LETTER

Lowly Expressed Human MicroRNA Genes Evolve Rapidly

Han Liang and Wen-Hsiung Li*†*

*Department of Ecology and Evolution, University of Chicago; and †Biodiversity Research Center, Academia Sinica, Taipei, Taiwan 115

To study the evolution of human microRNAs (miRNAs), we examined nucleotide variation in humans, sequence divergence between species, and genomic clustering patterns for miRNAs with different expression levels. We found that expression level is a major indicator of the rate of evolution and that ~30% of currently annotated human miRNA genes are almost free of selective pressure.

Introduction

MicroRNAs (miRNAs) play a key role in posttranscriptional regulation of gene expression (Ambros 2003, 2004; Bartel 2004; Zamore and Haley 2005). The miRNA genes are first transcribed as primary miRNA transcripts (pri-miRNAs), which are processed to form hairpin precursor molecules (pre-miRNAs), which are then processed to become ~22 nt mature miRNAs. A mature miRNA can repress target gene expression by binding to the 3′ untranslated region of the target gene. It has been suggested that more than 30% of human protein-coding genes are regulated by miRNAs (Lewis, Burge, and Bartel 2005). The number of miRNAs identified has been increasing steadily (Griffiths-Jones et al. 2006), and currently ~700 miRNAs have been annotated in the human genome (the miRBase), nearly three times the number 5 years ago (Lim et al. 2003). But a basic question remains, "what proportion of these miRNA genes is functionally important, thereby being maintained by natural selection?" Although it is well accepted that highly expressed miRNAs tend to be functionally important and evolutionarily conserved, the functional relevance of lowly expressed miRNAs is less clear. In particular, several deep sequencing studies have revealed a class of lowly expressed "miRNA precursor–like" hairpins that tend to be species specific (Berezikov et al. 2006; Fahlgren et al. 2007; Lu, Shen, et al. 2008). The purpose of this study is to investigate the possible effect of the expression level of an miRNA on the level of sequence variation in humans and the rate of nucleotide substitution between species.

We focused on the human miRNA genes in the miRBase database (Griffiths-Jones et al. 2006), which is widely accepted as the standard for miRNA annotation. Our analysis only included those miRNA genes that have a unique genome location, have available tissue expression data, and do not share a mature sequence with other miRNAs. We classified these miRNAs into different groups, using a mammalian miRNA expression atlas of many tissues and cell types (Landgraf et al. 2007), rather than using deep sequencing data, because the latter type of data may be biased toward a specific tissue. We considered two measures for the miRNA expression level: n, the total number of clones detected in the expression atlas, and p, the maximal composition percentage among the 172 small RNA libraries studied; the composition percentage of an miRNA is defined as the number of clones of the miRNA divided by the total number of miRNA clones. These two measures were almost identical in terms of expression level ranking (Spearman rank test $R\text{s} = 0.96, P < 9 \times 10^{-79}$), so we used only n in all subsequent analyses. In total, 383 human miRNAs were classified into four groups: the high, intermediate, low, and "0 expression" groups (see Methods). The classification allows each group to contain a similar number of miRNA genes, thereby avoiding any potential bias due to differences in group size in the analyses. In particular, the 0 expression group contains miRNAs that were not detected in any RNA library by the clone-based approach (~1,300 sequenced clones per library, corresponding to a detection limit of 0.1% of total small RNAs). To see if this clone-based approach, which has a relatively low sequencing depth, has a sampling bias, we also analyzed a set of Solexa deep sequencing data in human HeLa cells (Friedlander et al. 2008). We found that all the miRNAs in the 0 expression group have indeed an extremely low expression level (i.e., no hit or only several hits out of ~800,000 reads), confirming the validity of our miRNA group classification.

We first studied the sequence variation of these miRNAs in humans. The polymorphism level in a pre-miRNA gene relative to that in its flanking regions is taken as an index of selection intensity because the flanking regions are likely selectively nearly neutral (Saunders, Liang, and Li 2007). Using the polymorphism data in dbSNP128 (Sherry et al. 2001), for each group of pre-miRNAs, we calculated the ratio of average single-nucleotide polymorphism density in the pre-miRNAs to that in their same-length flanking regions. Figure 1*a* shows that the ratios for the high, intermediate, and low expression groups are much lower than 1, suggesting the presence of strong purifying selection. In contrast, the ratio for the 0 expression group is 1.06, indicating absence of purifying selection among most members in this group.

