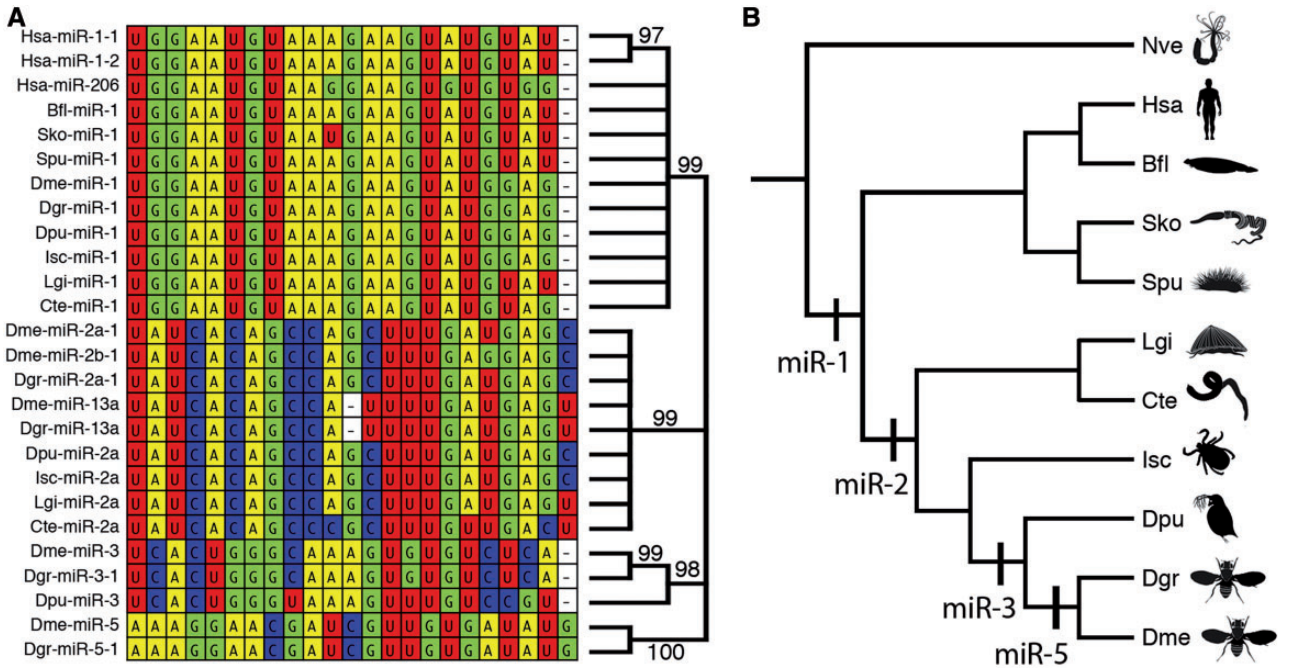

FIG. 2. Alignments of the mature sequences of the first four named miRNA families in *Drosophila* and their phylogenetic acquisition in metazoan evolution. (*A*) Representative examples of miR-1, miR-2, miR-3, and miR-5 aligned and analyzed phylogenetically (NJ tree using 1000 bootstrap replications; MacVector v.9.5.2). Note the clear monophyly of each named family, with the inclusion of paralogues from similarly named (e.g., miR-2 b-1) and differently named miRNA families (e.g., miR-13a) into their respective (in this case the miR-2) families. (*B*) Each of these four families characterize a particular node on the metazoan tree (see fig. 4) with no known subsequent losses: miR-1 is one of 31 families that characterize Bilateria (i.e., Protostomia + Deuterostomia); miR-2 is one of 12 families that characterizes Protostomia; miR-3 is one of 7 families that characterize Pancrustacea; and miR-5 is one of 10 families that characterizes the genus *Drosophila*. Nve, *Nematostella vectensis*; Dgr, *Drosophila grimshawi*; see figure 1 for other taxon abbreviations.

regions of mature sequences (Wheeler et al. 2009) because they are the most critical for target recognition (Grimson et al. 2007). The grouping of miRNAs into coherent families based on evolutionary descent is fundamental to their utility as phylogenetic markers, as each family represents the birth of a single gene independent of all other pre-existing miRNAs, similar to other types of genes such as transcription factor (*Hox*, *Fox*, *T-box*) or signaling ligand (*Wnt*, *hedgehog*, *TGF-beta*) families.

The Use of miRNAs as Phylogenetic Characters

miRNAs possess five properties that make them distinct from protein-coding genes, enhancing their utility as phylogenetic markers.

Processes of miRNA Biogenesis Enable the Identification of Novel miRNAs without Prior Knowledge of Sequence

Metazoan miRNAs are defined by their canonical secondary structures, which are the result of complementarity between the two arms within ~70 nucleotides of sequence (fig. 1 and supplementary fig. S1, Supplementary Material online). No specific sequence motif has been associated with miRNAs (Berezikov et al. 2006) aside from a propensity for the mature miRNA sequence to begin with a uracil (Lau et al. 2001). Rather, miRNAs are defined by their secondary structure, which require only ~16 bp of complementarity between the two arms, not any particular nucleotide sequence. Consequently, the cloning and sequencing of mature gene products does not require any information about the sequence itself either. This is exemplified in figure 3, which characterizes the pipeline of miRNA discovery programs such as miRMiner (Wheeler et al. 2009), miRDeep2 (Friedländer et al. 2011), mirDeep-P (Yang and Li 2011), miRtools (Zhu et al. 2010), MIREAP (Chen et al. 2009), and miRExpress (Wang et al. 2009). Of the four RNA sequences from a small RNA library made from the total RNA of a guinea pig, *Cavia porcellus*, one is the orthologous sequence of miR-1a (candidate 2), whereas the other three are unique sequences. Using a series of established criteria (see fig. 3), a second sequence appears to be a novel miRNA (candidate 3), whereas the other two are not miRNAs (candidates 1 and 4). Thus, miRNA discovery from deep sequencing does not rely on primary sequence data, but on a series of established criteria based on secondary structure and processing (Ambros et al. 2003; Meyers et al. 2008; Kozomara and Griffiths-Jones 2011; Tarver et al. 2012). Sequence similarity is only relevant for annotation when a new member of a known family is identified (whether it is an orthologue in a different species or a paralogue in the same species).

Continuous Addition of miRNA Families to Metazoan Genomes

Each of the first four miRNA families that were identified in *Drosophila* (fig. 2*B*) characterizes a different monophyletic

