Adaptive Evolution of Newly Emerged Micro-RNA Genes in *Drosophila*

Jian Lu, Yonggui Fu,† Supriya Kumar,* Yang Shen,† Kai Zeng,† Anlong Xu, Richard Carthew,‡ and Chung-I Wu*†*

*Department of Ecology and Evolution, University of Chicago; †State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen (Zhongshan) University, Guangzhou, People's Republic of China; and ‡Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University

How often micro-RNA (miRNA) genes emerged and how fast they evolved soon after their emergence are some of the central questions in the evolution of miRNAs. Because most known miRNA genes are ancient and highly conserved, these questions can be best answered by identifying newly emerged miRNA genes. Among the 78 miRNA genes in *Drosophila* reported before 2007, only 5 are confirmed to be newly emerged in the genus (although many more can be found in the newly reported data set; e.g., Ruby et al. 2007; Stark et al. 2007; Lu et al. 2008). These new miRNA genes have undergone numerous changes, even in the normally invariant mature sequences. Four of them (the *miR-310/311/312/313* cluster, denoted *miR-310*s) were duplicated from other conserved miRNA genes. The fifth one (*miR-303*) appears to be a very young gene, originating de novo from a non-miRNA sequence recently. We sequenced these 5 miRNA genes and their neighboring regions from a worldwide collection of *Drosophila melanogaster* lines. The levels of divergence and polymorphism in these miRNA genes, vis-à-vis those of the neighboring DNA sequences, suggest that these 5 genes are evolving adaptively. Furthermore, the polymorphism pattern of *miR-310*s in *D. melanogaster* is indicative of hitchhiking under positive selection. Thus, a large number of adaptive changes over a long period of time may be essential for the evolution of newly emerged miRNA genes.

Introduction

Micro-RNAs (miRNAs) are small endogenously expressed single-stranded RNAs, about 22 nt long, that regulate mRNAs posttranscriptionally (reviewed in Bartel 2004; Kim 2005; Zamore and Haley 2005). Each miRNA is derived from a larger stem-loop (hairpin) structure, 70–90 nt in length in animals and longer in plants. Here, the larger hairpin structure will be referred to as the miRNA gene and the mature 22-nt product will be designated as miR. In animals, a miR binds to the 3′UTR of complementary target transcripts, causing degradation (Bagga et al. 2005) or translation repression (Olsen and Ambros 1999; Doench et al. 2003). Because the most crucial part for miR:mRNA matching is only 7 nt long in animals (positions 2–8 on the miR, or the "seed," see Lewis et al. 2003; Lim et al. 2003), the number of potential targets for each miR is often more than a hundred (Lewis et al. 2003; Brennecke et al. 2005; Grun et al. 2005; Lewis et al. 2005; Rajewsky 2006; Wang 2006).

Because miRs interact broadly with many transcripts, the evolution of such a system is intriguing. The main evolutionary questions concern the origin of miRNA genes and their subsequent evolution. Where do the new miRNA genes come from? How often do they emerge? Do the new ones evolve faster than the older, more established, miRNA genes? If so, do many changes occur during the early life of a miRNA gene before it becomes fully integrated into the regulatory network? How often are these changes adaptive?

To answer these questions, we need to identify newly emerged miRNA genes. The recently finished genome sequences of 12 *Drosophila* species provide a means to analyze sequence evolution in miRNA genes (Drosophila Comparative Genome Sequencing and Analysis Consortium 2007). Their phylogenetic relationships are given in supplementary figure S1 (Supplementary Material online). In addition to *Drosophila melanogaster*, 6 of these species are of particular relevance to this report. (The rest provides redundant divergence information due to their phylogenetic positions.) These species and their approximate time of divergence from *D. melanogaster* are as follows: *Drosophila simulans* (5 Myr), *Drosophila yakuba* (11 Myr; both species are in the *D. melanogaster* subgroup), *Drosophila ananassae* (53 Myr), *Drosophila pseudoobscura* (55 Myr), *Drosophila willistoni* (63 Myr), and *Drosophila virilis* (63 Myr). The divergence time were taken from Tamura et al. (2004). Whenever necessary, we further used the genomic sequences from 2 species outside of *Drosophila*—mosquito and honeybee for comparison.

In this study, we surveyed the 78 miRNAs of *D. melanogaster* reported before 2007 in order to identify newly emerged miRNAs. We then analyzed the newly emerged miRNA genes for their origins and subsequent evolution. The analysis presented here may be pertinent to the study of most newly discovered miRNAs, many of which are likely to have emerged recently (e.g., Bentwich et al. 2005; Berezikov et al. 2006; Kloosterman et al. 2006; Ruby et al. 2007; Stark et al. 2007; Lu et al. 2008).

Materials and Methods

Genomic Sequences and Analyses

The sequences and genomic coordinate information of the 78 miRNA genes of *D. melanogaster* and 73 miRNA genes of *D. pseudoobscura* were downloaded from miRBase V8.0 (http://microrna.sanger.ac.uk/sequences/). The genomic sequences of the 12 *Drosophila* species were downloaded from the UCSC genome browser (http://genome.ucsc.edu). The miRNA gene sequence flanked with 500 bp at each side of *D. melanogaster* were used to Blast against the 11 insect genomes with *E*-cutoff value of 10^{-3} (Altschul et al. 1997). The homologous sequences of the 11 insects were extracted and Blast against *D. melanogaster* genome to obtain the reciprocal best Blast alignment. The orthologous sequences were aligned using the ClustalW1.83 program (Thompson et al. 1994). The

Key words: adaptive evolution, micro-RNAs, *Drosophila*, new gene.

E-mail: ciwu@uchicago.edu.

Mol. Biol. Evol. 25(5):929–938. 2008
doi:10.1093/molbev/msn040
Advance Access publication February 22, 2008

alignments of the 78 miRNA genes from the whole genome alignments of the 12 *Drosophila* species (http://genome.ucsc.edu) were also parsed out. The miRNA gene sequences from the 2 sources were pooled, and the synteny information (i.e., the location of miRNA genes relative to other genes) was used when discrepancy occurred between the 2 data sets. In case a miRNA sequence could not be identified from either database, the trace data of the original sequencing results of the 11 species (except *D. melanogaster*) were also used. All the results were manually checked whenever necessary.

The coding sequences (CDS) alignments of *D. melanogaster* and the other 11 species were downloaded from http://rana.lbl.gov/~venky/AAA/freeze_20061030/protein_coding_gene/GLEANR/alignment/. The Li (1993) method was used to calculate the K_s (synonymous substitutions per site) and K_a (nonsynonymous substitutions per site) of the protein-coding regions, and the Kimura's (1980) 2-parameter method was used to calculate the divergence of miRNA sequences.

The divergence statistics between *D. melanogaster* and the other 6 species (*D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, and *D. virilis*) were provided in the text while information about the other comparisons are available upon requests from authors.

