Testing hypotheses on the rate of molecular evolution in relation to gene expression using microRNAs

Yang Shen[a], Yang Lv[a], Lei Huang[a], Wensheng Liu[a], Ming Wen[a], Tian Tang[a], Rui Zhang[b], Eric Hungate[c], Suhua Shi[a], and Chung-I Wu[a,b,c,1]

[a]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen (Zhongshan) University, Guangzhou 510275, China; [b]Laboratory of Disease Genomics and Personalized Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China; and [c]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637

Edited by Masatoshi Nei, Pennsylvania State University, University Park, PA, and approved August 19, 2011 (received for review June 22, 2011)

There exists an inverse relationship between the rate of molecular evolution and the level of gene expression. Among the many explanations, the "toxic-error" hypothesis is a most general one, which posits that processing errors may often be toxic to the cells. However, toxic errors that constrain the evolution of highly expressed genes are often difficult to measure. In this study, we test the toxic-error hypothesis by using microRNA (miRNA) genes because their processing errors can be directly measured by deep sequencing. A miRNA gene consists of a small mature product (≈22 nt long) and a "backbone." Our analysis shows that (*i*) like the mature miRNA, the backbone is highly conserved; (*ii*) the rate of sequence evolution in the backbone is negatively correlated with expression; and (*iii*) although conserved between distantly related species, the error rate in miRNA processing is also negatively correlated with the expression level. The observations suggest that, as a miRNA gene becomes more highly (or more ubiquitously) expressed, its sequence evolves toward a structure that minimizes processing errors.

evolutionary rate | microRNA biogenesis | microRNA evolution

A host of factors, including expression level, tissue specificity, gene dispensability, number of interacting proteins, and local recombination rate, influence the rate of molecular evolution (1–7). Among them, gene expression level and ubiquity (or specificity) are of particular interest, because both are general characters and easily measurable. Both have been reported to be negatively correlated with the rate of molecular evolution (1, 2, 8, 9).

There are two classes of explanations for the negative correlations. The first class concerns the properties of the gene product (i.e., postprocessing functions). One conjecture is that the more highly or ubiquitously expressed genes might be functionally more important and, hence, might evolve more slowly (7, 10, 11). [However, protein-coding genes with essential functions, which include many housekeeping genes, do not always evolve slowly (3, 12, 13).] It is also possible that highly expressed genes may tend to have more interacting partners (14). The second class of explanations invokes the accuracy or efficiency in making the gene product (i.e., processing or preprocessing functions). For example, processing errors in protein synthesis are often assumed to result in toxic products (2, 15). Purifying selection acts strongly on highly expressed genes because of a greater quantity of erroneously synthesized products (2, 16).

The two classes of explanations are not mutually exclusive. Explanations of the first class, based on the properties of the mature products, have been extensively explored with mixed results (reviewed in refs. 12 and 15). A reason for the lack of a clear-cut conclusion may be the difficulties in separating the two classes of explanations using protein-coding sequences. Coding sequences determine the mature products but may often be important in processing as well. Furthermore, errors in processing, such as protein misfolding, are usually computationally inferred, rather than directly observed or measured (15).

In this study, we aim to test the second class of explanations concerning processing accuracy (or efficiency). Such tests would have to meet the following requirements. First, the portion of the gene that functions solely in processing can be separately analyzed from the portion that determines the mature product. This processing unit also needs to be evolutionarily conserved. Second, processing errors can be directly observed and measured. Because protein-coding genes often do not meet the two requirements, we propose to use microRNA (miRNA) genes to identify the connection between expression level and evolutionary rate.

The canonical structure of a miRNA gene is given in Fig. 1. The transcript is eventually processed into a mature miRNA (miR), which is ≈22 nt long (17–20). The imperfect complement of miR is denoted miR* (21, 22), which may sometimes be a functional product itself (*Discussion*). The sequence to the right of the miR:miR* stem is referred to as the loop end and, to the left, the stem extension. Hence, the miRNA gene can generally be divided into two parts, the mature miR and the backbone, consisting of miR*, the loop end, and the stem extension (Fig. 1).

The backbone, not being part of the mature product, maintains a proper hairpin structure for miR biogenesis (18, 23). By analogy, the stem extension of miRNAs is akin to the intron of mRNA; both are removed before the rest is exported to the cytoplasm. The loop region may be compared with the UTR of mRNA; both contribute to the processing but are not part of the final product. Furthermore, miRNAs are highly abundant in cells and the mature products, including those that are incorrectly processed, can be observed by deep sequencing. These properties make miRNAs suitable for investigating the relationship between evolutionary rate and gene expression.

Results

Evolutionary Conservation of miRNA Genes. Taking advantage of the available genomic sequences from the 12 species of *Drosophila* (24, 25), we calculated the evolutionary divergence in each part of the miRNA genes. Because the patterns are broadly similar across all levels of phylogenetic depth (2–65 million years; Table S1), we focused on the divergence between *D. virilis* and *D. melanogaster* for simplicity. The divergence between these two species is near the limit whereby synonymous distance can still be calculated.

