Molecular Evolution of a Primate-Specific microRNA Family

Rui Zhang,†‡ Yin-Qiu Wang,*† and Bing Su*†*

*State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China; †Kunming Primate Research Center, Chinese Academy of Sciences, Kunming, China; and ‡Graduate School of Chinese Academy Sciences, Beijing, China

Lineage-specific microRNA (miRNA) families may contribute to developmental novelties during evolution. However, little is known about the origin and evolution of new miRNA families. We report evidence of an Alu-mediated rapid expansion of miRNA genes in a previously identified primate-specific miRNA family, drawn from sequencing and comparative analysis of 9 diverse primate species. Evolutionary analysis reveals similar divergence among miRNA copies whether they are within or between species, lineage-specific gain and loss of miRNAs, and gene pseudolization in multiple species. These observations support a birth-and-death process of miRNA genes in this family, implicating functional diversification during primate evolution. In addition, both secondary structure conservation and reduced single nucleotide polymorphisms density attest to functional constraint of this family in primates. Finally, we observed preferential expression of miRNAs in human placenta and fetal brain, suggesting a functional importance of this family for primate development.

Introduction

The microRNAs (miRNAs) are a class of small single-stranded noncoding RNAs (20–24 nt) in both unicellular and multicellular organisms (Bartel and Chen 2004; Molnar et al. 2007). Previous studies showed that miRNAs play pivotal roles in regulating diverse developmental processes by targeting mRNAs for translational repression, cleavage, or destabilization (Bartel and Chen 2004; Miranda et al. 2006; Plasterk 2006). The miRNA genes are transcribed as primary miRNA transcripts and then processed into ~70-nt hairpin precursor miRNAs (pre-miRNAs). Finally, the pre-miRNAs are processed to produce mature miRNAs (Bartel and Chen 2004). Both animal and plant miRNAs can be grouped into distinct families containing 1 or more precursors whose structures range from forming genomic clusters to scattering across the genome (Li and Mao 2007). The within-family precursors are generally similar to each other and produce similar, if not identical, mature miRNAs. When comparing distant lineages, expansion or contraction of miRNA families is common (Houbaviy et al. 2003; Bartel and Chen 2004; Tanzer and Stadler 2004; Hertel et al. 2006), which was even observed in some closely related lineages (Zhang et al. 2007). Expansions of miRNA families usually occur through tandem duplications or segmental duplications (Tanzer and Stadler 2004; Maher et al. 2006; Zhang et al. 2007). In Drosophila, new miRNAs can also originate from non-miRNA sequences (Lu et al. 2008). In plants, novel miRNA families can sometimes be created through inverted duplications of protein-coding gene sequences (Allen et al. 2004; Fahlgren et al. 2007). However, little is known about the origin of new miRNA families in mammals. Recently, Bentwich et al. (2005) identified a primate-specific miRNA family. It contains 43 pre-miRNAs, forming a cluster spanning about 100 kb on human chromosome 19 and is preferentially expressed in placenta.

To dissect the origin and evolutionary dynamics of this miRNA family in primates, we screened bacterial artificial chromosome (BAC) libraries and sequenced this miRNA family in 3 nonhuman primates (siamang, *Symphalangus syndactylus*; Yunnan snub-nosed monkey [Xu et al. 2004], *Rhinopithecus bieti*; and black-handed spider monkey [Qian et al. 2004], *Ateles geoffroyi*). We also sequenced the amplicons of slow loris (*Nycticebus coucang*). These species cover the major lineages of primates, including apes, Old World monkey, New World monkey, and prosimian. The last common ancestor of these primate lineages can be traced back to about 63 million years ago (MYA) (Goodman 1999). Combined with the available data in human, chimpanzee, orangutan, rhesus monkey, and marmoset, we examined the evolutionary pattern of the Chr19-linked miRNA family and we observed an Alu-mediated rapid expansion of miRNA genes, along with birth and death of miRNA copies in primates.

Materials and Methods

BAC Library Screening and Sequence Assembly

The pooled polymerase chain reaction (PCR)–based method was used in screening 3 primate BAC libraries and identifying positive BAC clones. In particular, a 2-step strategy was adopted to design the specific primer pair for BAC screening in corresponding species. First, several primers were designed using sequences either from human pre-miRNA with flanking region or from intergenic region. Then we amplified the 3 primate genomic DNAs by all primers. If one primer pair worked in a species, we then sequenced the PCR product and designed the species-specific primers for BAC library screening. Primer sequences are shown in supplementary table S1 (Supplementary Material online). Three positive BAC clones (1 for each species: 130 kb, 140 kb, and 200 kb) were selected by end sequencing, then full length sequenced using shotgun sequencing method with 8× coverage at Beijing Genomics Institute, Chinese Academy of Sciences (CAS). The sequences were aligned and assembled by phred/phrap/consed package (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). The sequence gaps were finished by PCR-based sequencing on an ABI-3130 sequencer. Only one gap (about 200 bp) was not closed in spider monkey BAC clone due to the failure of PCR. The sequence quality of the putative miRNAs was checked by looking at the raw chromatograms.

Prosimian DNA Sequencing

Because the sequences of miR-512 or miR-498 are distantly related to the other members of this family, 2

Key words: microRNA, birth-and-death evolution, Alu, primates.

E-mail: sub@mail.kiz.ac.cn

Mol. Biol. Evol. 25(7):1493–1502. 2008
doi:10.1093/molbev/msn094
Advance Access publication April 15, 2008

degenerate primers (dp-512 and dp-498) were designed based on the consensus of miR-512 or miR-498 among primate species. Two another primers (dp3 and dp4) were designed using consensus sequences of all other miRNAs. All the primer sequences are shown in supplementary table S1 (Supplementary Material online). The PCR amplicons were purified using MinElute PCR Purification Kit (Qiagen, Valencia, CA) and cloned with TA cloning vectors. A total of 150 positive clones for dp3 and dp4 were sequenced, respectively. The quality of each sequence was checked by looking at the raw chromatograms. The miRNA sequences were confirmed in prosimian when the same sequence was present in at least 2 clones. Besides 6 confirmed miRNA copies, there were 4 pre-miRNA copies with only one nucleotide difference between each. Because we could not rule out the possibility of polymorphisms of the same copy, a range of 2–4 was used in counting the copy numbers, which resulted in the total copy number of 8–10 for slow loris.

