Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Kenny NJ, Sin YW, Hayward A, Paps J, Chu KH, Hui JHL. 2015 The phylogenetic utility and functional constraint of microRNA flanking sequences. *Proc. R. Soc. B* **282**: 20142983.
http://dx.doi.org/10.1098/rspb.2014.2983

Received: 7 December 2014
Accepted: 19 January 2015

Subject Areas:
evolution, taxonomy and systematics

Keywords:
microRNA, non-coding RNA, phylogenetics, flanking sequences, evolution

Author for correspondence:
Jerome H. L. Hui
e-mail: jeromehui@cuhk.edu.hk

†Present address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA.
‡Present address: Stockholm University Zoologiska Institutionen, Stockholm, Sweden.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2014.2983 or via http://rspb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

The phylogenetic utility and functional constraint of microRNA flanking sequences

Nathan J. Kenny[1], Yung Wa Sin[1,†], Alexander Hayward[2,‡], Jordi Paps[3], Ka Hou Chu[4] and Jerome H. L. Hui[1]

[1]School of Life Sciences and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong
[2]Department of Medical Biochemistry and Microbiology, Uppsala Universitet, Uppsala, Sweden
[3]Department of Zoology, University of Oxford, Oxford, UK
[4]School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong

MicroRNAs (miRNAs) have recently risen to prominence as novel factors responsible for post-transcriptional regulation of gene expression. miRNA genes have been posited as highly conserved in the clades in which they exist. Consequently, miRNAs have been used as rare genome change characters to estimate phylogeny by tracking their gain and loss. However, their short length (21–23 bp) has limited their perceived utility in sequenced-based phylogenetic inference. Here, using reference taxa with established phylogenetic relationships, we demonstrate that miRNA sequences are of high utility in quantitative, rather than in qualitative, phylogenetic analysis. The clear orthology among miRNA genes from different species makes it straightforward to identify and align these sequences from even fragmentary datasets. We also identify significant sequence conservation in the regions directly flanking miRNA genes, and show that this too is of utility in phylogenetic analysis, as well as highlighting conserved regions that will be of interest to other fields. Employing miRNA sequences from 12 sequenced drosophilid genomes, together with a *Tribolium castaneum* outgroup, we demonstrate that this approach is robust using Bayesian and maximum-likelihood methods. The utility of these characters is further demonstrated in the rhabditid nematodes and primates. As next-generation sequencing makes it more cost-effective to sequence genomes and small RNA libraries, this methodology provides an alternative data source for phylogenetic analysis. The approach allows rapid resolution of relationships between both closely related and rapidly evolving species, and provides an additional tool for investigation of relationships within the tree of life.

1. Introduction

Similarly to protein-coding genes, microRNAs (miRNA) are transcribed inside the nucleus as long primary transcripts before further processing. The transcript, known as a pri-miRNA, forms a pre-miRNA hairpin-loop structure and migrates to the cytoplasm for further processing, leading to the production of mature 5p and 3p miRNAs [1]. Subsequently, these mature miRNAs form RNA-induced silencing complexes, which usually target the 3′UTRs of mRNA molecules, leading to target gene expression repression or translational inhibition in animals. Previous studies have suggested that once a new microRNA is incorporated into the gene regulatory network of an animal, it is usually retained and becomes difficult to lose during evolution [2–5]. Given this special property, homoplasy has been proposed to be rare, and the presence and absence of mature miRNAs have been used in recent years as rare genomic change data to clarify the phylogenetic positions of many animal phyla [6,7]. Some authors disagree on the use of miRNAs as phylogenetic characters, especially with regard to their use in previous studies [8]. However, recent work has addressed many of the points raised in criticism of miRNAs as qualitative markers [9].

Despite the rising popularity of miRNA in qualitative phylogenetics, the use of miRNA as more traditional phylogenetic markers based on raw sequences is rare,

© 2015 The Author(s) Published by the Royal Society. All rights reserved.

Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

perhaps owing to their short lengths when compared with the ribosomal RNA or protein-coding genes traditionally employed for phylogenetic inference. With the exception of one approach examining pre-miRNA sequence data [9], the utility of the sequence of non-coding RNA in general and miRNA in particular in this fashion is untested, and could differ markedly from traditional characters in its utility to be used for reliably inferring phylogeny. However, miRNA sequence data have several potential advantages for use in estimating phylogenetic relationships. They are, as previously noted, rarely lost in the course of evolution, and their mutation rate is slow compared with many protein coding genes [3]. They could therefore be useful for recovering the phylogenies of problematic groups of taxa, such as fast-evolving radiations. Additionally, homology can be identified in both genomic and small RNA library sequencing datasets [3], although proper care should be taken to avoid mischaracterization of short reads [9]. Paralogy can also be discerned from syntenic landscapes (e.g. electronic supplementary material, file S1, table S3, genomic context), and subsequent alignment is straightforward. These features allow the rapid derivation of sequence data for alignment from *de novo* builds of such resources, easing analysis considerably. Unlike many other loci, back mutation is also less likely to hide true variation, especially in mature miRNA sequences, as mutation in miRNAs is tightly constrained by fidelity to miRNA binding sites in target transcripts [10].