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

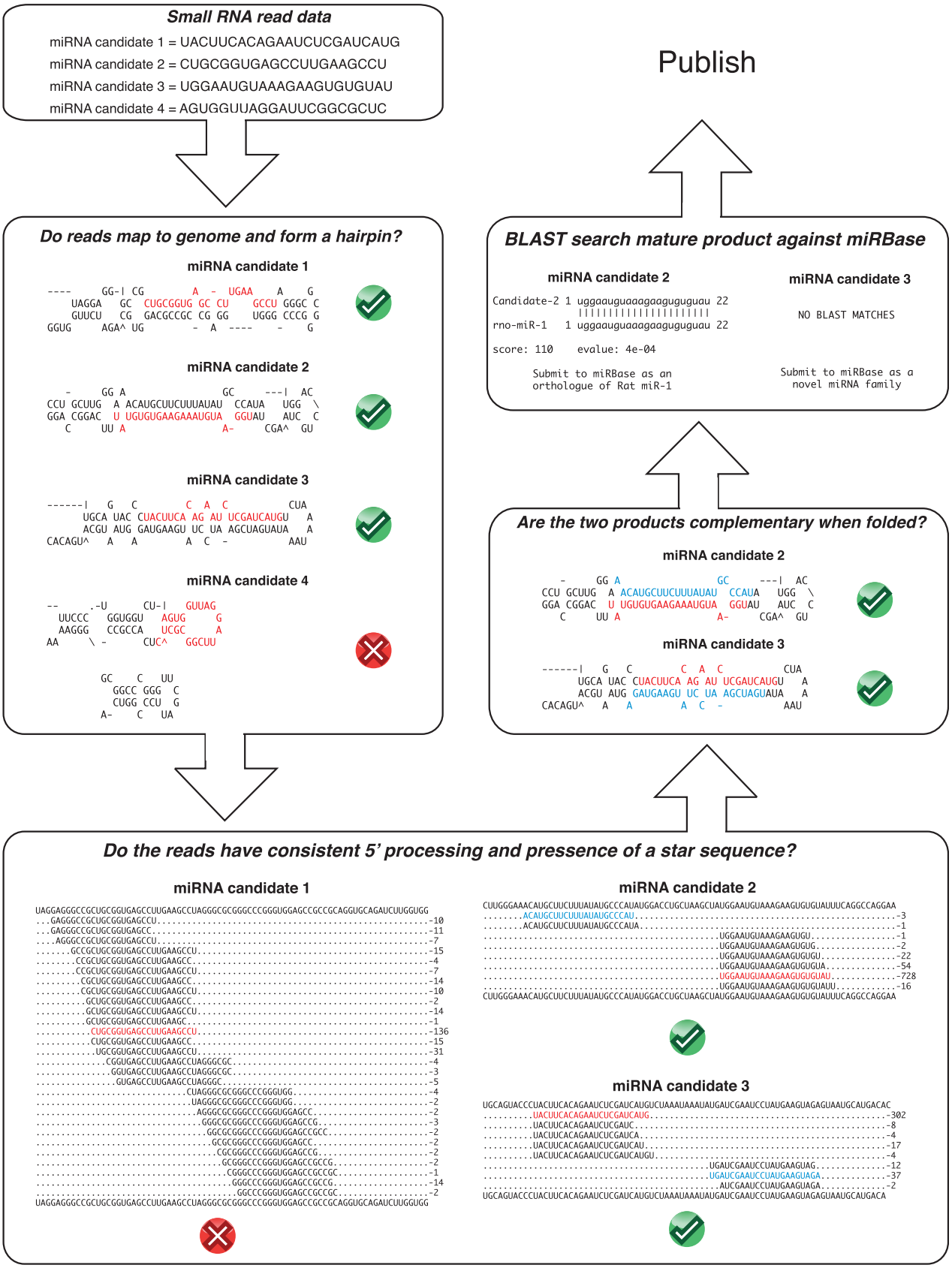

FIG. 3. An idealized protocol for identifying and validating miRNAs. Total RNA is extracted from the taxon of interest (here the guinea pig *Cavia porcellus*) and was sequenced using the Illumina Platform, generating thousands of candidate miRNAs, of which four are shown here. Reads are first mapped to the genome where an ~80 bp fragment from both upstream and downstream of the candidate miRNA is taken and folded using a program such as Mfold (Zuker 2003) to asses the folding properties of the candidate pre-miRNA using established criteria (Ambros et al. 2003; Tarver et al. 2012), which includes at least 16 bp complementary base pairing between the candidate miRNA and the opposing arm, and the absence of internal secondary

(continued)

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

group within Metazoa, proceeding in the order of their discovery, from Bilateria through Protostomia, Pancrustacea, and to *Drosophila*. The timing of their discovery was serendipitous, but it evidenced their gradual accrual through animal evolutionary history and, therefore, their potential as phylogenetic markers (Hertel et al. 2006; Sempere et al. 2006; Prochnik et al. 2007). Indeed, almost every clade investigated so far is characterized by the origin of at least one miRNA family (fig. 4). The utility of RGCs as phylogenetic markers relies on a rate of evolutionary origin, one that is quick enough to capture closely related speciation events. De novo noncoding RNA genes such as miRNAs appear to originate at a greater rate than de novo protein coding genes (Wu and Zhang 2013), thus providing greater opportunity for the use of miRNAs in phylogenetics. However, different evolutionary lineages exhibit evidence of vastly different rates of miRNA innovation and, thus, the phylogenetic utility of miRNAs may vary depending on the clade. For example, starting from the last common ancestor of Eumetazoa, 23 miRNA families were acquired in the lineage leading to the cnidarian *Nematostella vectensis*, in comparison to the 148 and 399 miRNA families acquired in the lineages leading to *Drosophila melanogaster* and *Homo sapiens*, respectively (fig. 4 and table 2).

Low Levels of Secondary Gene Loss

miRNAs, like all other phylogenetic markers, both molecular and morphological, exhibit evidence of loss, and if the rates of loss are high enough, any phylogenetic signal can be obscured. However, while all but one node in animal phylogeny presented in figure 4 is characterized by the gain of at least one miRNA family, the pattern of loss of these families is radically different. Seventeen nodes within the examined metazoan phylogeny (fig. 4) exhibit no loss of miRNA families, whereas overall there are 136 losses (red) compared with 1,047 gains (blue), a ratio of 1 loss for nearly 8 gains (table 2). Like gains, these losses are not evenly distributed across the phylogeny, with many metazoan taxa (e.g., polychaete annelids, gastropod molluscs, and cephalochordates) losing at most only a couple of miRNA families over their entire evolutionary history.

When losses do occur, they have been observed to follow one of two patterns. The typical pattern is to see losses that are mosaic in nature (Sperling and Peterson 2009) with no systematic trends within clades, although certain miRNAs tend to be lost more than others. Indeed, nearly one-third of the losses shown in figure 4 involve just eight miRNA families: miR-22 (six losses); miR-33 (five losses); miR-76 (six losses); miR-193 (five losses); 219 (seven losses); miR-315 (seven losses), miR-1993 (six losses), and miR-2001 (five losses).

The second pattern is of clade-specific losses that are usually associated with taxa including nematodes (Sperling and Peterson 2009) and acoels that appear to be secondarily miniaturized (Budd and Jensen 2000) and associated with the loss of primitive structures and cells types. Indeed, with respect to acoels, recent phylogenetic analyses identify them as deuterostomes (Philippe, Brinkmann, Copley, et al. 2011) and, along with *Xenoturbella*, a likely sister clade to Ambulacraria (i.e., echinoderms and hemichordates). As such, xenacoelomorphs should possess at least 32 miRNA families characteristic of a heritage shared with ambulacrarians. Although xenacoelomorphs demonstrate evidence of this shared heritage through the possession of the deuterostome miRNA family miR-103, they have otherwise lost many of those families that would have been primitively present as collectively they lost 10 miRNA families; the two acoel taxa *Symsagittifera* and *Hofstenia* lost an additional eight miRNA families, and even *Symsagittifera* has lost an additional seven families since splitting from *Hofstenia*. This large-scale loss of miRNA families cannot be rationalized by reinterpreting these organisms as the sister group of Bilateria (Philippe, Brinkmann, Copley, et al. 2011), but instead seems to be associated with considerable simplification of the xenacoelomorph bauplan (Erwin et al. 2011; Philippe, Brinkmann, Copley, et al. 2011).