The EvoNC program (Wong and Nielsen 2004) was used to detect signal of positive selection on the newly emerged miRNA genes. The precursors of *miR-310*s cluster and CDS alignments of the flanking genes (CG13432 and CG13434) across *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. ananassae* were tested for the neutral (M1) and positive selection (M3) models. The 2 times log likelihood values were calculated and tested for statistical significance using a χ^2 test with degrees of freedom = 2. For *miR-303*, the precursor and CDS alignments of the flanking genes (CG3620 and CG3626) between *D. melanogaster* and *D. simulans* were used for the tests.

Population Genetics Analysis

The polymorphism data of *D. simulans* were taken from the Drosophila Population Genomics Project (Begun et al. 2007). Briefly, the contig sequences and quality scores of 7 strains of *D. simulans* (sim6, simw50, MD106TS, NEWC48, C167, MD199S, and SIM4) were downloaded from http://genome.wustl.edu/pub/organism/Invertebrates. Any bases with quality score <30 were masked from further analysis. Genomic sequences of each *D. simulans* strain were aligned with the genome sequences of *D. melanogaster* using BlastZ (Schwartz et al. 2003) and axtBest (Kent et al. 2003). For each site in the miRNA or protein-coding gene, 4 strains of *D. simulans* were randomly selected.

In all, 27 strains of *D. melanogaster* were sequenced for *miR-310*s cluster and *miR-303*. The geographic location of origins of fly materials were implemented in supplementary table S1 (Supplementary Material online). The *miR-310*s cluster was amplified and sequenced using primers 5′-AGCTGTCATCTCGCTCACACCTA-3′ and 5′-ACTGTCATCCCGCTCTAAACCTC-3′, and *miR-303* was amplified and sequenced using primers 5′-ACAGAAACTGCATTCCCCGAAC-3′ and 5′-TGTCCAGGATCTAACATGATTTCG-3′. The polymorphism data were polarized using *D. simulans* sequence. The expected frequency spectrum of *miR-310*s cluster were calculated using the formula $\theta = S \Big/ \sum_{i=1}^{26} 1/i$ (Fu 1995), where S is observed number of segregating sites and i is the number of new mutations with frequency 1–26. The D test of Tajima and H tests of Fay and Wu were performed using the method as implemented in http://www.genetics.wustl.edu/jflab/htest.html. The *miR-310*s cluster is on band 57A and the estimated recombination rate is $4Nc$ = 45/kb (Comeron et al. 1999). Back-mutation rate is assumed 0.05. The simulation was performed 10 000 times. For *miR-303*, high-quality sequencing data were obtained in 25 lines of *D. melanogaster*. The polymorphism data were analyzed using the same methods as that of *miR-310*s.

The polymorphism data of *D. melanogaster* were deposited into GenBank with accession number DQ854750–DQ854801.

The miRNA Secondary Structure Prediction

The miRNA secondary structures were predicted using RNAfold (Hofacker et al. 1994) and mfold (Zuker 2003) program.

Detecting Expression of *miR-303* Using Reverse Transcriptase–Polymerase Chain Reaction

Total RNA from whole males or heads of the Iso-1 line of *D. melanogaster* and the sim6 strain of *D. simulans* was extracted using Trizol reagent (Invitrogen, Carlsbad, CA), according to the manufacturer's instructions, except that the 70% ethanol wash step was omitted. Small RNAs between 26 and 18 nt in length were isolated by elution from a 15% denaturing polyacrylamide gel. Adapters were ligated to the small RNAs using the method of Lau et al. (2001). After reverse transcriptase–polymerase chain reaction (RT-PCR) with primers specific to the 5′ and 3′ adapters, the PCR was diluted 164 times, and a second round of PCR was carried out using 0.1 μM primer specific to the 5′ adapter and 1uM primer specific to *miR-303* of each species. The control reaction involved only the primer specific to the 5′ adapter. The reactions were run on a 5% MetaPhor agarose gel (Cambrex Biosciences, Rockland, ME) containing Gel Star (Cambrex Biosciences, Rockland, ME); the 38-bp product of *D. melanogaster* was cloned and sequenced to verify that it was indeed *miR-303*.

Results

Newly Emerged versus ''Old miRNA Genes''

To identify new miRNA genes among the 78 known ones, we define those that are present in all the major branches of *Drosophila* species as "old miRNA genes." Newly emerged miRNA genes, on the other hand, cannot be defined simply by their absence in some species because "failure to find" is not the same as "absence." We thus define a newly emerged miRNA as one for which the source/ancestral sequence can be positively identified, and this source sequence can be analyzed for their miRNA properties (or lack of).

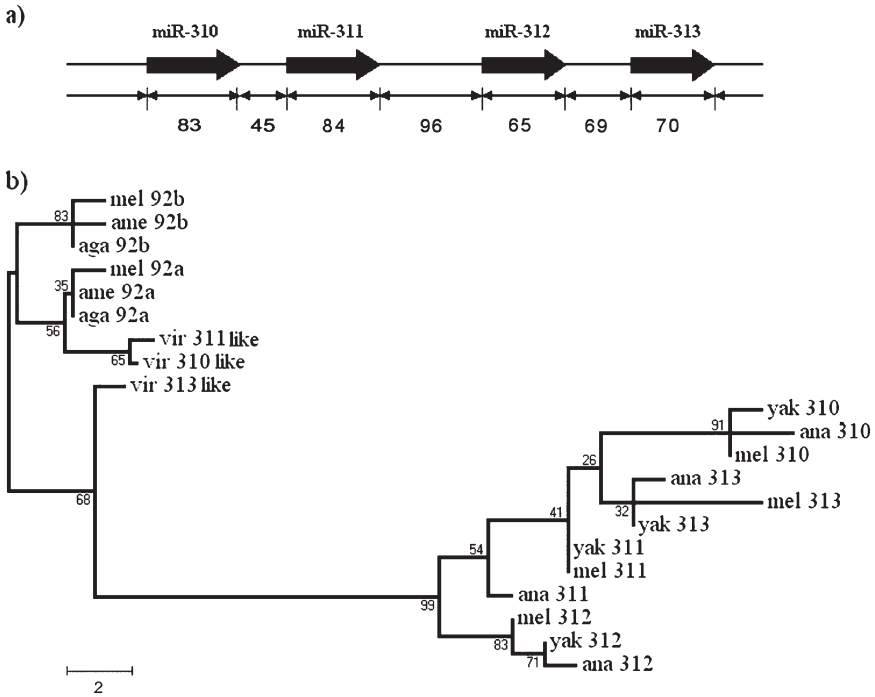

FIG. 1.—(*a*) The cluster of *miR-310/311/312/313* in *Drosophila melanogaster*. The region corresponds to position 16 098 628–16 099 078 on chromosome 2R (dme, V4.3). The numbers below each region are the sizes in base pairs. The entire region was sequenced from 27 lines of *D. melanogaster* as reported in Table 2. (*b*) The phylogeny of genes of the *miR-310/311/312/313* cluster in relation to *miR-92a/b*. The tree was reconstructed by the parsimony method, and bootstrapping values are given at each node. Only the mature miR sequences are used here. The root is set at the point of trifurcation among *miR-92a*, *miR-92b*, and *miR-310/311/312/313* cluster. Species names are used as the prefix of the gene. mel, *D. melanogaster*; yak, *Drosophila yakuba*; ana, *Drosophila ananassae*; vir, *Drosophila virilis*; ame, *Apis mellifera* (honeybee); and aga, *Anopheles gambiae* (mosquito). Only 3 members of the *miR-310*s cluster exist in *Drosophila virilis*. Each member in *D. virilis* is labeled vir 310-like because their orthologies with *D. melanogaster* are hard to determine. The line in the left lower corner represents the scale of 2 substitutions.

