nature genetics

The birth and death of microRNA genes in *Drosophila*

Jian Lu[1], Yang Shen[2], Qingfa Wu[3], Supriya Kumar[1], Bin He[1], Suhua Shi[2], Richard W Carthew[4], San Ming Wang[3] & Chung-I Wu[1,2]

MicroRNAs (miRNAs) are small, endogenously expressed RNAs that regulate mRNAs post-transcriptionally. The class of miRNA genes, like other gene classes, should experience birth, death and persistence of its members. We carried out deep sequencing of miRNAs from three species of *Drosophila,* and obtained 107,000 sequences that map to no fewer than 300 loci that were not previously known. We observe a large class of miRNA genes that are evolutionarily young, with a rate of birth of 12 new genes per million years (Myr). Most of these new miRNAs originated from non-miRNA sequences. Among the new genes, we estimate that 96% disappeared quickly in the course of evolution; only 4% of new miRNA genes were retained by natural selection. Furthermore, only 60% of these retained genes became integrated into the transcriptome in the long run (60 Myr). This small fraction (2.5%) of surviving miRNAs may later on become moderately or highly expressed. Our results suggest that there is a high birth rate of new miRNA genes, accompanied by a comparably high death rate. The estimated net gain of long-lived miRNA genes, which is not strongly affected by either the depth or the breadth (number of tissues) of sequencing, is 0.3 genes per Myr in *Drosophila*.

MicroRNAs are endogenously expressed, single-stranded RNAs, about 22 nucleotides (nt) long, that regulate mRNAs post-transcriptionally[1–4]. The mature miRNA is derived from a 70–90-nt-long stem-loop (hairpin) RNA structure. miRNAs are often highly conserved among even divergent animal species[3]. In fact, evolutionary conservation has been an oft-used criterion for identifying miRNA genes[5,6]. Because of this conservative approach to identifying this class of genes, the origin of miRNA genes and their subsequent evolution is largely unknown. Recently, reports have shown that the number of miRNA genes may be substantially larger than has been previously reported. A sizable fraction of these may not be conserved even between closely related species[7–11]. In this study we explore the birth, death and evolution of miRNA genes by comparing the miRNA transcriptomes of three *Drosophila* species.

Small RNAs, averaging 22 nt in length (range, 18–28 nt), were isolated from the heads of adult males of three *Drosophila* species, and cDNAs were sequenced by the 454 method[12]. In total, 16,436, 86,535 and 44,395 sequence reads were perfectly mapped onto the corresponding genome sequences from *D. melanogaster*, *D. simulans* and *D. pseudoobscura*, respectively (**Table 1**). After removal of reads matching rRNAs, tRNAs, small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and transposable elements, each of the remaining reads was computationally tested for its potential classification as an miRNA (see Methods and **Supplementary Methods** online). Briefly, the genome sequence corresponding to each read and the flanking sequences up to 150 nt on each side were analyzed by RNAFOLD[13]. The rationales for choosing the sizes of the flanking regions are given in **Supplementary Methods**. Within the 322-nt (= 22 + 150 × 2) segment, we searched for hairpin sequences for which the following were true: (i) the hairpin lacked multiple loops and was greater than

Table 1 Number of reads collected from male heads of three *Drosophila* species

	D. melanogaster	*D. simulans*	*D. pseudoobscura*
Genome size (Mb)	139	127	139
No. of reads (perfect match to the genome)	16,436	86,535	44,395
No. of non-miRNA reads	882	15,778	24,604
No. of miRNA reads[a]	15,554	70,757	19,791
Known miRNAs	14,318 (n = 56)[b]	65,818 (n = 63)	17,097 (n = 54)
% total miRNAs	92.1%	93.0%	86.5%
Putative new miRNAs			
High stringency	1,236 (n = 41)	4,939 (n = 119)	2,694 (n = 229)
Medium stringency	1,254 (n = 50)	5,185 (n = 187)	2,938 (n = 389)
Low stringency	1,268 (n = 58)	5,329 (n = 261)	3,145 (n = 598)

[a]Sum of known miRNA reads and new reads based on the high stringency criteria. [b]n denotes the number of miRNA loci.

[1]Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA. [2]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen (Zhongshan) University, Guangzhou, 510275, China. [3]Center for Functional Genomics, Division of Medical Genetics, Department of Medicine, Evanston Northwestern Healthcare Research Institute, Northwestern University Feinberg School of Medicine, Evanston, Illinois 60208, USA. [4]Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, Illinois 60208, USA. Correspondence should be addressed to C.-I.W. (ciwu@uchicago.edu) or S.M.W. (swang1@northwestern.edu).

