Estimating Phylogeny from microRNA Data: A Critical Appraisal

Short Title: Phylogenies from miRNA Data

Robert C. Thomson[1], David C. Plachetzki[2], D. Luke Mahler[3], and Brian R. Moore[3]

[1]*Department of Biology, University of Hawai'i at Mānoa, Honolulu, HI 96822, USA*
[2] *Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH 03824, USA*
[3]*Department of Evolution and Ecology, University of California, Davis CA 95616 USA*

Robert C. Thomson
Department of Biology
2538 McCarthy Mall, Edmondson Hall Rm. 216
University of Hawaii, Manoa
Honolulu, HI 96822
U.S.A.

Phone: (808) 956-6476
E-mail: thomsonr@hawaii.edu

Abstract

As progress toward a highly resolved tree of life continues to expose nodes that resist resolution, interest in new sources of phylogenetic information that are informative for these most difficult relationships continues to increase. One such potential source of information, the presence and absence of microRNA families, has been vigorously promoted as an ideal phylogenetic marker and has been recently deployed to resolve several long-standing phylogenetic questions. Understanding the utility of such markers for phylogenetic inference hinges on developing a better understanding for how such markers behave under suitable evolutionary models, as well as how they perform in real inference scenarios. However, as yet, no study has rigorously characterized the statistical behavior or utility of these markers. Here we examine the behavior and performance of microRNA presence/absence data under a variety of evolutionary models and re-examine datasets from several previous studies. We find that highly heterogeneous rates of microRNA gain and loss, pervasive secondary loss, and sampling error collectively render microRNA-based inference of phylogeny difficult, and fundamentally alter the conclusions for four of the five studies that we re-examine. Our results indicate that miRNA data have far less phylogenetic utility in resolving the tree of life than is currently recognized and we urge ample caution in their interpretation.

Keywords: miRNA, homoplasy, secondary loss, stochastic Dollo

Significance Statement

As progress toward a highly resolved Tree of Life continues, evolutionary relationships that defy resolution despite ongoing methodological improvements continue to be identified. Recently, the presence and absence of microRNA families have emerged as potentially ideal sources of information for these difficult phylogenetic problems, and have since been employed to resolve several long-standing problems in the metazoan tree of life. This study performs the first rigorous statistical assessment of the use of microRNAs for phylogenetic estimation and finds that a high incidence of homoplasy and sampling error render phylogenies based on microRNA data highly biased or uncertain. This study casts serious doubt on the central phylogenetic conclusions reached in several previous analyses of microRNA datasets.

As genomic tools and affordable DNA sequencing have become widely available, our ability to leverage molecular sequence data to estimate species phylogeny has rapidly increased. The flood of molecular data has, in turn, witnessed brisk progress in resolving the tree of life [1, 2]. Nevertheless, many relationships have resisted resolution despite repeated efforts using increasing amounts of sequence data. These challenging cases have motivated the search for new sources of (molecular) phylogenetic information, which places precedence on data that evolve by rare and nearly irreversible genomic changes. Patterns of gene rearrangement, duplication, insertion and deletion, as well as positional information for retrotransposons have all been promoted as candidate data with "ideal" phylogenetic properties (e.g., [3–6]). Although new types of phylogenetic data may hold promise in resolving difficult nodes in the tree of life, they require careful consideration in order to appropriately model the underlying evolutionary process by which they arose and to accommodate possible sampling biases associated with their collection.

One recently promoted class of putatively ideal phylogenetic data is the presence/absence of microRNA (miRNA) families [7, 8]. MicroRNAs are small regulatory RNA molecules that play a pervasive role in gene regulation and are understood to influence a variety of biological processes both in normal physiological and pathological disease contexts [9, 10]. Because of their widespread importance in regulating gene networks and their potential role in the evolution of complexity, miRNAs are currently the subject of considerable focus in developmental biology [11–13].

The justification for the phylogenetic utility of miRNA presence/absence data stems from the way that novel miRNA families arise. MicroRNAs originate from random hairpin sequences in intronic or intergenic regions (typically 60–80bp in length) of the genome that become transcribed into RNA [14, 15]. After transcription, the resulting primary miRNAs may fold into hairpins that serve as the substrate for a pair of enzymes—called *Drosha* and *Dicer*—involved in miRNA synthesis [16], culminating in a mature miRNA (typically 22bp in length).

The odds that any individual hairpin structure will acquire the requisite mutations to form a novel miRNA are exceedingly slim; however, genomes contain many thousands of these structures, such that novel miRNAs are likely to accumulate over deep time [14]. After the introduction of new functional miRNAs, strong purifying selection associated with their regulatory role can lead to both extraordinarily low rates of substitution within miRNA sequences, as well as long-term preservation of miRNAs in the genome [14]. This biological scenario is expected to lead to an evolutionary

pattern wherein new miRNAs—over long time scales—continually arise in genomes and experience a low rate of secondary loss [15]. Moreover, the origin of novel miRNAs involves the accumulation of random mutations to a relatively long sequence (60–80bp in animals), rendering it highly improbable that identical miRNAs will evolve convergently [17]. These considerations have led to the promotion of miRNAs as a new source of data that are ideal for parsimony inference of phylogeny: they should exhibit extraordinarily low levels of homoplasy (*i.e.*, they are not expected to arise convergently or to be lost secondarily) and thus provide unambiguous synapomorphies (shared-derived character states) that elevate miRNAs to "one of the most useful classes of characters in phylogenetics" [18].

The above reasoning has led to a recent proliferation of miRNA-based phylogenetic studies seeking to unequivocally resolve several recalcitrant relationships in the tree of life. At the time of our analysis, these include five formal* phylogenetic analyses of miRNA data focused on identifying the phylogenetic position of turtles within amniotes [24], acoelomorph flatworms within animals [25], lampreys within vertebrates [hagfish and jawed vertebrates; 18], myzostomidan worms within bilaterians [26], and to establish the monophyly of—and resolve relationships within—annelids [27].

These studies proceed by first identifying the set of miRNAs present in each study lineage using one of two general approaches: by searching for known or novel miRNAs either in existing genome assemblies and/or in novel data generated by sequencing small-RNA libraries. The identified miRNA families are then used to construct a data matrix in which each miRNA family is treated as an ordered binary character, where miRNA presence is the derived state. Finally, this data matrix is subjected to (Dollo or Wagner) parsimony analysis to estimate phylogenetic relationships.

Here, we critically examine the use of miRNA data for phylogeny estimation, focusing on three concerns: 1) the validity of claims related to the evolution of miRNA families (*i.e.*, that secondary loss is exceptionally rare); 2) limitations of parsimony methods used to infer phylogeny from miRNA presence/absence data; and 3) problems associated with the detection of miRNA families. We demonstrate that these concerns collectively render published phylogenetic conclusions based on miRNA data uncertain (obscured by their reliance on non-statistical methods) and/or strongly biased (owing to problems in miRNA detection and/or inference method). We illustrate these

*Several additional studies discuss the phylogenetic implications of miRNA data, but do not subject these data to a formal phylogenetic analysis. Typically in these studies, the phylogeny is first estimated from some other source of data, and then the correspondence of the inferred tree to select miRNA families is discussed (e.g., [19–23]).

concerns by reanalyzing five published phylogenetic studies of miRNA data.

Interpreting and analyzing microRNA data: Is miRNA absence evidence or absence of evidence?

In order to properly analyze and interpret miRNA presence/absence data, we must be explicit on the nature and meaning of *absence*. A microRNA family that is scored as absent in a particular lineage can, in principle, have one of three histories: 1) the miRNA family may have never arisen in or been inherited by that lineage ('true absence'); 2) the miRNA family may have previously been present in the lineage but subsequently lost from the genome ('secondary loss'); or 3) the miRNA family may actually be present in the genome but escaped detection during data collection ('sampling error'). If all (or nearly all) absences of miRNA families are true absences, then miRNA loss strictly does not occur (or occurs exceedingly rarely): this is the implicit assumption of miRNA studies. Accordingly, because the evolution of miRNA data involves minimal character change—miRNA families have a unique origin (bereft of convergence) with negligible/no secondary loss—the use of parsimony as an inference method might be justified.