Homology Search and Evolutionary Analysis

We reanalyzed the human Chr19-linked family using BlastN to human genome with previously identified 43 pre-miRNAs as query sequences. The matched results with *E* value $<10^{-5}$ and sequence length >40 bp were regarded as good hits. The edges of the presumable pre-miRNA sequences were decided by comparing the hits with the known pre-miRNAs in human. We searched the homologs of the miRNAs in nonhuman primates with BlastN, and all the 46 human pre-miRNAs were used as query sequences. In addition, because of the relative deep divergence between human and New World monkey, after the initial Blast, the identified New World monkey hits were used as query sequences in a second Blast to the spider monkey BAC sequence or marmoset genome sequences. We defined the criteria for the secondary structures of miRNA genes as 1) the structure does not have multiple loops and 2) the free energy (dG) is less than or equal to −15 kcal/mol. Based on this approach, we got hundreds of miRNAs, and the dG is −42.8 kcal/mol on average ranging from −18.5 to −60.9 kcal/mol. The miRNAs were named according to their genomic order in different species. All the pre-miRNA sequences are provided in FASTA format (supplementary table S2, Supplementary Material online). To detect potential highly divergent copies of this family, we repeated the homology search procedure using high sensitivity BlastN parameters (Blastall -p BlastN -G 1 -E 1 -W 9 -q -1 -r 1). No additional copies were found. We used the UCSC genome database (http://hgdownload.cse.ucsc.edu/downloads.html) in our sequence analysis, including the human March 2006 (hg18) assembly, the chimp March 2006 (panTro2) assembly, and the rhesus January 2006 (rheMac2) assembly. We downloaded 6X WGS preliminary assembly of Orangutan and Marmoset sequences from http://genome.wustl.edu/. Precursors and mature miRNAs of zebrafish were downloaded from miRBase (Griffiths-Jones et al. 2006). Gene sequences were aligned by ClustalW (Thompson et al. 1994) with manual adjustment. Phylogenetic trees were reconstructed using the Neighbor-Joining method in MEGA3 (Kumar et al. 2004) and evaluated by 1,000 bootstrap replications. The Kimura 2-parameter model was used to calculate the substitution rates (*K*).

Alu Distribution and Duplication Unit Identification Analysis

Repeatmasker (http://repeatmasker.org) was used to find repeat elements. The masked duplication unit definition is explained in supplementary figure S1 (Supplementary Material online).

SNP Distribution Analysis in Human

We downloaded data from the Single Nucleotide Polymorphism database (dbSNP; build 126) using ENSEMBL BioMart by querying given genome coordinates. For comparison, we identified SNPs in the pre-miRNAs and in regions flanking each miRNA gene spanning the windows with the same size of the given pre-miRNA (Saunders et al. 2007).

Fluorescence In Situ Hybridization

Chromosome metaphase and interphase nuclei of human and slow loris were kindly provided by Kunming Cell Bank, CAS. Minipreparation of plasmid DNA from the sequenced Yunnan snub-nosed monkey positive BAC clone was performed according to modified alkaline lysis method. DNA was labeled by Nick translation with Biotin-dCTP (Invitrogen, Carlsbad, CA). The fluorescence in situ hybridization (FISH) method was described previously (Xu et al. 2004). To eliminate the effect of cross-hybridization of common repeat sequences, probes were blocked by using repetitive DNA (C_{ot}) before hybridization. At least 10 independent metaphase or interphase nuclei were examined in determination of chromosomal band location.

Simulation

The pre-miRNAs from the same species were aligned by ClustalW, the consensus sequences were inferred according to majority rule. For example, when analyzing human pre-miRNAs, first, we inferred the consensus sequences of human pre-miRNAs by aligning all the human copies, then we calculated the substitution rate *K* between each of the 45 real miRNAs (not including pseudogene hsa-38) and the consensus sequence (not considering indels) by the Kimura's 2-parameter model and obtain 45 *K* values. Second, we randomly mutated the consensus sequence to obtain 45 mutated sequences that have substitution rates equal to the real miRNAs. The random mutation process was replicated for 1,000 times. Finally, the Initial dGs of the secondary structures of the 45 real miRNAs (length equal to consensus) were calculated by "mfold" (Mathews et al. 1999) and then the average Initial dG was calculated. For miRNA, which has several predicted secondary structures, the minimal Initial dG was chosen. The same process was performed for the randomly mutated DNA sequences to get the average dG. By 1-sample Kolmogorov–Smirnov test, the normal distribution of dG of the simulated data set was confirmed. The same analysis was performed in all other species.

RNA In Situ Hybridization and Reverse Transcriptase–Polymerase Chain Reaction

All human subjects were obtained from the local hospitals and approved by their internal review boards with informed consents. For RNA in situ hybridization experiment, we selected 3 miRNAs as the representatives, of which miR-516-5p is from the 5′ arm and miR-520e is from the 3′ arm. The sequence of miR-498 is distantly related to the other members of this family and, therefore, was also selected. The placenta was embedded in Tissue-Tek OCT compound and cryosectioned. Ten-micron cryosections were pretreated and hybridized with LNA digoxygenin-labeled probes (Exiqon, Copenhagen, Denmark) following the protocol of frozen section in situ hybridization (Exiqon) with some modifications. When performing experiment, for each probe, we used the scramble-miR as negative control. In addition, the testis sample was also used as another negative control (data not shown). For reverse transcriptase–polymerase chain reaction (RT-PCR) experiment, the total RNA was isolated by TRIzol (Invitrogen). After DNase treatment, we used gene-specific primers or oligo-dT for reverse transcription. For all samples, the control sets were performed in reverse transcription step without reverse transcriptase. The PCR amplicons were validated using gel electrophoresis, purified using MinElute PCR Purification Kit (QIAGEN), subcloned with TA cloning vectors, and sequenced. The list of primers is provided in supplementary table S3 (Supplementary Material online). For human placenta sample from a 5-week embryo, we performed the standard hematoxlyin and eosin (HE) staining to confirm its origin. For the normal term placenta, the total RNA was isolated from villus parenchyma.

Computational Analysis of Gene Family Evolution

Computational analysis of gene family evolution (CAFE) is based on the likelihood model of gene gain and loss across species tree, and it can identify gene families that have statistically accelerated rates of gain and loss (Hahn et al. 2005; De Bie et al. 2006). The null hypothesis of CAFE is that the gene family size changes are by chance.

Seed Number Estimation

Current miRNA prediction methods mostly focus on pre-miRNAs because it is difficult to predict the exact mature miRNA sequences, especially for nonconserved miRNAs. In the human Chr19-linked family, there are 48 distinct mature miRNAs that were confirmed (Bentwich et al. 2005). Sixty percent of the pre-miRNAs generate mature miRNAs in both arms and the start sites of the mature miRNAs are mostly located at the same position in the aligned pre-miRNAs. For a conserved seed number estimation, assuming that the start site of the mature miRNAs is only the most frequent start site in both 5′ arm and 3′ arm, we aligned the pre-miRNAs in each species and predicted the mature miRNAs. Using this method, we estimated that there are 25 seeds in human, slightly less than the real number 30, a suggestion of the feasibility of this approach.