Among the 78 known miRNA genes of *D. melanogaster*, 71 can be found in the 12 *Drosophila* lineages and are hence "old" miRNA genes (supplementary fig. S1, Supplementary Material online). Among them, 46 can be found in mosquito. (Two other miRNAs, *miR-289* and *miR-2B-1*, are somewhat ambiguous in their ages and will not be analyzed further; see Supplementary Material online.) These old genes have the known characteristics of miRNAs, including the distributions of ΔG in *D. virilis* and *D. pseudoobscura* and the levels of expression in *D. simulans* and *D. pseudoobscura* (Lu et al. 2008). In both cases, the characteristics in other species are very similar to those in *D. melanogaster* (supplementary fig. S2, Supplementary Material online).

The 5 remaining miRNA genes are newly emerged and fall into 2 categories—4 of them are duplications from existing miRNAs and 1 was created de novo from a hairpin-like non-miRNA sequence. Figure 1*a* shows the cluster of *miR-310/311/312/313* genes. The role of each member in embryonic development has been shown experimentally (Leaman et al. 2005). These 4 genes are homologous to the conserved *miR-92a* and *miR-92b*, both of which can be found in mosquito and honeybee. In contrast, none of the *miR-310*s cluster members can be found outside of *Drosophila*.

Although it is probably true that the formation of the *miR-310*s cluster occurred in *Drosophila*, the precise timing of their duplications cannot be accurately determined, due to the possible confounding effect of gene conversion. Nevertheless, from the phylogenetic analysis of figure 1*b*, we may conclude that there has not been extensive gene conversion since the separation between *D. ananassae* and the *D. melanogaster* more than 53 MYA (99% bootstrapping value). Note that the orthologous sequences are generally more closely related than the paralogous ones in the *D. ananassae–D. melanogaster* grouping in figure 1*b*. Hence, the formation of the *miR-310*s cluster must have been completed more than 53 MYA. We shall return to this relatively old age in the context of adaptive evolution of miRNAs later.

Tracking back deeper in the phylogeny, only 3 putative genes can be identified in this cluster in *D. virilis* of the *Drosophila* subgenus (the other *Drosophila* species are in the *Sophophora* subgenus). Because the sequences of *D. virilis* are quite distinct from those of the *Sophophora* species, the *miR-310*s cluster may be only partially formed at the time of the *Drosophila–Sophophora* split. However, the possibility of earlier origin followed by gene conversion until cluster members acquired their identities cannot be ruled out.

The youngest known miRNA gene in *Drosophila* is *miR-303* (fig. 2*a*). The orthologs of *miR-303* and the flanking sequences can only be identified within the *D. melanogaster* subgroup, which includes *D. simulans* and *D. yakuba* (fig. 2*a*). The expression of *miR-303* in *D. melanogaster* is not in doubt, both by cloning (supplementary

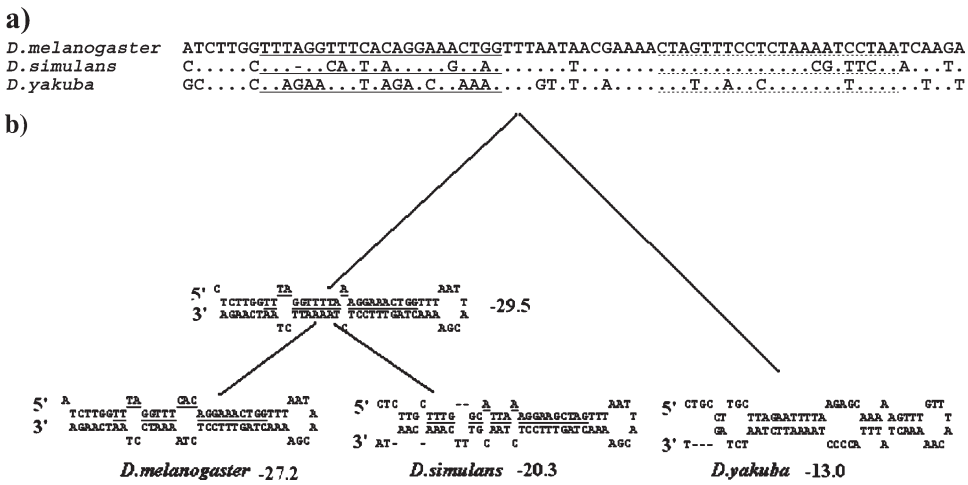

FIG. 2.—(*a*) The alignment of the *miR-303* gene in the 3 species where it can be identified. The mature miR sequences are solid underlined and miR* are broken underlined. (*b*) The inferred structure of the *miR-303* gene in each of the 3 species. The sequence at the ancestral node between *Drosophila melanogaster* and *Drosophila simulans* is also inferred. When the 2 species differ, *Drosophila yakuba* sequence is used to determine the ancestral nucleotides by parsimony. The inferred ancestral sequence has no ambiguity; hence, the hairpin structure can be determined. ΔG in Kcal/mole for each hairpin is also given.

table S2, Supplementary Material online) and by RT-PCR (see supplementary fig. S3, Supplementary Material online). In *D. simulans*, we were able to obtain a product of the right size by RT-PCR using the primer based on the *D. simulans* ortholog of *miR-303*. However, because the only miR sequencing work in *D. simulans* (Lu et al. 2008) may not be sufficiently deep to find *miR-303*, its expression profile in this species remains tentative.

Using the sequence of *D. yakuba* as an outgroup, *miR-303* sequence in the common ancestor of *D. melanogaster* and *D. simulans* can also be inferred. Figure 2*b* shows the hairpin structures of the *miR-303* gene in the 3 extant species and the ancestor. A stable hairpin appears to have emerged only in the lineages of *D. melanogaster* and *D. simulans* as well as their immediate common ancestor. On the contrary, the *D. yakuba* sequence is unlikely to form a stable precursor miRNA (pre-miRNA) structure as its ΔG, at −13.0 Kcal/mol, is far higher than all known pre-miRNAs (see supplementary fig. S2, Supplementary Material online). Because the homologous sequences cannot be found beyond the *melanogaster* subgroup, we suspect that the sequences, unconstrained by selection in those species lineages, have changed beyond recognition.