Next, we examined the miRNA sequence divergence between primate species. Using the University of California–San Cruz (UCSC) genome alignments (Blanchette et al. 2004), we identified the human–macaque orthologous sequences of pre-miRNAs and their same-length flanking regions (Liang, Lin, and Li 2008). For each group, we calculated the ratio of average sequence divergence in the pre-miRNAs to that in their flanking regions. Again, we found that the ratios for the high, intermediate, and low expression groups are much lower than 1, but the ratio for the 0 expression group is 1.09 (fig. 1*b*). In addition, using the 4-fold degenerate sites in the nearest protein-coding genes as another (nearly) neutral control, we obtained a similar ratio (0.95 ± 0.08) for the 0 expression group.

Key words: microRNA expression, sequence divergence, spatial clustering, selective pressure.

E-mail: whli@uchicago.edu.

Mol. Biol. Evol. 26(6):1195–1198. 2009
doi:10.1093/molbev/msp053
Advance Access publication March 19, 2009

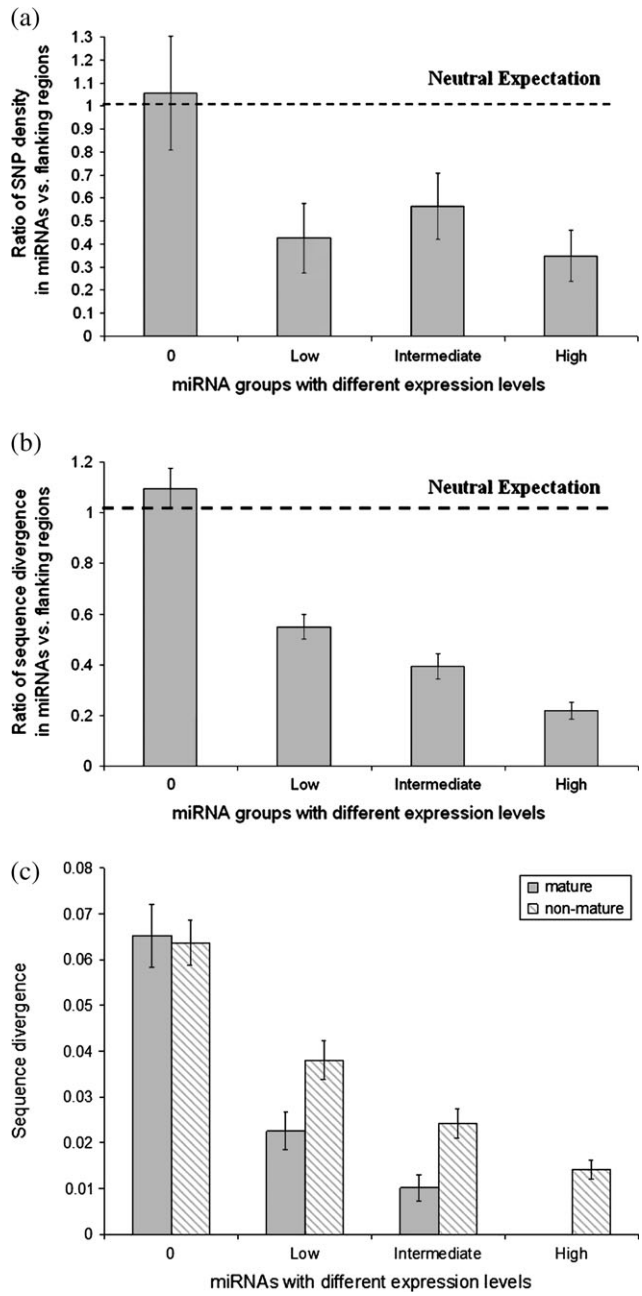

FIG. 1.—Nucleotide variation within and between species for human miRNAs with different expression levels. (*a*) Nucleotide variation of miRNA genes and their flanking regions in humans, (*b*) sequence divergence of miRNA genes and their flanking regions between human and macaque, and (*c*) sequence divergences in the mature regions and in the rest of miRNA genes between human and macaque. The gray bars represent the mature regions, and the striped bars represent the nonmature regions. The dot line represents the neutral expectation, that is, all the miRNAs in a group are selectively neutral. The error bars indicate ± the standard deviations calculated from 1,000 bootstrap samples.

If an miRNA gene is functional, we expect its mature region to be under stronger selective constraint than the rest of the gene (Lewis, Burge, and Bartel 2005; Chen and Rajewsky 2006). In the high, intermediate, and low expression groups, the sequence divergence in the mature regions is indeed significantly lower than that in the remaining regions of pre-miRNA (fig. 1*c*). However, in the 0 expression group, the divergence levels for the two types of regions are almost the same. These results strongly suggest that most of the miRNA genes in the 0 expression group have been subject to weak or no selection. This view is also supported by the sequence divergence between human and chimpanzee (data not shown). Moreover, using a genome-wide BLAT (Kent 2002) search, we found that only 88% of the human miRNAs in the 0 expression group have a homolog in macaque, which is significantly lower than that of the other three groups (94%, $P < 0.04$), suggesting that miRNAs in the 0 expression group tend to turn over quickly in evolution.