Target Gene Prediction

Target prediction was performed using 2 algorithms-*rna22* (Miranda et al. 2006) and Probability of Interaction by Target Accesibility (PITA) (Kertesz et al. 2007). In the first protocol, target genes of this family are predicted by *rna22* with default parameters. The full list of target genes can be downloaded from http://cbcsrv.watson.ibm.com/rna22.html. The target genes were then classified into inferred functional categories based on the Panther classification system (Thomas et al. 2003). To ask whether the target genes are overrepresented in some functional categories, we compared the target gene list with a reference list (National Center for Biotechnology Information: all human genes) using Panther gene analysis tool (Thomas et al. 2006). The *P* values are corrected for multiple comparisons by Bonferroni correction. In the second protocol, target genes predicted by PITA were downloaded from http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html. Target gene overrepresentation analysis was performed as above.

Accession Numbers

The sequences reported in this paper have been deposited in the GenBank database (accession numbers EU086470–EU086482).

Results

Sequence Structure of the Chr19-Linked miRNA Family in Primates

We reanalyzed the human Chr19-linked miRNA family region and identified 3 novel copies, which were not reported previously (supplementary fig. S2, Supplementary Material online). The expression of 2 novel copies was confirmed by cloning and sequencing. The third one is a pseudogene due to an Alu insertion. Through homology search, we identified varied numbers of pre-miRNAs (8–85 copies) and varied lengths of genomic regions (87.6–156.9 kb) covering the miRNA family in the 9 primate species (table 1), suggesting rapid evolution of this miRNA family. In addition, the intergenic regions between each miRNA also vary in length across species except for human–chimpanzee comparison (fig. 1). It was shown that mouse, rat, and dog do not have homologs of the miRNAs in this family; therefore, the emergence and rapid expansion of this miRNA family are restricted to primates (Bentwich et al. 2005).

Origin and Alu-Mediated Expansion of the Chr19-Linked miRNA Family

To understand the mechanism of origin and expansion of the miRNA family, we studied the sequence features. Prior study demonstrates that the miRNAs of this cluster are embedded in long (400–700 bp) sequences that are repeated along this cluster (Bentwich et al. 2005). In addition, the miRNAs are separated by Alus (the short interspersed element specific to primates) and the miRNAs and Alus account for over two-third of the ~100-kb sequences in human (Borchert et al. 2006). To test whether similar sequence structures are shown in nonhuman primates, we calculated the proportion of the repeat elements. Almost

Table 1
Copy Number and Genomic Length Variations of the Chr19-Linked miRNA Family in Primates

Species	Human	Chimp	Orangutan	Siamang	Rhesus	Ygm	Marmoset	Spider	Slow Loris
Number	46	46	63	42	44	44	81	85	8–10
Length (kb)	95.8	97.2	NA	87.6	100.3	93.2	156.9	146.8	NA

NOTE.—Ygm refers to Yunnan snub-nosed monkey. Rhesus refers to rhesus monkey. Spider refers to spider monkey. The copy numbers of orangutan and marmoset are rough estimation because the current assembly is not complete. The copy number of Yunnan snub-nosed monkey is not certain because the BAC clone sequenced did not cover the complete region of the miRNA family. The copy number of slow loris was estimated by sequencing PCR products amplified using degenerate primers. NA, not available.

all the repeat elements located within the family region are Alus, which account for about 50% in each species (supplementary table S4 and see supplementary fig. S3 [Supplementary Material online] for the detailed Alu and miRNA distributions in this region). In contrast, outside the family region, much of the repetitive sequences is comprised of non-Alu repeat elements, such as long terminal repeat and long interspersed elements (data not shown).

The Alu-mediated rearrangement events have long been recognized as a common source of local deletions and duplications associated with genome structure changes and genetic diseases (Smith et al. 1996; Deininger and Batzer 1999). Considering the high density of Alus in this region, we test whether the Alus are involved in the amplification of this cluster. If the miRNA family were expanded by Alu-mediated recombination event, the Alus would be physically located at the boundaries of the miRNA duplication units (Bailey et al. 2003). We first masked the cluster region sequences by RepeatMasker to identify the miRNA duplication units, then reinserted the repeat elements to reveal their distribution. We found that the masked duplication units in human range from 195 to 666 bp and nearly all units are adjacent to 1 or more Alus in the 5′ and/or 3′ ends, consistent with the predicted pattern of Alu-mediated recombination. Similar patterns were observed in other primates. Furthermore, the Alu-mediated recombination event would be expected to create a mosaic Alu element at the recombination junction (Bailey et al. 2003), and consequently, the Alus that cross the recombination junction should have increased sequence divergence when compared with the flanking duplicated sequence or internal Alu elements (Alus within the duplication) (Bailey et al. 2003). In all 7 traceable duplication events in human (see details about the definition of traceable duplication event in supplementary fig. S4, Supplementary Material online), 6 of them have Alus cross the junction. The divergence of the junction Alus is significantly higher than that of the flanking nonrepeat element sequences ($P = 0.025$, 2-tailed paired *t*-test). Three of the 6 duplicates contain internal Alus, and their divergence is significantly lower than the junction Alus ($P = 0.013$, 2-tailed paired *t*-test). In contrast, there is no significant divergence difference between the flanking nonrepeat element sequences and the internal Alus ($P = 0.22$, 2-tailed paired *t*-test). Several examples of such divergent Alu alignments are presented in supplementary figure S5 (Supplementary Material online), showing the junction Alu elements with increased sequence divergence, an indication of mosaic sequence structure. Hence, the sequence analysis at the junction region also supports that Alus are involved in the amplification of this miRNA family.