In fact, nearly all published miRNA studies (including all five re-examined here) have used some variant of the parsimony method to estimate phylogeny. The miRNA study by [27] used "standard" (Wagner) parsimony—in which gains and losses of miRNA families incur equal cost [28], and the remaining four studies [18, 24–26] employed Dollo parsimony [29]. Dollo parsimony allows for the unique evolution of a character and its subsequent loss (both with equal cost), but precludes re-evolution of the same character (with effectively infinite cost) once it has been lost.

Secondary loss of miRNA families is common

Here we explore the claim that secondary loss of miRNA families is exceedingly rare (e.g., [17, 20, 21]). We derived estimates of the prevalence of miRNA loss from analyses of published miRNA datasets. The prediction is quite simple: if loss of miRNA families is exceedingly rare, then the most parsimonious tree for a given miRNA dataset should be virtually free of homoplasy (implied secondary loss of miRNA families), given that Dollo parsimony does not permit convergent or parallel evolution.

To derive estimates of the implied prevalence of miRNA loss, we reanalyzed the miRNA datasets under Dollo parsimony with PAUP* v4b10 [30] by means of exhaustive searches, treating all characters as 'Dollo.up', which provides the parsimony score (*i.e.*, the total number of implied miRNA gains and losses) for the optimal tree. We then tabulated the number of miRNA losses using the 'dollop' function in Phylip v3.5c [31]. Finally, we estimated the prevalence of miRNA secondary loss in each of the five formal miRNA phylogenetic studies, which is simply calculated as the number of implied losses divided by the parsimony score (total number of implied changes).

[Table 1: miRNA loss rates]

Our survey of published studies suggests that secondary loss of miRNA families is apparently quite common (Table 1). In all but the amniote study ([24], addressed below), secondary miRNA losses constitute between 27–54%, with an overall average of 38%, of the implied evolutionary changes. These phylogenetic results accord well with those of molecular evolutionary studies, in which prevalent secondary loss of miRNA families have been inferred for various taxa [14, 32–35].

Although we suspect that the degree of secondary loss in published studies is somewhat inflated by miRNA sampling errors (see: *Sampling error in miRNA detection and its phylogenetic impact*, below), the complex character histories of miRNA evolution nevertheless suggest that the use of parsimony—which effectively places all of the probability on the single character history with the absolute minimal amount of change—is not a suitable method with which to infer phylogeny from miRNAs.

Statistical analysis of miRNA exposes considerable phylogenetic uncertainty

As discussed in the preceding section, the evolution of miRNA often appears to be complex, which raises concerns about the choice of parsimony as a method of inference. Stochastic models are available that are more appropriate for accommodating complex histories, as the likelihood of a given character (in this case, a miRNA family) is calculated by integrating over all possible character histories (in this case, patterns of miRNA gain and secondary loss that could give rise to the observations), weighting each history by its probability under the model. Furthermore, stochastic models are available that may be appropriate for the analysis of miRNA presence/absence data. For example, the binary stochastic Dollo model (SD: [36, 37]) appears to be well suited for the analysis

of miRNA presence/absence data. The SD model describes an immigration-death stochastic process for a set of observed binary characters where the origin of a character (miRNA family) is modeled as a homogeneous Poisson process with instantaneous rate λ, and its subsequent loss is modeled as a stochastic branching process (where the probability of loss is proportional to the branch length in which it persists toward the present) with an instantaneous rate of secondary loss, μ [37]. This allows the character to evolve a single time-, experience subsequent loss (possibly independently in multiple lineages), but prohibits regain of the character once it has been lost within a lineage [37]. Inference under stochastic models within a Bayesian statistical framework provides a natural means for assessing support/accommodating uncertainty in phylogenetic estimates. Because the majority of published miRNA studies to date have either ignored the issue of evidential support for estimates, or have relied on *ad hoc* support measures (such as the Bremer support index; [38]) which have no clear statistical interpretation [39], the availability of an inference framework that explicitly assesses support is particularly attractive.

Markov chain Monte Carlo (MCMC) simulation is used to approximate the joint posterior probability distribution of the phylogenetic parameters. A Markov chain is specified that has state space comprising all possible values for the phylogenetic model parameters, which has a stationary distribution that is the distribution of interest (*i.e.*, the joint posterior probability distribution of the model parameters). Samples drawn from the stationary Markov chain provide valid estimates of the joint posterior probability density, which can be queried marginally with respect to any parameter of interest. In the case of topology, the marginal posterior probability for a given clade is simply its frequency in the sampled trees.

Bayesian inference of phylogeny from miRNA datasets.—These considerations motivated us to re-analyze previously published miRNA datasets within a Bayesian statistical framework using a stochastic binary Dollo model [37] to describe the gain and loss of miRNA families. For each of the five miRNA datasets, we treated all characters as 'Dollo type' and approximated the joint posterior probability density via MCMC using BEAST v1.7.5 [40]. We specified a prior for the rate of miRNA loss, μ, using an exponential distribution with a small rate parameter ($\bar{x} = 1.0 \times 10^{-4}$) and specified a prior on the tree topology and node heights using a stochastic birth-death branching process.

Molecular studies have alternatively characterized the evolution of miRNAs as a gradual process

of continuous accumulation via mutation [14], or as an episodic process associated with major regulatory or developmental innovations [15]. Accordingly, we explored an array of (relaxed) clock models to describe the variation in rates of miRNA evolution across the tree or through time that range from stochastically constant to episodic. Selection among these alternative clock models yields ultrametric phylogenies that give us insight into the pattern of miRNA accumulation and loss as well as information about the placement of the root of the phylogeny. Specifically, for each dataset, we performed analyses under the strict-clock model, the random-local clock model (RLMK: [41]), and the uncorrelated lognormal (UCLN) and exponential (UCED) relaxed-clock models [42]. Inference of the joint posterior probability density for each composite phylogenetic model (*i.e.*, the binary stochastic Dollo model + one of the [relaxed] clock models) involved at least three independent MCMC analyses, running each chain for 100 million cycles and sampling every $10,000^{th}$ cycle.

In order to compare fit of the data to these four alternative clock models, we performed additional analyses targeting the marginal likelihood of the data under each of the four composite phylogenetic models. For each dataset, this entailed running the MCMC through a series of 50 power posteriors spanning from the prior to the posterior, with the powers spaced along a Beta(0.3, 1.0) distribution. We then estimated the marginal likelihood from this chain using both path and stepping stone sampling analyses [43–45]. These analyses were also each repeated at least three times to ensure stability of the marginal likelihood estimates. We then compared support for the alternative clock models by calculating Bayes factors as the ratio of the marginal likelihoods for each pairwise combination of candidate models. We interpret Bayes factors following Kass and Raftery [46]: viewing 2 ln BF values >10 as very strong support for the candidate model, between 6 and 10 as strong support, between 2 and 6 as positive evidence, and < 2 as essentially equivocal regarding the alternative models. We performed model comparison only for models where the analyses performed very well, judged by the MCMC mixing efficiently across the power posteriors and highly stable estimates of the marginal likelihood across replicated analyses with both stepping stone and path sampling.

In total, this analysis design entailed 180 MCMC analyses: each of the five miRNA datasets were analyzed under each of the four (relaxed) clock models, performing three independent MCMC analyses under each model, repeating analyses to target first the joint prior probability, then the joint posterior probability, and finally the marginal likelihood densities. We assessed the performance of

each MCMC analysis for all parameters (including the topology) using Tracer and AWTY [47, 48], which suggested that the chains mixed well and had converged prior to ~ 50 million cycles in nearly all cases. In the few instances where poor mixing or convergence was noted, we ran additional independent analyses until an adequate sample from the target density could be obtained, or it became clear that the MCMC could not adequately sample from the target distribution. Inferences under each model were based on the combined stationary samples from each of the independent chains, which provided adequate sampling for all parameters according to the effective sample size (ESS) [40].

Finally, we assessed support for the key phylogenetic findings of each published miRNA study using Bayes factors. This entailed a second round of analyses targeting the marginal likelihood density that were identical to our initial analyses under the best fitting clock model (as judged by the Bayes factor model comparisons above), but with the topology constrained to the relevant alternative hypothesis in each case (discussed in more detail below). These analyses allowed us to quantify the extent to which each miRNA dataset can decisively distinguish among alternative phylogenetic hypotheses.

