Primate MicroRNAs miR-220 and miR-492 Lie within Processed Pseudogenes

ERIC J. DEVOR

From Molecular Genetics and Bioinformatics, Integrated DNA Technologies, 1710 Commercial Park, Coralville, IA 52241.

Address correspondence to E. J. Devor at the address above, or e-mail: rdevor@idtdna.com.

MicroRNAs (miRNAs) are a new and abundant class of small, noncoding RNAs. To date, the evolutionary history of most of these loci appears to be marked by duplication and divergence. The ultimate origin of miRNAs remains an open question. A survey of the genomic context of more than 300 human miRNA loci revealed that two primate-specific miRNAs, miR-220 and miR-492, each lie within a processed pseudogene. In silico and in vitro examinations of these two loci suggest that this is a rare phenomenon requiring the juxtaposition of a specific combination of factors. Thus it appears that, while processed pseudogenes are good candidates for miRNA incubators, it is unlikely that more than a very small percentage of new miRNAs arise this way.

Materials and Methods

miRNAs hsa-miR-220 and hsa-miR-492 are two examples of a growing number of human miRNAs listed in RELEASE 7.0 of miRBase, the miRNA database (Ambros et al. 2003; Griffiths-Jones 2004), for which no ortholog can be found in mouse or rat genomes (http://microrna.sanger.ac.uk/sequences/index.shtml). Due in silico diligence will resolve some of these in either rodent or other eutherian genomes, but many are primate specific. Using the chromosome coordinates for hsa-miR-220 (X chromosome, 122421481–122421790, Xq25, minus strand) and for hsa-miR-492 (Chromosome 12, 93730642–93730757, 12q22, plus strand), these loci were found to be encoded within annotated processed pseudogenes. Locus hsa-miR-220 is transcribed on the opposite strand from and completely within the β-tubulin–processed pseudogene identified as LOC402422, GenBank accession no. NT_011786, and locus hsa-miR-492 is transcribed on the sense strand of a keratin-19–processed pseudogene identified as LOC160313, GenBank accession no. NG_002383 (Figure 1). This suggested that both miRNAs evolved after the pseudogenes were created. Genome annotation of LOC402422 identifies it as a TUBB4-processed pseudogene. Clustal alignments of numerous β-tubulin mRNAs with the pseudogene sequence indicate that TUBB5 is more likely to be the antecedent (data not shown). Thus, the complete mRNA sequence of human β-5-tubulin (TUBB5, GenBank accession no. AY890656) was used to estimate the age of the reverse transcription and retroposition event that created LOC402422. Similarly, the mRNA sequence of human keratin-19 (KRT19, GenBank accession no. NM_002276) was used to estimate the age of the reverse transcription and retroposition event that created LOC160313.
Hsa-miR-220 sequences from several nonhuman primates deposited in GenBank (see Berezikov et al. 2005) were then used to design polymerase chain reaction (PCR) primers to amplify and sequence this locus in additional nonhuman primate species. Nonhuman primate sequences were unavailable for hsa-miR-492, but the locus was found in the chimpanzee, orangutan, and rhesus macaque genome assemblies and these sequences were used to design PCR primers to amplify and sequence this locus in additional nonhuman primate species. All primers were designed with and assessed for melting temperature and secondary structures using PRIMERQUEST online software (available as part of the Integrated DNA Technologies (IDT) SCITOOLS software, www.idtdna.com/scitools/scitools.aspx). PCR amplifications were carried out against a genomic DNA panel composed of human (Homo sapiens), chimpanzee (Pan troglodytes), gorilla (Gorilla gorilla), orangutan (Pongo pygmaeus), siamang (Hylobates syndactylus), vervet monkey (Chlorocebus aethiops), olive baboon (Papio anubis), Assamese macaque (Macaca assamensis), rhesus macaque (Macaca mulatta), squirrel monkey (Saimiri boliviensis), white-fronted capuchin (Cebus albifrons), and brown lemur (Eulemur fulvus). Amplicons were sequenced in both directions on an Applied Biosystems Model 310 automated fluorescence DNA sequencer. Species for which miR-220 sequences were not previously deposited in GenBank by Berezikov et al. (2005) are orangutan (DQ088046), siamang (DQ088047), olive baboon (DQ088048), vervet monkey (DQ088049), and Assamese macaque (DQ289549). Comparative miR-492 sequences for chimpanzee, orangutan, and rhesus macaque were obtained via BLAST search of National Center for Biotechnology Information and ENSEMBL.