Alus can be categorized into 3 major subfamilies, which were active at different times during primate evolution: AluJ (65–40 MYA), AluS (45–25 MYA), and AluY (30 MYA to present) (Batzer and Deininger 2002; Price et al. 2004). Supplementary figure S6 (Supplementary Material online) demonstrates the Alu subfamily proportions of the miRNA family regions in the 6 primate species tested. Orangutan and marmoset were not analyzed due to the incomplete assembly of their draft genome sequences. The most abundant subfamily is AluS (57% of all Alus), which had a significant burst of activity after the split of New World monkeys and prosimians (Bailey et al. 2003), followed by AluJ (34%) and AluY (9%). Most of

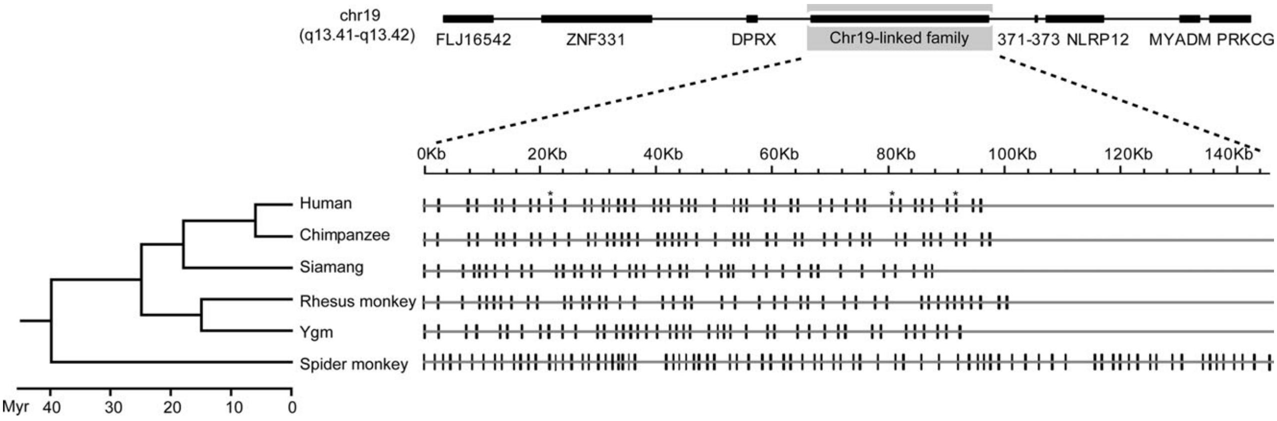

FIG. 1.—The miRNA distribution in 6 primate species. The 3 newly identified human miRNAs are labeled with asterisks. A phylogenetic tree of 6 primates with the time depth in million years (Myr) (Goodman 1999) is indicated. Ygm refers to Yunnan snub-nosed monkey.

Table 2
Nucleotide Differences per Site (*K* × 100) of Pre-miRNAs between and within Species

Species	Human	Chimp	Orangutan	Siamang	Rhesus	Ygm	Marmoset	Spider	Slow loris
Human	25								
Chimp	25	25							
Orangutan	25	24	24						
Siamang	28	27	27	30					
Rhesus	28	27	27	29	30				
Ygm	25	25	24	28	27	25			
Marmoset	32	31	31	34	33	31	34		
Spider	32	31	31	34	33	31	34	34	
Slow loris	30	28	28	31	30	28	30	31	20

the Alus located at the boundaries of the masked duplication units are AluS and AluJ. This observation suggests that the expansion of the miRNA family might start in the common ancestor of primates, predating the split of New World monkeys and prosimians, and that AluS, together with AluJ, leads to the further expansion of the miRNA family in Anthropoidea (New World monkey, Old World monkey, ape, and human).

To test whether this miRNA family exists in prosimian species, we performed BAC-FISH in slow loris. Using the Yunnan snub-nosed monkey BAC clone covering a 93-kb region of the miRNA family, we did not observe any signals in slow loris though signals were detected in the human sample control (supplementary fig. S7, Supplementary Material online). This result suggests that the sequence structure may not be conserved in prosimian species. We then designed degenerate primers to amplify the potential miRNA copies in slow loris. A total of ~8 to 10 miRNAs in this family were identified by cloning and sequencing, which is much fewer than the copy numbers observed in other primate species (table 1). As a control, we also conducted the same experiment using human sample, and 40 miRNA copies (87%, 40/46) were identified, therefore confirming the effectiveness of this approach. Thus, this observation supports the aforementioned early origin and later AluS-/AluJ-mediated expansion of the miRNA family.

It should be noted that several miRNAs of this family share one seed (positions 2–8 of the mature miRNAs, critical for target recognition [Brennecke et al. 2005]) with the neighbor mir-371,2,3 family (supplementary fig. S8, Supplementary Material online). The mir-371,2,3 family is conserved from human to rodents (Suh et al. 2004). Thus, it is possible that the Chr19-linked family may originate from one member of the mir-371,2,3 family.

Birth-and-Death Evolution of the Chr19-Linked Family

It is believed that most of protein multigene families may experience either concerted evolution or birth-and-death evolution or a mix of these 2 processes (Nei and Rooney 2005). Sequences evolving in a birth-and-death process usually have similar or higher between-species divergence when compared with within-species divergence. In contrast, sequences evolving in a concerted fashion result in much smaller within-species divergence (Piontkivska et al. 2002). To elucidate the evolutionary process of the Chr19-linked miRNA family, we calculated the nucleotide divergence between and within species. We observed similar divergence among copies whether they are within or between species (table 2). This pattern suggests that the birth-and-death evolution might be the major process during the expansion of this miRNA family.

Birth-and-death evolution also predicts pseudogenes as well as between-species orthologs and lineage-specific losses or gains. To see whether this is the case for the miRNA family, we used mfold (Mathews et al. 1999) to predict the secondary structure for each precursor. The result indicates that most precursors maintain stable hairpin and loop structures and, therefore, are potentially functional (supplementary fig. S2, Supplementary Material online). But we did observe disruptions of secondary structures in several copies resulting in pseudogenes. Four copies (hsa-38, ptr-38, orang-37, and ygm-37) were found disrupted by insertions of Alu sequences in the 3′ arm of stem regions (fig. 2*A*). The age-21 has an 18-bp deletion in the 5′ arm of the stem regions (fig. 2*B*). The hairpin structures of age-43 and cja-54 are disrupted because of multiple mutations (fig. 2*C*). It should be noted that the pseudo-miRNA prediction is conservative considering that some of the miRNA copies might be pseudogenes though no obvious secondary structure disruption was observed.

We also conducted phylogenetic analysis to test whether there are between-species orthologs and lineage-specific losses or gains. Through phylogenetic tree reconstruction using species pairs, we found that all the species pairs have lineage-specific copies except for the human–chimpanzee pair (data not shown), a suggestion of gene gains and/or losses in different lineages. Due to the uncertainty of orthology when analyzing all the copies (>400), we reconstructed the phylogenetic tree including copies from 4 species (human, siamang, Yunnan snub-nosed monkey, and rhesus monkey) and identified lineage-specific gene births and gene losses in multiple branches (supplementary fig. S9, Supplementary Material online).