Patterns and rates of miRNA evolution.—We used Bayesian model-comparison methods to assess the fit of the miRNA datasets to four (relaxed) clock models, which differ in their ability to accommodate rate variation across lineages. The strict clock makes the most stringent assumption of rate homogeneity, the random-local clock is intermediate, and the uncorrelated (exponential and lognormal) relaxed-clock models are able to capture the most extreme rate fluctuations across branches—rates on adjacent branches are modeled as independent and identically distributed random variables drawn from a common (exponential or lognormal) probability distribution (Drummond et al., 2006). Interestingly, the two uncorrelated relaxed-clock models had the highest marginal likelihood and were therefore the preferred model for every single dataset (Table 2). We were unable to perform a few of these comparisons due to poor mixing of MCMC that prohibited stable estimation of a marginal likelihood for some of the data + model combinations (the uncorrelated lognormal in particular, see Table 2). However, the uncorrelated exponential model was very strongly preferred ($2 \ln \mathrm{BF} > 10$) to the Strict model for four datasets, and was strongly preferred ($2 \ln \mathrm{BF} > 6$) for the fifth. These results, combined with the large coefficient of variation for rates among branches under the winning model (Table 2), imply substantial heterogeneity in the rate of miRNA evolution

across branches in these datasets, conditions in which parsimony inferences are more likely to be inconsistent (e.g., [49–51]). Finally, as in the case of the Dollo parsimony analyses, Bayesian estimates under the stochastic Dollo model indicate substantial rates of miRNA loss in all five miRNA datasets (Table 1).

[Table 2: miRNA Clock Models]

Evaluating support for key phylogenetic conclusions of published miRNA studies.—Bayesian analyses of miRNA data offered novel insight into several previously published studies. In three of the five cases, the Bayesian analysis recovers a result that disagrees in important respects from the parsimony result, but agrees with other published studies based on more-traditional phylogenomic analyses of molecular sequence datasets. Parsimony and Bayesian analyses recover congruent conclusions for the two remaining studies, although both of these cases remain problematic due to large uncertainty or sampling error. We briefly discuss key results for each of these analyses below.

Annelid dataset.—Sperling et al. [27] sought to evaluate the monophyly of and establish phylogenetic relationships within annelids. Based on the parsimony analysis of the miRNA dataset, they concluded that: 1) annelids are monophyletic (*Nereis*, *Lumbricus*, and *Capitella* form a clade); 2) the sipunculan species, *Phascolosoma*, is the sister group of annelids; and finally, 3) polychaete annelids are not monophyletic (*Nereis* and *Capitella* do not form a clade). Bayesian analysis of the miRNA data under the stochastic Dollo model infers the tree: ((*Nereis*, *Phascolosoma*), (*Lumbricus*, *Capitella*)) (Figure 1a). Accordingly, these results neither support annelid monophyly nor a sister-group relationship between sipunculans and annelids. Our finding that sipunculids (represented by *Phascolosoma*) are included within annelids—and thus, that annelids are paraphyletic—is consistent with most recent molecular phylogenetic/omic studies (e.g., [52–57]).

We assessed the decisiveness of support for these alternative topological models by performing analyses in which the topology was constrained alternatively to the parsimony estimate (Model M_1, Table 3) and the Bayesian estimate (Model M_0, Table 3) and compared the marginal likelihoods under the two models. A 2 ln BF of ~ 12 in favor of the Bayesian topology suggests that the data very strongly prefer the Bayesian estimate relative to the parsimony estimate.

[Table 3: miRNA Topology Models]

Bilaterian dataset.—Helm et al. [26] sought to resolve the phylogenetic affinity of myzostomid worms using an expanded version of the miRNA dataset from the Sperling et al. (2009) study, testing alternative hypotheses that either placed myzostomids within annelids or platyzoans. Their parsimony analysis of the miRNA data "strongly confirms a phylogenetic position of *Myzostomida*" as "deeply nested within the annelid radiation, as sister to *Capitella*." By contrast, Bayesian analysis of this miRNA dataset under the stochastic Dollo model implies that myzostomids are the sister group of annelids (with a clade probability of ~ 0.97–0.99), which agrees with estimates based on recent analyses of phylogenomic data (e.g., [55]) (Figure 1b).

We assessed the support for these alternative hypotheses by performing analyses in which the topology was constrained to the parsimony estimate (model M_1, Table 3), and compared the marginal likelihood of this model to that from analyses constrained to the Bayesian estimate (model M_0, Table 3). These analyses decisively reject the inclusion of *Myzostoma* within annelids ($2 \ln \text{BF} \sim 100$). It was not possible to perform a clear test of the alternative 'platyzoan' hypothesis, as Platyzoa was not inferred to be monophyletic in our unconstrained analyses.

Animal dataset.—Philippe et al. [25] sought to establish the phylogenetic placement of acoels and xenoturbellids within animals using three independent datasets: a large number of mitochondrial genes, a phylogenomic dataset comprising $38,330$ amino-acid positions, and a microRNA dataset. The phylogeny inferred from their Dollo parsimony analysis of the miRNA dataset implied that acoels (*Symsagittifera* and *Hofstenia*) and xenoturbellids (*Xenoturbella*) form a paraphyletic grade near the base of bilaterians: (*Symsagittifera*, (*Hofstenia*, (*Xenoturbella*, (remaining bilaterians))). The Bayesian analysis of this miRNA dataset under the stochastic Dollo model infers a very different tree in which acoels are monophyletic and sister to xenoturbellids: (((*Symsagittifera*, *Hofstenia*), *Xenoturbella*), remaining bilaterians) (Figure 1c). We assessed support for these hypotheses by performing additional analyses in which the topology was alternatively constrained to the parsimony estimate (topological model M_1, Table 3) and the Bayesian estimate (topological model M_0, Table 3) and compared the marginal likelihoods. In contrast to all the other studies, the Bayes factor suggests that the miRNA data favor the parsimony hypothesis in this case ($2 \ln \text{BF} \sim -12$). Thus, these contrasting results give no clear guidance on which alternative is the more reliable topology. However, the extensive phylogenomic analysis that was paired with the original miRNA analysis helps to clarify which topology is likely correct.

Notably, Philippe et al [25] favored a hypothesis that disagreed with the miRNA parsimony result. The central phylogenetic finding in Philippe et al. [25] is the close relationship between *Xenoturbella* and (a monophyletic) Acoela (*Symsagittifera*, *Hofstenia*). Although this result strongly conflicts with their parsimony analysis of miRNA data, they prefer it based on their rigorous Bayesian analyses of large-scale molecular datasets. In fact, in discussing the conflicting estimates based on their Bayesian analyses of the phylogenomic data and their parsimony analysis of the miRNA data, Philippe et al. [25] were skeptical of the miRNA phylogeny, attributing this discrepancy to the effects of pervasive secondary loss of miRNA families in acoels. Interestingly, our Bayesian analysis of the miRNA dataset recovers the same monophyletic Acoela sister to *Xenoturbella*. However, both Bayesian and parsimony analyses of the miRNA data conflict with the preferred tree from Philippe et al. [25] in other respects, suggesting that secondary loss has strongly obscured any phylogenetic signal in these data.

Vertebrate dataset.—Heimberg et al. [18] sought to resolve the phylogenetic position of lampreys within vertebrates using miRNA data, testing alternative hypotheses that either placed lampreys as sister to hagfish (the 'cyclostome' hypothesis) or to jawed vertebrates (the 'vertebrate' hypothesis). Analysis of the vertebrate miRNA dataset using Dollo parsimony supported the cyclostome hypothesis: the two lampreys, *Lampetra* and *Petromyzon*, form a clade that is sister to the hagfish species, *Myxine*: ((*Lampetra*, *Petromyzon*), *Myxine*)). Bayesian analysis of the vertebrate miRNA dataset under the stochastic Dollo model also supported the cyclostome hypothesis, albeit weakly (i.e., with a clade probability of ~ 0.79) (Figure 1d).