In *Drosophila*, mature miRs are highly conserved. Most of the moderately to highly expressed miRNAs are completely conserved in this genus. The extreme conservation of miRs, noted many times before (26–28), may be explained by the large num-

Author contributions: Y.S., S.S., and C.-I.W. designed research; Y.S., Y.L., L.H., W.L., M.W., T.T., R.Z., E.H., S.S., and C.-I.W. performed research; Y.S., Y.L., L.H., W.L., M.W., T.T., R.Z., E.H., S.S., and C.-I.W. analyzed data; and Y.S. and C.-I.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ciwu@uchicago.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1110098108/-/DCSupplemental.

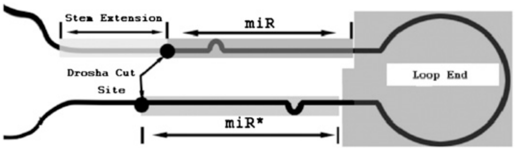

Fig. 1. The canonical structure of the miRNA gene. Mature miRNA and its complement are denoted miR and miR*, respectively. Loop end is the interval between miR and miR*. Stem extension is a stem structure, averaged to be ≈11 bp long, beyond the Drosha cut site. The four regions together constitute a miRNA gene. Although a small region beyond stem extension is often conserved as well, this region, having no obvious structure, is unsuitable for evolutionary analysis.

ber of targets (>50) each miR regulates. Likewise, proteins that have many interacting partners also tend to evolve slowly (12).

Interestingly, the observed divergence in the backbone (K_b) of miRs is only 11% of synonymous substitution (K_s), and is even lower than nonsynonymous substitution (13% of K_s). Because the backbone is not part of the mature product (with the possible exception of some miR*s; *Discussion*), its conservation cannot be explained by the functions of miRs. We hence hypothesize that the structure of the precursor that governs miR processing may be more relevant to the conservation. We carried out computer simulations in which the backbone is allowed to evolve while the structure of the miRNA precursors is preserved. The observed rates of divergence are then compared with the simulated values.

We completed two sets of simulations. The first set was based on thermodynamic constraints (29), permitting nucleotide substitutions only when they did not decrease the thermodynamic stability of the hairpin. In the second and more stringent set, we constrained the secondary structure to be invariant. In these simulations, wobbled pairing (G:U pairing) was allowed, because matched pairs (A:U or G:C pairing) would never change without wobbling.

We designate the simulated divergence in the backbone as K_{sim} and the neutral rate (i.e., mutational input in the simulation) as K_{neu}. K_s is used for K_{neu} for the observed set. In Fig. 2, we show that the thermostability model (the first set) is insufficient to explain the slow evolution of all parts of the backbone. Furthermore, the simulated rates under strict structural conservation (the second set) are still significantly higher than the observed (Kruaskal–Wallis test, $P < 0.001$; see Fig. 2). Thus, even complete conservation of the secondary structure fails to provide sufficient dampening of the evolutionary rate.

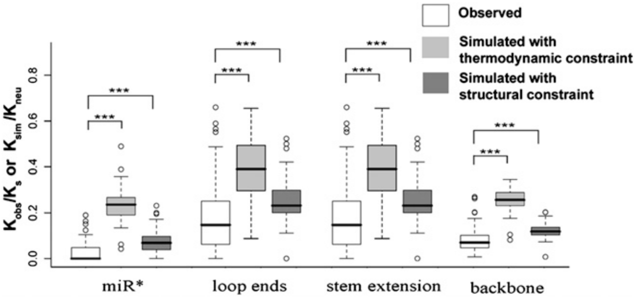

Fig. 2. Observed and simulated substitutions (K_{obs} vs. K_{sim}) between *D. melanogaster* and *D. virilis* for different parts of the miRNA gene. Backbone refers to the miRNA gene excluding miR. K_{obs}/K_s and K_{sim}/K_{neu} ratios are shown, where K_s is the average number of synonymous substitutions of the three genes flanking each miRNA and K_{neu} is the corresponding neutral divergence. Simulations were done under two different constraints—thermodynamic and structural (see text). Kruaskal–Wallis Test was performed to determine significance (***$P < 0.001$). Note that $K_{obs} < K_{sim}$ in all parts of the backbone. (Because we use K_s as a proxy for K_{neu}, the conclusion is somewhat conservative.)

Table 1 provides some details on sequence conservation. In simulations, we did not allow changes that disrupt the pairing configuration. In contrast, the unpaired sites change only half as often as simulations would allow (row 2). Fig. S1 presents four examples of conservation in the backbone. In the first two miRNA genes, there are 27 unpaired sites on the premiRNA (including the loop region and bulge sites). During the 200 million years of evolution in the four species, these unpaired sites never experienced nucleotide substitutions. Although misinferences in the secondary structure might account for some of the patterns, the number of sites involved suggests that the conservation of mispaired sites is not uncommon. For example, one might have expected some "U/U" mismatches to occasionally change to "U/C" mismatches (transitions) without affecting the secondary structure, but that was not observed. Apparently, unpaired sites are not all the same and identical secondary structures may not be functionally equivalent.