Functional Constraint on the Chr19-Linked miRNA Family

To understand whether the sequence substitution pattern of the pre-miRNAs is neutral or under secondary structure constraint, we performed computer simulation. The consensus sequence of the human pre-miRNAs has a nearly perfect hairpin structure, suggesting strong secondary structure constraint (fig. 3*A*). Next, we performed simulation based on the consensus sequence and we found that the average dG of the pre-miRNAs is significantly lower than the

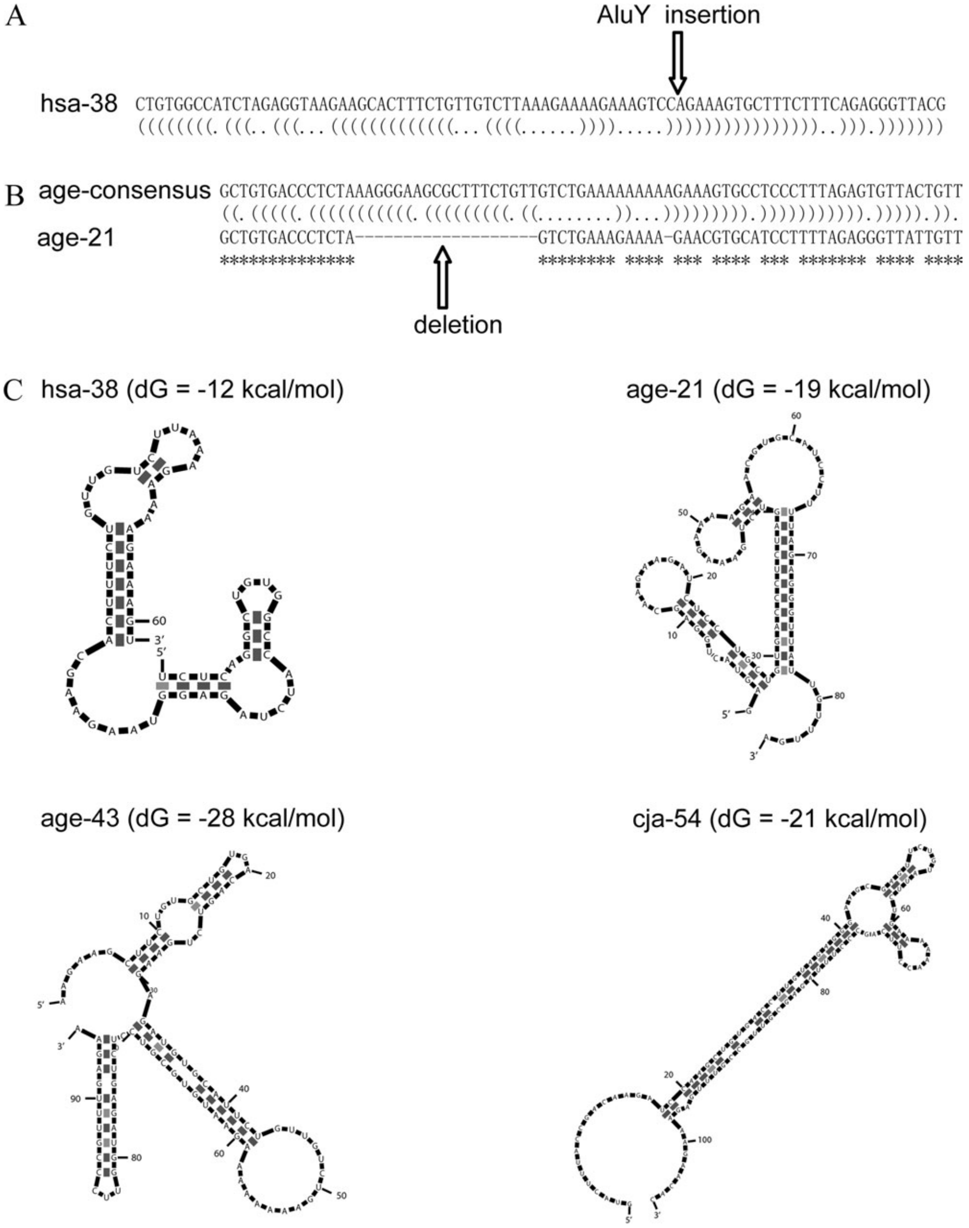
A
AluY insertion
hsa-38 CTGTGGCCATCTAGAGGTAAGAAGCACTTTCTGTTGTCTTAAAGAAAAGAAAGTCCAGAAAGTGCTTTCTTTCAGAGGGTTACG
(((((((.(((..(((...((((((((((((...((((......)))).....)))))))))))))))..))).)))))))
B age-consensus GCTGTGACCCTCTAAAGGGAAGCGCTTTCTGTTGTCTGAAAAAAAAAGAAAGTGCCTCCCTTTAGAGTGTTACTGTT
((.(((((.((((((((((.((((((((.((........))...)))))))).)))))))))).))))).)).
age-21 GCTGTGACCCTCTA------------------GTCTGAAAGAAAA-GAACGTGCATCCTTTTAGAGGGTTATTGTT
************** ******** **** *** **** *** ******* **** ****
deletion
C hsa-38 (dG = -12 kcal/mol)
age-21 (dG = -19 kcal/mol)
age-43 (dG = -28 kcal/mol)
cja-54 (dG = -21 kcal/mol)

FIG. 2.—The identified pseudogenes in the Chr19-linked miRNA family. (*A*) AluY inserted into the 3′ arm of pre-miRNAs (hsa-38 as an example). (*B*) The alignment of age-21 and spider monkey miRNA family consensus, the deletion of 18 bp is indicated. (*C*) The secondary structures of the predicted pseudogenes in hsa-38, age-21, age-43, and cja-54.

randomly mutated sequences ($P = 1.8 \times 10^{-15}$, probability of the observed dG compared with the normal distribution of the simulated data set) (fig. 3*B*), again an indication of secondary structure constraint of the miRNAs in the family.

Using dbSNP data (dbSNP build 126), we identified SNPs in the pre-miRNAs and the flanking regions in human populations. A total of 8 SNPs (including indel polymorphisms) were observed in the 45 pre-miRNAs (not including pseudogene hsa-38) (fig. 3*C*). The SNP density (2.1 SNPs per kilobase) in the pre-miRNA region is significantly lower than that of the flanking region (6.2 SNPs per kilobase) ($P < 0.02$, 2-tailed Fisher's exact test) (fig. 3*C*). As the flanking regions are likely under weak or no functional constraint, the observed lower SNP density in the pre-miRNAs suggests selective constraint on the evolution of this family and its functional importance. In addition, copy number variants in human populations were not observed by querying the database of genomic variants (http://projects.tcag.ca/variation/) (Iafrate et al. 2004) though this is not conclusive given the limited power of current high-throughput methods to identify small structure variations in the human genome.