Finally, we investigated the spatial clustering pattern of miRNAs in the human genome. It is well known that miRNAs tend to be clustered in chromosomal regions

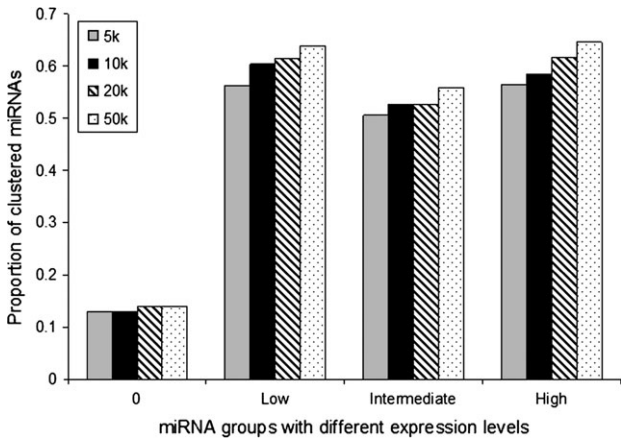

FIG. 2.—Spatial clustering patterns of human miRNAs with different expression levels. The gray bars indicate the proportions at a clustering distance limit of 5 kb, the black bars for 10 kb, the striped bars for 20 kb, and the dotted bars for 50 kb.

(Bartel 2004). This nonrandom spatial distribution may facilitate the functional coordination among miRNAs so that neighboring miRNAs tend to have similar expression profiles (Baskerville and Bartel 2005). As shown in figure 2, the proportions in the high, intermediate, and low groups are all very high (~60%), but the proportions of miRNAs in the 0 expression group are very low (~13%). This contrasting pattern supports the notion that the vast majority of miRNAs in the 0 expression group are randomly distributed over the genome, rather than being clustered with established miRNA genes.

Given the sharply contrasting patterns between the 0 expression group and the other three groups, one may want to examine the target sites of these miRNA groups (e.g., selective pressure or target abundance). Unfortunately, such analyses are not feasible at this moment. Currently, miRNA target prediction not only requires the sequence match to the seed of an miRNA but usually also relies on evolutionary conservation as a filter. Unlike highly expressed miRNAs, many miRNAs in the 0 expression group have a recent origin (e.g., in primates), and thus, conservation cannot be invoked as a filter to remove false sites. Therefore, it is difficult to identify the target sites of these young miRNAs with a reasonable accuracy. In the future, when many experimentally determined target sites are available, it will be interesting to conduct such an analysis.

Taken together, our results strongly suggest that the vast majority of miRNA genes in the 0 expression group, which comprise 26% of our miRNA data set, have been subject to little selective constraint. Most miRNAs in this group may only occasionally enter the small RNA biosynthesis pathway. How these miRNAs were derived in evolution remains unclear. Of course, our group-based results do not rule out the possibility that an individual 0 expression miRNA may play an important functional role in human, such as miRNA lsy-6, which is only expressed in one neuron in *Caenorhabditis elegans* but plays an important functional role (Johnston and Hobert 2003).

Recently, several deep sequencing studies have revealed a large number of miRNA precursor–like hairpins in primates, Arabidopsis, and Drosophila (Berezikov et al. 2006; Fahlgren et al. 2007; Lu, Shen, et al. 2008); and evolutionary analyses on the birth-and-death process of these hairpins suggest that most of them are evolutionarily neutral. These results and ours are largely compatible. We speculate that the 0 expression miRNAs annotated in current miRBase and many others (not included in the database or to be discovered by future deep sequencing) form a pool of raw materials for evolution. Due to the extremely low expression level, their effect on potential target mRNAs would, in most cases, be weak or negligible. Occasionally, a lowly expressed miRNA is selectively favored (Zhang et al. 2007; Lu, Fu, et al. 2008), integrated into the miRNA target regulatory network and then maintained in long-term evolution.

Our work also calls attention to the criteria of miRNA identification, which is an important issue in this fast-moving field. In the current practice, given a piece of evidence on the expression of a candidate small RNA, the major criterion for calling it an miRNA gene is whether the corresponding genomic sequence can be folded into a stable hairpin structure according to folding energy. However, a hairpin structure may not mean function. If we want to focus on the miRNAs with functional relevance, a criterion on the relative expression level should be added.