Divergence in miRNA Genes in Relation to Expression Level and Tissue Specificity. The strong conservation of the backbone makes the processing explanations plausible. In that case, a negative correlation between conservation and expression is expected. Using published data (refs. 30–32; see Table S2 for details) on genomic sequences and expression patterns in *Drosophila*, we analyzed the correlation between the rate of evolution and the level of expression (Fig. 3 *A–C*) and between the evolutionary rate and the ubiquity of expression (Fig. 3 *D–F*). Fig. 3*A* confirms the negative correlation between protein-coding transcript and their expression level (Spearman's rank correlation $\rho = -0.30$, $P < 0.001$), as reported (3). For the backbone of miRNAs, the correlation is similarly negative (Fig. 3*B*; $\rho = -0.29$, $P = 0.005$). The subregion of the backbone that shows the strongest negative correlation is the stem extension (Fig. 3*C*; $\rho = -0.26$, $P = 0.02$). The very slow evolution of miRNA genes follows the same trend as protein-coding genes (3). Because miRNAs are, on average, much more abundant than protein-coding transcripts (21), their extreme conservation appears to extend the range of this negative correlation.

We next examined the trend in tissue specificity. A tissue-specificity measure, "τ" (33), was calculated across the developmental stages of *Drosophila*. We use "τ^{-1}" in Fig. 3 *D–F* to denote the ubiquity of expression; the larger the value of τ^{-1}, the more ubiquitous the expression. Fig. 3*D* shows that, for protein coding genes, the expression level is negatively correlated with the ubiquity measure (Fig. 3*D*; $\rho = -0.42$, $P < 0.001$). Again, the correlation is similarly negative for the backbone of miRNAs (Fig. 3*E*; $\rho = -0.28$, $P = 0.008$). The subregion of the backbone that shows the strongest negative correlation between the evolutionary rate and expression ubiquity is the loop (Fig. 3*F*; $\rho = -0.27$, $P = 0.01$). We also carried out partial correlation analysis and principal component regression, and found that the conclusion holds (Table S3 and Fig. S2). Thus, miRNAs, like protein coding genes, evolve more slowly as the expression level increases or the tissue distribution becomes broader. Importantly, miRNAs show this pattern in the backbone, which is not part of the mature product.

Testing the Toxic-Error Hypothesis. The results suggest that the slower evolution of the more highly expressed genes is likely influenced by miR processing. Errors in processing either reduced the number of functioning molecules or created toxic effects on their own. Although errors in protein processing are difficult to study directly (34), errors in miR processing are measurable by deep sequencing, because mature miRs with a terminus off by even 1 bp can be observed (32, 35–38).

Error rate in miR processing is defined in this study as the proportion of reads with a 5′ (or 3′) end that is different from the most abundant miR sequence in a given sample (see *Methods* for

Table 1. Observed vs. simulated changes in the backbone of the 87 miRNA genes with an invariant miR in *Drosophila*

Structure	Observed	Simulated	Obs/Simulated
Paired sites in both species	152	162	0.94
Unpaired sites in both species	313	591	0.53
Paired vs. unpaired sites in the two species	262	0	—

Observed changes are between *D. melanogaster* and *D. virilis*. Changes are classified into 3 types—paired-to-paired, unpaired-to-unpaired, and paired-to-unpaired. In simulation, paired sites are not allowed to change to unpaired sites and vice versa. The salient observation is that unpaired sites change only half as often as expected (second row).

detail). To measure the error rate, we pooled available public sequence data. In total, 48 small RNA libraries from *Drosophila melanogaster*, and three each from *D. simulans* and *D. pseudoobscura*, were used to quantify the error rate with a grand total of ≈28 million reads mapped to premiRNA. A complete set of error rates is given in Table S4.

An important consideration is whether the "error" miRs, which differ from the dominant ones by at least 1 nt at the terminus, are functional. We hence counted miRs eluted from Argonaute1 (AGO1) antibody precipitates, which should more accurately represent the pool of miRs loaded onto RISC. In Fig. S3, it can be seen that the proportions of error miRs in the standard RNA samples, prepared without immunoprecipitation, are not different from those eluted from the immunoprecipitated samples. The error miRs appear to be loaded onto RISC in proportion to the canonical ones. In Fig. S4, we showed that the inferred target genes of the error miRs overlap with the target gene pool of the canonical miRs by only 55–60%.

Because the evolutionary rate of miRNA sequences may depend on the expression level and the error rate, we need to know the conservation of the two latter quantities. The correlation coefficient between *D. simulans* and *D. pseudoobscura* is 0.94 for the expression level (Fig. 4*A*) and 0.76 for the error rate (Fig. 4*B*). Although the evolutionary conservation in expression level is not unexpected (39), it is interesting to see that the error rate is also conserved.