Expression Pattern and Target Gene Prediction

With the use of mature miRNA cloning and microarray analysis, previous studies examined multiple tissues (testis, thymus, brain, prostate, fetal brain, and normal term

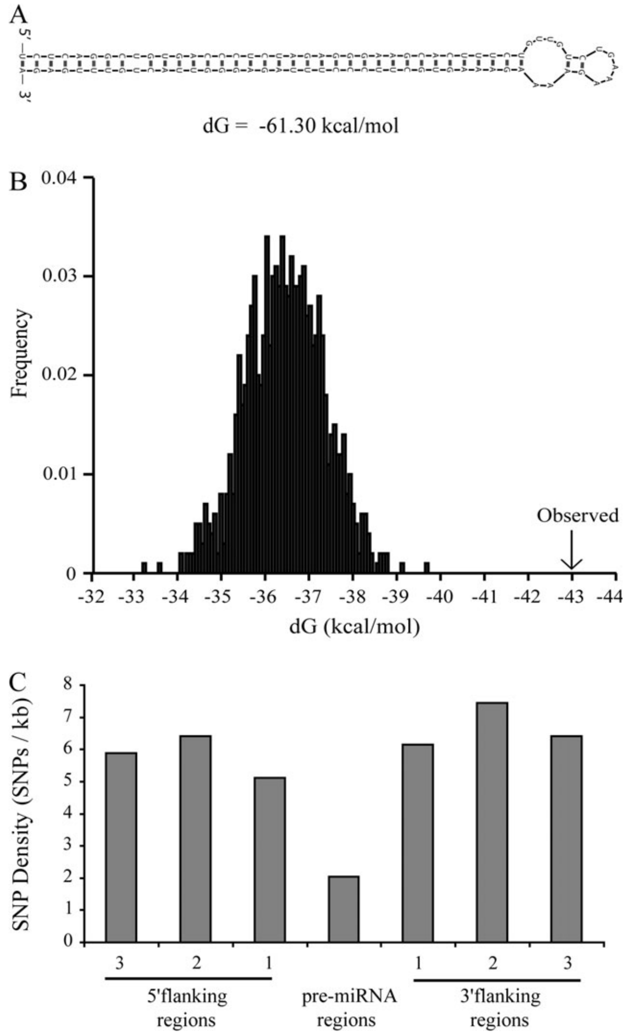

FIG. 3.—The results of simulation and SNP density analysis. (*A*) The nearly perfect hairpin structure of the human consensus. The dG value is indicated. (*B*) The observed average dG of the pre-miRNAs in human is significantly lower than the randomly mutated sequences generated by simulation ($P = 1.8 \times 10^{-15}$, probability of the observed dG compared with the normal distribution of the simulated data set). The bars show the frequency distribution of the dG values in a 1,000 simulated data set. The arrow indicates the average dG of real pre-miRNAs. (*C*) The SNP density comparison between the pre-miRNAs and their flanking regions in human. The flanking regions 1–3 represent successive, nonoverlapping windows with equal lengths to the given pre-miRNAs. The SNP density of the pre-miRNA region is significantly lower than the flanking regions (compared with the 5′ flanking region 1[5f1], $P = 0.019$; with 5f2, $P = 0.0013$; with 5f3, $P = 0.0051$; with 3f1, $P = 0.0033$; with 3f2, $P = 0.0008$; and with 3f3, $P = 0.0021$. 2-tailed Fisher's exact test).

placenta) and found that the miRNAs of this family are expressed in fetal brain and placenta (Bentwich et al. 2005; Berezikov et al. 2006). To further dissect the expression pattern, using RNA in situ hybridization and RT-PCR, we analyzed the human placenta and fetus samples (frontal brain, mid brain, posterior brain, lung, liver, heart, and brainstem). The RT-PCR analysis detected the expression of 21 out of the 46 pre-miRNAs (supplementary table S5 and fig. S10, Supplementary Material online) in one human placenta from a 5-week embryo (Carnegie stage 13 or 14 embryo). In fetus (20 weeks), only the expression of miR-498 was detected in the brain though previous report detected 14 mature miRNAs of this family (Berezikov et al. 2006). The difference is likely due to either different developmental stages or the low expression of those miRNAs. Furthermore, our in situ data indicate that all 3 miRNAs tested (miR-498, miR-516-5p, and miR-520e) are preferentially expressed in the cytoplasm of syncitiotrophoblasts in normal term placenta (fig. 4).

To identify potential target genes of this family in the human genome, we used *rna22* (Miranda et al. 2006) and PITA (Kertesz et al. 2007) to predict miRNA-binding sites. Both methods do not rely upon cross-species conservation, therefore, suitable for nonconserved miRNA families. Most of the predicted target genes are involved in signal transduction and nucleic acid binding. When the genes are grouped based on their functional classification, there are significant overrepresentations in several functional categories including signal transduction, development, and transcription regulation. The full list of the overrepresented groups is shown in supplementary table S6 (Supplementary Material online). This is consistent with the expression pattern of this family in fetal brain and placenta, suggesting a regulatory role during embryo development.

Discussion

Previous studies reported that animal miRNAs could be derived from repeat elements (Smalheiser and Torvik 2005; Piriyapongsa and Jordan 2007; Piriyapongsa et al. 2007). Our study expands this notion by showing that the repeat elements surrounding the miRNAs could also facilitate the emergence of novel miRNAs. It is well known that the expansion of Alu elements in the common ancestor of New World monkeys and Old World monkeys facilitated the expansion of segmental duplications through recombination (Bailey et al. 2003; Enard and Paabo 2004). Our study shows that the Alu expansion can also facilitate frequent local duplications of short units, that is, miRNAs. It is noteworthy that on average, the Chr19-linked miRNA family has 4.85 Alus per miRNA in the 6 primate species tested. This suggests that besides of the Alu elements adjacent to the masked miRNA duplication units, the Alu elements are widely dispersed among this region. Thus, we propose that during the early stage of primate evolution (65–25 MYA), the AluJ and AluS were inserted into this genomic region and the Alus were continuously accumulated in this specific region, and the enrichment of Alus in this region might promote rearrangement through misalignment and subsequent nonallelic homologous recombination, which eventually lead to the formation of the large miRNA family.