Here, we have successfully tested the use of concatenated miRNA sequence, concatenated miRNA flanking sequence and a combined dataset, including both miRNA and flanking region sequence in resolving the phylogeny of a group of closely related insect species. By comparing miRNA sequence from the 12 published drosophilid genomes, regions flanking the hairpin structure were found to be highly conserved in these sequences. Construction of phylogenetic trees using these regions from all miRNAs conserved across the 12 drosophilids, along with the beetle *Tribolium castaneum* as outgroup, recovered a tree topology matching that of previous work based on large concatenated nuclear alignments with strong support. This approach also proved its utility when used to reconstruct rhabditid nematode and primate phylogeny. Consequently, the sequence of miRNAs and their flanking regions represents suitable characters for phylogenetic reconstruction at the intragenus and intrafamily level, as well as potential *cis*-regulatory regions for controlling miRNA expression. Conversely, further analysis of flanking sequences in widely distributed *Drosophila melanogaster* populations showed little intraspecific variation in these regions, even in geographically distant populations, underlining the constraints placed on them compared with the wider non-coding genomic landscape. This study establishes a new approach for resolving animal species relationships, building markedly on ideas first noted in Field *et al.* [9], and suggests that the flanking sequences of miRNAs are under strong functional constraint after speciation events.

2. Material and methods

(a) miRNA identification and sequence recovery

The complete miRNA complements of all 12 sequenced *Drosophila* genomes and that of the beetle *T. castaneum* were downloaded from miRBase for comparison. Twenty-five miRNAs were found to be present as single copies in all 13 genomes, as listed in the electronic supplementary material, file S1. Using a shell script (written by the authors and available on request), 300 bp of sequence immediately up- and downstream from the stem–loop sequence of each miRNA was extracted from the genomes of all species in figure 1. For the species shown in figure 2, nine (nematode) and 10 (mammal) stem/loop sequences from miRNA present in single copy in all genomes examined were downloaded from miRBase, with 300 bp of surrounding sequence sourced from Wormbase/Ensembl's genome browsers as noted in the electronic supplementary material, file S1.

(b) Sequence alignment and phylogenetic inference

Sequences were analysed, and strand orientation was checked and corrected where necessary (see §3*a*). Genomic context for flanking regions was established using the UCSC Genome Browser, with additional FlyBase, RefSeq and GenBank tracks. miRNA gene sequences were aligned using MAFFT [14] with the Q-INS-i strategy individually gene by gene, with all 25 alignments concatenated to produce a final alignment. jModelTest [15] was used to select the best-fitted model of nucleotide substitution, the general time reversible (GTR) + I + 4G model. For figure 1*c*, the original alignment used in figure 1*a* was taken and stem/loop sequence manually excised in the Sequence Data Explorer of MEGA5.2 [16] to prevent alignment artefacts along the stem/loop excision boundary. A total alignment length of 28 952 bp (stem loop and flanking regions), 26 218 bp (flanking regions only) and 2463 bp (stem loop only) resulted from the concatenation of all data before manual exclusion of regions with one or more gaps (leaving 8176, 6071 and 1835 sites, respectively). This approach was mirrored for the species shown in figure 2 (figure 2*a*: total alignment lengths 9567/8701/1045 bp; gap-free alignments: 3265/2617/520 bp; figure 2*b*: total alignment lengths 7431/6464/987 bp; gap-free alignments: 6228/5270/762 bp). Alignments can be found in the electronic supplementary material, file S1 and sequences and alignments in alternate formats can be found in the electronic supplementary material, file S3.

PHYML 3.1 [12] was used under the GTR + I + G model (four categories) to infer a maximum-likelihood phylogeny. Only fully informative, gapless sites were used in the analysis, with alignment positions with any missing data removed by complete deletion. Bayesian analysis was performed using both MrBayes [11] and PhyloBayes [13]. MrBayes was run using the following settings: 4by4 nst = 6 rates = gamma. All MrBayes analyses were run until convergence was indicated when the average standard deviation of split frequencies was less than 0.01, after 1 000 000 generations (all and up/downstream flanking region trees) or 1 500 000 generations (stem/loop only tree). Convergence was confirmed by plotting likelihood scores against generations, and after determination of chain stationarity the initial 25% of sampled generations were discarded as 'burn-in'.

PhyloBayes was run using the pb automatic stopping rule, two chains, the CAT-GTR model, four discrete gamma categories, maximum discrepancy 0.1 and minimum effective size 100. readpb was used to discard 20% of sampled points as 'burn-in' and remaining samples were used to generate averages for display. The final consensus trees were visualized in FigTree 1.4.0. MrBayes-derived Bayesian trees are shown in figures, with ML/PhyloBayes-derived topology indicated using an asterisk at nodes and dashed lines to represent topology in the few cases where differences existed.

For partition analyses, to determine the best-fitting models and estimate likelihoods for the resulting trees, PartitionFinder [17,18] was run using PHYML, a greedy search scheme and BIC model selection.

(c) Population variability assay

Wild-collected *D. melanogaster* fly stocks were obtained from Ehime-fly, National Bio-Resource Project, Japan. Genomic DNA

Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

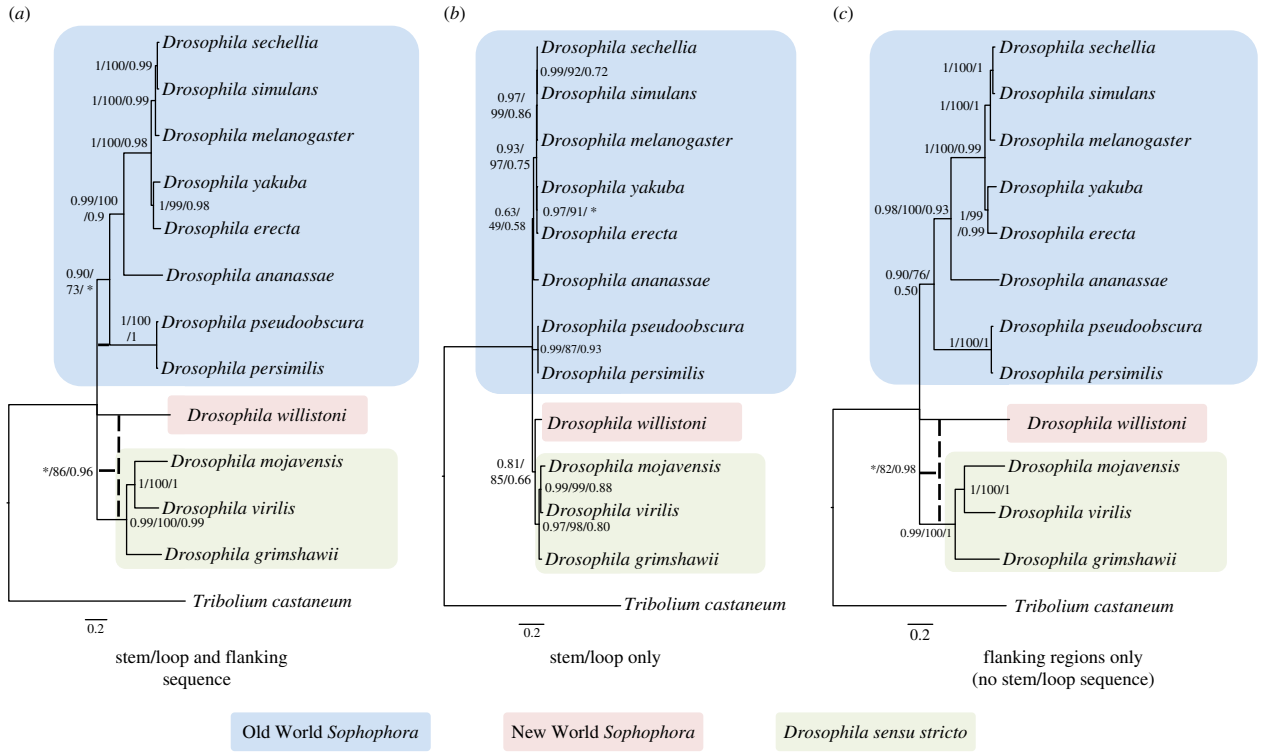

Figure 1. Phylogenetic trees show the results of inference based on alignments of (*a*) stem/loop regions together with flanking regions, (*b*) stem/loop regions and (*c*) flanking regions only. Trees shown are the result of Bayesian analysis in MrBayes [11], and where topologies differ under maximum-likelihood/PhyloBayes analysis [12,13], the topology recovered is indicated with an asterisk at the node in question and a dotted line representing the difference in topology. Numbers at the base of nodes represent posterior probabilities (MrBayes, GTR + 4G + I)/bootstrap proportions expressed as a percentage (1000 replicates)/posterior probabilities (PhyloBayes, CAT-GTR). Sequences used in phylogenetic analysis, along with alignment, can be found in the electronic supplementary material, files S1 and S3. Coloured boxes represent major drosophilid clades as indicated at the base of the figure. Scale bars represent substitutions per site.

samples were prepared using a DNeasy blood and tissue kit (Qiagen) following the manufacturer's protocol. Primers miR-993 F: CAGGACATCTGCTCG/miR-993 R: GTATACGCCGCATGGT GTTTGCC and miR-iab-4/8 F: GGCAACAAAGGGTGATTAT CG/miR-iab-4/8 R: CAAATGAAAGGCTTTCTGTG were used with Genesys Ltd. Taq polymerase and standard PCR conditions to amplify the relevant loci from samples, with resulting fragments cloned into pMD 18-T vector (Takara). Sanger sequencing was performed by BGI Hong Kong. Alignments were visualized in Jalview, coloured according to ClustalX identity [19], and haplotype maps were constructed using TCS [20].

3. Results and discussion

(a) Sequence curation and alignment

Initial assessment of miRNA conservation across the 12 sequenced drosophilid species and beetle *T. castaneum* revealed a large number of potential miRNA sequences for phylogenetic analysis. However, some species possess multiple copies of miRNAs as a result of lineage-specific duplications [21]. These paralogues, which are likely to be under varying selection pressures when compared with single-copy genes, could cause problems for phylogenetic analysis. We therefore used only the 25 miRNA genes that are found as clear single copies in these insect genomes (see the electronic supplementary material, file S1).

The stem/loop sequence and stem/loop with plus/minus 300 bp flanking region sequences of these 25 miRNAs were aligned on a miRNA-by-miRNA basis, with all 25 resulting alignments concatenated to form the final dataset for analysis. For the flanking region only analysis (figure 1*c*), the stem/loop sequences were excised leaving only the 600 bp of flanking region alignment as shown in figure 1*a*. This ensured the best possible alignment near the stem/loop region, without creating artefacts across the artificial 'boundary' that would result if the stem/loop was excised before alignment.

While the gain and loss of indels across clades can be a source of phylogenetically useful data, and although sites containing gaps can provide additional data for analysis in some software, we excluded all sites with one or more gaps from our analysis to avoid miscoding owing to alignment errors. Some miRNA genes have altered orientation in the genome relative to the inferred ancestral state. We corrected orientation in all instances, and recommend curation of data to check for their possible occurrences in future work. Where strand-specific sequencing of RNA has been performed, this task will be made especially simple, but it is generally straightforward in any case, and provides a further rare genomic change for future phylogenetic inference. Visual curation of alignments is generally sufficient to correct errors in this process, as stem/loop sequence is relatively symmetrical, but flanking regions differ greatly from the stem/loop containing the 5p and 3p mature miRNA sequences.