In Fig. 4*B*, data points that deviate from the 45-degree line represent miRNAs with an evolving error rate, which may be influenced by small differences in miRNA sequences. Different kinds of bulges and mismatches may yield slightly different tertiary structures that would impact miRNA processing. For example, U/U and U/C mismatches are overrepresented near the Drosha and Dicer processing sites and provide a stable reference for scissile phosphates (23, 40). Also, "G/A" mismatches in the stem may interfere with Dicer binding (40). In Fig. 4*B*, filled circles denote those miRNAs with unconserved nucleotides adjacent to the mature miRs and open circles denote those with fully conserved sites. By the Kolmogorov–Smirnov test, the filled circles deviate from the slope line more than the open circles ($P < 0.05$). The observation shows that a small difference in miRNA sequence can affect the error rate.

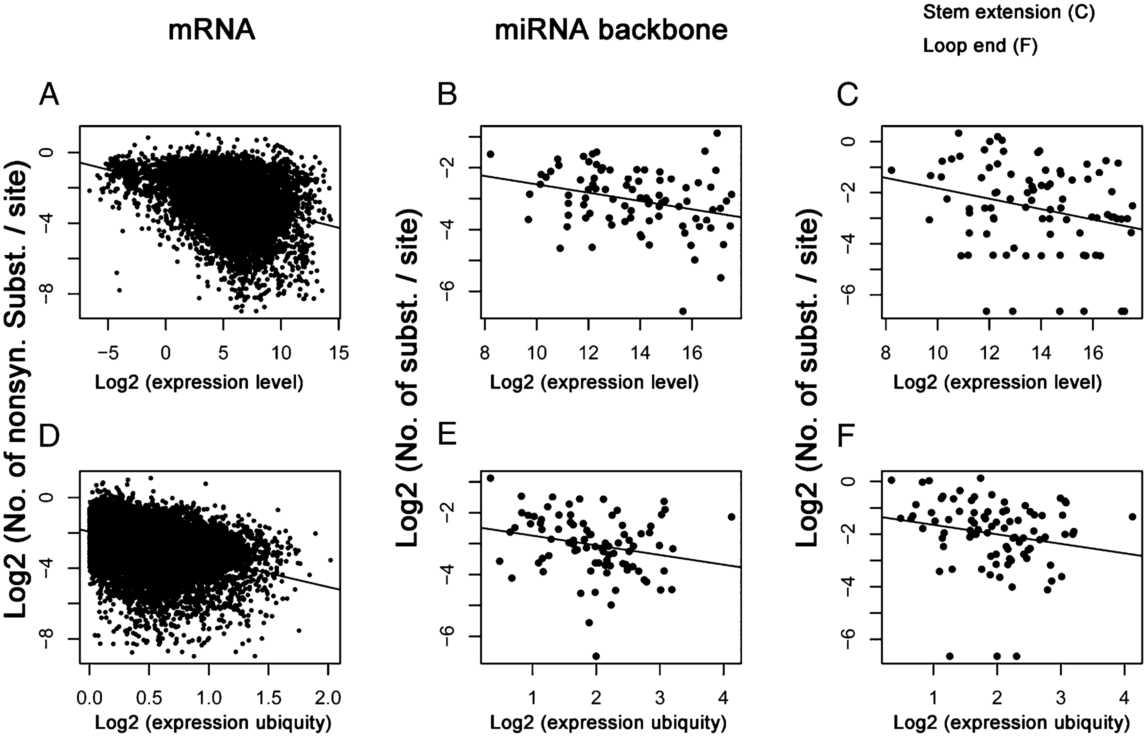

Fig. 3. Correlation between evolutionary divergence and expression level (*A*–*C*) or expression ubiquity (*D*–*F*) in mRNA or in miRNA backbone. Evolutionary divergence is measured between *D. melanogaster* and *D. virilis*. See *Materials and Methods* for the measure of expression level and ubiquity (τ^{-1}).