Evolutionary Rates of New versus Old miRNAs

We now ask whether these newly emerged miRNA genes have undergone rapid sequence evolution. Note that new emergence and rapid evolution are different phenomena. Ancient genes could in principle be rapidly evolving, whereas young genes may be highly conserved. Nevertheless, it seems plausible that a newly emerged miRNA may be evolving rapidly as it may require a period of coevolution with its target transcripts before the interactions become fully integrated into the transcriptome.

In figure 3, we plot the evolutionary rates of the 71 old and 5 new miRNA genes between *D. melanogaster* and 6 other species. These comparisons span the full range of *Drosophila* divergence (see supplementary fig. S1, Supplementary Material online). *K* is the number of changes per nucleotide site; thus, *K*(miRNA gene) is *K* of the entire gene, *K*(miR) is for the mature sequence and so on (see fig. 4*a*). The degree of sequence conservation of the 71 old miRNA genes is striking (the 2 lower solid lines; for *K*(miR) and *K*(miRNA gene), respectively), in comparison with K_s (number of synonymous changes per site; the thin top line). The mature miRs are usually less than 1% as divergent as the synonymous sites but even the entire miRNA genes are no more than 10% as divergent as the latter. The patterns suggest that the secondary structure of the miRNA gene, in addition to the mature miR, is also highly constrained. In comparison, the 5 newly emerged miRNA genes evolve much faster than the old ones, often by 1–2 orders of magnitude, in both the mature miRs and the entire gene.

Given the large number of changes that have accrued on the newly emerged *miR-310*s cluster and *miR-303*, one wonders if the rate is higher than the neutral rate in a manner akin to the K_a/K_s analysis. The maximal likelihood method of (Wong and Nielsen 2004) which compares rates of noncoding substitutions and synonymous substitutions in nearby coding regions is for such a purpose. The substitution rate in *miR-303* is higher than the nearby synonymous rate (albeit only marginally so, with $P = 0.05$) but that in the *miR-310*s cluster is not ($P = 0.48$). Thus, the divergence rate in these miRNA genes is not higher than the synonymous rate, which is often slightly below the neutral rate.

To contrast the evolutionary pattern of the old and new miRNA genes further, we divided each miRNA gene into 4 parts—miR, miR*, loop-end, and stem extension (see fig. 4*a*). The mature miR, about 22 nt long, is the functional product that interacts with target transcripts but all 4 parts are necessary for the successful production of the mature miR (Zeng et al. 2002, 2005; Han et al. 2006). In figure 4*a*, a miRNA gene includes the pre-miRNA and the stem extension (Zeng et al. 2005; Han et al. 2006). In figure 4*b*, we show the distribution of the evolutionary distance, *K*, for

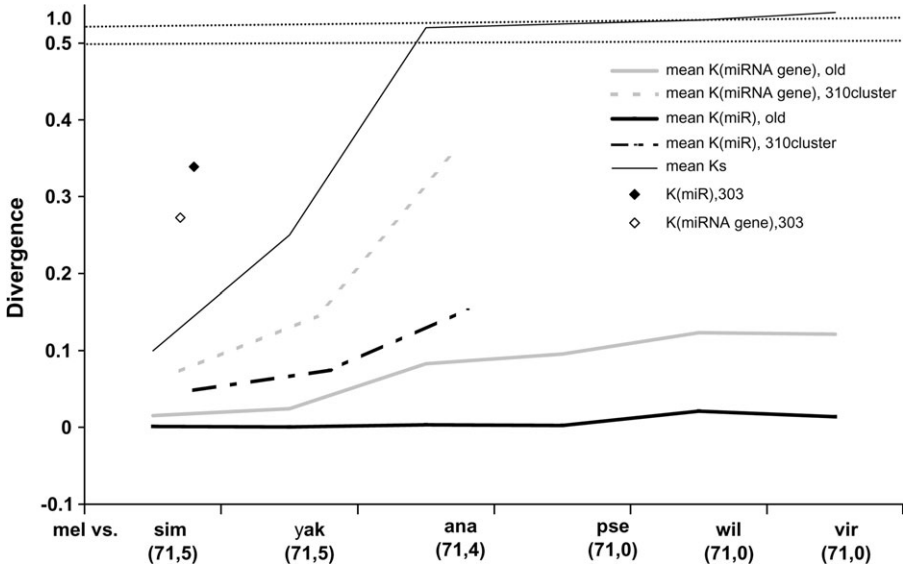

FIG. 3.—Evolutionary divergences for the 5 newly emerged miRNA genes versus the 71 old ones between *Drosophila melanogaster* and 6 other *Drosophila* species. The divergence in the mature miR region (*K*(miR)) and the whole miRNA gene (*K*(miRNA gene)) for each of the 5 new miRNA genes are presented. *K* denote the number of substitutions per site. The mean divergence of the 71 old genes in each comparison as well as the genomic average in K_s are also given. (The standard errors are usually less than 1/10 of the mean value and neglected in the figure.) Numbers of old and newly emerged miRNA genes used in the comparisons are given in the parenthesis on the *x* axis. Numbers of protein-coding genes in each species are always greater than 9000. To emphasize the difference in divergence across categories, the points in each species are not lined up on the *x* axis.

each of the 4 parts of the miRNA gene and for the entire gene between *D. melanogaster* and *D. yakuba*.

For a large fraction of the old miRNAs, the entire miRNA gene, including miR*, loop-end, and stem extension, is conserved (see *K*(miRNA gene) of fig. 4*b*). It is possible that all 3 parts are vital for the production of miR, maintaining a stem-loop structure recognizable by the Drosha–Pasha complex (Han et al. 2006). Indeed, a single nucleotide change could strongly affect the production of the functional miR (Gottwein et al. 2006). Based on our analysis, the general rate of evolution is in the order of loop-end > stem extension > miR* > miR. This can be seen in the size of the nonevolving class ($K = 0$) in the 4 lower panels of figure 4*b*. The corresponding average substitution rates are 0.072 ± 0.013, 0.043 ± 0.007, 0.022 ± 0.006, and 0.016 ± 0.012 in the descending order for the 4 regions. The same patterns are observed in other pairwise species comparisons (supplementary fig. S4, Supplementary Material online).

Adaptive Evolution in New miRNA Genes between Closely Related Species

A widely used method for detecting positive selection is the McDonald and Kreitman test (1991) (MK test for short). For coding sequences, MK test contrasts the levels of divergence (*D*) and polymorphism (*P*) for nonsynonymous and synonymous changes. If all changes are strictly neutral, the *D*/*P* ratios for nonsynonymous and synonymous sites should not be statistically different. For coding regions in *Drosophila*, there is often an excess in the *D*/*P* ratio for nonsynonymous sites over that for synonymous sites. Such results in *D. melanogaster* can best be attributed to the action of positive selection (Fay et al. 2002; Smith and Eyre-Walker 2002; Shapiro et al. 2007).