(b) Phylogenetic inference

Drosophilid trees recovered by MrBayes-based Bayesian analyses (GTR + 4G + I model) are presented in figure 1, with ML(GTR + 4G + I)/PhyloBayes (CAT-GTR model) topology indicated with dashed lines where differences exist. All the trees shown in figure 1 recover drosophilid phylogeny reliably as presented by Seetharam & Stuart [22]. Slight

Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

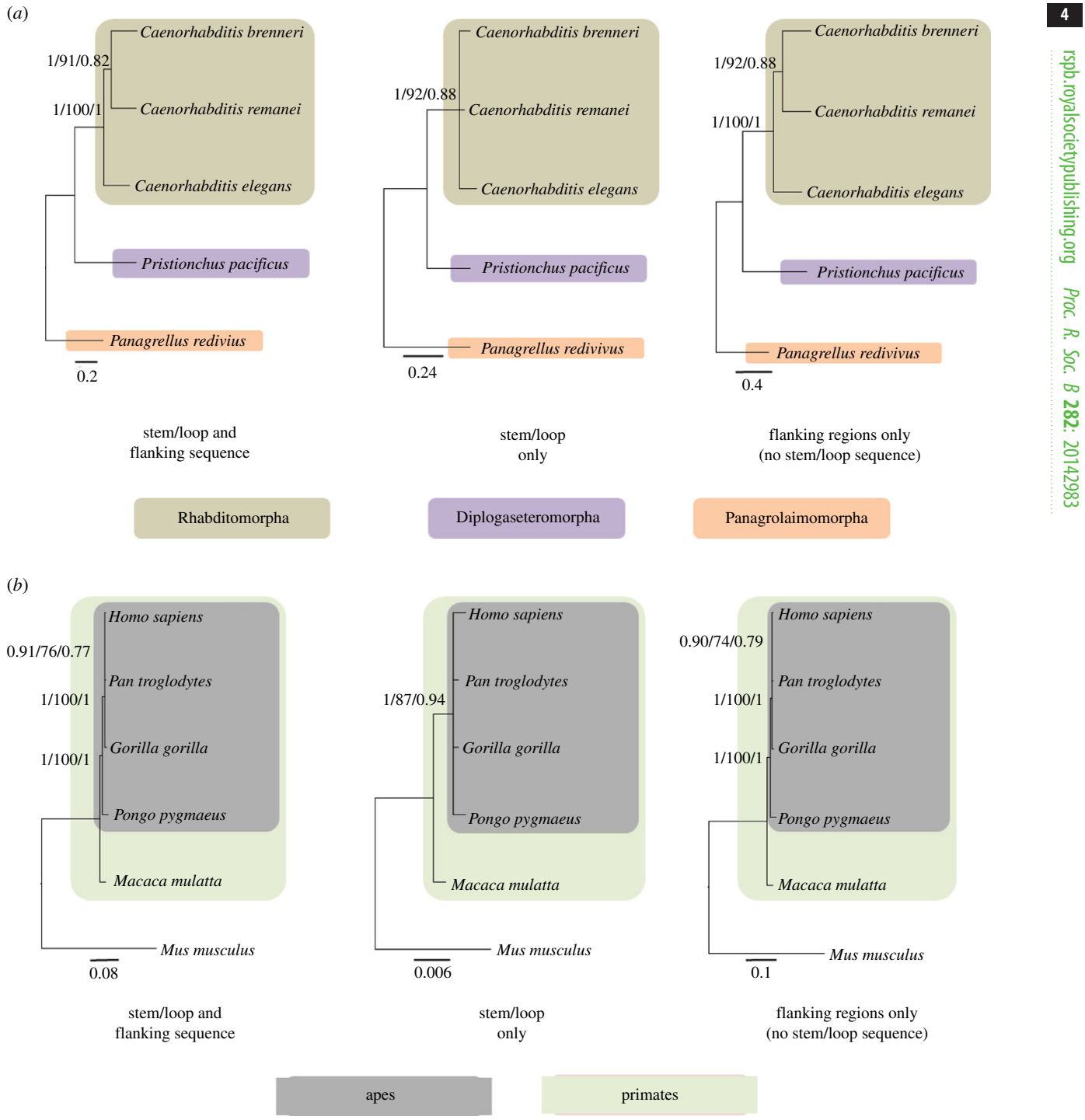

Figure 2. Phylogenetic trees showing relationships of (*a*) rhabditid nematodes and (*b*) primates (with mouse outgroup), as recovered using alignments of stem/loop regions and flanking regions considered together (left), of stem/loop regions alone (centre) and flanking regions only (right). MrBayes-derived trees shown [11]. Numbers at base of nodes represent posterior probabilities (MrBayes, GTR + 4G + I)/bootstrap proportions expressed as a percentage (1000 replicates) [12]/posterior probabilities (PhyloBayes, [13], CAT-GTR). Alignments used in phylogenetic analysis, along with sequences, can be found in the electronic supplementary material, file S3. Coloured boxes represent clades as indicated. Scale bars represent substitutions per site at given unit distance.

differences include the exact position of *D. willistoni*, which is possibly the sister group of the *Drosophila* subgenus, although this remains to be confirmed [23]. This species's position is not recovered consistently in our trees, but collapses to a polytomy with Old World *Sophophora* and *Drosophila sensu stricto* in some MrBayes-recovered trees. This may arise as a consequence of a long-branch effect, and discerning the true position of this clade would be aided by greater taxon sampling, as New World drosophilids are markedly under-sampled compared with their Old World and sophophorid counterparts. The topology recovered by PhyloBayes (under the CAT-GTR model, which is robust to long-branch attraction) suggests that it is, indeed, more closely related to the *Drosophila* subgenus than to the Old World *Sophophora*. The monophyly of the *D. sechellia, simulans, melanogaster, yakuba, erecta* and *ananassae* clade is correctly recovered but has relatively weak support (0.63/49/0.58) in our stem/loop only phylogeny (figure 1*b*)—these support values are more robust in the other two phylogenies, where more residues are available for bootstrap sampling. The 'obscura' group (*D. pseudoobscura* and *persimilis*) and the *Drosophila* subgenus are always recovered with almost maximal support.

Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

Table 1. Results of partitioning analysis, performed in PartitionFinder [17,18] using PHYML, a greedy search scheme and BIC model selection. All models implemented in PartitionFinder were tested for best fit with the exception of the GTR + G + I individual partition test (final row).

partitions	no of partitions	models selected	log-likelihood: (2 dp)	BIC (2 dp)
none	1	GTR + G + I	−64 614.40	129 312.66
all flanking regions versus all stem/loop sequences	2	TrN + G—flanking regions HKY + G—stem/loop regions	−63 540.29	127 395.89
mir-by-mir (each mir flanking and stem/loop seq together)	25	HKY + I + G—12 (all mir not listed below) HKY + G—12 (mir 11, mir 13a, mir 277, mir 283, mir 33, mir 92a, mir 92b, mir 14, mir 2c, mir 7, mir bantam, mir iab4) TrNef + G—1 (mir 308)	−64 384.35	129 174.10
each mir flanking and stem/loop sequence considered individually	50	HKY + G—40 (all partitions not named below) K80—5 (mir 124 stemloop, mir 14 stemloop, mir 277 stemloop, mir 283 stemloop, mir 34 stemloop) TrNef + G—4 (mir 10 stemloop, mir 275 stemloop, mir 303 stemloop, mir 308 flanking) JC—1 (mir iab4 stemloop)	−62 959.80	126 451.12
each mir flanking and stem/loop sequence considered individually	50	GTR + G + I (partitions considered individually)	−62 976.34	126 646.36

All other nodes are supported with high (0.9/>90) posterior probability and bootstrap values respectively, giving us firm confidence in the utility of this method. We gain consistent and strong topological, posterior probability and bootstrap support for the same phylogeny and branching order for all other species, suggesting that phylogenetic signal is consistent throughout all regions of the alignment. Furthermore, and vitally for ongoing work using these phylogenetic characters, phylogenetic relationships are reliably recovered by both ML and Bayesian methods of analysis. This confirms the utility of these characters and demonstrates this means of inferring phylogeny is robust to the application of different algorithms of phylogenetic reconstruction. We suggest the use of both stem/loop and flanking sequence, as trees using both sources of data provide marginally better resolution of relationships in our study.

To confirm the broader utility of these markers in the resolution of phylogeny, we have investigated both rhabditid nematode (figure 2*a*) and primate (figure 2*b*) phylogeny using the same methodology as used for the broader drosophilid example above. In both cases, we robustly recovered the well-catalogued inter-relationships of these organisms, with the best resolution and best supported trees resulting from the use of both flanking and stem/loop sequence—in the case of primates, stem/loop only sequence is so well conserved that there is little resolution in the tree using only that data. Even in the case of rhabditid nematodes, a notably fast-evolving clade, phylogeny is recovered using a small sample of miRNA sequences, and in the case of primates, where a deliberately restricted sample of conserved miRNA was chosen for investigation (10 conserved miRNA genes found in single copy, of the more than 25 catalogued examples of these), the true phylogeny was nonetheless recovered. We therefore believe both fast-evolving and slow-evolving clades can be investigated using these characters, although in slow-evolving clades the use of flanking sequence as well as stem/loop sequence is necessary.

While sufficient signal can be gained from as few as nine genes for the robust resolution of phylogenetic inter-relationships, in other taxa, additional data may be required to separate nodes. Many programs commonly used for phylogenetic inference can account for missing data in alignments, and many genes (for example, *miR-12* or *miR-29/285* in the insect species shown here) are secondarily absent in only one taxon in any given sample of species. Such programs could therefore be used in concert with wider, albeit 'gappier' samples of miRNA genes, providing additional phylogenetic signal in more recalcitrant cases. Furthermore, the ease of identification of miRNA sequences [3] means that less well-catalogued genomes, or even orphan sequences, such as those found in the NCBI Trace Archive, could be used to provide raw material for phylogenetic inference. Such possibilities are beyond the scope of this study, but will be of great interest to biologists working on less well-sequenced organisms.

(c) Partitioned sequence analysis

To evaluate the degree of diversity between the molecular signals at the different miRNA loci that were used for phylogenetic analysis, PartitionFinder [17,18] was used to evaluate the best-fitted model of molecular evolution for a variety of partitioned forms of our overall dataset. Table 1 summarizes these data, showing the models of molecular substitution chosen for *D. melanogaster* and *T. castaneum* data when partitioned by location (flanking versus stem/loop), by miRNA, and by miRNA with flanking and stem/loop data considered separately.

With no partitions, the GTR + 4G + I model was used to construct an initial tree with a log-likelihood of −64 614.40 (to 2 decimal points (dp)). This was improved upon even with

Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

the addition of a single partition, and when flanking regions and stem/loop regions were considered separately under the Tamura–Nei (TN) and Hasegawa–Kishino–Yano (HKY) models the log-likelihood of the best tree was calculated to be −63 540.29 (2dp).

The addition of extra partitions generally continued to raise the calculated log-likelihood of the resulting tree, with a 50 partition dataset calculated to have a likelihood of −62 959.80 (2 dp). This is itself unsurprising—when additional partitions are added, the number of parameters in the model increases, and thus the likelihood score is likely to improve simply because the model is more complex and therefore provides better fit with the data. However, the log-likelihood of the miRNA-by-miRNA analysis (−64 384.35) is markedly lower than the likelihood of the simplest of partitioned models where flanking regions and stem/loop sequences are considered separately as only two separate partitions (−63 540.29)—this indicates that flanking regions and stem/loop regions are better modelled apart, as they are under slightly different evolutionary pressures. This is suggested further by comparison of BIC values—at 127 395.89 for the two partition scheme, versus 129 174.10 for 25 partitions, partitioning stem/loop and flanking regions separately is improved using to miRNA-by-miRNA approaches. BIC values continue to improve when individual miRNA stem and flanking regions are themselves partitioned. When possible, the addition of extra partitions, particularly those that split flanking and stem/loop sequence, clearly allows more correct modelling of molecular evolution for miRNA alignment data, although we note no changes in tree topology result from the implementation of partitioned models in our experimentation when compared with figure 1.