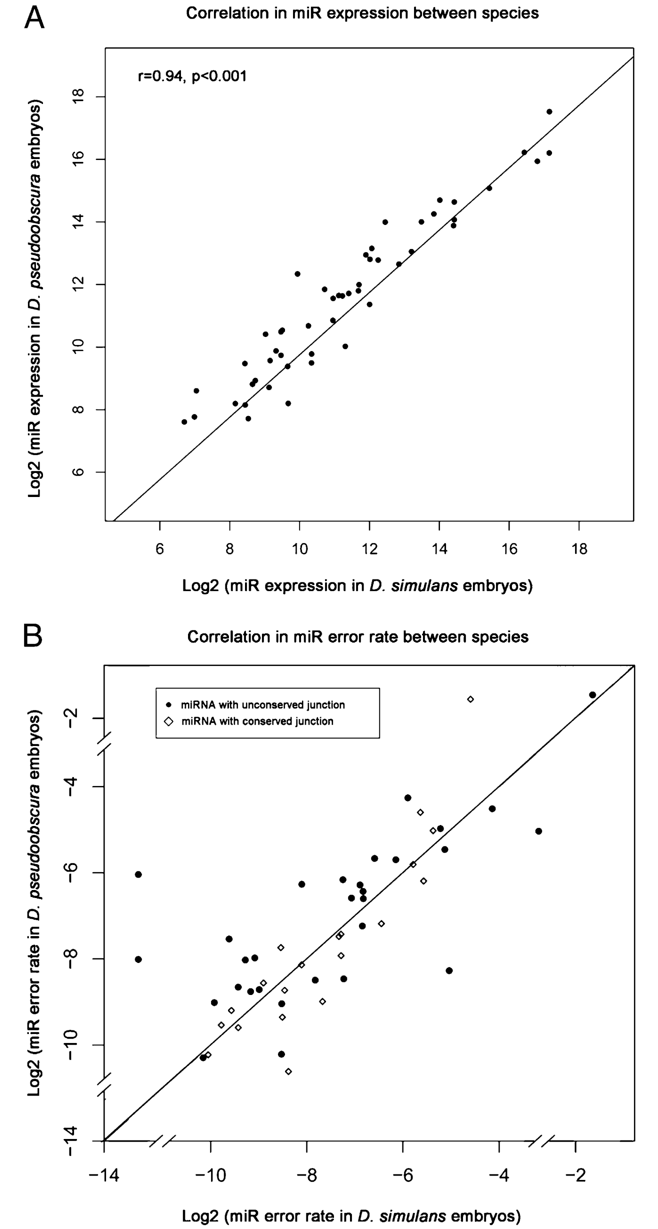

Fig. 4. (*A*) Correlation in miR expression between species. See *Materials and Methods* for the source of data and measure of expression level. (*B*) Correlation in the miR error rate at the 5′ terminus between species. miRNAs with no fewer than 1,000 reads were used to estimate the error rate. Open circles denote those miRNAs with fully conserved nucleotides adjacent to the mature miRs, and filled circles denote those with unconserved sites. (See *Materials and Methods* for the delineation of the adjacent sites.)

Knowing that the error rate is conserved, influenced by slight changes in the backbone sequence and that the error products are likely functional, we can test the toxic-error hypothesis directly. Under this hypothesis, miRNA processing error is constrained by selection and the selective pressure is stronger for more highly expressed genes (32). Fig. 5 shows the correlation between the error rate and the level of expression for the 5′ and 3′ ends, separately. The error rate at the 3′ end is, on average, 30 times greater than the error rate at the 5′ end. The high error rate at the 3′ end is hardly surprising because an incorrect 3′ end does not have as much an impact on miRNA targeting as an incorrect 5′ end. For the 3′ end error, the correlation with the expression level is negative, but not significant (Spearman's rank correlation $\rho = -0.11$, $P = 0.28$). In contrast, the correlation between the 5′ end error and miR expression is significantly negative (Spearman's rank correlation $\rho = -0.32$, $P < 0.001$). A similar trend was observed between *D. simulans* and *D. pseudoobscura* (Fig. S5).

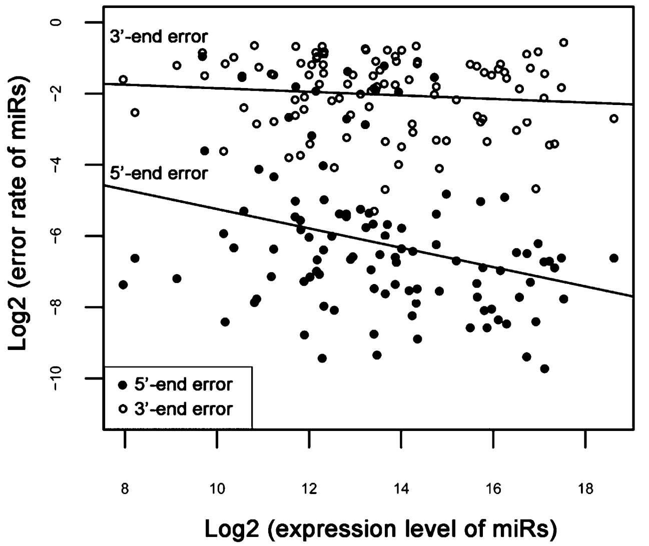

Fig. 5. Correlation between miR error rate and miR expression for all miRs with at least 1,000 reads. Errors at the 5′ and 3′ ends are treated separately.

Discussion

In this study, we found that miRNAs evolve more slowly as the expression level becomes higher (or the tissue distribution becomes broader), similar to the pattern seen in protein coding genes. The error rate, especially in the 5′ end where accuracy in miRNA processing is crucial, also becomes lower as the expression level increases. Because the trend is observed in the backbone, the explanation for the negative correlations should be based on miR processing, rather than the properties of the mature product.