Results

LOC402422, referred to herein as TUBB5Ψ, lies in Xq25 about 3 kb 3′ from another processed pseudogene (NDUFA4Ψ, AL030996) and 36.7 kb 5′ from the gene encoding the transcription/export complex member THOC2 (AL030996, NM_020449) (Figure 1). REPEATMASKER (http://www.repeatmasker.org) shows that the region between NDUFA4Ψ and TUBB5Ψ is composed almost entirely (94%) of repetitive sequences including Alu and L1 elements and most (68.6%) of the region between TUBB5Ψ and THOC2 as well. In particular, the 10-kb region immediately 3′ is completely (98%) composed of L1 elements. Alignment of the LOC402422 sequence with the mRNA of human TUBB5 reveals a 7.9% sequence divergence (P = .079) composed of 168 nucleotide changes (136 transitions and 32 transversions) and 10 indels. These changes introduce a number of frameshifts into the coding region of the pseudogene sequence along with a total of 17 stop codons of which 11 are in-frame. A 12-base insertion site repeat (TTAATTAA-TAG-5′ and TTAATAAAATAG-3′) flanks the TUBB5Ψ sequence. LOC160313, referred to herein as KRT19Ψ, lies in a sparsely populated region of 12q22, 184 kb 5′ of KIAA1147 and 136 kb 3′ of DAP13 (Figure 1). REPEATMASKER shows that the region immediately surrounding KRT19Ψ is also repeat rich (70%). Alignment of KRT19Ψ with the KRT19 mRNA shows
that the pseudogene is missing 234 bp of the 5' end, including the 5' untranslated region (UTR) and start codon. Within the remaining aligned sequence, there is a 10.8% sequence divergence \((P = 0.108)\) composed of 122 nucleotide changes (95 transitions and 27 transversions) and nine indels. These changes introduce 22 stop codons into the sequence. The large 5' deletion appears to have happened at the time of retroposition as the remaining sequence is flanked by a 15-base insertion site repeat (AGAAAAGTTCCAGTC). Thus, these loci display all the hallmarks of classical processed pseudogenes (Devor and Moffat-Wilson 2003).

Using the age estimation expression \(T = K/2r\), where \(r\) is taken to be 1.5 \(\times\) 10\(^{-9}\) sequence changes per position per year (Li 1997) and \(K\) is the Jukes-Cantor correction \(-3/4 Ln(1 - 4/3p)\) (Jukes and Cantor 1969), an estimated age of 27.8 million years is obtained for TUBB and 38.9 million years for KRT19. Though this method of age estimation must be considered approximate because there are numerous instances of the volatile CpG dimer in each sequence (cf. Labuda and Striker 1989) and the large deletion in the miR-492 sequence, subsequent PCR amplifications are consistent as only human, ape, and OWM samples yielded ampli cons containing miR-220 or miR-492. Therefore, both loci were reverse transcribed and retroposed into the primate genome after the divergence of OWM and new world monkey, an event estimated to have taken place between 35 and 40 million years ago, but prior to divergence of OWM and apes, an event estimated to have taken place between 20 and 25 million years ago (Szalay and Delson 1979).

