A uniform system for microRNA annotation

VICTOR AMBROS,[1] BONNIE BARTEL,[2] DAVID P. BARTEL,[3] CHRISTOPHER B. BURGE,[4] JAMES C. CARRINGTON,[5] XUEMEI CHEN,[6] GIDEON DREYFUSS,[7] SEAN R. EDDY,[8] SAM GRIFFITHS-JONES,[9] MHAIRI MARSHALL,[9] MARJORI MATZKE,[10] GARY RUVKUN,[11] and THOMAS TUSCHL[12,13]

[1]Dartmouth Medical School Department of Genetics, Hanover, New Hampshire 03755, USA
[2]Department of Biochemistry and Cell Biology, Rice University, Houston, Texas 77005, USA
[3]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Department of Biology, Cambridge, Massachusetts 02139, USA
[4]Massachusetts Institute of Technology, Department of Biology, Cambridge, Massachusetts 02142, USA
[5]Center for Gene Research and Biotechnology, and Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331, USA
[6]Waksman Institute, Rutgers University, Piscataway, New Jersey 08854, USA
[7]Howard Hughes Medical Institute and Department of Biochemistry & Biophysics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104-6148, USA
[8]Howard Hughes Medical Institute, Washington University Department of Genetics, Saint Louis, Missouri 63110, USA
[9]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[10]Institute of Molecular Biology, Austrian Academy of Sciences, A-5020 Salzburg, Austria
[11]Department of Genetics, Harvard Medical School, and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
[12]Max-Planck-Institute for Biophysical Chemistry, Department of Cellular Biochemistry, Göttingen, Germany

ABSTRACT

MicroRNAs (miRNAs) are small noncoding RNA gene products about 22 nt long that are processed by Dicer from precursors with a characteristic hairpin secondary structure. Guidelines are presented for the identification and annotation of new miRNAs from diverse organisms, particularly so that miRNAs can be reliably distinguished from other RNAs such as small interfering RNAs. We describe specific criteria for the experimental verification of miRNAs, and conventions for naming miRNAs and miRNA genes. Finally, an online clearinghouse for miRNA gene name assignments is provided by the Rfam database of RNA families.

Keywords: microRNA; noncoding RNA; Rfam; Dicer; genome annotation

MicroRNAs (miRNAs) are small noncoding RNA gene products about 22 nt long that are found in diverse organisms, including animals (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001; Mourelatos et al. 2002) and plants (Llave et al. 2002a; Park et al. 2002; Reinhart et al. 2002). The founding members of this class of noncoding RNAs are the *lin-4* and *let-7* gene products of *Caenorhabditis elegans* (Lee et al. 1993; Reinhart et al. 2000; for review, see Pasquinelli and Ruvkun 2002). Although *lin-4* and *let-7* were identified by standard positional cloning of genetic loci, most miRNA genes are discovered through cDNA cloning of sequences from size-fractionated RNA samples. One pitfall of miRNA gene hunting by cDNA cloning is that fragments of other noncoding RNAs (such as rRNAs, tRNAs, and snRNAs), as well as messenger RNAs, are also cloned from size-selected RNA samples. Candidate miRNA sequences can be readily tested using sequence-matching searches to annotated databases to flag most of these contaminating non-miRNA sequences. A more difficult task is differentiating miRNAs from other classes of small RNAs that might be present in the cell, particularly endogenous small interfering RNAs (siRNAs). Not only is the misclassification of the siRNAs undesirable, but because innumerable siRNAs are generated from each double-stranded RNA (dsRNA) precursor, it would be problematic to give each endogenous siRNA a different gene name. Another factor to consider when annotating miRNA genes is that many miRNA sequences are evolutionarily conserved (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001; Mourelatos et al. 2002; Park et al. 2002; Reinhart et al. 2002). Thus, giving the same name to obvious orthologs is useful, and giving the same name to unrelated sequences from different organisms causes confusion. For these reasons, we suggest the adoption of a consistent set of guide-

Reprint requests to: Victor Ambros, Dartmouth Medical School Department of Genetics, 7400 Remsen, Hanover, NH 03755, USA; e-mail: vambros@dartmouth.edu; or David P. Bartel, Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Department of Biology, 9 Cambridge Center, Cambridge, MA 02142, USA; e-mail: dbartel@wi.mit.edu; or Thomas Tuschl, [13][Present address] Rockefeller University, 1230 York Avenue, New York, NY 10021, USA; fax: (212) 327-7652.

Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2183803.

lines for the classification of small RNAs as bona fide miRNAs, and provide a mechanism for the assignment of miRNA gene names.