Received 13 August 2007; accepted 6 December 2007; published online 17 February 2008; doi:10.1038/ng.73

Table 2 Phylogenetic distributions of newly discovered miRNA genes among six *Drosophila* species

	Observed A_i (A_i with high expression)		Background B_i	
Phylogenetic pattern (m-s-y-a–p-v)	Medium stringency	High stringency	Medium stringency	High stringency
1-0-0-0-0-0	8 (1)	4 (1)	144,611	89,979
0-1-0-0-0-0	62 (6)	42 (5)	–	–
0-0-0-0-1-0	309 (32)	179 (19)	–	–
0-1-1-0-0-0	11 (3)	3 (1)	–	–
1-1-0-0-0-0	22 (3)	17 (3)	46,343	16,014
1-1-1-0-0-0	25 (4)	11 (2)	17,740	4,542
1-1-1-1-0-0	10 (3)	6 (6)	966	210
1-1-1-1-1-0	8 (4)	3 (2)	506	193
1-1-1-1-1-1	24 (19)	19 (17)	149	79
Other patterns	46 (7)	15 (5)	17,912	6,345
Total	525 (82)	299 (61)	228,227	117,362

m-s-y-a-p-v stands for the first letters of the following species names: *D. melanogaster, D. simulans, D. yakuba, D. ananassae, D. pseudoobscura* and *D. virilis*. High expression means read number ≥ 3 in *D. melanogaster*, ≥ 5 in *D. simulans* or *D. pseudoobscura*, or ≥ 1 in at least two species. A_i is the number of observed miRNA genes with the phylogenetic pattern i; B_i is the number of genomic background hairpins with the same pattern i. Observed numbers of miRNA genes with high expression are given in parentheses.

55 nt in length; (ii) the hairpin had a free energy less than or equal to –15 kcal/mol; (iii) the read was within one arm of the hairpin. These three criteria comprise the baseline of miRNA identification in this study (low stringency criteria). Two additional sets of criteria, for medium and high stringency, based on the RANDFOLD[14] and RNASHAPES[15] programs were then applied to these hairpin structures (see Methods for details).

Based on these criteria, about 95%, 82% and 45% of the sequence reads were considered miRNA-like for *D. melanogaster*, *D. simulans* and *D. pseudoobscura*, respectively (**Table 1**). Non-miRNA reads included rRNAs, tRNAs, snoRNAs, snRNAs and transposable elements. The different efficiencies in recovering miRNAs from samples of different species are probably a reflection of technical issues, such as those relating to the exclusion of small rRNAs from the preparation, rather than to fundamental differences in the transcriptomes.

Of the 78 characterized miRNA genes of *D. melanogaster* (miRbase V9.2)[16], nearly all can be found in the genomes of the 12 *Drosophila* species sequenced[17]. In our sequence analysis, reads were observed for 56 of the 78 miRNA genes in *D. melanogaster*, accounting for 92% of the miRNA-like reads (**Table 1**). The corresponding numbers in the other two species from which reads were derived were comparable (**Table 1**). Thus, this was a reasonable approach to finding miRNAs in the *Drosophila* head transcriptome. The remaining miRNA-like reads were classified into groups according to their genomic locations. For example, 1,236 previously unknown miRNA-like reads fell into 41 loci within the *D. melanogaster* genome (**Table 1** and **Supplementary Table 1** online). In *D. simulans* and *D. pseudoobscura*, the number of previously unknown loci found was significantly higher: 119–598, depending on the stringency of the definition.

The fivefold larger number of reads sequenced from *D. simulans* might account for the greater number of miRNA-like loci observed in that species, and it suggests that a large class of rare loci exist. The greatest numbers of previously unknown loci were observed in *D. pseudoobscura*, even though fewer reads were sequenced in that species than in *D. simulans*. Some of these loci might be unknown and uncommon transposable elements that escaped masking (see **Supplementary Methods**), or there may have been a relatively recent expansion of miRNA genes in *D. pseudoobscura*. There was no significant correlation between the genome size and the number of putative miRNA loci in the three species (the Pearson correlation coefficient $r = 0.10$ and $P = 0.90$). Across species, the expression levels of orthologous miRNA genes (55 known and 40 newly identified with ≥ 5 reads in *D. simulans*) were highly correlated. r ranged from 0.81 to 0.90 ($P < 0.0001$), although some miRNA genes varied substantially in expression between species (**Supplementary Table 1**; see also our website).