In this application of the MK test, we substituted changes in miRNA genes for nonsynonymous changes. We first compared the divergence and polymorphism patterns for the old and new miRNAs, calibrated against the genomic average for synonymous sites as shown in table 1. Divergence is between *D. melanogaster* and *D. simulans* and polymorphism data are from *D. simulans* as determined by the Drosophila Population Genomics Project (Begun et al. 2007).

In table 1, the *D*/*P* ratio for the old miRNA genes is significantly lower than that for the synonymous sites (1.69 vs. 5.96, $P < 0.001$, χ^2 test). The old miRNAs are apparently strongly constrained such that few mutations ever became fixed. On the contrary, the *D*/*P* ratio for the new miRNA genes is slightly higher (but statistically insignificant) than that for the synonymous sites (7.0 vs. 5.96, $P = 0.8$, χ^2 test). Because polymorphisms come from only 4 strains of *D. simulans*, the statistical power of the MK test is limited. To test whether the newly emerged miRNA genes bear the signature of positive selection (i.e., having a high *D*/*P* ratio), we sequenced these genes and the flanking regions from 27 lines of *D. melanogaster* (see Materials and Methods). In the *miR-310*s cluster and the *miR-303* region, the *D*/*P* ratio is significantly higher in the miRNA genes than in the neighboring sequences (table 2; $P = 0.002$ and 0.015, respectively, by Fisher's exact tests). Such a pattern is often a sign of positive selection during the divergence between the 2 species.

Population Genetics of New miRNA Genes in *D. melanogaster*

We further ask whether the signature of positive selection can be detected in the polymorphisms of

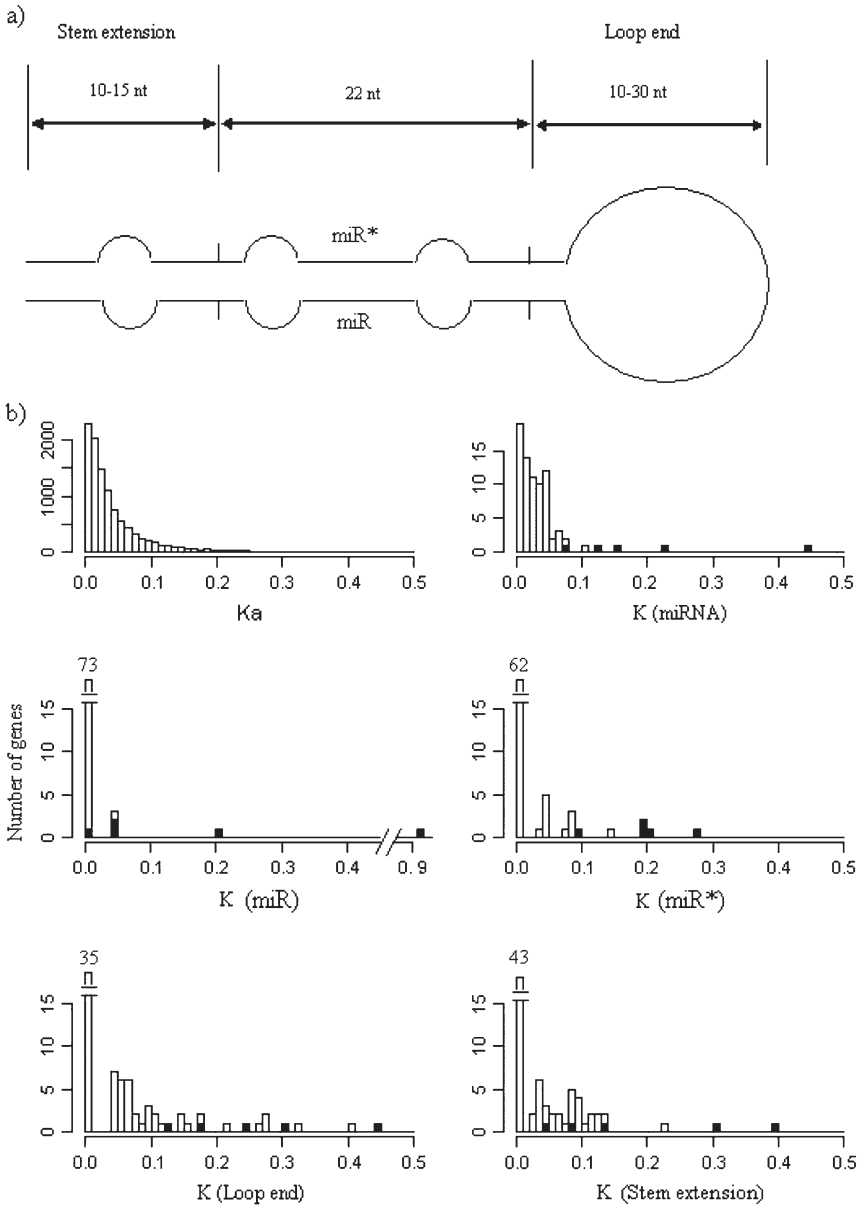

FIG. 4.—(*a*) The structure of a miRNA gene. The miR and miR* are the mature product and its complement, respectively. Loop-end is defined as the region between miR and miR*, including any loop or stem structure. The size of stem extension has been suggested to be exactly 11 bp long (Han et al. 2006) but appears to vary between 10–15 bp among naturally occurring miRNA genes. (*b*) Distributions of miRNA evolutionary distances between *Drosophila melanogaster* and *Drosophila yakuba* for each part of the miRNA gene in (*a*) and for the entire gene. *K* is the number of substitutions per nucleotide site, corrected for multiple changes. *K*(miRNA gene) is hence the distance for the whole gene, *K*(miR) is for the miR region, and so on. K_a is the number of nonsynonymous changes per site in coding regions. The *y* axis is the number of miRNA genes or coding genes. Dark bars denote the 5 newly emerged miRNA genes in *Drosophila* (see text). The position of *miR-303* is marked. The orthologous sequence of *miR-303* in *D. yakuba*, although might not be a true miRNA gene, is also used in the comparison here.

D. melanogaster. A hallmark of positive selection and the associated hitchhiking is the excess of high frequency, and the deficit of intermediate frequency, variants. Figure 5 shows the frequency spectrum of derived variants in the *miR-310*s and the neighboring regions. In total, there are 34 polymorphic sites—4 in the miRNA genes and 30 in the adjacent regions. This pattern could be (partially) informative about the focal sites of selection (see fig. 6 below). The excess on the right and the deficit in the middle of the spectrum of figure 5 are significant by both *D* of Tajima (1989) and *H* of Fay and Wu (2000) tests ($P = 0.005$ and 0.0008, respectively).