A critical distinction that needs to be made when considering losses of miRNA families is the one between secondary absence and apparent absence as a consequence of incomplete genome and/or small RNA sequencing. The most effective way to characterize the miRNA repertoire of any taxon is to combine small RNA sequencing with genomic screening, as few genomes are completely sequenced, and not all miRNAs are expressed in all tissues at all times. However, in several cases miRNA repertoires have been reconstructed solely from genomic screening, but this can be problematic as an accurate assessment of the organism's conserved miRNA repertoire is dependent directly on the amount of genomic coverage—genomes sequenced at low coverage would be expected to be

Fig. 3. Continued

structure. Using such criteria, candidate 4 is rejected as a potential miRNA because it contains internal secondary structure—all other candidate miRNAs form acceptable hairpin loops and are still considered potential miRNAs at this point. Once the putative precursor structures have been identified, all reads generated from the deep sequencing libraries are mapped to the precursor sequences to assess the 5′ homogeneity of the candidate miRNAs. Although not originally identified as a criterion for miRNA annotation (Ambros et al. 2003), the current use of deep sequencing technologies allows it to be an invaluable tool to distinguish miRNAs from siRNAs and other types of noncoding RNAs (Berezikov et al. 2011; Tarver et al. 2012) as miRNAs have 5′-ends with high level of homogeneity, often with the same start nucleotide accounting for >90% of all reads, whereas siRNAs have no consistent 5′-end. Using this criterion, candidates 2 and 3 are potential miRNAs, but candidate 1 is rejected. Although not essential when identifying orthologous miRNAs a key criteria for establishing novel miRNAs is the presence of reads from both the 5′ and 3′ arms (often referred to as the mature and star sequences). Ideally such sequences should show a two-nucleotide offset, resulting from the sequential RNAse III processing. Both candidate miRNAs express both arms with the requisite offset, and are thus considered bona fide miRNAs and deposited in miRBase as Cpo-miR-1a (an orthologue of miR-1a) and Cpo-Novel-1, a previously unidentified miRNA as yet only known in the guinea pig, where miRBase will provide a formal name in due course.

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

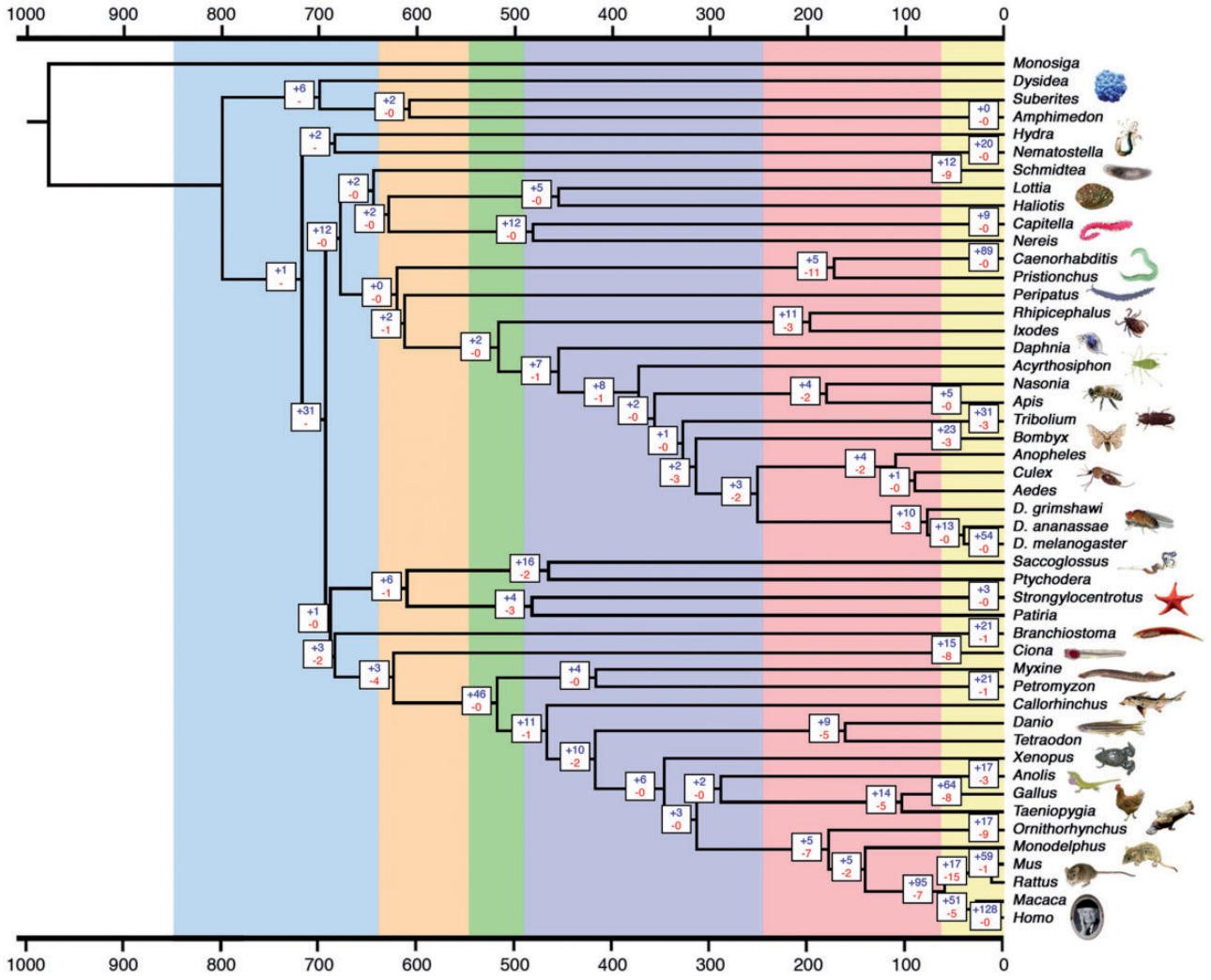

Fig. 4. The acquisition (blue) and loss (red) of miRNA families for 48 metazoan taxa. Note that almost every node is characterized by the addition of at least one new miRNA family, and that high rates of secondary loss are only associated with morphologically secondarily simplified taxa like nematodes and the flatworm *Schmidtea*. Further, there are four instances of a relatively high rate of miRNA family acquisition, once at the base of Bilateria, once at the base of the vertebrates, once at the base of Eutheria, and once at the base of the primates. See table 2 for the explicit gains and losses of miRNAs for each node.

missing many more miRNAs than highly sequenced genomes. Indeed, this is what we encountered when we compared 10 mammalian species representing three different levels of genome coverage to assess real versus perceived absence due to incomplete genome sampling: complete coverage (human); high (~6–7×) coverage (macaque, marmoset, dog, horse, and rabbit); and low (~1.5–2×) coverage (pika, kangaroo rat, cat, and sloth) (supplementary fig. S2, Supplementary Material online). When considering only the presence of miRNA families reconstructed as present in the last common ancestor of all crown eutherian mammals (to avoid the obfuscating gain and loss of lineage-specific miRNAs within Eutheria), only five miRNA families are missing in the human, representing a lineage-specific loss of 2.6%. These losses are probably real given the depth of genome coverage, the amount of deep sequencing studies that have been done in human, and the fact that these loci are also not found in the genomes of other primates including *Macaca* (fig. 4). The high-coverage genomes (approximately 7× coverage) had an average of 5.8 (3.0%) families missing, which is effectively comparable to the results from complete coverage. In contrast though, the low-coverage genomes (approximately 2× coverage) were missing an average of 37.25 (19.3%) families, the majority of which are likely false negatives. Such artefacts can, of course, be rectified, through further sequencing of the genome, combined with small RNA sequencing, keeping in mind though that the required depth of sequencing will vary with the sequencing strategy used—all taxa here were analyzed by large-scale consortia using a combination of platforms, including BAC cloning, and thus 6–7× coverage for these taxa will produce genomes that are more complete than if generated using 6× coverage from only an Illumina platform.

Thus, miRNAs, unlike retroposons, are not lost at an increasing rate with time, but are instead largely retained in most, and sometimes in all, descendant lineages (fig. 4), as might be expected of functional elements within gene regulatory networks (Sumazin et al. 2011). The evidential

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

Table 2. Evolutionary Acquisitions and Losses of miRNA Families.