To analyze the evolution of these newly identified miRNA-like genes, we identified the orthologous regions from six *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura* and *D. virilis*). The orthologous sequences were then examined for miRNA-like features, as described in the previous section. In addition, the orthologous sequences were also required to have a secondary structure similar to those of the corresponding read-producing genes (with a similarity score greater than 0.3 by RNAFORESTER[18]).

The new miRNA genes are grouped by their phylogenetic distribution in **Table 2**. For example, under the medium stringency criteria, 25 miRNA genes are present in *D. melanogaster, D. simulans* and *D. yakuba* but absent in the other three species. Parsimonious inference indicates that these 25 miRNA genes arose in the common ancestor of *D. melanogaster*, *D. simulans* and *D. yakuba*. A straightforward way to visualize the evolutionary dynamics of **Table 2** is to place the emergence of miRNA genes on a phylogenetic tree. We used the parsimony method to assign the rise of each new miRNA gene to the appropriate phylogenetic interval (**Fig. 1**). This analysis was performed by using the medium and high stringency criteria only. The phylogenetic tree can be divided into a main trunk and external branches (**Fig. 1**). Because emergence can be inferred only

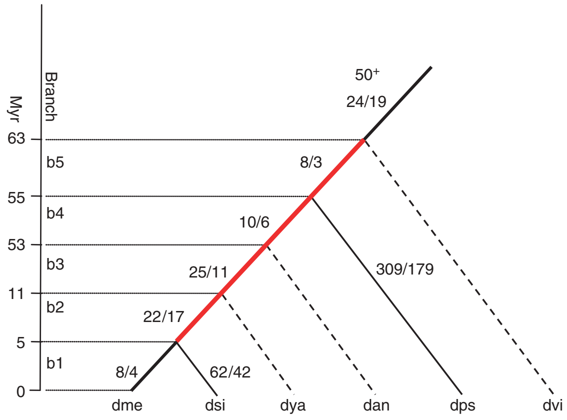

Figure 1 Gains of miRNA genes inferred by the parsimony method. Along each solid branch, the numbers of miRNAs inferred to originate in that period are given. The two numbers are based on the medium/high stringency criteria for inferring the existence of miRNAs. All numbers are new miRNA genes discovered in this study except for "50+" on the top branch, which is the number of miRNA genes in our data (high stringency criteria) and in the literature. Branch designation and the time depth in Myr[17,26] are given on the left of the figure. Only those branches that are ancestral to *D. melanogaster*, *D. simulans* and *D. pseudoobscura* can be assigned these gains. The other branches for which no information is available are shown by dashed lines. The main trunk (b2–b5), marked by the red line, is most informative about the evolution of putative functional miRNA genes. dme, *D. melanogaster*; dsi, *D. simulans*; dya, *D. yakuba*; dan, *D. ananassae*; dps, *D. pseudoobscura*; dvi, *D. virilis*.

with expression data, external branches to *D. yakuba*, *D. ananassae* and *D. virilis*, from which we did not collect samples, do not show miRNA emergence.

On the main trunk, we observed 37–65 new miRNA genes. Because the four branches span 58 Myr, the frequency of emergence is about 0.6–1.1 gene per Myr. In contrast, on the long external branch to *D. pseudoobscura*, there were 179–309 new miRNA genes in a period of 55 Myr, or about 3.3–5.6 new miRNA genes every Myr. On the short external branch to *D. simulans*, there were 42–62 new miRNA genes, or about 8.4–12 new genes per Myr. We used *D. simulans*, instead of *D. melanogaster*, because the collection from the former is much larger than that from the latter. The numbers shown on the internal branches depend mainly on the larger collection.

In **Figure 1**, the numbers on the external branches are about tenfold greater than the rate of emergence on the older branches. This pyramid-like age structure suggests a birth-and-death process for miRNA genes. Although the phylogenetic assignment is possible only in species groups with many fully sequenced genomes, the general trend has been reported in other taxa. In primate, mouse, zebrafish and *Arabidopsis thaliana*[7–11], there is also a preponderance of miRNA genes that are species specific and weakly expressed. Given this general trend, it seems reasonable to postulate that many newly emerged miRNA genes are evolutionarily transient, emerging and disappearing at a high rate. In this view, most recently emerged miRNA genes have not had time to degenerate, accounting for the high proportion of new genes on the external branches, and especially on the short ones.