On a miRNA-by-miRNA basis, the HKY model is almost universally that which provides the best fit, with 12 miRNAs best modelled by HKY + G + I and 12 by HKY + G (no invariant sites). Only miR-308, modelled by TrNef + G, differs from the other loci used in the present phylogenetic analysis. This is remarkably consistent for a diverse set of genetic markers involved in a variety of molecular processes.

The miRNA stem/loop and flanking sites, when split apart on a by-gene basis, remained remarkably homogeneous in the best-fit model. Of 50 flanking and stem/loop regions considered separately as partitions, 40 were best modelled using the HKY + G model of molecular evolution. Flanking regions in particular are similar, with only miR-308's flanking region best fitted to an alternative model of molecular evolution (TrNef + G). Why flanking regions, when considered as an 'all flanking region' dataset, were best modelled with the TrN + G model while individually they were almost universally best modelled with HKY + G seems to be the result of a very slight contribution from the miR-308 sequence—HKY + G was the second-best performing model for the 'all flanking' dataset (BIC: TrN + G 110 493.76 versus HKY + G 110 497.96, 2 dp).

This suggests that the flanking regions surrounding a diverse set of miRNA sequences are under relatively consistent evolutionary pressure; indeed, slightly more consistent pressure than that of the miRNA stem/loop sequences they flank. This further supports the status of both miRNA and flanking region sequence as useful sources of phylogenetic character information for further analysis—contrary to protein-coding genes, which may be under a diverse range of pressures depending on their functional roles and cellular locations, miRNA flanking sequences seem to be relatively consistent in their molecular signal.

Partitioning undoubtedly results in more adequate models of molecular evolution across concatenated miRNA sequence trees. When the correct model for each partition is unknown, GTR + G + I adequately models partitioned sequences. In the example shown here, performing analyses with each partition under the GTR + G + I model results in a very similar log-likelihood to individually selected models for each partition (−62 976.34 versus −62 959.80). As computational power has markedly increased in availability in recent years, running the GTR + G + I model on individual partitions seems an adequate compromise when further information is not available.

(d) Robustness to population-level variation

As a further test of our method's robustness to naturally existing variation, we assayed *D. melanogaster* samples drawn from different parts of the world to check the diversity of sequences at such loci. Numerous studies have underlined potential variability at the miRNA gene level, with medical and developmental consequences [24–26]. Any great variability in miRNA stem/loop and flanking sequences between drosophilid intraspecies populations would potentially weaken our results.

Loci surrounding two miRNA, miR-993 and miR-iab-4/8, the latter of which is shared ancestrally and was part of the alignment used for phylogenetic analysis, were cloned and sequenced from *D. melanogaster* samples taken from across their worldwide distribution. The alignment of these sequenced loci, along with strain number, collection site and collection year, are provided in the electronic supplementary material, file S2*a*,*b*. Very little variability was seen between samples from different collection sites, with no differences whatsoever observed in stem/loop sequence. With the exception of 6/7 bp indels (miR-993 at residue 642 in Shanghai strain, and miR-iab-4 in Zambia and Tokyo strains), maximum differences of 3 bp are observed between samples and the consensus over 700–800 bp of sequence. This suggests that these loci are highly constrained, even in widely separated, fast-evolving lineages.

Analysis of haplotype diversity (electronic supplementary material, file S2c,d) suggests that this sequence conservation may limit the suitability of miRNA for biogeographic inference. The limited number of extant mutations and relative stochasticity of topological relationships that result from the mapping of this data (electronic supplementary material, file S2b) markedly differ from that of the 'out of Africa' habitat expansion known to have occurred in the *D. melanogaster* lineage [27]. We therefore do not recommend the use of miRNA-linked traits for biogeographic inference. However, these results suggest that even such rapidly evolving and widely spread species as drosophilids are subjected to strong stabilizing selection at miRNA loci. This further implies that miRNA can provide excellent complementary means for quantitative phylogenetic inference in a range of contexts.

(e) Functional constraints outside the miRNA stem/loop region

Regions flanking the miRNA stem/loop sequences show a high degree of conservation among species. For example, figure 3 shows an alignment for the regions flanking miR-133. To confirm that conserved flanking sequences shown here do not represent exonic fragments of surrounding genes, we investigated their genomic context (electronic supplementary material,

Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

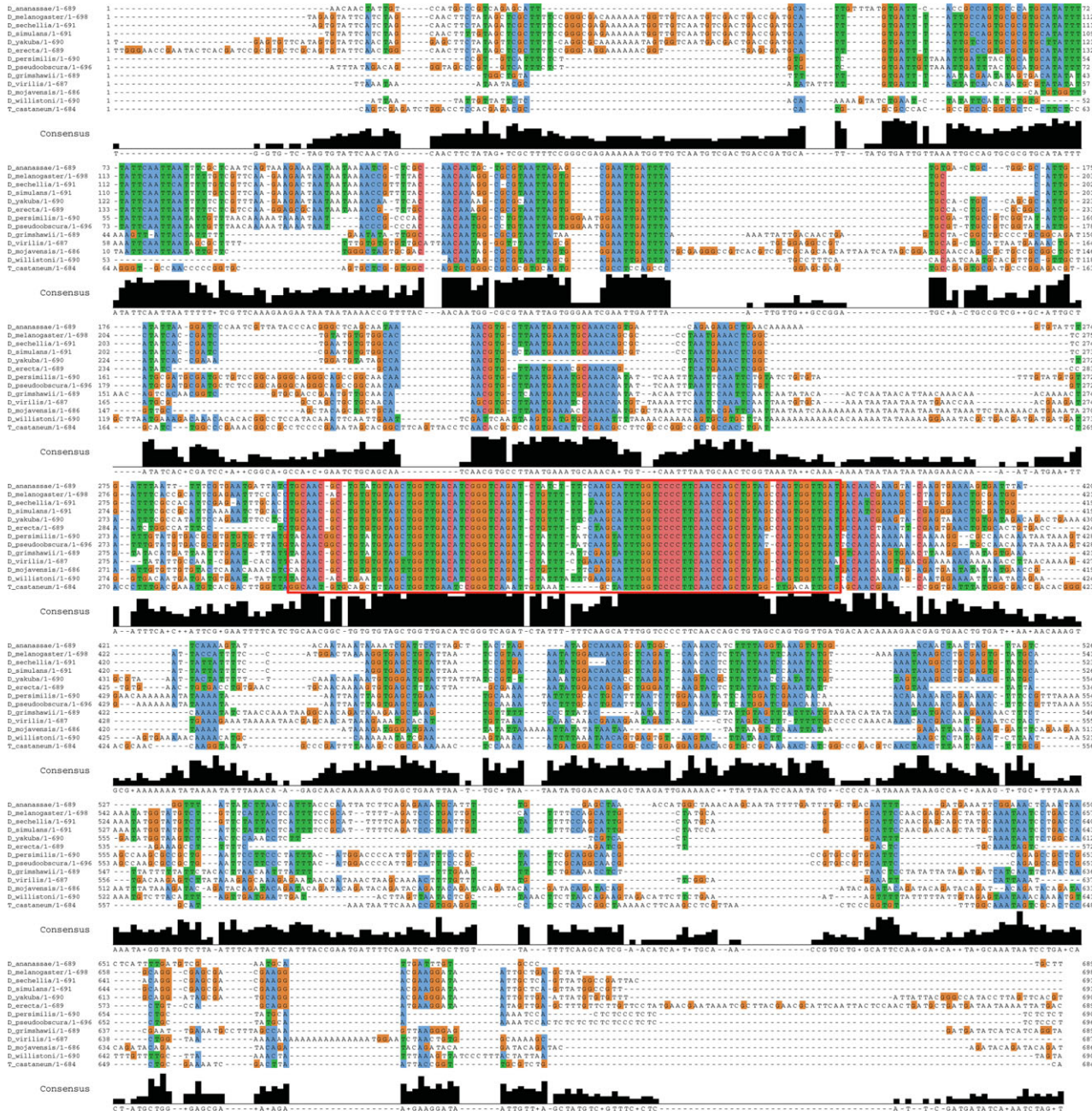

Figure 3. Alignment of miR-133 stem loop and flanking sequence. Boxed in red is the miR-133 stem/loop sequence, with boundaries taken from the *D. ananassae* miR-133 gene. This alignment shows the high levels of conservation found surrounding miRNA genes.

file S1). The majority of flanking sequences used in our analyses are composed of non-coding intergenic DNA, suggesting that conservation of these hairpin-loop flanking sequences is independent of either the presence of exonic sequence or protein-coding gene regions.

Aspects of the regulation of miRNA genes have already been the subject of some investigation, with findings suggesting developmentally and clinically relevant effects [28,29]. Given the high degree of conservation of miRNA flanking sequence among species, it is possible that flanking sequence may play an important regulatory role for controlling expression of miRNAs. Previous studies have shown that the expression levels of 5p and 3p miRNAs of homologous miRNAs differ across developmental stages and species [1,28]. In the case of some miRNAs, such as miR-10 in *T. castaneum* and *D. melanogaster*, the dominance of 5p and 3p arms differ even when they have identical duplex sequences, suggesting arm usage is encoded in the primary miRNA sequence. As 5p and 3p miRNAs target different mRNAs, changes in their expression levels/dominance provide another way for gene regulation to evolve in animals, termed 'arm switching'. The highly conserved regions identified in this study may represent new targets for the investigation of *cis*-regulation of miRNA arm usage, which provides a mechanism by which the function of a miRNA locus and its target gene network can evolve.

4. Conclusion

Both miRNAs and their flanking sequences provide phylogenetic signals suitable for the inference of phylogeny with high levels of accuracy, when sufficient numbers of this type are concatenated. As detailed here, the clear identity and easy alignment of these sequences makes them good candidates for estimating phylogeny, and they can reliably be found and identified across all members of a clade of interest.

Downloaded from http://rspb.royalsocietypublishing.org/ on February 18, 2015

Their relatively slow evolution [3] also means that they can easily be identified in *de novo* assemblies of genomes. Such alignments exhibit strong conservation across populations, which can add utility for inference of relationship above the species level, but limits the use of miRNA sequences for biogeographic inference. Despite historic issues regarding their use for phylogenetic inference [8], miRNAs can be employed as both qualitative [9] and quantitative markers, with the latter demonstrated clearly here. Our investigation demonstrates the utility of miRNA sequences as classical phylogenetic markers, and shows this usage is robust to different algorithms of phylogenetic analysis and the analysis of fast-evolving lineages. Such a method provides novel characters for assessing phylogenetic relationships that will be of use in a range of contexts for resolving branches across the tree of life.