How much of the observed processing errors could be attributed to experimental artifacts, such as sequencing errors? For example, additional nucleotides might be added to the 5′ or 3′ end during library construction. For three reasons, this artifact should not be a serious concern. First, the extra nucleotide in the 5′ or 3′ end usually corresponds to the one in the genomic sequence, hence arguing against the addition of a random nucleotide. (We estimated ≈20% of the reads have a nontemplated extranucleotide, mostly A or U, because it is known that some miRs are naturally modified at their 5′ or 3′ end with a uridyl or adenyl residue (41, 42). The modification of miRNA ends is a known biological mechanism rather than experimental artifacts.)

Second, the error rates in the 5′ end of miR and the 3′ end of miR* are positively correlated (Pearson correlation $r = 0.32$, $P < 1e^{-5}$). This correlation can be more easily explained by processing inaccuracy, which creates the two types of error products in one reaction, than by the random addition of nucleotides to the ends of miRs and miR*s. Third, the inferred errors in processing are comparable between results from either the "Roche 454" or "Illumina Solexa" platform, even though the reaction chemistry (and presumably the error pattern) varies greatly from one platform to the other.

Another potential bias created by the analytical procedure is the use of only perfectly matched reads. To test possible biases in this procedure, which ignores sequencing errors, we also used reads with up to two mismatches against the reference genome.

Error rate estimates from the two methods are highly concordant (Pearson correlation $r = 0.96$, $P < 1e^{-10}$). Thus, the error rate estimated from perfectly mapped reads does not appear to be biased.

Although the negative correlation between expression and evolution has been explored by studying coding genes, the analysis of miRNA evolution has provided direct evidence beyond what coding genes can offer. For protein coding genes, the negative correlation can be explained either by the properties of the mature gene products or by the accuracy in processing. For miRNA genes, different parts of the miRNA precursor offer different types of information. The sequence of miR, like the coding sequence, is presumably important for both function and processing. The role of miR* is sometimes ambiguous. Some miR* products can be functional (43, 44), whereas the rest are unlikely to be so. The remaining two parts of the miRNA precursor, the loop end and the stem extension, are very unlikely to give rise to functional products. In our own data of several million reads, no more than two copies of small RNAs were observed from the loop end of each miRNA gene.

For stem extensions and loop ends, and for the majority of miR*s, the evolutionary conservation can best be explained by the need to preserve a secondary and higher-order structure conducive to miR processing. Hence, the backbone of miRNA can be viewed as the structure that ensures the production of a correct mature gene product. A remaining question is whether the evolutionary constraint is driven by the need for a sufficient number of correct mature products (processing efficiency) or by the need to minimize the number of error products (processing accuracy).

There are reasons why processing efficiency might be important. For example, many motifs of RNA binding protein are located in the backbone, and highly or ubiquitously expressed miRNAs may require intensive regulation, relative to that for lowly expressed miRNAs. Such regulation could be in the form of RNA editing or biogenesis tuning (reviewed in ref. 45). However, it is unclear whether this explanation could take into account the negative correlation between error rate and expression level. If processing efficiency is crucial, one might have expected the less highly expressed genes to be less error-prone (and more slowly evolving), contrary to the observations.

Among the explanations suggested, the toxic-error hypothesis seems quite reasonable when all observations are considered together. In the biogenesis of miRNAs, the error products would be more numerous as the level of expression increases. Products with errors in the 5′ end would contain a new seed and target a different set of mRNAs (Fig. S4). Phenotypic changes can be strong when “novel miRNAs” are introduced (46). Selection would favor a reduction in error rate in any miRNA as it becomes more highly expressed. Therefore, we expect the sequence of highly expressed miRNAs to evolve into a “local minimum” in processing errors. This local minimum may constrain the sequence evolution of these genes.

Materials and Methods

Collection of miRNAs. Precursor sequences of 152 miRNAs of *D. melanogaster* were downloaded from miRBaseV14.0 (47) or from previous literature (30, 34, 48–50). The intronic miRNAs that bypass the Drosha processing pathway (51, 52) were not included in our analysis. All reads used in the analysis could be uniquely mapped to known miRNA genes.

Comparative Genomics of miRNAs. Based on the genomic coordinates of the 152 miRNAs in *D. melanogaster*, the orthologous sequences in the other 11 *Drosophila* species (24) were parsed out from the whole genome sequence alignments of the 12 *Drosophila* species (http://genome.ucsc.edu). The 87 miRNAs that are invariant in mature sequence across the 12 *Drosophila* species were used in studying the relationship of expression level and evolutionary rate. Divergence measured by nucleotide substitutions (K_{obs} for miRNAs, K_s and K_a for coding regions) was calculated by using the same methods as in Lu et al. (28).