We developed a birth-and-death model to account for the evolution of new miRNA genes. Briefly, the birth rate of miRNA genes is assumed to be constant along lineages and the death rate of miRNA genes is modeled by a survivorship function (that is, the probability that a new miRNA survives for a time period of *t* Myr). The birth and death of miRNA genes on branch b1–b5 (**Fig. 1**) would yield 32 possible phylogenetic distributions among *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura* (with *D. virilis* as outgroup). Given any survivorship function, the probabilities of the 32 patterns can be calculated. Based on the observed phylogenetic distributions, a maximum-likelihood method is used to find the best-fit parameters of the survivorship function (see **Supplementary Note** online for detailed description).

We considered new miRNA genes as comprising two distinct classes —a putative functional class and a background class. It has been shown that a large fraction of the genome is transcribed[19,20]. Hence, by most known criteria for defining miRNA genes, there are a large number of potential miRNA-encoding regions in the genome[7–9]. Many such candidate miRNA genes are likely to be neutral in fitness, although our model does not strictly rely on neutrality. These are referred to as the background class. (We show in **Supplementary Fig. 1** online that the survivorship function of this class is indeed close to the one obtained by neutral simulations.) In contrast, the putative functional class consists of miRNA genes that are under significant selective pressure.

To estimate the survivorship of background miRNA genes, we used two types of data: (i) genomic hairpin structures not associated with miRNA reads of **Table 1**, which we label "background hairpins" in **Table 2** and **Supplementary Table 2** online, and (ii) miRNA reads not discovered in this study but retrieved by other deep sequencing efforts[21]. miRNAs in type (ii) dataset are generally very weakly expressed (see our website). Their analysis is presented in **Supplementary Methods**. Analyzing the two types of data, we found that their phylogenetic distributions are highly congruent (Pearson correlation coefficient $r = 0.969$, $P < 10^{-10}$). In other words, if a particular phylogenetic pattern accounts for 20% of the type (i) data, it also accounts for ~20% of the type (ii) data. Hence, we describe only the analysis of type (i) data below.

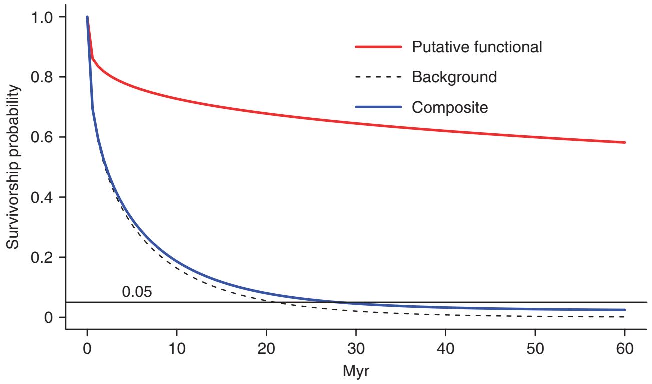

Figure 2 The survivorship probabilities of background (dashed line) and putative functional (red line) miRNA genes as a function of time in Myr, obtained by the maximum-likelihood method. Only miRNA genes inferred by the high stringency criteria are used here. The survivorship is approximated by a gamma function. For background miRNA genes, the MLE estimates of the parameters of the gamma function are rate = 0.082 and shape = 0.42. The corresponding numbers for the putative functional class are rate = 0.0003 and shape = 0.24. The MLE of the proportion of the putative functional class, *f*, is 0.04. The composite survivorship with 4% from the functional class and 96% from background miRNA genes is given by the blue line.

There are more than 110,000 background hairpins in the genome of *D. melanogaster*, and 76% of them show a hairpin structure only in this species. Fewer than 0.5% of them are present in more than four of the six *Drosophila* species as hairpins. We used a gamma distribution to approximate the probability density of death of the background miRNA genes at time *t*. The survivorship function of the background miRNA genes obtained by the ML method according to the phylogenetic patterns of **Table 2** and parameters defined in **Supplementary Tables 3** and **4** is given in **Figure 2**. In that figure, a large proportion of the background miRNA genes degenerate rapidly. We determined the half-life of the neutral background miRNA genes to be 1.96 Myr. The median number of mutations needed to destroy the secondary structure of a background miRNA is only 1.57.