Data accessibility. All data used in this manuscript are available as supplementary files to this manuscript and available at Dryad: http://dx.doi.org/10.5061/dryad.49q24.

Acknowledgements. The authors are grateful for the constructive suggestions by the editor and the three anonymous referees.

Funding statement. This work was supported by a Direct Grant (4053034) from the Research Committee, The Chinese University of Hong Kong, to J.H.L.H.

References

1. Griffiths-Jones S, Hui JH, Marco A, Ronshaugen M. 2011 MicroRNA evolution by arm switching. *EMBO Rep.* **12**, 172–177. (doi:10.1038/embor.2010.191)
2. Peterson KJ, Dietrich MR, McPeek MA. 2009 MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *Bioessays* **31**, 736–747. (doi:10.1002/bies.200900033)
3. Tarver JE, Sperling EA, Nailor A, Heimberg AM, Robinson JM, King BL, Pisani D, Donoghue PCJ, Peterson KJ. 2013 miRNAs: small genes with big potential in metazoan phylogenetics. *Mol. Biol. Evol.* **30**, 2369–2382. (doi:10.1093/molbev/mst133)
4. Heimberg AM, Sempere LF, Moy VN, Donoghue PCJ, Peterson KJ. 2008 MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl Acad. Sci. USA* **105**, 2946–2950. (doi:10.1073/pnas.0712259105)
5. 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. (doi:10.1111/j.1525-142X.2008.00302.x)
6. Heimberg AM, Cowper-Sal 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. USA* **107**, 19 379–19 383. (doi:10.1073/pnas.1010350107)
7. Campbell LI *et al*. 2011 MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc. Natl Acad. Sci. USA* **108**, 5690–5695. (doi:10.1073/pnas.1012675108)
8. Thomson RC, Plachetzki DC, Mahler DL, Moore BR. 2014 A critical appraisal of the use of microRNA data in phylogenetics. *Proc. Natl Acad. Sci. USA* **111**, E3659–E3668. (doi:10.1073/pnas.1407207111)
9. Field DJ, Gauthier JA, King BL, Pisani D, Lyson TR, Peterson KJ. 2014 Toward consilience in reptile phylogeny: miRNAs support an Archosaur, not Lepidosaur, affinity for turtles. *Evol. Dev.* **16**, 189–196. (doi:10.1111/ede.12081)
10. Barbash S, Shifman S, Soreq H. 2014 Global coevolution of human microRNAs and their target genes. *Mol. Biol. Evol.* **31**, 1237–1247. (doi:10.1093/molbev/msu090)
11. Ronquist F, Huelsenbeck J. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)
12. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. (doi:10.1093/sysbio/syq010)
13. Lartillot N, Lepage T, Blanquart S. 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288. (doi:10.1093/bioinformatics/btp368)
14. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. (doi:10.1093/molbev/mst010)
15. Darriba D, Taboada GL, Doallo R, Posada D. 2012 jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772. (doi:10.1038/nmeth.2109)
16. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011 MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739. (doi:10.1093/molbev/msr121)
17. Lanfear R, Calcott B, Ho SY, Guindon S. 2012 PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701. (doi:10.1093/molbev/mss020)
18. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014 Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.* **14**, 82. (doi:10.1186/1471-2148-14-82)
19. Clamp M, Cuff J, Searle SM, Barton GJ. 2004 The jalview java alignment editor. *Bioinformatics* **20**, 426–427. (doi:10.1093/bioinformatics/btg430)
20. Clement M, Posada D, Crandall K. 2000 TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1660. (doi:10.1046/j.1365-294x.2000.01020.x)
21. Lu J, Fu Y, Kumar S, Shen Y, Zeng K, Xu A, Carthew R, Wu CI. 2008 Adaptive evolution of newly emerged microRNA genes in *Drosophila*. *Mol. Biol. Evol.* **25**, 929–938. (doi:10.1093/molbev/msn040)
22. Seetharam A, Stuart G. 2012 Whole genome phylogenies for multiple *Drosophila* species. *BMC Res. Notes* **5**, 670. (doi:10.1186/1756-0500-5-670)
23. Obbard DJ, Maclennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. 2012 Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol. Biol. Evol.* **29**, 3459–3473. (doi:10.1093/molbev/mss150)
24. Saunders MA, Liang H, Li WH. 2007 Human polymorphism at microRNAs and microRNA target sites. *Proc. Natl Acad. Sci. USA* **104**, 3300–3305. (doi:10.1073/pnas.0611347104)
25. Carbonell J *et al*. 2012 A map of human microRNA variation uncovers unexpectedly high levels of variability. *Genome Med.* **4**, 62. (doi:10.1186/gm363)
26. Linhares JJ, Azevedo M, Siufi AA, de Carvalho CV, Wolgien MDCGM, Noronha EC, Bonetti TCdS, da Silva IDCG. 2012 Evaluation of single nucleotide polymorphisms in microRNAs (hsa-miR-196a2 rs11614913 C/T) from Brazilian women with breast cancer. *BMC Med. Genet.* **13**, 119. (doi:10.1186/1471-2350-13-119)
27. Nunes MD, Neumeier H, Schlötterer C. 2008 Contrasting patterns of natural variation in global *Drosophila melanogaster* populations. *Mol. Ecol.* **17**, 4470–4479. (doi:10.1111/j.1365-294X.2008.03944.x)
28. 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 microRNA locus in insects. *Nucleic Acids Res.* **41**, 3352–3361. (doi:10.1093/nar/gks1445)
29. Lu J, Clark AG. 2012 Impact of microRNA regulation on variation in human gene expression. *Genome Res.* **22**, 1243–1254. (doi:10.1101/gr.132514.111)
30. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008 miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**(Suppl. 1), D154–D158.