Taxon	miRNA Family Gains[a]	miRNA Family Losses[b]
Demospongiae	6: 2014, 2015, 2016, 2017, 2019, 2021	?
Haplosclerida	2: 2018, 2020	0
Eumetazoa	1: (10, 51, 99, *100*, 125/lin4, 993, 1991, 2003, 3588[as])	?
Cnidaria	2: 2022, 2030	?
Bilateria	31: (*let7*, 48, 84, 98, 241, 3596, 4510, 5991), (*1*, 206, 3571[as]), (*7*, 3529 [as]), (*8*, 141, 200, 236, 429, 3548 [as], 3575[as]), (*4*, *9* = 79*, 244), (22, 745, 980, 3600[as]), (29, 83, 285, 746), (*31*, *72*), 33, (*34*, 449, 4933), 71, (*76*, 981), (25, 92, 235, 310, 311, 313), (96, 182, 183, 228, 263, 3553[as], 3983[as]), 124, (*133*, 3582[as]), (*137*, 234), (*153*, 2163), (*184*, 748), (50, *190*), (*193*, 365, 2788, 3549 [as], 4817, 5309), (*210*, 2164, 3286, 3574[as]), (*216*, 283, 304, 747, 3477), (*219*, 2964[as]), 242, 252, 278, (46, 47, *281*), (*315*, 1820), 375, 2001	?
Protostomia	12: (*Bantam*, 58, 81, 82), (*2*, 13, 43, 250), (*12*, 2157, 2158), (35, *36*, 37, 38, 39, 40, 41, 42), (*67*, 307), 87, (*277*, 4989), (44, 45, 61, 279, 286, 996, 2945), 317, 750, (958, *1175*), (*1993*, 60, 2162)	0
Lophotrochozoa	2: (*1989*, 2154), 1992	0
Neotrochozoa	2: 1990, 1994	0
Annelida	12: 1987, 1995, 1996, 1997, 1998, 1999, 2000, 2685, 2688, 2689, 2691, 2692	0
Mollusca	4: 1984, 1985, 1986, 2722	0
Gastropoda	1: 1988	0
Ecdysozoa	0	0
Nematoda	5: (*54*, 55, 56), (*63*, 2268), 86, 239, 240	11: 12, 22, 33, 153, 193, 210, 219, 277, 278, 317, 2001
Ascaris + Caenorhabditis	4: 49, 57, 791, 1822	0
Panarthropoda	2: 276, 305	1: 242
Arthropoda	2: 275, (*iab4*, iab8[as])	0
Chelicerata	1: 3931	1: 216
Ixodidae	10: 5305, 5306, 5307, 5308, 5310, 5311, 5312, 5313, 5314, 5315	2: 22, 31
Mandibulata	3: 282, 316, 965	0
Myriapoda	1: 3930	0
Pancrustacea	4: (3, 309, 318), (995, 998), 2765, (3791, 3478[as])	1: 2001
Pterygota	8: 14, 306, 927, 929, 971, 1000, 2796, 3049	1: 153
Endopterygota	2: 11, 932	0
Apocrita	4: 6001, 6038, 6039, 6067	2: 995, 1993
Coleoptera + Lepidoptera-Diptera	1: 970	0
Lepidoptera + Diptera	2: 274, 308	3: 36, 3049, 3791
Obtectomera	8: 2755, 2756, 2763, 2766, 2767, 2768, 3327, 3338	3: 219, 309, 315
Diptera	3: 957, 988, 999	2: 750, 1993
Culicidae	4: 1889, 1890, 1891, 2942	2: 274, 971
Culicinae	1: 2941	0
Drosophila	10: 5, 6, 955, 962, 969, 976, 987, 994, 1006, 1010	3: 71, 2765, 2796
D. melanogaster + *D. ananassae*	13: 312, 959, 960, 961, 964, 968, 974, 975, 978, 986, 1003, 1011, 1014	0
Deuterostomia	1: (*103*, 107, 2013)	0
Ambulacraria	6: 2006, 2007, 2008, 2009, 2011, 2012	1: 216
Echinodermata	3: 2002, 2004, 2010	3: 190, 281, 315
Eleutherozoa	1: 2005	0
Echinacea	3: 4847, 4850, 4854	0
Hemichordata	4: 4818, 4828, 4829, 4834	1: 22
Enteropneusta	12: 4819, 4820, 4821, 4824, 4825, 4826, 4830, 4831, 4835, 4838, 4839, 4841	1: 76
Chordata	3: 129, 135, 217	2: 76, 2001
Olfactores	3: (*15*, 16, 195, 322, 424, 457, 497), 126, 196	4: 71, 242, 252, 278

(continued)

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

Table 2. Continued

Taxon	miRNA Family Gains[a]	miRNA Family Losses[b]
Vertebrata	46: (17, 18, 20, 93, 106, 324), 19[c], 21, 23, (24, 3074[as]), 26, 27, 30, (122, 3591[as]), 128, (130, 301, 3590[as]), (132[d], 212), 138, 140, 142, 143, 144, 145, 146, 147, (148, 152, 2957[as]), 155[d], (181[d], 3570[as]) (192, 215) 194, (199[d], 3604) 202, 203, (204, 211), (205, 760), (208, 736, 3546 [as]), (214, 3120[as]), 218, (221, 222), (338, 3065[as]), (290, 291, 292, 293, 294, 295, 302, 371, 372, 373, 427, 430, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 1283), 451, 455, 456, (459, 802), 499, 551, 875, 1329, 1788, 4541	0
Cyclostomata	4: 4542, 4543, 4544, 4545	0
Petromyzontiformes	18: 4546, 4547, 4548, 4549, 4550, 4551, 4552, 4554, 4556, 4557, 4558, 4559, 4560, 4561, 4562, 4563, 4564, 4565	0
Gnathostomata	11: 32, 101[d], 139, 150, 191, (223, 599), (425, 731), 454, 1388, 2188, 3618	1: 315
Osteichthyes	10: 187, 363, 458, (460, 730), 489, 726, 727, 737, 1306, 2184	2: 281, 4541
Teleostei	9: 462, 722, 723, 724, 725, 728, 733, 734, 2187	5: 32, 191, 551, 875, 1329
Tetrapoda	6: 367[e], 383, 1662, 1805, 2970, 3064	0
Amniota	3: 490, 1397, 1416	0
Reptilia	2: 1641, (1620, 1677, 1784, 1803)	0
Archosauria	3: 1720, 1791, 2984	0
Aves	11: 1451, 1467, 1550, 1552, 1559, 1655, 1729, 1781, 1782, 2131, 2954	5: 208, 459, 727, 737, 875
Mammalia	5: 186, 325, 590, 671, 873	7: 456, 726, 727, 737, 1662, 2184, 2188
Theria	5: 340, 885, 1251, 3613, 3661	2: 458, 1416
Marsupialia	6: 1540, 1542, 1546, 1547, 1548, 1549	2: 459, 489
Eutheria	93: (28, 151, 708)[f], (95, 421, 545, 1264), 105, 127, 134, (136, 3071[as]), 149, (154, 323, 369, 376, 377, 381, 382, 410, 453, 487, 494, 496, 539, 655, 656, 1185, 3576[as], 3578[as], 3581[as], 3595[as], 3958), 185, (188, 532, 660), 197, 224, 296, 299, 320, 326, 328, (329, 495, 543), 330, 331, 335, (337, 3544[as]), 339, 342, 345, 346, 350, 361, (362, 500, 501, 502, 3560[as]), 370, 374, (378, 422, 3557[as]), (379, 380, 411, 758, 1197, 3959, 3579[as]), 384, 423, 431, 432, 433, 448, 450, 452, 483, 485, (486, 3107[as]), 488, 491, 493, 503, 504, (505, 3589[as]), (201, 463, 465, 470, 471, 506, 507, 508, 509, 510, 513, 514, 547, 741, 742, 743, 871, 878, 880, 881, 883, 888, 890, 892, 3551, 3585), 541, (542, 3601[as]), 544, 574, 582, 592, 615, 628, 652, 653, 654, 670, 672[d], 675, 676, 744, 767, 769, 872, 874, 876, 889, 1193, 1247, 1249, 1271, 1296, 1298, 1301, 1307, 1343[g], 1468, 1839, 1842, 1912, 2114, 2355, 2387, 2483, 3059, 3085, 3106	6: 460, 1329, 1397, 1788, 1805, 2970
Boreoeutheria	2: 511, 1911	
Euarchontoglires	0	3: 2483
Muridae	17: 298[h], 344, 351, 434, 540, 667, 673, 674, 879, 1188, 3072, 3075, 3099, 3109, 3112, 3572, 5103	15: 197, 432, 454, 769, 885, 889, 1296, 1307, 1343, 1388, 1468, 1842, 2114, 2355, 2387
Simiiformes	44: 512, 550, 552, 557, 562, 576, 577, 580, 581, 584[i], 586, 587, 589[j], 600, 601, 605, 609, 612, 616, 618, 642, 887, 891, 934, 937, 939, 940, 942, 944, 1180, 1182, 1230, 1253, 1256, 1262, 1269, 1278, 1293, 1323, 1915, 2117, 3672, 3937, 4423	3: 1388, 1842, 3106
Catarrhini	7: 625, 627, 1245, 3927, 4446, 4667, 4768, 4803	2: 672, 872