We next estimated the survivorship of putative functional miRNA genes. The observed miRNA genes of **Table 2** are assumed to be a mixture of the putative functional class and the background class. The survivorship function of the latter class has been estimated (**Fig. 2**). We assumed that the death rate of the putative functional class also follows a gamma distribution. We then fitted the two survivorship functions to the observed phylogenetic patterns of new miRNA genes, identified by reads from *D. simulans* (**Table 2**). The maximum-likelihood estimate solution for survivorship of the emerging functional class was obtained and is given in **Figure 2**. The death rate (that is, the slope of the survivorship function) was substantial when miRNA genes are young, and it decreased as these emerging functional miRNAs age. The death rate became relatively low if the emerging functional miRNAs managed to survive 10 Myr. In the long run (>60 Myr), about 60% of these miRNA genes survive. The MLE solution also gives the proportion of emerging functional miRNAs among all new miRNAs. At 4%, the emerging functional class is a small minority. Nevertheless, after 30 Myr, almost all surviving miRNA genes are in the functional class. In other words, among all new miRNA genes in *Drosophila*, only about 2.5% (~4% × 60%) avoid death and become 'immortal'.

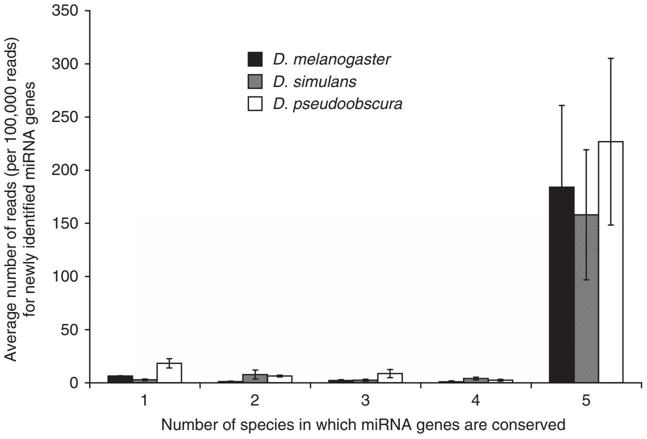

Figure 3 Average number of reads for newly identified miRNAs as a function of their phylogenetic distribution. These numbers were normalized to a total of 100,000 observed reads and are given for new miRNA genes observed in each of the three species separately. Note the marked increase in expression among miRNAs that are found in five (out of six) species or more.

We then used *D. simulans*–specific genes to estimate the birth rate of new miRNAs. 42 *D. simulans*–specific miRNA genes were identified by the high stringency criteria (**Fig. 1**). Because some of these could have emerged on branches b2–b5 (**Fig. 1**), we needed to estimate the proportion that emerged since *D. melanogaster* and *D. simulans* diverged 5 million years ago. The number was estimated to be 30 (see **Supplementary Methods**). Integrating the composite survivorship function of **Figure 2**, we calculated that 50% of the miRNA genes that emerged in the last 5 Myr are still observable now (see **Supplementary Note** for detail). Thus, the total number of miRNA genes emerged in that period is estimated to be 60, or 12 new miRNA genes per Myr. This high birth rate was accompanied by a comparably high death rate, such that only 2.5% of the new miRNAs survived in the long run. The estimated long-term gain of 0.3 miRNAs per Myr (12 × 2.5%) is corroborated by the parsimony estimates of **Figure 1**. In that figure, 20 miRNAs were gained in the 52 Myr along branches b3–b5. The observed rate of 0.38 (20/52) miRNA genes per Myr agrees well with the estimated difference between the birth and death rate.

We expect our estimates of the birth and death rate of miRNAs to be low minima of the true genomic values as our samples came from only one organ. Nevertheless, we do not expect the estimated net gain of 0.3 miRNA gene per Myr to be too far from the values obtained when more tissues are sampled. Many of the older, more established miRNAs tend to be moderately to highly expressed (**Fig. 3**), and the expression of those miRNAs generally lacks tissue specificity[22]. Therefore, we expect the numbers on the top branches (branch b4 and above) not to be strongly dependent on the depth or breadth (number of tissues) of sampling. To test that prediction, we examined the phylogenetic patterns of the ~1,900 hairpins that match at least one read in the samples of ref. 21, which contain ~2 million small RNA reads collected from ten different tissues of *D. melanogaster* (**Supplementary Table 2**). On branch b4 or above, the observed number of miRNAs in our sample is 79 (by high stringency criteria), whereas the corresponding number in the larger dataset is 99. Hence, we missed about 20% of the established miRNAs by sampling only one organ.