Novel miRNAs can emerge through gene duplications. In the Chr19-linked family, the estimated miRNA seed numbers range from 6 to 67 in 9 primate species. For example, in human, there are 48 different mature miRNAs (Bentwich et al. 2005) and 30 distinct seeds. Therefore, the abundant miRNA seeds in the Chr19-linked family indicate that novel miRNAs have been rapidly generated during primate evolution that may lead to functional

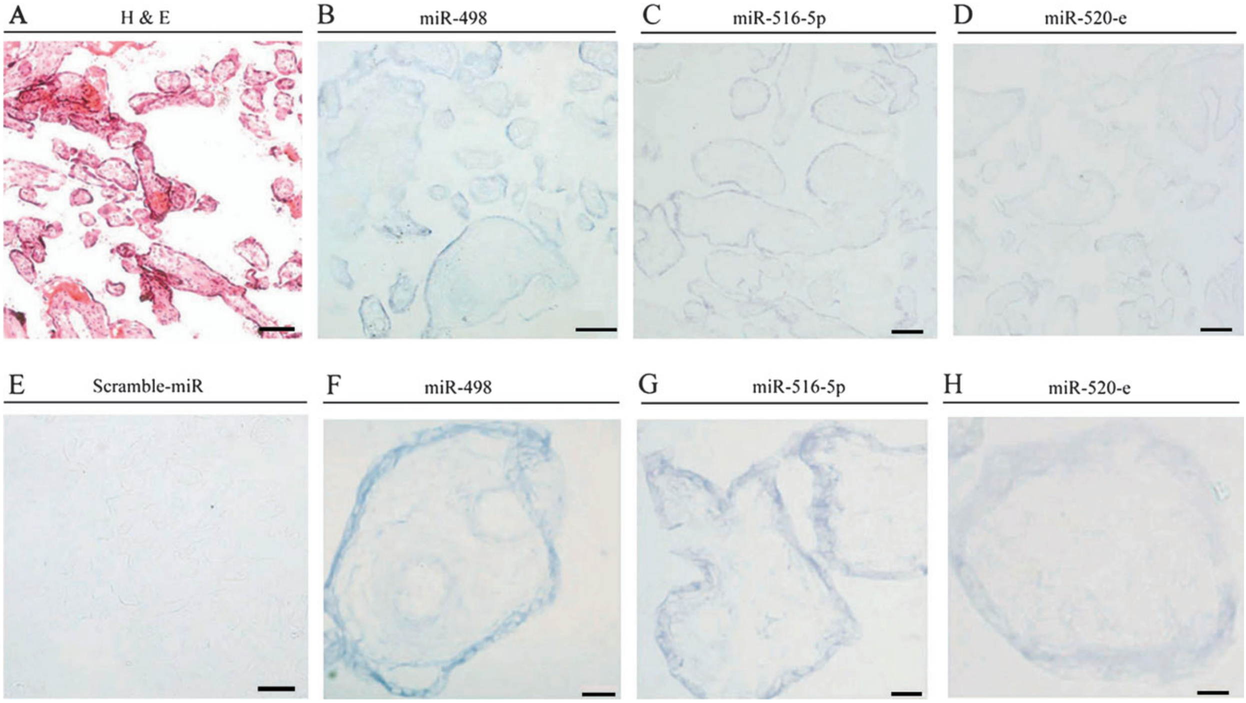

FIG. 4.—Expression of miR-498, miR-516-5p, and miR-520-e in human placenta. (*A*) The standard HE staining of placenta showing the tissue structure. (*B–H*) The preferential expression of miR-498, miR-516-5p, and miR-520-e in cytoplasm of syncitiotrophoblasts. (*F–H*) are the zoomed-in versions of (*B–D*). (*E*) Negative control. All photomicrographs were obtained and processed with identical software and hardware settings. Scale bars for (*A*), 150 microns; (*B–E*), 100 microns; and (*F–H*), 17 microns.

diversification. However, due to the complex homologous and orthologous relationship of this family among the primates, it is difficult to infer all the exact mature miRNA sequences directly through computational methods, thus preventing us from predicting the possible lineage-specific target genes in different primate species and probing into the consequences of miRNA diversification.

Because large differences in gene family size are generally attributed to the effects of natural selection (Hahn et al. 2005), we used a statistical approach—CAFE—to analyze the Chr-19 linked family (Hahn et al. 2005; De Bie et al. 2006). The result reveals that this family is subject to a significantly high rate of expansion or contraction among primates ($P = 0$), a strong indication of natural selection on the family size change, which is consistent with the proposed functional diversification. However, we cannot rule out the possibility that some of the newly emerged miRNAs on this cluster are slightly deleterious, but still persist in the genomes, and their fate was likely dictated by genetic drift.

Concerted evolution and birth-and-death evolution are 2 important mechanisms for multigene families. For protein-coding genes, multigene families whose members have the same function in a species are generally believed undergone concerted evolution that homogenizes the DNA sequences of the gene members. In contrast, protein gene families, particularly those producing variable gene products, are usually subject to birth-and-death evolution, such as major histocompatibility complex (MHC) genes (Nei et al. 1997), immunoglobulin genes (Nei et al. 1997), and MADS-box genes (Nam et al. 2004) (for review, see Nei and Rooney 2005). Our data on the Chr19-linked miRNA family support the birth-and-death evolution in primates; thus, functional diversification might be one of the major outcomes from the expansion of this family. The observed abundant miRNA seeds agree with this notion, implicating functional divergence among members of this family. However, some precursors generate same mature miRNAs and the seed numbers are still much less than the copy numbers, suggesting that dosage effect might be also important for the function of this miRNA family though we cannot rule out the possibility of recently duplicated nonfunctional copies. Hence, both the sequence diversity of mature miRNAs and the dosage effect of copy number variations within and between species could allow fine-tuning of target gene repression during development, which is consistent with the hypothesized evolutionary pattern of miRNAs (Bartel and Chen 2004).

Why do primate genomes maintain so many copies of miRNAs in this family? With the use of small RNA library sequencing strategy, Landgraf et al. (2007) set up an “atlas” of miRNA expression in 26 different organs and cell types. According to the atlas, the Chr19-linked family is expressed mostly in placenta but not in embryonic stem cell as observed in the mir-371,2,3 family (Suh et al. 2004). Thus, the function of the Chr19-linked family might be different from the mir-371,2,3 family. Also, the individual miRNA of the Chr-19 family is not the most highly expressed miRNA in placenta. But the family in total are the most abundantly expressed miRNAs in placenta, implicating that the functional requirement in placenta facilitates the presence of abundantly expressed miRNAs, just like the miR-430 gene family reported in zebrafish (Giraldez et al. 2005), though their evolutionary patterns might be different.