To gauge whether the significance in the *D* and *H* tests might be due to local sampling or demographical factors, we applied the same tests to the 419 genes sequenced from 24 lines of *D. melanogaster* as reported in Shapiro et al. (2007). These lines came from a more restricted sampling in geography than the sample used in this study (see supplementary table S1, Supplementary Material online). Supplementary figure S5*a* and *b* (Supplementary Material online) shows that only 1 of the 419 genes is significant at $P < 0.01$ for both *D* and *H* tests. Therefore, by both *D* and *H* tests, the spectrum in figure 5 deviates more

Table 1
Numbers of Nucleotide Changes between *Drosophila melanogaster* and *Drosophila simulans* (divergence) and among 4 *D. simulans* Lines (polymorphism) for Newly Emerged and Old miRNA Genes and for Synonymous Sites

	Divergence (*D*)	Polymorphism (*P*)	*D*/*P* Ratio
Old miRNAs	44	26	1.69
New miRNAs[a]	21	3	7.00
Synonymous changes	240 594	40 346	5.96

[a] Including *miR-310*s cluster and *miR-303*.

Table 2
Numbers of Nucleotide Changes between *Drosophila melanogaster* and *Drosophila simulans* (divergence) and among 27 *D. melanogaster* Lines (polymorphism) for miRNA genes and the Neighboring Sites

		Divergence	Polymorphism	*P* value[a]
*miR-310*s region[b]	miRNA genes	20	4	0.002
	Neighboring sites	26	30	
miR-303 region	miRNA genes	16	4	0.015
	Neighboring sites	64	67	

NOTE.—Data from *miR-310*s cluster and *miR-303* regions are more extensive than those of table 1, which were obtained from the DPGP project in order to be compared with the old miRNAs.

[a] See figure 1*a* for the region sequenced.

[b] The Fisher's exact tests were used to calculate the *P* values.

strongly from neutrality in the direction of positive selection than the 419 genes. Because our sample came from a broader geographical collection than that of Shapiro et al. (2007), the pattern is not likely to have resulted from local sampling or local demographical events, such as a recent population bottleneck.

A more effective application of the population genetic tests to detect positive selection has been proposed by Zeng et al. (2006), Zeng, Mano, et al. (2007), Zeng, Shi, Wu (2007). They pointed out that few population genetic tests used for detecting positive selection are sensitive to selection only. Most are sensitive to some of the other forces, such as population growth, population subdivision, background selection, and so on (see also Przeworski 2002 and Jensen et al. 2005). They suggested a number of compound tests that are sensitive, and reasonably specific, to positive selection. One such example is the *DH* test (a compound test of *D* test of Tajima and *H* test of Fay and Wu), which is moderately sensitive to selection but is quite insensitive to population growth, decline, and subdivision. Applying the *DH* test to the frequency spectrum of figure 5, we obtained $P < 0.001$, further hinting the possible influence of positive selection.

A more detailed analysis near the putative site of selection could be informative about the possible sites of positive selection and can be done in a number ways. The extended haplotype homozygosity (EHH) test (Sabeti et al. 2002) and its extensions (Voight et al. 2006) are good examples. However, as pointed out in Zeng, Shi, Wu (2007), test of the EHH type are powerful only when the selected variant has not reached fixation. Given the highly significant result from the MK test in the *miR-310*s region, we suspect that many of the beneficial mutations may have been fixed between species. In such cases, Fay and Wu (2000) pointed out that a focal site under selection sometimes has 2 observable characteristics: 1) the immediate vicinity of the site has very low genetic diversity; 2) this small region is flanked by 2 peaks of high diversity, if measured by θ_H (Fay and Wu 2000; Zeng et al. 2006). The polymorphism pattern of the *miR-310*s region is hence shown in figure 6, together with the divergence in the form of the MK test.

There are several noteworthy trends in figure 6. 1) *miR-310* has very low diversity (shown by θ_H but also true by other measures). 2) There are 2 peaks of θ_H surrounding *miR-310* (see captions for the sliding window parameters). The peak to the right is about 200 bp wide, consistent with previous reported cases in *Drosophila* (e.g., Fay and Wu 2000). Unfortunately, our sequencing effort did not cover far enough to inform about the width of the peak to the left. 3) The MK test shows that the *D*/*P* ratio in *miR-310* is 4/1, whereas, in the vicinity coextensive with the 2 peaks, the *D*/*P* ratio is 8/15 (the test is not statistically significant, probably because the regions compared are too short). 4) The larger surrounding region does not have any other known genes that might be the focus of selection. The closest 2 neighbors are *CG13432* and *CG13434*, respectively. Because each neighbor is more than 1.5 kb away, they seem unlikely to be the focal sites of selection responsible for the diversity pattern of figure 6.

In summary, a parsimonious explanation for the polymorphism shown in figures 5 and 6 is that a selective sweep was associated with the adaptive mutations in, or very near, *miR-310*. In contrast, the spectrum in the *miR-303* region does not deviate significantly from neutrality. Because the signature of positive selection in the frequency spectrum is transient, significant deviation from neutrality is not always expected even in regions of frequent selective sweeps (Fay and Wu 2000; Przeworski 2002).

Discussion

Among the 5 newly emerged miRNA genes, the 4 in the *miR-310*s cluster play a role in embryonic development (Leaman et al. 2005). The function of *miR-303* remains elusive. They are all expressed in *D. melanogaster* although *miR-313* and *miR-303* are only about 5% as abundantly transcribed as the other 3 (Ruby et al. 2007, see also supplementary table S3, Supplementary Material online).

These newly emerged miRNAs underwent a surprisingly long period of rapid and adaptive evolution. Among these 5 newly emerged genes, there are 36 nucleotide changes between *D. melanogaster* and *D. simulans*. We estimate that nearly 80% of these changes $\{[(20 - 26 \times 4/30) + (16 - 4 \times 64/67)]/(20 + 16) \approx 0.8$, see table 2$\}$ were adaptive (McDonald and Kreitman 1991). Although *miR-303* is very new and many adaptive changes are expected, it is intriguing that *miR-310*s are still evolving adaptively after the split of *D. melanogaster* and *D. simulans*. Because this cluster came into existence in the last ~55 Myr by duplication from the old miRNA genes