In short, the birth and death rates of miRNAs are both high. With samples from male heads only, the estimated birth rate is 12 miRNAs per Myr. The death rate is correspondingly high and the long-term net gain is only 0.3 miRNAs per Myr. Both the birth and death rates may be found to be considerably higher than those given here with deeper and broader (more tissues) sampling. Nevertheless, the estimated long-term gain (0.3 miRNAs per Myr) will only be 20% higher when the depth and breadth of sampling increase by tenfold. In several recent papers connected to the sequencing of the 12 *Drosophila* genomes, the numbers of newly discovered miRNAs in the *Drosophila* genomes were given as >41 (ref. 23) and 59 (ref. 21). In these studies, new miRNAs have to be at least moderately expressed and are conserved. (Species-specific miRNAs in these studies often originated on or above branch b3 of **Figure 1** because their definition of species specificity is based on sequence identity, not time of origination.) In our analysis, the number of miRNAs originating on those branches is 67 or 39, by the medium or high stringency criteria, respectively. Our analysis is thus in broad agreement with other studies concerning the net gain (= birth – death) of miRNAs. In our conclusion, the birth and death rates are both much greater than the rate of net gain.

Another intriguing question is how new miRNA genes evolve, given the large number of potential target transcripts. We observed that most young miRNA genes are weakly expressed whereas many older miRNA genes are highly expressed (**Fig. 3**). A new miRNA is likely to survive only when its expression is low and hence its fitness effect small. Subsequent mutations may allow a miRNA to shed target genes that should not be repressed. While the target pool is being shuffled, the expression level of the regulating miRNA may gradually increase. In *Arabidopsis*, newly emerged miRNAs generally have weaker regulatory effects than older ones[11]. It may take new miRNA genes in *Drosophila* 50 Myr to become highly expressed (**Fig. 3**). Indeed, the population genetic analysis of four young miRNAs suggests continual adaptive evolution more than 50 Myr after their emergence (J.L., S.K., Y.S., R.W.C. and C.-I.W. *et al.*, unpublished data).

Finally, we address the origin of miRNA sequences. In plants, duplication of the whole miRNA gene and inverted duplication of short segments are common mechanisms for the generation of new miRNAs[11,24,25]. In *Drosophila*, the predominant source of these appears to be non-miRNA sequences that accumulate nucleotide changes to become miRNA genes. Only 1.7% of the newly identified miRNA genes in our study have paralogous miRNA loci, and none of them originated by inversion of duplicated stems (see **Supplementary Methods**). Thus, *Drosophila* and plants seem quite different with respect to miRNA origination. In the **Supplementary Methods** and **Supplementary Figure 2** online, we further show that the potential rate of miRNA origination by single-nucleotide substitution is greater than 1,000 per Myr in the *Drosophila* genomes. How much of this potential is realized is an empirical question, but it is not difficult to see the roles that point mutations may play in miRNA origination. In conclusion, miRNAs stand as a most dynamic class of genes in terms of birth and death during evolution.

METHODS

Small RNA cloning, sequencing and bioinformatics analyses. The sequenced iso-1 strain of *D. melanogaster*, the sim6 strain of *D. simulans* and the MV2-25 strain of *D. pseudoobscura* were used in this study. Small RNAs were extracted from the heads of males of these three strains. RNA extraction, small RNA library construction, sequencing, and genome mapping of the small RNA reads are described in details in the **Supplementary Methods**.

Defining miRNA genes. We defined three levels of criteria for a hairpin structure. (i) Low stringency: the structure is greater than 55 nt in length, the structure does not have multiple loops, the free energy (ΔG) of the structure is less than or equal to –15 kcal/mol and the read is on one arm of the hairpin. (ii) Medium stringency: in addition to the low stringency criteria, the RANDFOLD probability of the hairpin structure is ≤ 0.2 and the probability of folding into that structure is ≥ 0.5 as determined by RNASHAPES. (iii) High stringency: in addition to the low stringency criteria, the RANDFOLD probability of the hairpin structure is ≤ 0.05 and the probability of folding into that structure is over 0.8. If two hairpin-forming sequences overlapped by ≥ 40 nt in the genomic location and matched all the reads used for identifying the two hairpins, the one with the more stable structure (that is, lower free energy) was adopted.