[a]Families are designated parenthetically and are underlined; family names are given in italics. In some cases, the same gene is given at least two different names (e.g., miR-22 = miR-745 = miR-980), whereas in other cases there were gene duplications generating at least two copies of the gene in an individual taxon's genome (e.g., miR-10 family, miR-252 family, miR-96 family). Families were derived from miRBase v. 19. Every entry was checked for validity using standard criteria (e.g., Tarver et al. 2012) and only those showing positive evidence for miRNA processing and expression were counted as valid.

[b]Question marks indicate that it is not possible at the moment to reconstruct losses for this node.

[c]miRBase entry for *Branchiostoma* is bogus.

[d]miRBase entry for *Ciona* is bogus.

[e]miRBase entry for *Ciona* is a real miRNA, but not the named miRNA.

[f]miRBase entry for *Monodelphis* is bogus.

[g]miRBase entry for *Ornithorhynchus* is bogus.

[h]The rodent members are valid and have been confirmed with small RNA sequencing, but the primate ones are based on low similarity to the rodents and have never been found in a small RNA library.

[i]miRBase entry for *Bos* is bogus.

[j]miRBase entry for *Canis* is bogus.

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

correlation between widespread secondary absence of ancestral miRNAs and organismal simplification underlines this expectation. Hence, concerns about large-scale homoplasy caused by secondary absence (loss) appear to be unfounded, with studies showing higher rates of loss in taxa such as sloth, cat, and pika (Guerra-Assunção and Enright 2012), reflecting incomplete genome sequencing rather than large-scale secondary loss.

Rarity of Substitutions to the Mature miRNA Sequence

In addition to this order of magnitude difference between miRNA gains versus losses, miRNAs are some of the most conserved genetic elements in the genome (Sempere et al. 2006). Wheeler et al. (2009) analyzed 16,525 nucleotides from the mature miRNA sequences from 14 metazoan taxa and showed that the substitution rate of all miRNAs across these 14 taxa, whose lineages represent over 7800 million years of independent evolutionary history, is only 3.4% (567 total substitutions). In comparison 18S rDNA, one of the most conserved genes in the metazoan genome, has a substitution rate of 7.3% even when the unalignable regions are ignored (Wheeler et al. 2009). Hence, the primary nucleotide sequence of the mature miRNA gene product evolves more than twice as slowly as the most conserved positions of a gene that has long been used to reconstruct the deepest nodes in the tree of life.

The result of such slow sequence divergence means that the mature miRNA product can be identified easily when analyzing small RNA sequence data or when BLAST searching against an animal's genome. Even in taxa with high levels of loss, the remaining miRNAs still show relatively high levels of sequence conservation, allowing for the remaining miRNA repertoire to be discovered easily in both a small RNA library and genomic sequence.

Small Probability of the Independent Evolution of the Same miRNA

Rokas and Holland (2000) argued that RGCs have the distinct advantage over other types of molecular data in that independent RGCs can be distinguished easily from one another, which reduces dramatically errors caused by misinterpretations of homology. Although it has been argued previously that the statistical chance of two miRNAs evolving convergently is exceedingly small (Sperling and Peterson 2009), this remains a major concern, given that the mature gene product (~22 nt in length) is so short. However, there is more to a miRNA than its mature sequence. The entire pre-miRNA must also fold into a hairpin with no large, and in particular asymmetrical, internal loops or bulges (see fig. 3, candidate 4). It must have a free energy value lower than approximately −19 kcal/mole, the result of the complementarity between the 5′ and 3′ arms, and this complementarity must occur within about 70–100 nucleotides of primary DNA sequence (figs. 1–3). In addition, the source arm of the mature miRNA sequence (i.e., whether it is on the 5′ arm or the 3′ arm) must be conserved, and any predicted mature gene must be processed, generating the expected read data from a small RNA library. All of these factors are standard annotating criteria for the identification of miRNAs (Ambros et al. 2003; Kozomara and Griffiths-Jones 2011; Tarver et al. 2012).

Exceptions to such rules have been suggested. For example, Li et al. (2010) took 3,861 miRNA sequences downloaded from miRBase (v. 13.0) and searched against the repeat-masked genome sequences of 56 animal species, purportedly identifying 300 miRNAs in taxa not known (or predicted) to possess these miRNAs. For instance, miR-430 is known currently only from vertebrates, but Li et al. (2010) reported a miR-430 orthologue in the mosquito *Anopheles* (but not in any other non-vertebrate animal including other dipterans) (fig. 5, "a"). Similarly, these same authors reported the protostome-specific miR-317 family in *Macaca* but not in any other deuterostome including human (fig. 5, "b"). These results might suggest much higher levels of homoplasy than have been observed previously, either through convergent evolution and/or horizontal gene transfer (HGT) and which if substantiated would nullify the use of miRNAs in phylogenetics.

To investigate this we reanalyzed each of their reported occurrences in 27 different animal species (for a total of over 28 billion sequenced nucleotides) using the criteria established by Ambros et al. (2003; see also Tarver et al. 2012) for identifying miRNAs. We found that every phylogenetically discordant claim was a false-positive result (fig. 5; data available upon request), arising because the putative miRNA precursor structure had 1) a putative mature sequence located on the wrong arm (fig. 5; green), and/or 2) a structure not meeting minimal free energy values (fig. 5; blue), and/or 3) internal secondary structure (fig. 5; red), and/or 4) a putative mature sequence that was not expressed in small RNA libraries (fig. 5; orange). Thus, despite the comprehensive nature of the bioinformatic survey of miRNAs in which Li et al. (2010) identified putative candidates through genomic homology searches, not a single instance of convergent evolution could be documented despite searching a total of more than 28 billion sequenced nucleotides over the taxonomic breadth of bilaterian evolution.

miRNAs and Phylogenetics

Because miRNA families have an evolutionary origin that is independent of one another, they should be treated as discrete characters within phylogenetic analyses. Such presence/absence data can be coded in a manner similar to morphological or other genome-content data (Rivera and Lake 2004) and analyzed phylogentically using the standard models of discrete character evolution. Further, given the pattern of miRNA families gains and losses that emerge from the observation of the distribution of the known miRNAs across the Metazoa, it follows that miRNA data seem to evolve under Dollo's law.

