# Vertebrate MicroRNA Genes

Lee P. Lim,[1,2]*† Margaret E. Glasner,[1,2]* Soraya Yekta,[1,2]*
Christopher B. Burge,[1]† David P. Bartel[1,2]†

MicroRNAs (miRNAs) are an abundant class of ~22-nucleotide (nt) noncoding RNAs, some of which are known to control the expression of other genes at the posttranscriptional level (*1*–*4*). We developed a computational procedure (MiRscan) to identify miRNA genes (*5*) and apply it here to identify most of the miRNA genes in vertebrates. MiRscan relies on the observation that the known miRNAs derive from phylogenetically conserved stem loop precursor RNAs with characteristic features. MiRscan evaluates conserved stem loops as miRNA precursors by passing a 21-nt window along each conserved stem loop, assigning a log-likelihood score to each window that measures how well its attributes resemble those of the first 50 experimentally verified *C. elegans* miRNAs with *C. briggsae* homologs (*2*, *3*, *5*).

Folding of aligned regions of the human and mouse genomes, with subsequent comparison to the pufferfish *Fugu rubripes* genome, identified ~15,000 human genomic segments that fell out-side of predicted protein coding genes, were predicted to form stem loops, and were at least loosely conserved among the three vertebrate species (*6*). MiRscan evaluation revealed a high-scoring set of 188 human loci, using a natural cutoff score of 10, defined by a dip in the distribution at this point (Fig. 1). This set included 81 of the 109 members of a reference set of known human miRNA loci, for a sensitivity of 0.74. The fact that a procedure developed and trained solely using nematode miRNAs could also identify most of the vertebrate miRNAs shows that the generic features of the miRNAs and their precursors are conserved broadly among diverse animals, even though the sequences of most miRNAs are not as broadly conserved.

Our analysis can be used to calculate an upper bound on the number of human miRNA genes. If all 188 candidates were authentic miRNA genes and these represented 74% of the total miRNA genes, then there are no more than 255 miRNA genes in the genome. Note that this calculation assumes that rare miRNAs—those expressed at low levels or in a limited set of conditions or cell types, which would be underrepresented in our reference set of cloned miRNAs—will have a distribution of scores and degree of conservation similar to the cloned miRNAs. This assumption is supported by our finding that in nematodes, there is no correlation between the number of times an miRNA was cloned and its MiRscan score (*5*). Furthermore, a tissue such as mouse brain, which might be expected to have miRNAs unique to mammals, is not a particularly rich source of miRNAs without *Fugu* homologs (*7*).

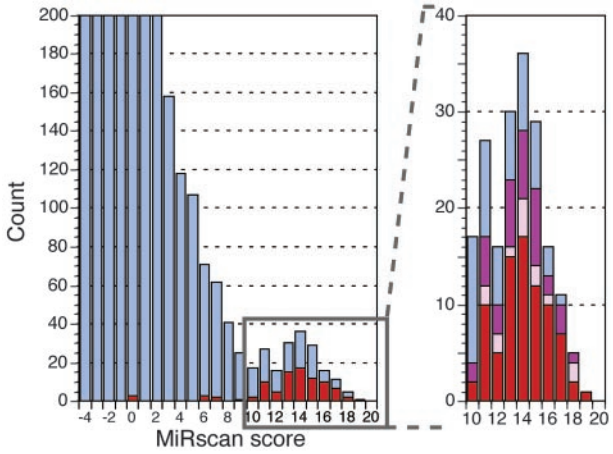


**Fig. 1.** Computational identification of vertebrate miRNA genes (*6*). The histogram represents the distribution of MiRscan scores for 15,133 human/*Fugu* consensus structures. Of the 109 reference-set loci, 91 were retained among these aligned segments (red), indicating that at least 80% of the human miRNAs are conserved in fish. The distribution peaks at the score of −4, with a count of 1198, but is truncated at a score of −4 and count of 200 to increase resolution at the high-scoring tail of the distribution. The 188 candidates with scores greater than 10.0 were examined further (expanded portion of the histogram): 81 were in the reference set of known loci (red), 14 were close paralogs of loci in the reference set (≤2 point substitutions within the miRNA) or represented cloned human miRNAs for which loci had not been previously reported (pink), and 38 were found in miRNA cDNA libraries made from zebrafish (purple) (*6*).

The estimate of 255 human genes is an upper bound implying that no more than 40 miRNA genes remain to be identified in mammals [~40 = ~255 – (109 known genes + 107 new candidates)]. The estimates for both the gene total and genes remaining to be identified would be lower if some of the 107 newly identified gene candidates were false positives. To evaluate this possibility, we sought to verify these new candidates. Of the 107 new candidates, 14 were close paralogs of loci in the reference set or represented cloned human miRNAs for which loci had not been previously reported. Another 38 were detected in zebrafish cDNA libraries constructed specifically to contain miRNA and siRNA sequences (*6*). (Zebrafish was chosen for this analysis to facilitate examination of a diverse range of tissues and developmental stages.) This leaves 55 of the 188 candidates as either false positives or authentic miRNAs expressed at levels too low to be detected. Even if all 55 were false positives, the specificity of our computational procedure would be 133/188 (= 0.71), at a score cutoff that identifies 74% of known loci. This minimum specificity value can be used to calculate a lower bound on the number of miRNA genes in mammals as (188 × 0.71)/0.74 = 180. When accounting for the sensitivity of our zebrafish experiments and the incomplete coverage of the genome assemblies used, the lower bound increases to about 200 genes (*6*).

The 200 to 255 miRNA genes represent nearly 1% of the predicted genes in humans, a fraction similar to that seen for other very large gene families with regulatory roles, such as those encoding transcription-factor proteins. There is no indication that miRNAs are present in single-celled eukaryotes such as yeast. It is tempting to speculate that the substantial expansion of miRNA genes in plants and animals (and the apparent loss of miRNA genes in yeast) is related to their importance in specifying cell differentiation and developmental patterning.

### References and Notes

1. M. Lagos-Quintana, R. Rauhut, W. Lendeckel, T. Tuschl, *Science* **294**, 853 (2001).
2. N. C. Lau, L. P. Lim, E. G. Weinstein, D. P. Bartel, *Science* **294**, 858 (2001).
3. R. C. Lee, V. Ambros, *Science* **294**, 862 (2001).
4. E. G. Moss, R. S. Poethig, *Cur. Biol.* **12**, 688 (2002).
5. L. P. Lim *et al.*, *Genes Dev.*, in press.
6. Supplemental material describing methods and sequences of the predicted miRNA loci and their validation in zebrafish is available on *Science* Online.
7. M. Lagos-Quintana *et al.*, *Curr. Biol.* **12**, 735 (2002).
8. We thank E. Wiellette and H. Sive for guidance in culturing and staging zebrafish embryos, J. Lange and D. Page for assistance and use of equipment and facilities, and Compaq for computer resources. Supported by grants from the NIH (C.B.B. and D.P.B) and a grant from the David H. Koch Cancer Research Fund (D.P.B).

**Supporting Online Material**
www.sciencemag.org/cgi/content/full/299/5612/1540/DC1
Materials and Methods
Tables S1 and S2
Fig. S1

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA.

*These authors contributed equally this work.
†To whom correspondence should be addressed.