Among the 78 miRNA genes of *D. melanogaster* deposited in miRbase (as of July 2007), 59 (76%) pass the high stringency criteria. By randomly shuffling the 78 characterized miRNA genes 1,000 times, and categorizing the shuffled sequences according to the above three levels of criteria, we obtained a maximum false-discovery rate of 1% for the high stringency criteria.

The procedure used to identify the phylogenetic patterns of the miRNA genes among six *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura* and *D. virilis*) is described in the **Supplementary Methods**.

Background hairpins and type (ii) data. There are about 110,000 background hairpins that meet the high stringency criteria in *D. melanogaster*. Out of these hairpins, 1,861 have read sequences matching with reads collected from ref. 21, and more than 95% of them have reads only observed once (the total read number is over 2 million). The detailed analysis procedure is implemented in the **Supplementary Methods**.

Computer programs. The settings of the computer programs were implemented as described in the **Supplementary Methods**. The statistical tests were performed using the R package.

URL. Alignments of the primary sequences and the secondary structures, as well as the read information of the miRNA genes newly identified in this study, can be found at our website, http://pondside.uchicago.edu/wulab/microRNA.

Accession numbers. Small RNA sequences were deposited in the GEO database with accession numbers GSM246084, GSM246085 and GSM246086.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank V. Ambros, G. Ruvkun, A. Clark, J. Borevitz, J. Liang, M. Long and H. Liang for helpful comments and discussion; W. Wu for statistical consultation; D. Turissini for technical assistance in database construction; and the Washington University School of Medicine Genome Sequencing Center, Agencourt Bioscience, the Broad Institute and the Drosophila Population Genomics Project for making the *Drosophila* genome sequence data publicly available. Y.S. and S.S. were supported by the 973 Program of China (2007CB815701). The main portion of the work was funded by the Chicago Biomedical Consortium with support from the Searle Funds at the Chicago Community Trust.

AUTHOR CONTRIBUTIONS

C.-I.W., S.M.W. and R.W.C. designed this study; Q.W., S.K. and S.M.W. performed the experiments; J.L., Y.S., B.H. and S.S. did the analyses; J.L., C.-I.W. and R.W.C. contributed to the writing of this paper.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
2. Ambros, V. MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* **113**, 673–676 (2003).
3. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
4. Zamore, P.D. & Haley, B. Ribo-gnome: the big world of small RNAs. *Science* **309**, 1519–1524 (2005).
5. Ruby, J.G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
6. Lim, L.P. *et al.* The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991–1008 (2003).
7. Kloosterman, W.P. *et al.* Cloning and expression of new microRNAs from zebrafish. *Nucleic Acids Res.* **34**, 2558–2569 (2006).
8. Berezikov, E. *et al.* Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.* **16**, 1289–1298 (2006).
9. Bentwich, I. *et al.* Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37**, 766–770 (2005).
10. Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38**, 1375–1377 (2006).
11. Fahlgren, N. *et al.* High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of miRNA genes. *PLoS ONE* **2**, e219 (2007).
12. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
13. Hofacker, I.L. *et al.* Fast folding and comparison of RNA secondary structures (the Vienna RNA package). *Monatsh. Chem.* **125**, 167–188 (1994).
14. Bonnet, E., Wuyts, J., Rouze, P. & Van de Peer, Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**, 2911–2917 (2004).
15. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. & Giegerich, R. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22**, 500–503 (2006).
16. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144 (2006).
17. Drosophila Comparative Genome Sequencing and Analysis Consortium. Genomics on a phylogeny: evolution of genes and genomes in the genus *Drosophila*. *Nature* **450**, 203–218 (2007).
18. Hochsmann, M., Voss, B. & Giegerich, R. Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **1**, 53–62 (2004).
19. Manak, J.R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet.* **38**, 1151–1158 (2006).
20. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**, 655–660 (2004).
21. Ruby, J.G. *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* **17**, 1850–1864 (2007).
22. Landgraf, P. *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).
23. Stark, A. *et al.* Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res.* **17**, 1865–1879 (2007).
24. Allen, E. *et al.* Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat. Genet.* **36**, 1282–1290 (2004).
25. Rajagopalan, R., Vaucheret, H., Trejo, J. & Bartel, D.P. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* **20**, 3407–3425 (2006).
26. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21**, 36–44 (2004).