We assessed the support for cyclostome monophyly by performing analyses in which the topology was constrained to the alternative phylogenetic hypothesis in which lampreys are sister to jawed vertebrates (model M_1, Table 3), and compared the marginal likelihoods of the constrained and unconstrained (model M_0, Table 3) analyses. Comparison of the marginal likelihoods under the constrained and unconstrained models suggests that the miRNA data are essentially equivocal regarding the phylogenetic affinity of lampreys ($2 \ln \mathrm{BF} \sim 1$).

Amniote dataset.—Lyson et al. [24] sought to resolve the phylogenetic placement of turtles within amniotes, using a miRNA dataset to test whether turtles were either sister to lizards + tuatara (the 'lepidosaur' hypothesis), or to birds + crocodilians (the 'archosaur' hypothesis). Analysis of

the miRNA dataset using Dollo parsimony supports the lepidosaur hypothesis, and this finding was also strongly supported by Bayesian analysis under the stochastic Dollo model (with a clade probability of ~ 1.0) (Figure 1e).

We further assessed support for the lepidosaur hypothesis by performing analyses of the amniote miRNA dataset in which the topology was constrained to the alternative phylogenetic hypothesis in which turtles are sister to archosaurs (model M_1, Table 3), and compared the marginal likelihoods to those from the lepidosaur hypothesis (model M_0, Table 3). In contrast to all other studies, comparison of the marginal likelihoods under the two models suggests that the miRNA data provide strong support for the originally published result ($2 \ln \text{BF} \sim 17$). However, we demonstrate below that this result is an artifact of sampling error in the detection of amniote miRNAs (see: *Sampling error in miRNA detection and its phylogenetic impact*).

Anomalous results from miRNA analyses.—Bayesian analysis of published miRNA datasets casts considerable doubt on the key phylogenetic conclusions of these previously published studies. In three of five cases (animals, annelids, and bilaterians), using a model that accounts for the uncertainty in character histories changes the key phylogenetic conclusion, often with strong support. In a fourth case (vertebrates), considering the uncertainty in character history leads to the conclusion that miRNAs are essentially silent on the relationship of interest. In only one case (amniotes) does accounting for uncertainty in character history leave the key conclusion unchanged, although this case reveals a second issue that we explore below. Moreover, our re-analyses of published miRNA datasets also supported some highly unusual phylogenetic results. For example, Bayesian analyses of the amniote miRNA dataset failed to support the (virtually incontrovertible) monophyly of archosaurs (Figure 1e), whereas analyses of the animal miRNA dataset supported (the very odd placement of) chordates as the sister to all other bilaterians (Figure 1c). We argue below that such remarkable findings likely have a more prosaic explanation.

Shortly after the present manuscript returned from an initial round of peer review, a paper appeared that further discussed the phylogenetic potential of miRNAs and demonstrated phylogenetic inference with miRNAs using the binary stochastic Dollo model [8]. This paper assembled a dataset of miRNA presence/absence for 29 metazoan taxa from subsets of the data matrices developed in previous studies (including those that we re-examine here) and analyzed it using the stochastic Dollo. This analysis recovers high posterior probabilities on all nodes except one and is congruent

with other phylogenies constructed from more traditional phylogenetic and phylogenomic analyses. Thus, the Tarver et al. [8] result appears to be in stark contrast with our results. The discrepancy appears to stem from the choice of taxa for inclusion in the Tarver et al. [8] data matrix. The dataset retains only a subset of the taxa reported in the original studies, while we analyze the original studies' data matrices in full. Further, the Tarver et al. [8] matrix is missing all the taxa that we identify as leading to problematic results above. For example, we identify low support and pervasive uncertainty associated with the relationship between the lamprey (*Lampetra* and *Petromyzon*) and the hagfish (*Myxine*)—the central taxa under study in the dataset of Heimberg et al. [18]. Tarver et al. [8] retain only one lamprey (and no hagfish) from this dataset and thus do not test the support for this clade. Similarly, the acoels (*Symsagittifera*, *Hofstenia*) and *Xenoturbella* are central to the study by Philippe et al. [25]. These taxa disagree strongly with traditionally constructed phylogenies but are not included in Tarver et al. [8]. The two birds (*Gallus* and *Taenopygia*) and lizard from the Lyson et al. [24] dataset are included in Tarver et al. [8], but the critical turtle and alligator data are not. Likewise, the key taxon *Myzostomida* from Helm et al. [26] is not included, nor are *Nereis* and *Phascolosoma* from Sperling et al. [27]. No details outlining the choice of taxa for this matrix are given, so we are unsure why only subsets of previous datasets were included, nor why certain taxa were included versus not. That said, the apparent discrepancy among our results appears to stem from our varying choices of taxa. Because the utility of miRNAs in phylogenetics lies in their purported ability to resolve particularly vexing phylogenetic relationships, our view is that including taxa that allow for tests of such vexing relationships is a critical part of studying these marker's phylogenetic utility.

Sampling error in miRNA detection and its phylogenetic impact

Sampling error can to lead to the (apparent) absence of miRNAs in phylogenetic datasets. This is of particular concern because most miRNA phylogenetic studies use a mixture of approaches to identify miRNAs in different lineages (namely, using a combination of bioinformatic scans of complete genomes and/or *de novo* sequencing of small-RNA libraries). If these approaches vary in their detection probabilities, then miRNAs are more likely to be discovered in some lineages than in others. As more and more data are collected under this biased detection scheme, certain lineages are likely to accumulate true presences while the remaining lineages will accumulate apparent absences.

Since the presence and absence of miRNAs are the direct source of phylogenetic information, this sampling artifact may lead to biased estimates of topology.

Here we demonstrate sources of sampling error in the detection of miRNA families, first focusing on the analysis of turtle relationships within amniotes as a detailed case study, and then assessing the generality of this sampling error by means of a more general survey.

Sampling bias in the detection of amniote miRNAs.—Lyson et al. [24] employed a mixture of miRNA detection methods in an attempt to resolve the phylogenetic position of turtles within amniotes. Specifically, their study searched for miRNAs using: 1) similarity searches against whole-genome assemblies for two birds—chicken (*Gallus*), zebra finch (*Taeniopygia*)—and four outgroup taxa; 2) a combination of similarity searches against the genome assembly for the lizard (*Anolis*) and *de novo* sequencing of an *Anolis* RNA library; and 3) *de novo* sequencing of RNA libraries for a turtle species—the painted turtle (*Chrysemys*)—and the American alligator (*Alligator*). At the time of their study, full genome assemblies for the painted turtle and alligator were not available. The authors identified 19 miRNA families unique to birds, one miRNA family unique to archosaurs (birds and crocodilians), but no miRNA families shared between archosaurs and turtles. Furthermore, the study identified four miRNA families that are shared between the anole and turtle. Taken at face value, these data appear to unequivocally support a turtle + lizard relationship, to the exclusion of archosaurs.

Draft genome assemblies for both the painted turtle and American alligator are now available [58, 59], which provide an independent check of the miRNAs detected—and the phylogenetic conclusions reached—in the Lyson et al. [24] study. We sought to confirm that each of the miRNA families that were identified by Lyson et al. [24] as unique to birds ($N = 19$) were in fact absent from the turtle and alligator genomes, and that the single archosaur-specific miRNA was absent from the turtle genome. We also assessed whether each of the miRNA families that were identified as being shared exclusively by turtles and lizards were in fact present in the turtle genome and absent from the alligator genome.

We downloaded both the longer stem-loop sequence (60–80 bp) and the shorter mature sequence (22 bp) for each relevant miRNA from miRBase [60] for each appropriate reference taxon (*Gallus* for the 19 bird-specific and the single archosaur-specific miRNA families; *Anolis* for the four miRNA families uniquely shared by turtle + lizard). We constructed local BLAST databases from the turtle

and alligator genome assemblies ($v3.0.3$ and $0.1d27$, respectively) and searched against them with each of the relevant miRNA stem-loop sequences using BLASTN ($v2.2.25$, minimum word size = 11, e-value cutoff = 10–2; [61]). We then predicted secondary structure for any putative miRNAs that we identified using mFold [62].