Precursor sequences (pre-miRNAs) of miR-220 in 11 primate species and of the orthologous reverse complement of human TUBB5 are shown in Figure 2. Also shown in Figure 2 are pre-miR-492 sequences for nine primate species and the orthologous region of human KRT19. MiRNAs are composed of a primary RNA transcript (pri-miRNA) up to several kilobases in length. Within this is the pre-miRNA transcript, usually 80–110 bases long, that forms a stable hairpin. This hairpin is excised from the pri-miRNA by a complex containing the enzyme DROSHA and its cofactor DGCR8 (aka. PASHA in Drosophila melanogaster and C. elegans). The hairpin structure is exported from the nucleus as a double-stranded RNA by exportin-5, whereon a mature miRNA sequence 21–23 bases long is processed by the same Dicer/RISC complex known to be responsible for RNA interference (cf. Bartel 2004; Berezikov and Plasterk 2005). It is the mature miRNA sequence that acts as a posttranscriptional regulatory element. One of the hallmarks of miRNAs in all species is the action of purifying selection, particularly in the mature miRNA sequence, such that even very ancient loci display little variation even among distantly related families (Floyd and Bowman 2004; Pasquinelli et al. 2000). The pre-miRNA sequence alignments presented in Figure 2 reveal a number of nucleotide changes throughout both miR-220 and miR-492. The usual pattern of interspecies nucleotide variation in miRNAs is marked by a high level of conservation in the mature miRNA and its complement, a lower level of conservation in both the stem and loop sequences, and a further decrease of conservation in the sequences flanking the pre-miRNA. This is the “camel-shaped” conservation profile described by Berezikov and Plasterk (2005). Berezikov et al. (2005) point out that nucleotide changes in pre-miRNAs can occur in unpaired sites or in paired sites in the hairpin. Among paired sites, the nucleotide substitution will either disrupt the pairing or not (e.g., their example G:U to A:U).
The nucleotide substitutions seen in Figure 2 represent all three types, including several in the mature miRNA, particularly in miR-220. In order to assess the effects of the observed sequence variations on hairpin structure and thermodynamic stability, pre-miRNA transcripts were evaluated using MFOLD (Zuker 2003, available online in IDT SCITools). Each transcript was evaluated as a linear RNA sequence at 37°C. Results of this analysis are shown in Table 1. Hairpin stability is measured by the thermodynamic parameter ΔG, the change in Gibbs free energy in kilocalories per mole. The expression ΔG = ΔH − TΔS, where ΔH is the total energy exchange between the system and its environment (enthalpy), ΔS is the energy spent by the system to organize itself (entropy), and T is the absolute temperature in Kelvin (°C + 273.15), will indicate the stability of a hairpin structure at a given temperature. The more negative the value of ΔG, the more stable the hairpin. In both miR-220 and miR-492, nucleotide differences relative to the human sequence are seen to have a negative impact on thermodynamic stability. That is, maxΔG increases becomes less negative. However, in only one case, that of miR-220 in gorilla, was the hairpin structure itself significantly altered. These results tend to support the view of Berezikov et al. (2005) that there is selective pressure on pre-miRNA secondary structure, but some amount of structural change is tolerated.

### Table 1. Maximum ΔG values for miRNA hairpin structures for various primate species

<table>
<thead>
<tr>
<th>Species</th>
<th>miR-220</th>
<th>miR-492</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em> (Hsa)</td>
<td>−40.9</td>
<td>−40.7</td>
</tr>
<tr>
<td><em>Pan troglodytes</em> (Pty)</td>
<td>−34.4</td>
<td>−39.4</td>
</tr>
<tr>
<td><em>Gorilla gorilla</em> (Ggo)</td>
<td>−32.9</td>
<td>−40.7</td>
</tr>
<tr>
<td><em>Pongo pygmaeus</em> (Ppy)</td>
<td>−39.4</td>
<td>−40.0</td>
</tr>
<tr>
<td><em>Hylabates syndactylus</em> (Hsy)</td>
<td>−33.8</td>
<td>−36.0</td>
</tr>
<tr>
<td><em>Chlorocebus aethiops</em> (Cae)</td>
<td>−35.3</td>
<td>−33.6</td>
</tr>
<tr>
<td><em>Papio anubis</em> (Pai)</td>
<td>−34.2</td>
<td>−33.3</td>
</tr>
<tr>
<td><em>Macaca mulatta</em> (Mur)</td>
<td>−33.6</td>
<td>−29.9</td>
</tr>
<tr>
<td>Ψ-source gene ortholog*</td>
<td>−41.9</td>
<td>−38.2</td>
</tr>
</tbody>
</table>

* For miR-220 this is the reverse transcript of the orthologous region of TUBB5, and for miR-492 this is the direct ortholog from KRT19.

**Discussion**

miRNAs miR-220 and miR-492 are unique to primates, specifically to OW, apes, and humans. Both were found to lie within processed pseudogenes estimated to have been created 27 and 39 million years ago, respectively. Pre-miR-220 and pre-miR-492 sequences were obtained for several primate species representing African and Asian OW, African and Asian apes, and humans. These pre-miRNA sequences display a number of sequence variants, including a total of seven variants within the mature miRNA itself. However, while these changes do impact hairpin stability, they do not affect hairpin structure.