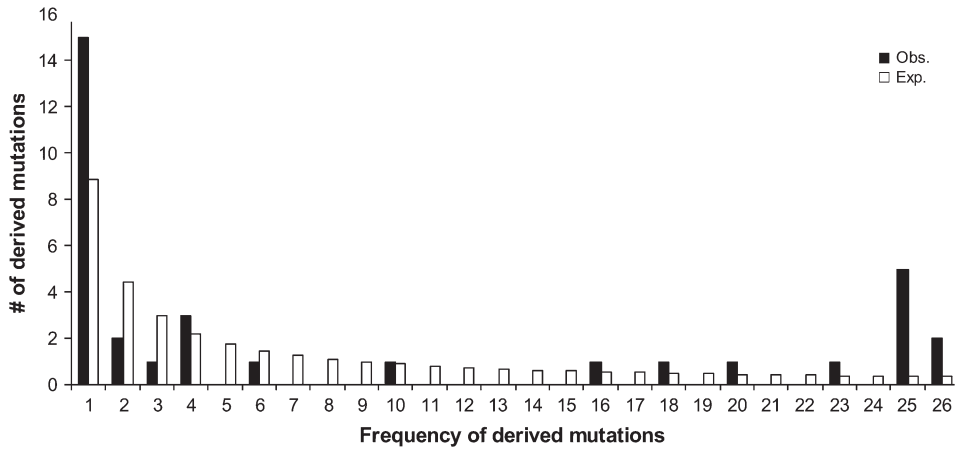

FIG. 5.—The frequency spectrum of derived mutations in the *miR-310*s gene regions and neighboring sites in *Drosophila melanogaster*. Four mutations in the miRNA gene regions and 30 mutations in the neighboring sites were included. The *x* axis is the number of occurrences of a mutation in a sample of 27, and the *y* axis is the number of sites with the designated occurrence. The expected is the spectrum under neutral equilibrium (θ /*i*, where *i* is the number of occurrence and θ is 4Neu; see Fu 1995).

(*miR-92a*/*b*), these new miRNAs must have accumulated many adaptive changes during this period.

What drives the rapid evolution of these newly emerged miR genes, especially in the mature miRs? From the pattern of figure 1, it appears that the *miR-310*s cluster members have been evolving away from the ancestral *miR-92a*/*b*. Experimentally, antisense-mediated depletion of any member of this cluster caused phenotypic defects that could not be compensated for by either *miR-92a* or *miR-92b*; if fact, compensation is lacking even among the miR310's cluster members (Leaman et al. 2005). These results suggest functional diversification among these miRs.

Because miRs function in repressing their target transcripts, functional diversification among miRNA genes should mainly be manifested in the divergence of their target pools. There are indeed many bioinformatic tools for inferring the target pool of each miRNA in each species (Enright et al. 2003; Stark et al. 2003; John et al. 2004; Rajewsky and Socci 2004; Brennecke et al. 2005; Grun et al. 2005; Lewis et al. 2005). However, these tools are of limited utility in addressing functional divergence between miRNAs, partly because functional conservation is often a key assumption in target prediction. In a separate study (S.K and C.W unpublished data), we address

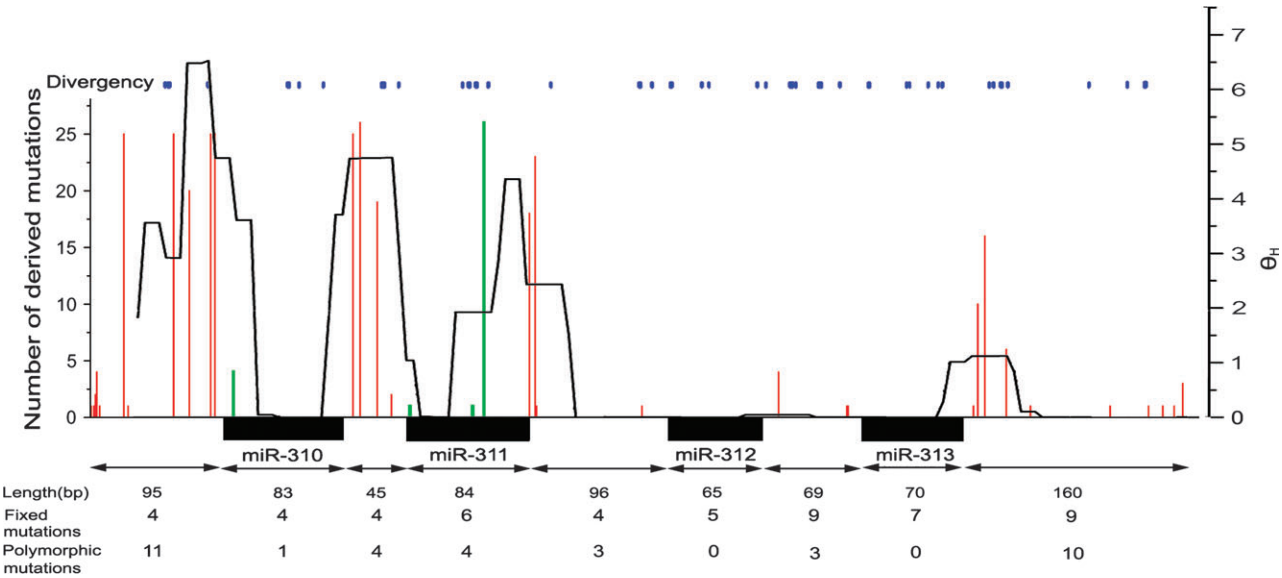

FIG. 6.—A sliding window of *H* test of Fay and Wu on the derived mutations in *miR-310*s cluster and neighboring regions in *Drosophila melanogaster*. The gene regions of *miR-310*, *311*, *312*, and *313* were marked by the black box under the *x* axis. The number of occurrences of a mutation in a sample of 27 was plotted against the genomic positions. Red bars, the number of derived mutations in the neighboring sites of miRNA genes; green bars, the number of derived mutations in the miRNA gene region. The black line is θ_H value obtained by a sliding window of *H* test of Fay and Wu. The window size is 50 nt, and the step size is 5 nt. The θ_H value was plot against the position of the middle point of each window. The mutations fixed were plotted on the top of the figure, according to the genomic locations. The length of miRNA gene and neighboring regions, as well as the number of mutations that are fixed and polymorphic in each region, were provided on the bottom of this figure.

this issue by using the transgenic technique to insert *miR-310*s from different species into *D. melanogaster* as well as by knocking out the resident *miR-310*s from this species. These transgenic lines can then be used in whole-genome expression studies to empirically identify the miRNA targets. Evidence for the coevolution of *miR-310*s and their targets, which has measurable fitness consequences, will be presented in that study.

As befits the youngest known miRNA gene, the rate of evolution for *miR-303* is unusually high. The difference of 17 substitutions (out of 70 sites, or 24%) between *D. melanogaster* and *D. simulans* is more than twice the average synonymous divergence ($K_s \sim 0.11$) between the 2 species. Furthermore, the mature sequence has diverged (7 substitutions out of 21 sites) more than the rest of the miRNA gene (33% vs. 20%). We have shown that such a high rate of molecular evolution might be driven by positive Darwinian selection (table 2). An intriguing observation is that the seed of the miR sequence (positions 2–8), which interacts directly with the targets and is almost always invariant over large evolutionary distance among miRNA genes, has 3 changes between the 2 species. All 3 of them, including a 1-bp deletion, occurred in *D. simulans* (fig. 2*a*). This suggests that *miR-303* in *D. simulans* might have been evolving away from interacting with the pool of transcripts which *miR-303* in *D. melanogaster* (and their common ancestor) target. Alternatively, because the expression of *miR-303* is still tentative, *miR-303* might have degenerated in *D. simulans*.