Computer Simulation. In the first set of computer simulations, we considered the constraints on a miRNA precursor imposed by thermodynamic stability. For each miRNA precursor, we generated random mutations on the sequence one at a time. For each mutation, the starting point is the sequence of extant miRNA genes of *D. melanogaster*. We folded the derived sequence by using RNAFOLD (53). A mutation was accepted if the derived sequence could form a hairpin that meets the following criteria: (*i*) the secondary structure has $\Delta G < -15$ kcal/mol; (*ii*) the mutation is not on the miR itself (as in the observations); and (*iii*) the secondary structure of the derived sequence is more stable than the original one. In other words, the cumulative change in ΔG is less than zero.

This process was reiterated, and the sequence with the new mutation, if accepted, was used as the starting sequence for the next cycle. The total number of cycles was made equal to the expected neutral mutations based on the K_s values of the five adjacent genes. The simulated evolutionary rate is thus equal to the number of accepted mutations divided by the total number of mutations introduced (K_{sim}/K_{neu}). For each miRNA gene, the process was repeated for 100 rounds and the average evolutionary rate was calculated.

The second set of simulations used the same protocol except criterion (*iii*). In this second set of simulations, the secondary structure is not allowed to change in each cycle.

Expression Level and Tissue Specificity. We collected 48 small RNA from public *Drosophila* deep sequencing datasets from National Center for Biotechnology Information Gene Expression Omnibus (GEO) database. The complete lists of GEO accession number are in Table S1. We used small RNA data from 28 libraries (as shown in Table S2) that were generated by Illumina from different development stages or tissues of *D. melanogaster* to estimate the expression level and tissue specificity. The raw reads were mapped to premiRNA with BOWTIE (54), and we only kept perfect matches. The expression level of miRNA in each library was measured as normalized counts of reads mapped to premiRNA. The reads were normalized by the number of mapped reads divided by the total number of mapped reads and multiplied by 1 million. The expression level of each miRNA was measured by the maximum expression level across multiple libraries. The tissue ubiquity was measured as “$\frac{1}{\tau}$” (32), which is given as the following formula:

$$\tau = \frac{\sum_{i=1}^{n} 1 - \left(\frac{\log_2(N_i)}{\log_2(N_{\max})}\right)}{n-1},$$

where N is the normalized reads count and n is the number of tissues. To measure the expression level and tissue ubiquity of mRNA, we adopted the procedure that was used in analysis of miRNA by using the transcriptome of different developmental stages of *D. melanogaster* (55).

Error Rate in the Biogenesis of miRs. All 48 libraries of small RNA generated by Illumina or Roche 454 were used to quantify the processing error rates of miRNAs. This collection comprises ≈20 million, 2.9 million, and 5.8 million miRNA reads that can be perfectly mapped to the premiRNA sequences of *D. melanogaster, D. simulans,* and *D. pseudoobscura,* respectively. The parameters are used in mapping is “BOWTIE -n 0 -v 0 -f -k 50 -a -p 10”. For each miRNA in each *Drosophila* species, the most abundant sequence was defined as the “correct” miR. Error rate is the proportion of those reads with a 5′ or 3′ end that is different from the correct ones, among the total reads from the same arm of the miRNA gene.

The following formula is used to calculate the error rate. In this example, miR is located on the 5′ arm of premiRNA.

$$error_{miR} = \frac{\sum_{i\in[1,m], i\neq j} x_i}{\sum_{i\in[1,m]} x_i},$$

where x_i is the total number of reads with the 5′ end at the *i*th position in premiRNA, the correct mature miRNA starts on the *j*th position, and the 5′ arm of premiRNA is from the first to the *m*th position. Only reads with unique and perfect mapping are used to calculate error rate. Moreover, miRNAs with identical mature products, such as dme-mir-281-1 and dme-mir-281-2, are excluded from our analysis. We also compared reads with zero to two mismatches to evaluate the error rate (with parameters: “BOWTIE -n 2-v 2 -f -k 50 -a -p 10”) and the results are highly correlated (Pearson correlation, $r = 0.96$, $P < 1e^{-10}$; *Discussion*). To evaluate the error rate more precisely, only miRs with at least 1,000 reads were used. Instead of using small RNA profiles from *D. melanogaster*, we used embryonic small RNA profiles from

D. simulans (GSM343915) and *D. pseudoobscura* (GSM343916), which were collected from identical development stage (0–24 h embryo), to estimate the divergence of error rate and level of expression. The correlation of error rate between species was measured by Pearson correlation coefficient (*r*) using R (www.r-project.org). To define the miRNA with conserved miR and miR* adjacent sites, we considered three neighboring nucleotides of miR or miR* ends at stem extension or loop ends region. The miRNA has conserved adjacent sites if a total of 12 sites are conserved between *D. simulans* and *D. pseudoobscura*.