Although pairwise distance measures could be used, we chose to focus on parsimony and Bayesian phylogenetic analyses using Dollo parsimony and the stochastic Dollo model (Alekseyenko et al. 2008), respectively, using a data set comprising 29 taxa and 565 characters (343 parsimony

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

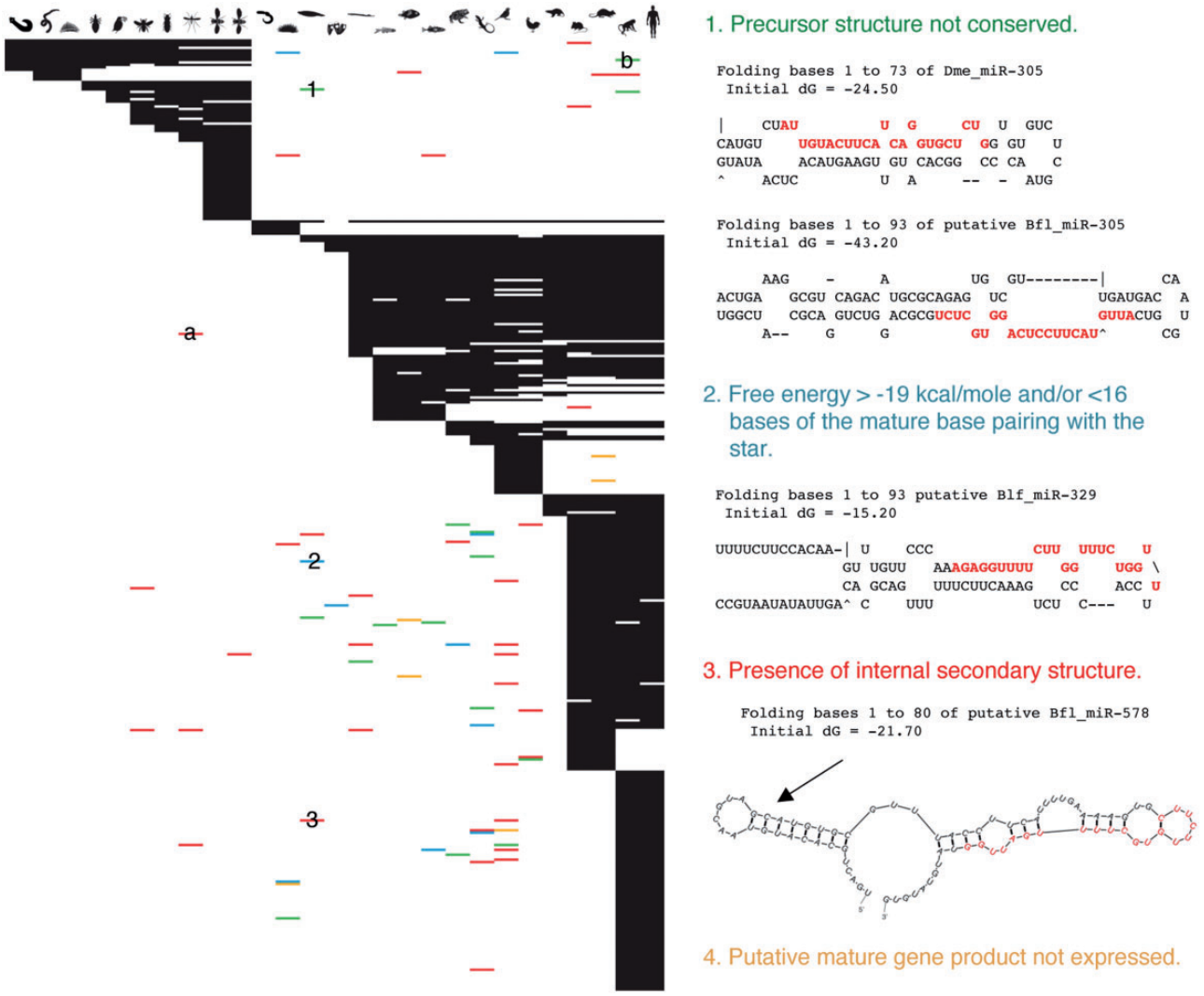

Fig. 5. The hierarchical evolution of metazoan miRNAs with no known incidences of convergent evolution. Shown across the top are 27 animal species that have had their miRNA complements ascertained with genomic screens and small RNA libraries. Shown in black are the known occurrences of 389 miRNA families in these 27 species. Note the strong hierarchical clustering, consistent with the documented pattern of continued acquisition with little secondary loss (fig. 4). Shown in color are the purported occurrences of these miRNAs in these taxa according to Li et al. (2010) who used a bioinformatic approach to identify miRNAs in animals with sequenced genomes (available on request). On closer inspection though none of these represent bona fide miRNAs as they fail to meet established standards for miRNA structure and expression because they 1) have the mature sequence on the wrong arm (green); 2) do not meet minimum free energy values for precursor structure and/or have <16 nucleotides base pairing between the mature (shown in red) and the star (blue); 3) have internal secondary structure in the precursor (red; see arrow); and/or 4) are not expressed in small RNA preparations (orange). Note that most of these putative miRNAs fail multiple criteria (and indeed none are known to be expressed in the purported taxon where deep sequencing has been done). The two purported occurrences labeled "a" and "b" are the purported occurrence of the vertebrate-specific miR-430 in the mosquito *Anopheles* and protostome-specific miR-317 in the primate *Macaca*, respectively (see text for details).

informative, supplementary file S3, Supplementary Material online). Parsimony analyses were performed using PAUP* 4.0 (using the Dollo up criterion) with all characters being unordered and having equal weight. The analyses were performed using the branch and bound algorithm, and the results obtained are thus exact rather than heuristic (fig. 6*A*). For the sake of convention, support was estimated using the bootstrap (1,000 repetitions), although this is not the optimal mode of estimating support as support for individual nodes would not be expected to be spread homogenously though the data set. The resulting phylogeny is congruent with contemporary phylogenies based on traditional molecular data, although some nodes have low bootstrap support values. The Bayesian analysis was performed using BEAST version 1.6.2 (Drummond and Rambaut 2007). Two independent chains of 100 million generations were run sampling every 100 generations. Convergence was tested using Tracer and a majority rule consensus tree was built after excluding trees sampled before convergence (i.e., during the burnin period). Support for the nodes in the tree represents posterior probabilities. The resulting tree (fig. 6*B*) is fully congruent with that obtained from the Dollo parsimony analysis, but support values are higher with posterior probabilities of 1 on all but one of the nodes. The key difference between the two trees