As with the vast majority of miRNAs, the specific regulatory targets of miR-220 and miR-492 are unknown. However, there is evidence that these loci are being transcribed, at least in the human genome (Bentwich et al. 2005; Lim et al. 2003). On the other hand, their transcriptional status in other primates is yet to be determined.

The observation of miRNAs evolving from inside processed pseudogenes raises the question of whether such a mechanism might explain the origin of at least some other miRNAs that are not clearly due to duplications. It has already been demonstrated that one subset of miRNAs is derived from LINE-2 transposable elements and other genomic repeat features (Smallheizer and Torvik 2005). Several features of processed pseudogenes make them potential candidates as miRNA antecedents (Devor and Moffat-Wilson 2003). First, they are reasonably common occurrences in many genomes. Further, the genes from which they arise are most often those that are suitable candidates for miRNA regulation such as housekeeping genes and other genes expressed at fairly high levels. Second, while not essential for miRNA formation, processed pseudogenes are created from reverse transcribed miRNAs, which usually result in the presence of an intact sequence from 5’ UTR to 3’ UTR. Thus, any resulting hairpin structure would be guaranteed to contain sequence that would be present in a target transcript. Finally, they are almost always free of selection pressure. This would permit changes affecting the sequence to occur at will.

There are two different ways to approach an answer to the question of the potential role of processed pseudogenes as miRNA incubators. The most straightforward is to simply look. Using chromosome coordinates listed in RELEASE 7.0 of miRBase, genome context of more than 300 human miRNAs was evaluated. Among these loci about 40% were seen to be located in introns and the remainder in intergenic space. However, only the two loci reported here were found within an annotated processed pseudogene. This is not to say that more such loci will not be found as estimates of the ultimate number of miRNAs in the human genome as high as 1,000 have been forwarded.

The apparent rarity, at least for now, of *hsa-miR-220* and *hsa-miR-492* leads to the second approach to answering the question. How likely is it that a processed pseudogene will contain a hairpin structure suitable for forming a miRNA? The preliminary answer is that it is very likely. In silico RNA transcripts from 14 human processed pseudogenes, selected solely because they were about the same size as TUBB5Ψ and KRT19Ψ (2,302 and 1,153 bp, respectively), were submitted to MFOLD analyses with the result that every one presented one or more pre-miRNA–suitable hairpin structures (i.e., length between 70 and 110 continuous bases with an estimated ΔG of −30.0 kcal or greater). If, therefore, it is so apparently easy for processed pseudogene sequences to have potential miRNA hairpins, why are they not more common? The answer to this lies in the fact that miRNAs are transcribed, and the appropriate transcription machinery is not carried within pseudogenes themselves. Thus, *hsa-miR-220* and *hsa-miR-492* not only possessed an appropriate hairpin structure but also were fortuitously retrofitted to a position where a cis-acting RNA polymerase II
transcription site (cf. Cai et al. 2004; Lee et al. 2004) was available within a reasonable distance. While the precise location of these sites must await identification of the pri-miRNA transcript for both these loci, a PROMOTER 2.0 (Knudsen 1999) scan of some 10 kb of upstream human genomic sequence did indicate that several candidate transcription sites are present.

Finally, accepting for the moment that processed pseudogenes and juxtaposed L2 or other repeats will prove to be rare origins for new miRNAs, the question remains as to the ultimate origin of this important class of gene expression regulators. Allen et al. (2004) offered a tantalizing glimpse from Arabidopsis of miRNAs evolving from inverted duplications of what ultimately becomes the target site for regulation, but this, too, appears to be a rare occurrence. On the other hand, perhaps the identification of three very different albeit rare mechanisms for miRNA origins is, in fact, the answer. miRNAs may have evolved opportunistically and took advantage of cellular mechanisms that were already present, such as the Dicer/RISC complex, and there is no one ultimate source for these loci. This possibility is not out of the question, and it could explain why there are no consistent features among miRNAs apart from the fact that all of them have a pre-miRNA hairpin of some sort.

References


Received December 5, 2005
Accepted January 9, 2006

Corresponding Editor: William Modi