ACKNOWLEDGMENTS. We thank Jian Lu, Yong Zhang, Bin He, Han Liang, Jianrong Yang, and two anonymous reviewers for helpful comments and discussions. This work was supported by National Science and Technology Major Project 2009ZX08010-017B; National Program on Key Basic Research Project 973 Program Grant 2007CB815701; National Natural Science Foundation Grants 31130069, 31170308, and 40976081; Chinese Academy of Sciences Grant KSCX1-YW-22; State Key Laboratory of Biocontrol Grant SKLBC2011A06,SKLBC2010B01; Natural Science Foundation of Guangdong Province Grant 10451027501005628; and the National Institutes of Health.

1. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.
2. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.
3. Larracuente AM, et al. (2008) Evolution of protein-coding genes in Drosophila. *Trends Genet* 24:114–123.
4. Betancourt AJ, Presgraves DC (2002) Linkage limits the power of natural selection in Drosophila. *Proc Natl Acad Sci USA* 99:13616–13620.
5. Wall DP, et al. (2005) Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA* 102:5483–5488.
6. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* 296:750–752.
7. Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* 12:962–968.
8. Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 17:68–74.
9. Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 168:373–381.
10. Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. *Nature* 411: 1046–1049.
11. Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. *Annu Rev Biochem* 46: 573–639.
12. Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7:337–348.
13. Wang Z, Zhang J (2009) Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet* 5:e1000329.
14. Fraser HB, Hirsh AE (2004) Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol Biol* 4:13.
15. Drummond DA, Wilke CO (2009) The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* 10:715–724.
16. Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* 6:421.
17. Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* 17:3011–3016.
18. Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. *Science* 303:95–98.
19. Kim VN (2005) MicroRNA biogenesis: Coordinated cropping and dicing. *Nat Rev Mol Cell Biol* 6:376–385.
20. Lee Y, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425:415–419.
21. Lim LP, et al. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev* 17: 991–1008.
22. 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.
23. Han J, et al. (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* 125:887–901.
24. Clark AG, et al.; Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–218.
25. Stark A, et al.; Harvard FlyBase curators; Berkeley Drosophila Genome Project (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 450:219–232.
26. Berezikov E, et al. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 120:21–24.
27. Grad Y, et al. (2003) Computational and experimental identification of C. elegans microRNAs. *Mol Cell* 11:1253–1263.
28. Lu J, et al. (2008) Adaptive evolution of newly emerged micro-RNA genes in Drosophila. *Mol Biol Evol* 25:929–938.
29. Bonnet E, Wuyts J, Rouzé P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* 20:2911–2917.
30. Stark A, et al. (2007) Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res* 17:1865–1879.
31. Czech B, et al. (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature* 453:798–802.
32. Berezikov E, et al. (2011) Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. *Genome Res* 21: 203–215.
33. Yanai I, et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21:650–659.
34. Kramer EB, Farabaugh PJ (2007) The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA* 13:87–96.
35. Seitz H, Ghildiyal M, Zamore PD (2008) Argonaute loading improves the 5′ precision of both MicroRNAs and their miRNA* strands in flies. *Curr Biol* 18:147–151.
36. Wu H, Ye C, Ramirez D, Manjunath N (2009) Alternative processing of primary microRNA transcripts by Drosha generates 5′ end variation of mature microRNA. *PLoS ONE* 4:e7566.
37. Lee LW, et al. (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA* 16:2170–2180.
38. Park JE, et al. (2011) Dicer recognizes the 5′ end of RNA for efficient and accurate processing. *Nature* 475:201–205.
39. Lu J, et al. (2008) The birth and death of microRNA genes in Drosophila. *Nat Genet* 40: 351–355.
40. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452:51–55.
41. Burroughs AM, et al. (2010) A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res* 20:1398–1410.
42. Kim YK, Heo I, Kim VN (2010) Modifications of small RNAs and their associated proteins. *Cell* 143:703–709.
43. Okamura K, et al. (2008) The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nat Struct Mol Biol* 15:354–363.
44. Yang JS, et al. (2011) Widespread regulatory activity of vertebrate microRNA* species. *RNA* 17:312–326.
45. Krol J, Loedige I, Filipowicz W (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet* 11:597–610.
46. Tang T, et al. (2010) Adverse interactions between micro-RNAs and target genes from different species. *Proc Natl Acad Sci USA* 107:12935–12940.
47. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 34(Database issue):D140–D144.
48. Ruby JG, et al. (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res* 17:1850–1864.
49. Lai EC, Tomancak P, Williams RW, Rubin GM (2003) Computational identification of Drosophila microRNA genes. *Genome Biol* 4:R42.
50. Aravin AA, et al. (2003) The small RNA profile during Drosophila melanogaster development. *Dev Cell* 5:337–350.
51. Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature* 448:83–86.
52. Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* 130:89–100.
53. Hofacker IL, et al. (1994) Fast Folding and Comparison of Rna Secondary Structures. *Monatsh Chem* 125:167–188.
54. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
55. Graveley BR, et al. (2011) The developmental transcriptome of Drosophila melanogaster. *Nature* 471:473–479.