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

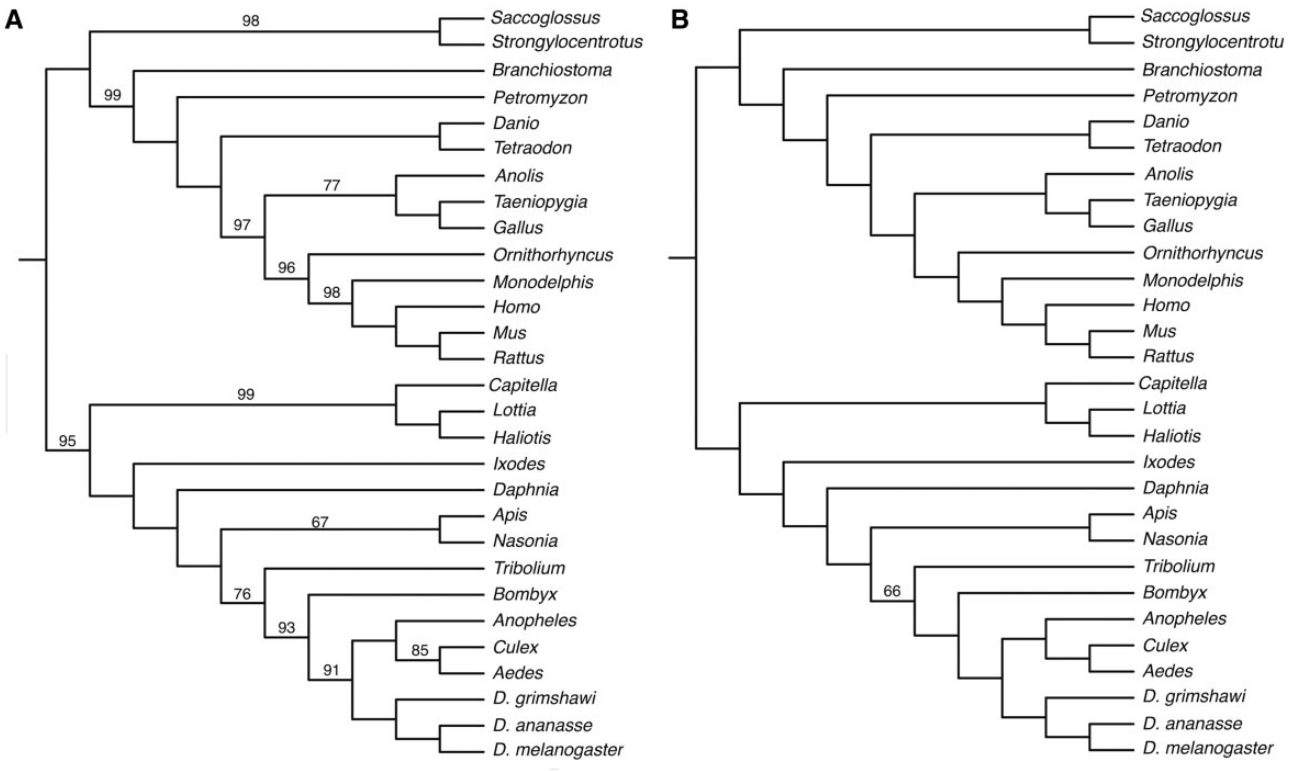

FIG. 6. The maximum parsimony tree generated using PAUP (*A*) and the Bayesian tree generated using BEAST (*B*) for the miRNA data set (supplementary file S3, Supplementary Material online). The two trees are in full agreement with one another, although support values differ between the two methods with the Bayesian tree appearing to have greater levels of support with only one node having a posterior probability <1. This discrepancy occurs due to the way support values are calculated: when bootstrapping the dataset individual characters are resampled and so nodes that are supported by only a few characters will have low bootstrap support values. However, because there is very little homoplasy in the data set, the same phylogeny will almost always be returned even for nodes supported by relatively few characters, and hence the 100% posterior probability values in the Bayesian analysis.

is the level of support with the parsimony tree exhibiting far lower levels of (bootstrap) support. Nodes within the Bayesian tree such as the *Apis*/*Nasonia* split have a posterior probability of 1, whereas in the parsimony tree it had a bootstrap support value of 68%. Even though posterior probabilities are expected to exceed the corresponding bootstrap values, this discrepancy also occurs because nodes in the phylogeny are characterized by variable numbers of acquired miRNAs. Nodes supported by few miRNAs are less likely to have high bootstrap support values, because the characters supporting them will be less likely to be present in the resampled data sets. This is not expected to be a problem for standard sequence data where the signal for each node is expected to be abundant (if the data are adequate to answer the question at hand) and quite homogeneously distributed across the sites. Yet, miRNAs are rare genomic changes, which, by definition, are not expected to be abundant nor homogeneously distributed. Accordingly, we suggest that bootstrap might not be an ideal support measure for miRNA data sets. Indeed, alternative statistical measures to support the significance provided by LINEs and SINEs were introduced by Waddell et al. (2001) and here we suggest that Bayesian posterior probabilities, which are not based on character resampling, should be preferred to measure the support provided by miRNA data sets.

Future Directions

To date, the use of miRNAs in phylogenetics has been limited to the use of presence/absence data of individual miRNA families. However, with changes in both small RNA and genomic sequencing platforms, there are opportunities to expand and refine the use of miRNAs in phylogenetics. Two principal areas for future research include 1) the use of individual miRNAs rather than families and 2) the development of model-based approaches to miRNA phylogenetics.

Use of Individual miRNA Genes

The low rate of loss of miRNA families in comparison to individual miRNA genes has led to their preferential use in phylogenetics due to concerns about homoplasy. However, next-generation sequencing has led to increases in sequencing read counts from the hundreds of thousands to the hundreds of millions, allowing the identification of individual lowly expressed miRNA genes, which would have been missed using older sequencing technologies. Furthermore, the plummeting cost of genome sequencing means that it has become economically viable to sequence the genome in conjunction with the small RNA read data, allowing the identification of individual paralogues.

However, before individual miRNA genes can be used, the classification of miRNA genes into families must be refined.

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

Individual families should represent fundamental units of innovation in miRNA evolution—any miRNAs that share ancestry should be classified within the same family, regardless of whether they originated as a consequence of whole genome, chromosomal, segmental, or merely tandem duplication, and irrespective of any particular difference in nucleotide sequence, including changes to the seed sequence. This will require revision of existing miRNA annotation. Specific algorithms are being developed for the identification and grouping of individual miRNAs into distinct families (Huang and Gu 2007), and these should be refined and implemented to revise the taxonomy and ontology so that gene orthology is clear.

In conjunction with the presence/absence of families, previous studies have employed individual point mutations on the mature miRNA products as further evidence of phylogenetic relationships (Heimberg et al. 2010). This has led to the suggestion that the individual sequences of miRNA genes themselves could be used in phylogenetic analysis. Although accurate phylogenies can be resolved using small numbers of taxa in such a manner, the small number of phylogenetically informative sites (~70 nt per miRNA) means that a large number of taxa will saturate the data and remove any phylogenetic signal. Thus, sequence-based analyses are most applicable to the identification of gene (or whole genome) duplication events and may prove particularly informative in polyploid taxa such as plants. Alternatively, the number of phylogenetically informative sites could be increased by concatenating individual miRNAs into larger alignments. However, care would be needed in correctly identifying paralogy groups, again highlighting the need for a coherent scheme of ontology for miRNAs. Future increases in the number of annotated genomes with high-quality synteny maps will facilitate more accurate identification of gene orthology than simple homology searches, making a concatenation approach more feasible. Syntenic information will also provide critical information for the classification of both miRNA families and individual paralogous genes.

Developing New Phylogenetic Models

Currently, miRNA data are analyzed under a Dollo model using either parsimony or Bayesian approaches (fig. 6), but there is scope for alternate models to be developed. Differential weighting strategies could be used to reflect the varying likelihood of gene loss, such that a loss of an entire miRNA family is considered more unlikely than the loss of individual miRNA genes. Furthermore, some miRNAs exhibit empirical evidence of a greater likelihood of evolutionary loss, such as miR-315, which could also be incorporated into the model. Such a modelling approach could also be applied to the individual sequence data in a manner that would accommodate different likelihoods of substitution rate in the different structural elements of a presequence. Thus, a weighting strategy which ranks the mature > star > loop (see fig. 1) could be used, reflecting the differential likelihood of substitutions within the different regions of the pre-miRNA.

Conclusions

miRNAs have the potential to become an invaluable resource for phylogenetic analyses, especially when used as an additional line of evidence, separate from either protein coding genes or morphological data sets. They have been utilized to resolve previously intractable phylogenetic debates within the tree of life, whether at the species or the phyla level (table 1). Every phylogenetic method and data set exhibits some level of homoplasy, and miRNAs are no exception. However, the level of homoplasy observed in miRNA data seems lower than many other class of phylogenetic data, with few losses and no known instances of convergence between miRNAs.

Furthermore, miRNAs are an ideal complement to data sets comprising protein-coding genes, entire genomes, or morphological characters, allowing researchers to use congruence (Miyamoto and Fitch 1995; Pisani et al. 2007; Leigh et al. 2008), between different data sets, which can now include miRNAs, to assess support for competing topologies (Heimberg et al. 2010; Campbell et al. 2011; Wiegmann et al. 2011). The decreasing cost of sequencing means that miRNA library construction and sequencing are becoming more viable for smaller grants, whereas the minimal computational power means that the phylogenetic analyses can be done quickly and efficiently without the prolonged use of supercomputers, which is necessary for phylogenomic-scale analyses.

Supplementary Material

Supplementary figures S1 and S2 and file S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank the editor and three anonymous reviewers for thoughtful comments on an earlier draft of this manuscript. We acknowledge funding from the Biotechnology and Biological Sciences Research Council, the Irish Research Council, Marie Curie actions of EU FP7, the National Aeronautics and Space Administration, and the National Science Foundation.