Our result revealed there is a link between the rate of evolution in the miR sequence and the rest of the miRNA gene. Indeed, the 5 new miRNA genes with fast-evolving miRs have relatively high substitution rates in miR*, loop-end, and stem extension (fig. 4b). For these fast-evolving miRNA genes, changes in the miR sequence are likely dictated by the interactions with the target sequences. These changes may not be compatible with a stem-loop structure conducive for miR maturation. As a result, ensuing changes in other parts of the miRNA gene may be needed to restore the biochemical properties of the hairpin. This may be an example of compensatory evolution in which a feature altered by a genetic mutation is later restored by a different mutation.

Our analysis of a small number of newly emerged miRNA genes among the known ones in *Drosophila* could be applicable to the large number of newly discovered miRNAs, many of which are younger than the 78 miRNAs reported earlier (Bentwich et al. 2005; Berezikov et al. 2006; Kloosterman et al. 2006; Ruby et al. 2007; Stark et al. 2007; Lu et al. 2008). Although it will take great efforts to elucidate their functions, it should be feasible to find out their adaptive significance by population genetic means. Positive answers could bolster our confidence in their functionality. This independent knowledge is important in the search for function as many newly emerged miRNAs are likely to have neutral fitness consequences (Lu et al. 2008).

Supplementary Material

Supplementary figures S1–S5 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Washington University School of Medicine Genome Sequencing Center, Agencourt Bioscience, the Broad Institute, and the Drosophila Population Genomics Project project for making the Drosophila genome sequence data publicly available. We thank Drs Blake C. Meyers, Ilya Ruvinsky, Jianzhi Zhang, Victor Ambros, and Casey Bergman for critical comments and discussions. We also thank Dr Anthony J. Greenberg for technical assistance in sequencing. C.-I.W. is supported by National Institutes of Health grants, a Chicago Biological Consortium seed grant, and additional support from Sun Yat-sen University.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli AE. 2005. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. Cell. 122:553–563.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 116:281–297.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). Forthcoming. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila* simulans. PLoS Biology. 5:2534–2559.

Bentwich I, Avniel A, Karov Y, et al. (13 co-authors). 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet. 37:766–770.

Berezikov E, van Tetering G, Verheul M, et al. (14 co-authors). 2006. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res. 16:1289–1298.

Brennecke J, Stark A, Russell RB, Cohen SM. 2005. Principles of microRNA-target recognition. PLoS Biol. 3:e85.

Comeron JM, Kreitman M, Aguade M. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics. 151:239–249.

Doench JG, Petersen CP, Sharp PA. 2003. siRNAs can function as miRNAs. Genes Dev. 17:438–442.

Drosophila Comparative Genome Sequencing and Analysis Consortium. 2007. Genomics on a phylogeny: evolution of genes and genomes in the genus *Drosophila*. Nature. 450:203–218.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2003. MicroRNA targets in *Drosophila*. Genome Biol. 5:R1.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155:1405–1413.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature. 415:1024–1026.

Fu YX. 1995. Statistical properties of segregating sites. Theor Popul Biol. 48:172–197.

Gottwein E, Cai X, Cullen BR. 2006. A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects Drosha processing. J Virol. 80:5321–5326.

Grun D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. PLoS Comput Biol. 1:e13.

Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho Y, Zhang BT, Kim VN. 2006. Molecular basis for the

recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell. 125:887–901.
Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures (the Vienna RNA package). Monatsh Chem. 125:167–188.
Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 170: 1401–1410.
John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human microRNA targets. PLoS Biol. 2:e363.
Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. PNAS. 100: 11484–11489.
Kim VN. 2005. MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol. 6:376–385.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. J Mol Evol. 16:111–120.
Kloosterman WP, Steiner FA, Berezikov E, de Bruijn E, van de Belt J, Verheul M, Cuppen E, Plasterk RH. 2006. Cloning and expression of new microRNAs from zebrafish. Nucleic Acids Res. 34:2558–2569.
Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science. 294:858–862.
Leaman D, Chen PY, Fak J, Yalcin A, Pearce M, Unnerstall U, Marks DS, Sander C, Tuschl T, Gaul U. 2005. Antisense-mediated depletion reveals essential and specific functions of microRNAs in *Drosophila* development. Cell. 121: 1097–1108.
Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 120:15–20.
Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. 2003. Prediction of mammalian microRNA targets. Cell. 115:787–798.
Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36: 96–99.
Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. 2003. Vertebrate microRNA genes. Science. 299:1540.
Lu J, Shen Y, Wu Q, Kumar S, He B, Carthew R, Wang SM, Wu C-I. Forthcoming 2008. The birth and death of microRNA genes in *Drosophila*. Nat Genet.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature. 351:652–654.
Olsen PH, Ambros V. 1999. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol. 216:671–680.
Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics. 160:1179–1189.
Rajewsky N. 2006. microRNA target predictions in animals. Nat Genet. 38(Suppl 1):S8–S13.
Rajewsky N, Socci ND. 2004. Computational identification of microRNA targets. Dev Biol. 267:529–535.
Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. Genome Res. 17:1850–1864.
Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature. 419:832–837.
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. Genome Res. 13:103–107.
Shapiro JA, Huang W, Zhang C, et al. (12 co-authors). 2007. Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci USA. 104:2271–2276.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature. 415:1022–1024.
Stark A, Brennecke J, Russell RB, Cohen SM. 2003. Identification of *Drosophila* microRNA targets. PLoS Biol. 1:E60.
Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M. 2007. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. Genome Res. 17:1865–1879.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.
Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol. 21:36–44.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.
Wang X. 2006. Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res. 34:1646–1652.
Wong WS, Nielsen R. 2004. Detecting selection in noncoding regions of nucleotide sequences. Genetics. 167:949–958.
Zamore PD, Haley B. 2005. Ribo-gnome: the big world of small RNAs. Science. 309:1519–1524.
Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 174:1431–1439.
Zeng K, Mano S, Shi S, Wu CI. 2007. Comparisons of site- and haplotype-frequency methods for detecting positive selection. Mol Biol Evol. 24:1562–1574.
Zeng K, Shi S, Wu CI. 2007. Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol. 24: 1898–1908.
Zeng Y, Wagner EJ, Cullen BR. 2002. Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol Cell. 9:1327–1333.
Zeng Y, Yi R, Cullen BR. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. Embo J. 24:138–148.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406–3415.

Marta L. Wayne, Associate Editor

Accepted February 4, 2008