The miR-430 family in zebrafish has more than 90 copies spanning 120 kb in the genome (Giraldez et al. 2005). It is the most abundant miRNA family expressed during zebrafish early development, and it is shown to facilitate the deadenylation and clearance of maternal mRNAs during early embryogenesis and regulate brain morphogenesis in zebrafish (Giraldez et al. 2005, 2006). In the miR-430 family, there is only one miRNA seed though more than 90 copies, suggesting functional conservation of this miRNA gene family. Interestingly, this seed is also present in the mir-371,2,3 family and the Chr19-linked family. In fact, this seed belongs to an miRNA superfamily that includes the vertebrate miR-17–miR-20 family, the miR-371,2,3 family, the miR-302 family, and the zebrafish miR-430 family. It was suggested that the miR-430 family might share evolutionary origins with some of the miRNAs expressed specifically in embryonic stem cells, including miR-302 and miR-372, which have the same seed nucleotide sequences and derive from the same arm of the hairpin (Giraldez et al. 2005). But there are many other miRNA seeds in the Chr19-linked family, suggesting that the functional outcomes of expansion of these 2 miRNA families are probably different. Furthermore, the sequence substitution rates in the mature miRNA (*K*m) and the precursor (*K*p) regions are also different between the miR-430 and Chr19-linked families. The Chr19-linked family has a much higher *K*m/*K*p ratio (0.36 vs. 0.23 in human) compared with miR-430 (0.18 vs. 0.46 in zebrafish). Taken together, the difference between the Chr19-linked family and the miR-430 family suggests that the high copy number in zebrafish likely results in dosage-dependent regulation, whereas the multiple copies in primates might lead to functional diversification.

In mammalian species including primates, placenta has a short life span but is crucial for survival and development of embryo and fetus. The syncytiotrophoblast in placenta synthesizes the steroid hormones, that is, progesterone and estrogen, and it is the primary source of the glycoprotein hormone, which plays a pivotal role during embryo development (Page 1993). Hence, the preferential expression of the Chr19-linked family in syncytiotrophoblast suggests its critical involvement in embryo development of primates, which is also supported by the predicted target genes overrepresenting the functional categories of signal transduction, development, and transcription regulation.

In conclusion, we provide evidence of rapid expansion of a primate-specific miRNA family, mediated by the rich Alu elements in this region. We also demonstrate the birth-and-death evolution of this family, suggesting functional diversification during primate evolution.

Supplementary Material

Supplementary figures S1–S10 and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Scott W. Roy for critical reading of the manuscript. We also thank the Washington University School of Medicine, Genome Sequencing Center for the use of unpublished marmoset and orangutan genome data. We thank Shuang Yang in Beijing Genomics Institute, CAS for his help in BAC sequencing. We thank Jun Che and Wenhui Nie for BAC-FISH experiment. We also thank Aolei Niu, Ruo-lin Yang, and Xiao-Jing Yu for their help in this study. We are grateful to Xiao-na Fan and Hui Zhang for their technical assistance in this study. This study was supported by grants from the National 973 project of China (*2007CB947701*, *2007CB815705*), the CAS (*KSCX1-YW-R-34*), the National Natural Science Foundation of China (*30525028*, *30630013*, and *30623007*), and the Natural Science Foundation of Yunnan Province of China.

Literature Cited

Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nat Genet. 36:1282–1290.

Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 73:823–834.

Bartel DP, Chen CZ. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet. 5:396–400.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. Nat Rev Genet. 3:370–379.

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

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

Borchert GM, Lanier W, Davidson BL. 2006. RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol. 13:1097–1101.

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

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 22:1269–1271.

Deininger PL, Batzer MA. 1999. Alu repeats and human disease. Mol Genet Metab. 67:183–193.

Enard W, Paabo S. 2004. Comparative primate genomics. Annu Rev Genomics Hum Genet. 5:351–378.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175–185.

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.

Giraldez AJ, Cinalli RM, Glasner ME, Enright AJ, Thomson JM, Baskerville S, Hammond SM, Bartel DP, Schier AF. 2005. MicroRNAs regulate brain morphogenesis in zebrafish. Science. 308:833–838.

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF. 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science. 312:75–79.

Goodman M. 1999. The genomic record of Humankind's evolutionary roots. Am J Hum Genet. 64:31–39.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195–202.

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.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 15:1153–1160.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF. 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics. 7:25.

Houbaviy HB, Murray MF, Sharp PA. 2003. Embryonic stem cell-specific microRNAs. Dev Cell. 5:351–358.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet. 36:949–951.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. 2007. The role of site accessibility in microRNA target recognition. Nat Genet. 39:1278–1284.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

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.

Li A, Mao L. 2007. Evolution of plant microRNA gene families. Cell Res. 17:212–218.

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.

Maher C, Stein L, Ware D. 2006. Evolution of Arabidopsis microRNA families through duplication events. Genome Res. 16:510–519.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 288: 911–940.

Miranda KC, Huynh T, Tay Y, Ang Y-S, Tam W-L, Thomson AM, Lim B, Rigoutsos I. 2006. A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell. 126:1203–1217.

Molnar A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC. 2007. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. Nature. 447: 1126–1129.

Nam J, Kim J, Lee S, An G, Ma H, Nei M. 2004. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci USA. 101:1910–1915.

Nei M, Gu X, Sitnikova T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci USA. 94:7799–7806.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121–152.

Page KR. 1993. The physiology of the human placenta. London: Taylor & Francis.

Piontkivska H, Rooney AP, Nei M. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. Mol Biol Evol. 19:689–697.

Piriyapongsa J, Jordan IK. 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE. 2:e203.

Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. Genetics. 176:1323–1337.

Plasterk RH. 2006. Micro RNAs in animal development. Cell. 124:877–881.

Price AL, Eskin E, Pevzner PA. 2004. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res. 14:2245–2252.

Qian Y, Jin L, Su B. 2004. Construction and characterization of bacterial artificial chromosome library of black-handed spider monkey (*Ateles geoffroyi*). Genome. 47:239–245.

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

Smalheiser NR, Torvik VI. 2005. Mammalian microRNAs derived from genomic repeats. Trends Genet. 21:322–326.

Smith TM, Lee MK, Szabo CI, Jerome N, McEuen M, Taylor M, Hood L, King MC. 1996. Complete genomic sequence and analysis of 117 kb of human DNA containing the gene BRCA1. Genome Res. 6:1029–1049.

Suh MR, Lee Y, Kim JY, et al. (12 co-authors). 2004. Human embryonic stem cells express a unique set of microRNAs. Dev Biol. 270:488–498.

Tanzer A, Stadler PF. 2004. Molecular evolution of a microRNA cluster. J Mol Biol. 339:327–335.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 13:2129–2141.

Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B. 2006. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res. 34:W645–W650.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

Xu HL, Qian YP, Nie WH, Chi JX, Yang FT, Su B. 2004. Construction, characterization and chromosomal mapping of bacterial artificial chromosome (BAC) library of Yunnan snub-nosed monkey (*Rhinopithecus bieti*). Chromosome Res. 12:251–262.

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

Jianzhi Zhang, Associate Editor

Accepted April 10, 2008