MicroRNAs and siRNAs cannot be distinguished on the basis of their functions. Although siRNAs trigger the cleavage of target mRNAs (Zamore 2002), and the canonical miRNA, *lin-4*, is a translational repressor that does not influence target mRNA abundance (Olsen and Ambros 1999), other miRNAs recognize their targets and direct RNA cleavage as if they were siRNAs (Hutvagner and Zamore 2002; Llave et al. 2002b; Rhoades et al. 2002; Zeng et al. 2002). Even if miRNAs could be distinguished from siRNAs based on function, there is often the need to name a gene before the function is elucidated. The biochemical compositions of miRNA and siRNA molecules are also indistinguishable, both having the characteristic features of Dicer products (20–25 nt length, 5′-phosphate, and 3′-hydroxyl). Thus, miRNAs are most readily distinguished from siRNAs based on unique aspects of their biogenesis. Biological siRNAs come from long exogenous or endogenous dsRNA molecules (very long hairpins or bimolecular duplexes), processed such that numerous siRNAs accumulate from both strands of the dsRNA. MicroRNAs come from endogenous transcripts that can form local hairpin structures, which ordinarily are processed such that a single miRNA molecule accumulates from one arm of a hairpin precursor molecule. Sometimes the primary transcript contains multiple hairpins, and different hairpins give rise to a different miRNAs. These are considered polycistronic miRNA transcripts, and each hairpin is given a unique gene name.

Given the desire to avoid designating siRNAs or fragments of other RNAs as miRNAs, miRNAs are identified using a combination of criteria for both their expression and biogenesis.

Expression criteria

A. Detection of a distinct ~22-nt RNA transcript by hybridization to a size-fractionated RNA sample (ordinarily by the Northern blotting method).
B. Identification of the ~22-nt sequence in a library of cDNAs made from size-fractionated RNA. Such sequences must precisely match the genomic sequence of the organism from which they were cloned (except as noted below).

Biogenesis criteria

C. Prediction of a potential fold-back precursor structure that contains the ~22-nt miRNA sequence within one arm of the hairpin. In this criterion, the hairpin must be the folding alternative with the lowest free energy, as predicted by mfold (Mathews et al. 1999) or another conventional RNA-folding program, and must include at least 16 bp involving the first 22 nt of the miRNA and the other arm of the hairpin. It should not contain large internal loops or bulges, particularly not large asymmetric bulges. In animals, these fold-back precursors are usually about 60–80 nt, whereas in plants, they are more variable, and may include up to a few hundred nucleotides.
D. Phylogenetic conservation of the ~22-nt miRNA sequence and its predicted fold-back precursor secondary structure. The conserved hairpin should meet the same minimal pairing requirements as in criterion C, but need not be the lowest free energy folding alternative.
E. Detection of increased precursor accumulation in organisms with reduced Dicer function.

None of the above criteria on its own is sufficient for a candidate gene to be annotated as a novel miRNA, because evidence of expression alone (A or B) would not exclude siRNAs, and a hairpin structure (C or D) is not a unique characteristic of miRNA biogenesis, nor is Dicer processing (E). Therefore, evidence of both expression and biogenesis characteristic of miRNAs are required. Ideally, a miRNA would be identified based on strong evidence that an asymmetric ~22-nt product accumulates in vivo, and that the product is processed from a phylogenetically conserved hairpin precursor by Dicer (A + D + E). However, in the absence of processing data, A + D is sufficient. Note that criterion A, which involves data about abundance of a ~22-nt RNA, is a stronger expression criterion than is B. Similarly, D is a stronger criterion for a hairpin precursor than is C, because D incorporates phylogenetic conservation to support the precursor structural prediction. Therefore, criterion C should be supported by strong expression data (A + C), and criterion B should be supported by strong biogenesis data (B + D). In a case where the ~22-nt miRNA is not detected or cloned, a candidate gene can still be annotated as an miRNA gene if a phylogenetically conserved hairpin precursor can be detected and is shown to accumulate in organisms depleted in Dicer function (D + E). However, this latter situation should be approached with special care, as other classes of hairpin-containing RNAs might be metabolized by Dicer.