Methods

miRNA Expression Classification

From the miRBase (version 11.0) annotation and a recent mammalian miRNA expression atlas of small RNA library sequencing (Landgraf et al. 2007), we obtained a data set of 383 human miRNAs, each of which has a unique genome location and has tissue expression data. The miRNA genes that share a same mature miRNA sequence were excluded from our analysis because their expression level was hard to define. The miRNAs were classified into four groups, each containing a similar number of members: 1) $n \geq 100$, the high expression group, which contains 94 miRNAs; 2) $100 > n \geq 15$, the intermediate expression group, which contains 93 miRNAs; 3) $15 > n > 0$, the low expression group, which contains 96 miRNAs; and 4) $n = 0$, the 0 expression group, which contains 100 miRNAs that were not detected in the clone sequencing–based survey. (The miRNAs in the 0 expression group were mainly detected by deep sequencing techniques.) The classification data set is available in supplementary file (Supplementary Material online).

We obtained the Solexa sequencing data in human HeLa cells from the NCBI GEO database (GenBank accession number GSE10829). The correspondence between annotated miRNAs and sequence reads was established by the Blast program with an E value cutoff of 0.001.

miRNA Sequence Analysis

Human polymorphism data were obtained from dbSNP128 and mapped to the pre-miRNAs as well as the same-length flanking regions according to the miRBase annotation. The human–macaque orthologous sequences of miRNAs and their flanking regions were constructed from

the UCSC Human/Rhesus (hg18/rheMac2) and Rhesus/Human (rheMac2/hg18) pairwise genome alignments, and the pairwise alignments of 350 miRNAs were obtained. The mature and nonmature regions were classified based on the miRBase annotation. For each group, the standard deviations were calculated from 1,000 bootstrap samples.

Clustering Analysis of Human miRNAs

If two miRNA genes are located in the same chromosome/strand and their distance is smaller than a given distance threshold, they are classified into a cluster. Using different distance limits (5, 10, 20, and 50 kb), we clustered all the annotated miRNAs into clusters (the total numbers of miRNA clusters varied from 494 at 5 kb to 461 at 50 kb), and we then calculated the proportions of clustered miRNAs in each group. For each group, the standard deviations were calculated from 1,000 bootstrap samples.

Supplementary Material

Supplementary file is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jian Lu and Yang Shen for helpful discussion and three reviewers for valuable comments. This work was supported by National Institutes of Health grants (GM30998 and GM081724) to W.H.L.

Literature Cited

Ambros V. 2003. MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell. 113:673–676.

Ambros V. 2004. The functions of animal microRNAs. Nature. 431:350–355.

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

Baskerville S, Bartel DP. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 11:241–247.

Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH. 2006. Diversity of microRNAs in human and chimpanzee brain. Nat Genet. 38: 1375–1377.

Blanchette M, Kent WJ, Riemer C, et al. (12 co-authors). 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Chen K, Rajewsky N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet. 38:1452–1456.

Fahlgren N, Howell MD, Kasschau KD, et al. (11 co-authors). 2007. High-throughput sequencing of arabidopsis microRNAs: evidence for frequent birth and death of miRNA genes. PLoS ONE. 2:e219.

Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. 2008. Discovering microRNAs from deep sequencing data using mirdeep. Nat Biotechnol. 26: 407–415.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. MiRbase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34:D140–D144.

Johnston RJ, Hobert O. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. Nature. 426: 845–849.

Kent WJ. 2002. Blat–the blast-like alignment tool. Genome Res. 12:656–664.

Landgraf P, Rusu M, Sheridan R, et al. (51 co-authors). 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell. 129:1401–1414.

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.

Liang H, Lin YS, Li WH. 2008. Fast evolution of core promoters in primate genomes. Mol Biol Evol. 25:1239–1244.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. 2003. Vertebrate microRNA genes. Science. 299:1540.

Lu J, Fu Y, Kumar S, Shen Y, Zeng K, Xu A, Carthew R, Wu CI. 2008. Adaptive evolution of newly emerged microRNA genes in drosophila. Mol Biol Evol. 25:929–938.

Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI. 2008. The birth and death of microRNA genes in Drosophila. Nat Genet. 40:351–355.

Saunders MA, Liang H, Li WH. 2007. Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci USA. 104:3300–3305.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. DBSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29:308–311.

Zamore PD, Haley B. 2005. Ribo-gnome: the big world of small RNAs. Science. 309:1519–1524.

Zhang R, Peng Y, Wang W, Su B. 2007. Rapid evolution of an x-linked microRNA cluster in primates. Genome Res. 17: 612–617.

Takashi Gojobori, Associate Editor

Accepted March 10, 2009