References

Alekseyenko AV, Lee CJ, Suchard MA. 2008. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Syst Biol.* 57:772–784.

Ambros V, Bartel B, Bartel DP, et al. (13 co-authors). 2003. A uniform system for microRNA annotation. *RNA* 9:277–279.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233.

Berezikov E. 2011. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet.* 12:846–860.

Berezikov E, Cuppen E, Plasterk RHA. 2006. Approaches to microRNA discovery. *Nat Genet.* 38:S2–S7.

Berezikov E, Robine N, Samsonova A, et al. (14 co-authors). 2011. Deep annotation of Drosophila melanogaster MicroRNAs yields insights into their processing, modification, and emergence. *Genome Res.* 21: 203–215.

Budd GE, Jensen S. 2000. A critical reappraisal of the fossil record of the bilaterian phyla. *Biol Rev Camb Philos Soc.* 75:253–295.

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

Cai Q, Zhang X. 2010. MiRNAs as promising phylogenetic markers for inferring deep metazoan phylogeny and in support of Olfactores hypothesis. 2010 IEEE International Conference on Bioinformatics and Biomedicine. p. 101–104.
Campbell LI, Rota-Stabellli O, Marchioro T, Longhorn SJ, Edgecombe GD, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D. 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest the velvet worms are the sister group of Arthropoda. *Proc Natl Acad Sci U S A.* 108:15920–15924.
Chen X, Li Q, Wang J, et al. (16 co-authors). 2009. Identification and characterization of novel amphioxus microRNAs by Solexa sequencing. *Genome Biol.* 10:R78.
Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.
Erwin DH, LaFlamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ. 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* 334: 1091–1097.
Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2011. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 40:37–52.
Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell.* 27:91–105.
Guerra-Assunção JA, Enright AJ. 2012. Large-scale analysis of microRNA evolution. *BMC Genomics* 13:218.
Hallström BM, Janke A. 2010. Mammalian evolution may not be strictly bifurcating. *Mol Biol Evol.* 27:2804–2816.
Heimberg AM, Cowper-Sal-lari R, Sémon M, Donoghue PCJ, Peterson KJ. 2010. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci U S A.* 107:19379–19383.
Hejnol A, Obst M, Stamatakis A, et al. (17 co-authors). 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc Lond B Biol Sci.* 276:4261–4270.
Helm C, Bernhart SH, Siederdissen CHZ, Nickel B, Bleidorn C. 2012. Deep sequencing of small RNAs confirms an annelid affinity of Myzostomida. *Mol Phylogenet Evol.* 64:198–203.
Hertel J, Lendemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* 7:25.
Huang Y, Gu X. 2007. A bootstrap based analysis pipeline for efficient classification of phylogenetically related animal miRNAs. *BMC Genomics* 8:66.
Hui JHL, Marco A, Hunt S, Melling J, Griffiths-Jones S, Ronshaugen M. 2013. Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods. *Nucleic Acids Res.* 41:3352–3361.
Huntzinger E, Izaurralde E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet.* 12:99–110.
Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39: D152–D157.
Krol J, Loedige I, Filipowicz W. 2010. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet.* 11: 597–610.
Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294:858–862.
Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. *Syst Biol.* 57:104–115.
Li S-C, Chan W-C, Hu L-Y, Lai C-H, Hsu C-N, Lin W-c. 2010. Identification of homologous microRNAs in 56 animal genomes. *Genomics* 96:1–9.
Lyson TR, Sperling EA, Heimberg AM, Gauthier JA, King B, Peterson KJ. 2012. microRNAs support a Testudines-Lepidosaur clade. *Biol Lett.* 8: 104–107.
Marco A, Hooks KB, Griffiths-Jones S. 2012. Evolution and function of the extended miR-2 microRNA family. *RNA Biol.* 9:1–7.
Meyers BC, Axtell MJ, Bartel B, et al. (21 co-authors). 2008. Criteria for Annotation of Plant MicroRNAs. *Plant Cell* 20:3186–3190.
Miyamoto MM, Fitch WM. 1995. Testing the covarion hypothesis of molecular evolution. *Mol Biol Evol.* 12:503–513.
Peterson KJ, Su Y-H, Arnone MI, Swalla B, King BL. 2013. MicroRNAs support the monophyly of enteropneust hemichordates. *J Exp Zool B Mol Dev Evol.* 9999:1–7.
Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* 470:255–258.
Philippe H, Brinkmann H, Lavrov D, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9:e1000602.
Pisani D, Benton MJ, Wilkinson M. 2007. The congruence of molecular and morphological phylogenies. *Acta Biotheor.* 55:269–281.
Pisani D, Feuda R, Peterson KJ, Smith AB. 2012. Resolving phylogenetic signal from noise when divergence is rapid: a new approach to the old problem of echinoderm class relationships. *Mol Phylogenet Evol.* 62:27–34.
Prochnik SE, Rokhsar D, Aboobaker AA. 2007. Evidence for a microRNA expansion in the bilaterian ancestor. *Dev Genes Evol.* 217:73–77.
Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152–155.
Rokas A, Holland PWH. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol.* 15:454–459.
Rota-Stabellli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford M. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc R Soc Lond B Biol Sci.* 278:298–306.
Sempere LF, Cole CN, McPeek MA, Peterson KJ. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol.* 306B:575–588.
Sperling EA, Peterson KJ. 2009. microRNAs and metazoan phylogeny: big trees from little genes. In: Telford MJ, Littlewood DTJ, editors. Animal evolution—genomes, trees and fossils. Oxford: Oxford University Press. p. 157–170.
Sperling EA, Pisani D, Peterson KJ. 2011. Molecular paleobiological insights into the origin of the Brachiopoda. *Evol Dev.* 13:290–313.
Sperling EA, Robinson JM, Pisani D, Peterson KJ. 2010. Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200-Myr missing Precambrian fossil record of siliceous-sponge spicules. *Geobiology* 8:24–36.
Sperling EA, Vinther J, Moy VN, Wheeler BM, Sémon M, Briggs DEG, Peterson KJ. 2009. MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. *Proc R Soc Lond B Biol Sci.* 276:4315–4322.
Starega-Roslan J, Koscianska E, Kozlowski P, Kryzyzosiak WJ. 2011. The role of the precursor structure in the biogenesis of microRNA. *Cell Mol Life Sci.* 68:2859–2871.
Sumazin P, Yang X, Chiu H-S, et al. (11 co-authors). 2011. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 147: 370–381.
Tarver JE, Donoghue PCJ, Peterson KJ. 2012. Do miRNAs have a deep evolutionary history? *Bioessays* 34:857–866.
Telford MJ, Copley RR. 2011. Improving animal phylogenies with genomic data. *Trends Genet.* 27:186–195.
Waddell PJ, Kishino H, Ota R. 2001. A phylogenetic foundation for comparative mammalian genomics. *Genome Inform.* 12:141–154.
Wang W-C, Lin F-M, Chang W-C, Lin K-Y, Huang H-D, Lin N-S. 2009. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* 10:328.
Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ. 2009. The deep evolution of metazoan microRNAs. *Evol Dev.* 11:50–68.
Wiegmann BM, Trautwein MD, Winkler IS, et al. (26 co-authors). 2011. Episodic radiations in the fly tree of life. *Proc Natl Acad Sci U S A.* 108:5690–5695.

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014

Wu DD, Zhang YP. 2013. Evolution and function of de novo originated genes. *Mol Phylogenet Evol.* 67:541–545.
Yang XZ, Li L. 2011. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 27: 2614–2615.
Zhu E, Zhao F, Xu G, Hou H, Zhou L, Li X, Sun Z, Wu J. 2010. mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.* 38:W392–W397.
Zucker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.

Downloaded from http://mbe.oxfordjournals.org/ at National Library of Health Sciences on May 5, 2014