We scored a miRNA family as being present in the turtle and/or alligator genome if it met three criteria: 1) We observed a highly significant hit (*i.e.*, with a minimum e-value of 10^{-20}) for the reference stem-loop sequence against the relevant genome assembly; 2) The matching sequence in the genome contained a nearly perfect match to the mature ~22bp miRNA sequence (*i.e.*, containing no more than one substitution in the mature miRNA sequence); 3) The matching sequence in the turtle or alligator genome folded into the expected hairpin secondary structure and this structure was similar to the predicted secondary structure published for the reference sequence.

Our search confirmed that the single archosaur-specific miRNA (miRNA 1791) was present in the alligator genome, as expected. However, we discovered that this miRNA is also present in the turtle genome (for sequences and predicted secondary structure, see Figure S1). Furthermore, we discovered three additional miRNA families present in both the alligator and turtle genomes that were reported by Lyson et al. [24] as being unique to birds (miRNA families 1641, 1743, and 2964). All four families exhibited very high sequence similarity with the known miRNA from the reference taxon, highly conserved stem-loop structures with similar free energies to that predicted from the reference taxon, and mature sequences that were identical (two families) or nearly identical (two families) to the reference (see Figure S1 for sequence alignments and predicted structures). This sampling error may be inherent to miRNA-detection approaches that rely on RNA sequencing. For example, Sperling et al. [27] observed a similar pattern in the polychaete worm, *Capitella*. They discovered five additional miRNAs from the genome of this organism that were not detected in the sequences derived from an RNA library. MicroRNAs are frequently expressed only in certain tissues, at certain stages of development, or expressed at low levels [27, 63–66]. In these cases, it is likely that miRNAs actually present in the genome will be missed because they are not being transcribed (or only being transcribed at low levels) in the tissue that was used to make the RNA library.

Finally, we sought to confirm that the four miRNA families identified by Lyson et al. [24] as uniting a lizard + turtle clade were, in fact, present in the turtle genome and absent in the alligator genome (miRNA families 5390, 5391, 5392, and 5393). Our search confirmed that all four miRNA

families were absent from the alligator genome, as expected. However, we were only able to find one of the four reported miRNA families (miRNA 5391) in the turtle genome. We found no significant BLAST hits to any of the other three expected miRNAs, even under relaxed search settings (word size = 4, e-value cutoff = 10). We then assessed whether we could identify these miRNAs in the *Anolis* genome and found all four families, as expected. At present, the cause of this discrepancy is unclear. Our failure to detect these sequences could be a false negative, indicating that the turtle genome assembly is incomplete and missing these three sequences. Alternatively, their previous detection could be a false positive in the Lyson et al. [24] study, stemming from contamination between the *Anolis* and *Chrysemys* sequencing libraries or from another source of error. The turtle genome assembly has 18x coverage and is estimated to be 93% complete, which suggests that the former explanation is unlikely [59]. Nevertheless, we can not formally distinguish between these possibilities at present.

We then revised the Lyson et al. [24] data matrix to correct this sampling error and subjected the revised matrix to Bayesian phylogenetic analysis under the stochastic Dollo model (analyses performed as detailed above). Rather than supporting a strong relationship between lizards and turtles, the corrected miRNA dataset supports a relationship between turtles and archosaurs, albeit weakly (i.e., with a clade probability of ~ 0.54) (Figure 2). This result is consistent with several recently published studies that examine the phylogenetic placement of turtles using large DNA sequence datasets [59, 67–69].

We assessed support for the 'archosaur' hypothesis by performing analyses of the corrected amniote miRNA dataset in which the topology was constrained to the alternative 'archosaur' and 'lepidosaur' hypotheses (models M_0 and M_1 in Table 3, respectively). Comparison of the marginal likelihoods under the alternative models indicate that the miRNA data provide positive evidence in favor of the archosaur hypothesis ($2 \ln \text{BF} \sim 5$). This analysis illustrates that miRNA detection is prone to strong sampling error, to a degree that can fundamentally alter the conclusions of phylogenetic inferences based on these data.

General survey of sampling bias in miRNA detection.—Our ability to provide a detailed description of the miRNA detection bias in the amniote study largely rests on the serendipitous availability of two new genome assemblies. Accordingly, it is not possible to perform a comparably detailed analysis of the potential sampling errors in the other four published miRNA phylogenetic studies. However,

we can make a more general comparison of alternative miRNA detection strategies. To do so, we compiled information from the literature of cases in which the total miRNA complement of various organisms had been estimated both by means of *de novo* sequencing of small-RNA libraries and also by means of bioinformatic searches of DNA sequence resources. If no sampling error exists, identical sets of miRNA families should be identified using alternative strategies. In stark contrast to this expectation, however, we see a high degree of variation in the miRNA complement identified under the two strategies (Table 4). Although this comparison does not directly replicate the alternative methods employed in published phylogenetic studies, it clearly indicates the prevalence of variation in total miRNA complement detection and, as we have shown, this type of sampling error has the potential to impact estimates of phylogeny.

[Table 4: miRNA Comparison]

Conclusions

The current wealth of molecular data will continue to resolve relationships in the tree of life, but not all nodes will acquiesce with equal effort. Predictably, the variously recalcitrant, enigmatic, inscrutable and impenetrable relationships will continue to be identified. Ultimately, resolution of these problematic cases may require the discovery of new and improved phylogenetic data (and/or the elaboration and careful application of more realistic models that better describe important aspects of the processes that give rise to conventional genomic data). Accordingly, it is predictable that the addition of a putative silver bullet—such as miRNA presence/absence data—to our phylogenetic arsenal will be greeted with enthusiasm. We would argue, however, that this enthusiasm should be tempered with careful consideration of how to appropriately accommodate the correspondingly novel processes by which these new data evolved and/or new procedures by which they are collected.

We have demonstrated that the evolution of miRNA families is complex. Contrary to repeated claims, secondary loss of miRNA appears to be quite prevalent, and miRNA evolution typically exhibits substantial variation in rate across branches through time. Consequently, the complex character histories associated with miRNA evolution suggest that parsimony—which effectively places all of the probability on the character history with the minimal change—is not a defensible

method with which to infer phylogeny from these new data. We have demonstrated that, in principle, it is both possible and preferable to estimate phylogeny from miRNA data within a Bayesian statistical framework using stochastic evolutionary models. Adopting a statistical approach for estimating phylogeny from miRNA (or other) data confers many benefits: this approach allows us to choose objectively among models, to perform formal tests of competing hypotheses, promotes a richer study of the evolutionary process, and enables us to gauge and accommodate uncertainty in our estimates. We have established the importance of adopting a more appropriate statistical approach: Bayesian analyses of published miRNA datasets qualitatively altered key phylogenetic conclusions and/or revealed considerable phylogenetic uncertainty in these estimates in four of the five cases that we examined.

Finally, we have demonstrated that the detection of miRNA families is prone to error—especially when using a mixture of detection methods—and this sampling error can substantially bias estimates of phylogeny. Accordingly, it is critical that we either extend existing stochastic models to accommodate this ascertainment bias, or take precautionary measures to minimize it. For example, models used to analyze both SNP data in population genetics [70] and discrete-morphological data in phylogenetics [71] explicitly model the associated ascertainment strategies in order to reduce the associated biases. The stochastic Dollo model might be similarly extended to accommodate the documented miRNA ascertainment bias. While the complexity of the mixed genomic/RNA-library detection strategy would make such an extension challenging, the intense focus on miRNA detection methods (e.g., [72]) gives reason for optimism that these extensions may be possible. Alternatively, studies seeking to estimate phylogeny from miRNA presence/absence data should strictly employ identical, genome-based detection methods in all lineages. This may not always eliminate sampling error, but it should reduce bias arising from differential detection probabilities of the various miRNA discovery methods.

Although our appraisal of miRNA as a novel source of phylogenetic information is admittedly critical, we clearly recognize the potential of these data to inform phylogeny: inferences based on miRNA data often correspond broadly to those based on more conventional gene/omic data. We take issue, however, with the recent promotion of miRNA data as a phylogenetic panacea. New data are attended by new issues that need to be carefully resolved in order to realize their full potential.

Acknowledgments

We thank Artyom Kopp and the members of a phylogenetic reading group at UC Davis for helpful discussion and advice during the development of this project. We also thank the Turtle Genome Sequencing Consortium and the International Crocodilian Genomes Working Group for providing pre-publication access to the genome assemblies used in this study. Support for this work was provided by University of Hawaii research funds to RCT and by National Science Foundation grants DEB-0842181 and DEB-0919529 to BRM.