Homologs of previously validated miRNAs need not meet as stringent criteria to be annotated as additional miRNA loci. Very close homologs in other species can be annotated as miRNA orthologs without experimental validation, provided they satisfy criterion D. In cases where a cDNA sequence does not match the available genomic sequences, if that precise cDNA was nevertheless cloned multiple times (criterion B) and is close in sequence to a known miRNA, it can be annotated as a variant form of the known miRNA. Candidates can also be identified based on sequence similarity to authenticated miRNAs from the same species. When such a candidate is so similar in sequence to the known miRNA that the probe designed to detect it would surely cross-hybridize with the known miRNA, criterion A need not be satisfied, and the candidate can be annotated as a variant form of the known miRNA, provided

that the candidate itself meets criterion C, and there is also very high confidence that one of these paralogs is a confirmed miRNA (i.e., its classification is supported by at least three of the listed criteria). When there are two or more mismatches between the miRNA paralogs such that probes could differentiate between the species on Northern blots, criterion A must be satisfied experimentally. Sometimes partially overlapping cDNA sequences are identified from the same locus; these are assumed to represent differentially processed products from the gene.

MicroRNAs come from a single-molecule fold-back (hairpin) structure; RNAs that are processed by Dicer from a hybrid between two antiparallel transcripts are siRNAs and not miRNAs. Thus, an observation that ~22-nt RNAs are produced from both a sense and an antisense transcript is evidence of siRNAs. Because essentially all miRNA precursors described so far produce a mature microRNA preferentially from either the 5′ or 3′ fold-back strand, cases where both strands accumulate as ~22-nt RNAs should warrant caution. Similarly, no known microRNA precursors have yet been found that produce multiple nonoverlapping mature miRNAs from the same arm of the foldback precursor. Therefore, such cases should also be carefully considered as possible siRNAs; at any rate, it is recommended that each hairpin structure be considered a single gene, and all mature miRNAs from that hairpin should be annotated as alternatively processed gene products.

MicroRNAs are named using the "miR" prefix and a unique identifying number (e.g., miR-1, miR-2, . . . miR-89, etc.). The genes that encode the miRNA are also named using the same three-letter prefix, with capitalization, hyphenation, and italics according to the conventions of the organism (for example, *mir-1* in *C. elegans* and *Drosophila*, *MIR156* in *Arabidopsis* and rice). The identifying numbers are assigned sequentially, with identical miRNAs having the same number, regardless of organism. Nearly identical orthologs can also be given the same number, at the discretion of the researcher. For example, miR-1 of *Drosophila* differs by a single nucleotide from miR-1 of *C. elegans* and humans. Identical or very similar miRNA sequences within a species can also be given the same number, with their genes distinguished by letter and/or numeral suffixes, according to the convention of the organism (e.g., the ~22-nt transcripts of *Drosophila mir-13a* and *mir-13b* are slightly different in sequence, whereas those of *mir-6-1* and *mir-6-2* are identical; Lagos-Quintana et al. 2001).

An online clearinghouse for miRNA gene name assignments (http://www.sanger.ac.uk/Software/Rfam/mirna/) is provided by the Rfam database of RNA families (Griffiths-Jones et al. 2003). The primary purpose of the clearinghouse is to assign unique gene names to distinct miRNAs while maintaining complete confidentiality for unpublished data. To avoid accidental overlap, Rfam will assign a name only after a paper describing the sequence has been accepted for publication. This resource also provides a searchable database of published miRNAs and aims to facilitate the evaluation of candidate sequences according to the above guidelines.

ACKNOWLEDGMENTS

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Received November 12, 2002; accepted December 4, 2002.

REFERENCES

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* **31:** 439–441.

Hutvagner, G. and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297:** 2056–2060.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.

Lee, R.C., Feinbaum, R.L., and Ambros, V. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75:** 843–854.

Lee, R.C. and Ambros, V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862–864.

Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002a. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14:** 1605–1619.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. 2002b. Cleavage of *Scarecrow*-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297:** 2053–2056.

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

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. 2002. miRNPs: A novel class of ribonucleoproteins containing numerous microRNAs. *Genes & Dev.* **16:** 720–728.

Olsen, P.H. and Ambros, V. 1999. The *lin-4* regulatory RNA controls developmental timing in *C. elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **2:** 671–680.

Park, W., Li, J., Song, R., Messing, J., and Chen, X. 2002. CARPEL FACTORY, a Dicer Homolog, and HEN1, a Novel Protein, Act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12:** 1484–1495.

Pasquinelli, A.E. and Ruvkun, G. 2002. Control of developmental timing by microRNAs and their targets. *Annu. Rev. Cell. Dev. Biol.* **18:** 495–513.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. 2000. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403:** 901–906.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. 2002. MicroRNAs in plants. *Genes & Dev.* **16:** 1616–1626.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. 2002. Prediction of plant microRNA targets. *Cell* **110:** 513–520.

Zamore, P.D. 2002. Ancient pathways programmed by small RNAs. *Science* **296:** 1265–1269.

Zeng, Y., Wagner, E.J., and Cullen, B.R. 2002. Both natural and designed microRNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol. Cell* **9:** 1327–1333.