References

[1] Sanderson, M. J. (2008) *Science* **321**, 121–3.

[2] Thomson, R. C & Shaffer, H. B. (2010) *BMC Biology* **8**, 19.

[3] Hillis, D. (1999) *Proceedings of the National Academy of Science, USA* **96**, 9979–9981.

[4] Rokas, A & Holland, P. W. H. (2000) *Trends in Ecology & Evolution* **15**, 454–459.

[5] Boore, J. L. (2006) *Trends in Ecology & Evolution* **21**, 439–446.

[6] Boore, J. L & Fuerstenberg, S. I. (2008) *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **363**, 1445–1451.

[7] Dolgin, E. (2012) *Nature* **486**, 460–462.

[8] 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) *Molecular Biology and Evolution* **30**, 2369–82.

[9] Lu, J, Getz, G, Miska, E. A, Alvarez-Saavedra, E, Lamb, J, Peck, D, Sweet-Cordero, A, Ebert, B. L, Mak, R. H, Ferrando, A. A, Downing, J. R, Jacks, T, Horvitz, H. R, & Golub, T. R. (2005) *Nature* **435**, 834–8.

[10] Alvarez-Garcia, I & Miska, E. A. (2005) *Development (Cambridge, England)* **132**, 4653–62.

[11] Berezikov, E. (2011) *Nature Reviews Genetics* **12**, 846–60.

[12] Peterson, K. J, Dietrich, M. R, & McPeek, M. A. (2009) *BioEssays* **31**, 736–47.

[13] Heimberg, A. M, Sempere, L. F, Moy, V. N, Donoghue, P. C. J, & Peterson, K. J. (2008) *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2946–50.

[14] Nozawaet, M, Miura, S, & Nei, M. (2010) *Genome Biology and Evolution* **2**, 180–189.

[15] Campo-Paysaa, F, Sémon, M, Cameron, R. A, Peterson, K. J, & Schubert, M. (2011) *Evolution & Development* **13**, 15–27.

[16] Krol, J, Loedige, I, & Filipowicz, W. (2010) *Nature Reviews Genetics* **11**, 597–610.

[17] Sperling, E. A & Peterson, K. J. (2009) *Exchangeability and related topics.* (Oxford Univ Press), pp. 157–170.

[18] Heimberg, A. M, Cowper-Sal-lari, R, Sémon, M, Donoghue, P. C. J, & Peterson, K. J. (2010) *Proceedings of the National Academy of Science, USA* **107**, 19379–19383.

[19] Rota-Stabelli, O, Campbell, L, Brinkmann, H, Edgecombe, G. D, Longhorn, S. J, Peterson, K. J, Pisani, D, Philippe, H, & Telford, M. J. (2011) *Proceedings of the Royal Society: Biological Sciences* **278**, 298–306.

[20] Sempere, L. F, Martinez, P, Cole, C, Baguñá, J, & Peterson, K. J. (2007) *Evolution & Development* **9**, 409–415.

[21] Wheeler, B. M, Heimberg, A. M, Moy, V. N, Sperling, E. A, Holstein, T. W, Heber, S, & Peterson, K. J. (2009) *Evolution & Development* **11**, 50–68.

[22] Campbell, L. I, Rota-Stabelli, O, Edgecombe, G. D, Marchioro, T, Longhorn, S. J, Telford, M. J, Philippe, H, Rebecchi, L, Peterson, K. J, & Pisani, D. (2011) *Proceedings of the National Academy of Science, USA.*

[23] Sperling, E. A, Pisani, D, & Peterson, K. J. (2011) *Evolution and Development* **13**, 290–303.

[24] Lyson, T. R, Sperling, E. A, Heimberg, A. M, Gauthier, J. A, King, B. L, & Peterson, K. J. (2012) *Biology Letters.*

[25] Philippe, H, Brinkmann, H, Copley, R. R, Moroz, L. L, Nakano, H, Poustka, A. J, Wallberg, A, Peterson, K. J, & Telford, M. J. (2011) *Nature* **470**, 255–258.

[26] Helm, C, Bernhart, S. H, zu Siederdissen, C. H, Nickel, B, & Bleidorn, C. (2012) *Molecular Phylogenetics and Evolution* **64**, 198–203.

[27] Sperling, E. A, Vinther, J, Wheeler, B. M, Sémon, M, Briggs, D. E. G, & Peterson, K. J. (2009) *Proceedings of the Royal Society: Biological Sciences* **276**, 4315–4322.

[28] Kluge, A. G & Farris, J. S. (1969) *Systematic Zoology* **18**, 1–32.

[29] LeQuesne, W. J. (1974) *Systematic Zoology* **23**, 513–517.

[30] Swofford, D. L. (1998) *PAUP*: Phylogenetic analysis using parsimony and other methods.* (Sinauer Associates, Inc., Sunderland, Massachusetts).

[31] Felsenstein, J. (1993) PHYLIP: (Phylogeny Inference Package) (Distributed by author).

[32] Guerra-Assunçáo, J. A & Enright, A. J. (2012) *BMC genomics* **13**, 218.

[33] Meunier, J, Lemoine, F, Soumillon, M, Liechti, A, Weier, M, Guschanski, K, Hu, H, Khaitovich, P, & Kaessmann, H. (2013) *Genome research* **23**, 34–45.

[34] Lyu, Y, Shen, Y, Li, H, Chen, Y, Guo, L, Zhao, Y, Hungate, E, Shi, S, Wu, C.-I, & Tang, T. (2014) *PLoS genetics* **10**, e1004096.

[35] Fromm, B, Worren, M. M, Hahn, C, Hovig, E, & Bachmann, L. (2013) *Molecular biology and evolution* **30**, 2619–28.

[36] Nicholls, G & Gray, R. (2008) *Journal of the Royal Statistical Society, B* **70**, 545–566.

[37] Alekseyenko, A. V, Lee, C. J, & Suchard, M. A. (2008) *Systematic Biology* **57**, 772–784.

[38] Bremer, K. (1988) *Evolution* **42**, 795–803.

[39] DeBry, R. (2001) *Systematic Biology* **50**, 742–752.

[40] Drummond, A. J, Suchard, M. A, Xie, D, & Rambaut, A. (2012) *Molecular Biology and Evolution* **29**, 1969–1973.

[41] Drummond, A. J & Suchard, M. A. (2010) *BMC Biology* **8**, 114.

[42] Drummond, A. J, Ho, S. Y, Phillips, M. J, & Rambaut, A. (2006) *PLoS Biology* **4**, e88.

[43] Gelman, A & Meng, X. (1998) *Statistical Science* **13**, 163–185.

[44] Xie, W, Lewis, P. O, Fan, Y, Kuo, L, & Chen, M.-H. (2011) *Systematic Biology* **60**, 150–60.

[45] Baele, G, Lemey, P, Bedford, T, Rambaut, A, Suchard, M, & Alekseyenko, A. V. (2012) *Molecular Biology and Evolution* **29**, 2157–67.

[46] Kass, R. E & Raftery, A. E. (1995) *Journal of the American Statistical Association* **90**, 773–795.

[47] Rambaut, A & Drummond, A. J. (2007) Tracer v1.4 (http://beast.bio.ed.ac.uk/Tracer).

[48] Nylander, J, Wilgenbusch, J. C, Warren, D. L, & Swofford, D. L. (2008) *Bioinformatics* **24**, 581.

[49] Felsenstein, J. (1978) *Systematic Zoology* **27**, 401–410.

[50] Huelsenbeck, J. P & Hillis, D. M. (1993) *Systematic Biology* **42**, 247–264.

[51] Huelsenbeck, J. P. (1995) *Systematic Biology* **44**, 17–48.

[52] Colgan, D. J, Hutchings, P. A, & Braune, M. (2006) *Organismal Diversity & Evolution* **6**, 220–235.

[53] Hausdorf, B, Helmkampf, M, Meyer, A, Witek, A, Herlyn, H, Bruchhaus, I, Hankeln, T, & andB. Lieb, T. S. (2007) *Molecular Biology and Evolution* **24**, 2723–2729.

[54] Rousset, V, Pleijel, F, Rouse, G. W, Erséus, C, & Siddall, M. E. (2007) *Cladistics* **23**, 41–63.

[55] Struck, T. H, Schult, N, Kusen, T, Hickman, E, Bleidorn, C, McHugh, D, & Halanych, K. M. (2007) *BMC Evolution Biology* **7**, 57.

[56] Dunn, C. W, Hejnol, A, Matus, D. Q, Pang, K, Browne, W. E, Smith, S. A, Seaver, E, Rouse, G. W, Obst, M, Edgecombe, G. D, Sorensen, M. V, Haddock, S. H. D, Schmidt-Rhaesa, A, Okusu, A, Kristensen, R. M, Wheeler, W. C, Martindale, M. Q, & Giribet, G. (2008) *Nature* **452**, 745–749.

[57] Xin, S, X. Ma, J. R, & Zhao, F. (2009) *BMC Genommics* **10**, 36.

[58] St John, J, Braun, E, Isberg, S, Miles, L, Chong, A, Gongora, J, Dalzell, P, Moran, C, Bed'Hom, B, Abzhanov, A, Burgess, S, Cooksey, A, Castoe, T, Crawford, N, Densmore, L, Drew, J, Edwards, S, Faircloth, B, Fujita, M, Greenwold, M, Hoffmann, F, Howard, J, Iguchi, T, Janes, D, Khan, S, Kohno, S, de Koning, A. J, Lance, S, McCarthy, F, & McCormack, J. (2012) *Genome Biology* **13**, 415.

[59] Shaffer, H. B, Minx, P, Warren, D. E, Shedlock, A. M, Thomson, R. C, Valenzuela, N, Abramyan, J, Amemiya, C. T, Badenhorst, D, Biggar, K. K, Borchert, G. M, Botka, C. W, Bowden, R. M, Braun, E. L, Bronikowski, A. M, Bruneau, B. G, Buck, L. T, Capel, B, Castoe, T. A, Czerwinski, M, Delehaunty, K. D, Edwards, S. V, Fronick, C. C, Fujita, M. K, Fulton, L, Graves, T. A, Green, R. E, Haerty, W, Hariharan, R, Hernandez, O, Hillier, L. W, Holloway, A. K, Janes, D, Janzen, F. J, Kandoth, C, Kong, L, de Koning, A. J, Li, Y, Literman, R, McGaugh, S. E, Mork, L, O'Laughlin, M, Paitz, R. T, Pollock, D. D, Ponting, C. P, Radhakrishnan, S, Raney, B. J, Richman, J. M, St John, J, Schwartz, T, Sethuraman, A, Spinks, P. Q, Storey, K. B, Thane, N, Vinar, T, Zimmerman, L. M, Warren, W. C, Mardis, E. R, & Wilson, R. K. (2013) *Genome biology* **14**, R28.

[60] Kozomara, A & Griffiths-Jones, S. (2011) *Nucleic acids research* **39**, D152–7.

[61] Zhang, Z, Schwartz, S, Wagner, L, & Miller, W. (2000) *Journal of Computational Biology* **7**, 203–214.

[62] Zuker, M. (2003) *Nucleic Acids Research* **31**, 3406–3415.

[63] Landgraf, P, Rusu, M, Sheridan, R, Sewer, A, Iovino, N, Aravin, A, Pfeffer, S, Rice, A, Kamphorst, A. O, Landthaler, M, Lin, C, Socci, N. D, Hermida, L, Fulci, V, Chiaretti, S, FoÃă, R, Schliwka, J, Fuchs, U, Novosel, A, MÃijller, R.-U, Schermer, B, Bissels, U, Inman, J, Phan, Q, Chien, M, Weir, D. B, Choksi, R, Vita, G. D, Frezzetti, D, Trompeter, H.-I, Hornung, V, Teng, G, Hartmann, G, Palkovits, M, Lauro, R. D, Wernet, P, Macino, G, Rogler, C. E, Nagle, J. W, Ju, J, Papavasiliou, F. N, Benzing, T, Lichter, P, Tam, W, Brownstein, M. J, Bosio, A, Borkhardt, A, Russo, J. J, Sander, C, Zavolan, M, & Tuschl, T. (2007) *Cell* **129**, 1401–1414.

[64] Powder, K. E, Ku, Y.-C, Brugmann, S. A, Veile, R. A, Renaud, N. A, Helms, J. A, & Lovett, M. (2012) *PLoS ONE* **7**, e35111.

[65] Darnell, D. K, Kaur, S, Stanislaw, S, Konieczka, J. K, Yatskievych, T. A, & Antin, P. B. (2006) *Developmental Dynamics* **235**, 3156–3165.

[66] Wienholds, E, Kloosterman, W. P, Miska, E, Alvarez-Saavedra, E, Berezikov, E, de Bruijn, E, Horvitz, H. R, Kauppinen, S, & Plasterk, R. H. A. (2005) *Science (New York, N.Y.)* **309**, 310–1.

[67] Crawford, N. G, Faircloth, B. C, McCormack, J. E, Brumfield, R. T, Winker, K, & Glenn, T. C. (2012) *Biology Letters*.

[68] Shen, X.-X, Liang, D, Wen, J.-Z, & Zhang, P. (2011) *Molecular Biology and Evolution* **28**, 3237–3252.

[69] Chiari, Y, Cahais, V, Galtier, N, & Delsuc, F. (2012) *BMC Biology* **10**, 65.

[70] Clark, A, Hubisz, M. J, Bustamante, C. D, Williamson, S. H, & Nielsen, R. (2005) *Genome Research* **15**, 1496–1502.

[71] Ronquist, F, Teslenko, M, van der Mark, P, Ayres, D. L, Darling, A, Höhna, S, Larget, B, Liu, L, Suchard, M. A, & Huelsenbeck, J. P. (2012) *Systematic Biology* **61**, 539–542.

[72] Pritchard, C. C, Cheng, H. H, & Tewari, M. (2012) *Nature Reviews Genetics* **13**, 358–369.

[73] Gerlach, D, Kriventseva, E. V, Rahman, N, Vejnar, C. E, & Zdobnov, E. M. (2009) *Nucleic acids research* **37**, D111–7.

[74] Chen, X, Yu, X, Cai, Y, Zheng, H, Yu, D, Liu, G, Zhou, Q, Hu, S, & Hu, F. (2010) *Insect molecular biology* **19**, 799–805.

[75] Marco, A, Hui, J. H. L, Ronshaugen, M, & Griffiths-Jones, S. (2010) *Genome biology and evolution* **2**, 686–96.

[76] Ruby, J. G, Stark, A, Johnston, W. K, Kellis, M, Bartel, D. P, & Lai, E. C. (2007) *Genome research* **17**, 1850–64.

[77] de Wit, E, Linsen, S. E. V, Cuppen, E, & Berezikov, E. (2009) *Genome research* **19**, 2064–74.

[78] Friedländer, M. R, Adamidi, C, Han, T, Lebedeva, S, Isenbarger, T. A, Hirst, M, Marra, M, Nusbaum, C, Lee, W. L, Jenkin, J. C, Sánchez Alvarado, A, Kim, J. K, & Rajewsky, N. (2009) *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11546–51.

[79] Lu, Y.-C, Smielewska, M, Palakodeti, D, Lovci, M. T, Aigner, S, Yeo, G. W, & Graveley, B. R. (2009) *RNA (New York, N.Y.)* **15**, 1483–91.

[80] Xue, X, Sun, J, Zhang, Q, Wang, Z, Huang, Y, & Pan, W. (2008) *PloS one* **3**, e4034.

[81] Simões, M. C, Lee, J, Djikeng, A, Cerqueira, G. C, Zerlotini, A, da Silva-Pereira, R. A, Dalby, A. R, LoVerde, P, El-Sayed, N. M, & Oliveira, G. (2011) *BMC genomics* **12**, 47.

[82] Dai, Z, Chen, Z, Ye, H, Zhou, L, Cao, L, Wang, Y, Peng, S, & Chen, L. (2009) *Evolution & development* **11**, 41–49.

[83] Soares, A. R, Pereira, P. M, Santos, B, Egas, C. a, Gomes, A. C, Arrais, J, Oliveira, J. L, Moura, G. R, & Santos, M. A. S. (2009) *BMC genomics* **10**, 195.

[84] Li, S.-C, Chan, W.-C, Ho, M.-R, Tsai, K.-W, Hu, L.-Y, Lai, C.-H, Hsu, C.-N, Hwang, P.-P, & Lin, W.-c. (2010) *BMC genomics* **11 Suppl 4**, S8.

[85] Grimson, A, Srivastava, M, Fahey, B, Woodcroft, B. J, Chiang, H. R, King, N, Degnan, B. M, Rokhsar, D. S, & Bartel, D. P. (2008) *Nature* **455**, 1193–7.

Figure Legends

Figure 1. Comparison of phylogenetic hypotheses for each dataset: A) Annelids, B) Bilaterians, C) Animals, D) Vertebrates, and E) Amniotes. The left column is the originally published parsimony result and the right column is the maximum clade credibility tree from the stochastic Dollo re-analysis under the winning clock model. Red branches highlight topological differences between the trees, and dots on nodes signify nodal posterior probabilities for the Bayesian trees.

Figure 2. The maximum clade credibility tree for the Amniote dataset before (left) and after (right) correcting for sampling error. Red branches highlight topological differences between the trees.

Figure S1. Sequence alignment and predicted secondary structure for four microRNA families that were detected in the *Alligator* and *Chrysemys* genomes via BLAST similarity searches. The mature miRNA sequence from miRBase is underlined in the sequence and secondary structure of the reference species (*Gallus*). Substitutions relative to the reference sequence are highlighted in red. miRNA 1743 sits at the end of a contig in the *Chrysemys* genome assembly and is truncated by 10 bases on the 5' end as a result. We represent these as ambiguous bases and make no attempt to predict secondary structure in this region.

Table 1: Prevalence of miRNA loss inferred under Dollo parsimony and the stochastic Dollo model.

miRNA study	Reference	# Parsimony informative	Optimal parsimony score	# implied miRNA losses	Proportion of secondary loss	Estimated rate of miRNA loss (mean, [HPD])
Amniotes[a]	[24]	34	36	1	0.03	$1.99\times10^{-4}, [3.48\times10^{-6}, 4.75\times10^{-4}]$
Animals	[25]	115	158	43	0.27	$2.01\times10^{-4}, [4.05\times10^{-6}, 4.79\times10^{-4}]$
Annelids[b]	[27]	71	113	42	0.37	$1.99\times10^{-4}, [9.15\times10^{-6}, 4.75\times10^{-4}]$
Bilaterians	[26]	71	147	79	0.54	$2.01\times10^{-4}, [2.73\times10^{-6}, 4.82\times10^{-4}]$
Vertebrates	[18]	172	249	84	0.34	$2.04\times10^{-4}, [1.08\times10^{-5}, 4.87\times10^{-4}]$

[a]The number of implied miRNA losses calculated here (and reported in the original study) is an underestimate. The original study indicates that additional miRNAs were detected that entailed secondary losses (see supplementary Table 1, Lyson et al. [24]), but these data were excluded from the dataset.

[b]The original study for this dataset used 'standard' (Wagner) parsimony, which implies a greater degree of secondary miRNA loss.

Table 2: Marginal likelihoods of miRNA datasets under four different clock models ranging from strictly clock-like to highly variable evolutionary rates.

miRNA Study	Marginal Likelihood[a]				CV[b]
	Strict	Random local	Uncorrelated exponential	Uncorrelated lognormal	
Amniotes	-128.44 (0.06)	-127.93 (0.14)	-124.00* (0.02)	-125.22 (0.53)	0.996
Animals	-649.83 (0.15)	-639.60 (0.36)	-605.37* (0.08)	–	1.117
Annelids	-454.75 (0.16)	–	-433.34* (0.21)	–	1.105
Bilaterians	-622.88 (0.25)	-602.47 (0.89)	-583.58* (0.31)	–	1.107
Vertebrates	-1107.98 (0.17)	-1054.95 (0.06)	-1034.07* (0.17)	-1037.44 (0.38)	1.043

[a]The marginal log probability of miRNA datasets under the stochastic Dollo and (relaxed) clock models estimated using path sampling. Values are means and standard error of three independant runs. The winning models are denoted with an asterisk. Empty cells denote the model-dataset combinations for which poor MCMC mixing prevented a stable estimate of the marginal likelihood.

[b]Coefficient of Variation in evolutionary rate among branches of the phylogeny for the winning model.

Table 3: Selection of topology models (tests of phylogenetic hypotheses) for miRNA datasets based on Bayes factor comparisons of estimated marginal likelihoods.

miRNA study	Topology model[a]	$\ln P(X\|M_i)$[b]	$2 \ln BF_{ij}$[c]	Description	Resulting from[d]
Amniotes	M0	-114.98 (0.03)	17.52	Lepidosaur hypothesis: turtles + lizards	B & P
	M1	-123.74 (0.14)		Archosaur hypothesis: turtles + archosaurs	
Amniotes-corrected[e]	M0	-126.21 (0.22)	5.17	Archosaur hypothesis: turtles + archosaurs	B & P
	M1	-128.80 (0.08)		Lepidosaur hypothesis: turtles + lizards	
Animals	M0	-574.67 (0.44)	-11.69	(((Acoel 1, Acoel 2), Xenoturbella), (remaining Bilateria))	B
	M1	-568.83 (0.17)		(Acoel 1 (Acoel 2 (Xenoturbella (remaining Bilateria)))	P
Annelids	M0	-414.65 (0.11)	11.97	((Phascolosoma, Nereis), (Lumbricus, Capitella))	B
	M1	-420.64 (0.23)		(Phascolosoma (Nereis (Lumbricus, Capitella)))	P
Bilaterians	M0	-552.63 (0.02)	100.92	myzostomids sister to annelids	B
	M1	-603.09 (0.26)		myzostomids nested within annelids	P
Vertebrates	M0	-994.81 (0.14)	0.91	cyclostome hypothesis: lampreys sister to hagfish	B & P
	M1	-995.26 (0.13)		jawed vertebrate hypothesis: lampreys sister to jawed vertebrates	

[a]Topology models refer to various phylogenetic hypotheses corresponding to the description column; see text for details.

[b]The marginal log probability of miRNA datasets (and standard error) under the stochastic Dollo model and the preferred relaxed-clock model estimated using path sampling as described in the text.

[c]Two times the natural log of the Bayes factor is twice the difference between the natural log marginal likelihoods estimated under the alternative topological models. Our interpretation follows [46].

[d]An indicator for which unconstrained analysis type recovered each topological model. B= Bayesian Stochastic Dollo, P = Parsimony

[e]This is a version of the amniote miRNA dataset from the study of Lyson et al. [24] that has been corrected for sampling error; see text for details.

Table 4: Comparison of empirical and computationally derived estimates of miRNA complements for selected taxa.

Species	Number of miRNA orthologs obtained empirically[a]	Number of miRNA orthologs accessioned in: miROrtho[b]	Number of miRNA orthologs accessioned in: miRBase[c]
Apis mellifera	267 [74]	52	222
Tribolium castaneum	203 [75]	35	430
Drosophila melanogaster	148 [76]	147	426
Caenorhabditis elegans	112 [77]	130	368
Schmidtea mediterranea	122 [78]; 66 [79]	38	257
Schistosoma japonicum	227 [80]	-	78
Schistosoma mansoni	211 [81]	-	29
Petromyzon marinus	267 [15]	40	302
Branchiostoma floridae	152 [15]; 32 [82]	-	187
Saccoglossus kowalevskii	90[15]	-	115
Strongylocentrotus purpuratus	58 [15]	12	70
Danio rerio	198 [83]	113	255
Oryzias latipes	599 [84]	-	146
Nematostella vectensis	40 [85]	-	78

[a]miRNA counts in this column are derived from studies that used small RNA isolation followed by deep sequencing to estimate miRNA complements per species; see citations.

[b]miRNA counts in this column were predicted by combining orthology with a vector support machine for each sequenced genome as described in Gerlach et al. [73].

[c]miRNA counts in this column are derived from the public repository for all published miRNA sequences and includes data from small RNA sequencing and computational predictions [60].