Cell, Vol. 120, 21–24, January 14, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.cell.2004.12.031

Letter to the Editor

Phylogenetic Shadowing and Computational Identification of Human microRNA Genes

We sequenced 122 miRNAs in 10 primate species to reveal conservation characteristics of miRNA genes. Strong conservation is observed in stems of miRNA hairpins and increased variation in loop sequences. Interestingly, a striking drop in conservation was found for sequences immediately flanking the miRNA hairpins. This characteristic profile was employed to predict novel miRNAs using cross-species comparisons. Nine hundred and seventy-six candidate miRNAs were identified by scanning whole-genome human/mouse and human/rat alignments. Most of the novel candidates are conserved also in other vertebrates (dog, cow, chicken, opossum, zebrafish). Northern blot analysis confirmed the expression of mature miRNAs for 16 out of 69 representative candidates. Additional support for the expression of 179 novel candidates can be found in public databases, their presence in gene clusters, and literature that appeared after these predictions were made. Taken together, these results suggest the presence of significantly higher numbers of miRNAs in the human genome than previously estimated.

MicroRNAs (miRNAs) are noncoding RNAs that regulate the expression of genes at the posttranscriptional level (reviewed in Bartel [2004]). Although only recently discovered, they have been found to play key roles in a wide variety of biological processes, including cell fate specification, cell death, proliferation, and fat storage (reviewed in Ambros [2004]). More than 200 different miRNAs have now been described for mouse and human (Griffiths-Jones, 2004). The molecular requirements and mechanism by which miRNAs regulate gene expression are currently being clarified (Bartel, 2004), but individual biological functions remain largely unknown. Temporal and spatial expression of miRNAs may be key features driving cellular specificity.

Cross-species sequence comparison is a powerful approach to identify functional genomic elements, but its sensitivity decreases with increasing phylogenetic distance, especially for short sequences. In addition, taxon-specific elements may be missed. To overcome the limitations of classical phylogenetic footprinting methods, we applied the phylogenetic shadowing approach (Boffelli et al., 2003), allowing unambiguous sequence alignments and accurate conservation determination at single nucleotide resolution level. This approach is based on the alignment of phylogenetically closely related species; since these show only few sequence differences, many different (but related) genomes need to be aligned to identify invariant (conserved) positions. We have sequenced 700 bp regions surrounding 122 miRNAs in ten different primate species, including orangutan, gorilla, two chimpanzee and two macaque species, tamarin, spider monkey, wooly monkey, and lemur. A strong correlation between the fraction of successfully amplified regions and evolutionary distance was observed since the design of primers for PCR was based solely on human sequences (Supplemental Table S1 at http://www.cell.com/cgi/content/full/120/1/21/DC1/). We used VISTA-like plots (Mayor et al., 2000; Ovcharenko et al., 2004) for the representation and visual investigation of the 376 kb of the informative sequence obtained. An example miRNA conservation profile is shown in Figure 1A (the complete set of profiles and all underlying sequence alignment data are provided in Supplemental Figure S1 and Supplemental Dataset 1, respectively. Raw sequences have been deposited in NCBI GenBank under accession numbers AY865825–AY866368). For 68 miRNAs, we observed variations in the pre-miRNA precursor region, but only three were variable in the mature miRNA (mir-211, mir-220, and mir-198). In total, we observed 154 variations that do not affect the secondary structure of a pre-miRNA sequence and 118 variations that lead to disruption of base-pairing in a hairpin (Supplemental Figure S2), indicating that there is a selective pressure to stabilize the secondary structure of a pre-miRNA but that some structural changes are tolerated. Most of the observed variations occur in terminal loop sequences and at the ends of a precursor, suggesting that there is no functional constraint on the primary sequence of these regions of miRNA gene, which is in agreement with experimental data on sequence requirements for miRNA processing (Lee et al., 2003; Zeng and Cullen, 2003). We investigated the occurence of compensatory changes in helix-forming regions and observed only two such cases: an A:U to G:C change in mir-19b-2 lemur sequence and a G:C to A:U change in mir-220 rhesus and pigtailed macaque sequences. Interestingly, the latter resides in the mature miRNA sequence. Although compensatory changes are relatively rare, they may reflect a mechanism of miRNA evolution.

Besides the region spanning the pre-miRNA, no additional conserved regions common to different miRNAs could be found, suggesting that, in contrast to *C. elegans* (Ohler et al., 2004), no common *cis*-acting elements can be immediately recognized in mammalian miRNAs. Interestingly, there is a prominent drop of conservation immediately flanking pre-miRNA regions (Figures 1A and 1B). We hypothesized that this characteristic conservation pattern could be used to identify novel miRNA genes. Indeed, visual inspection of shadowing profiles revealed regions with patterns of conservation similar to that of known miRNAs, but more importantly, such patterns could also be recognized in pairwise alignments between more diverged species like human and mouse, clearly revealing positions of both known and novel miRNAs (Figure 1C). Next, we extended this discovery strategy to a genome-wide scale (Figure 1D) by screening mouse/human and rat/human whole-genome sequence alignments for this typical conservation pro-

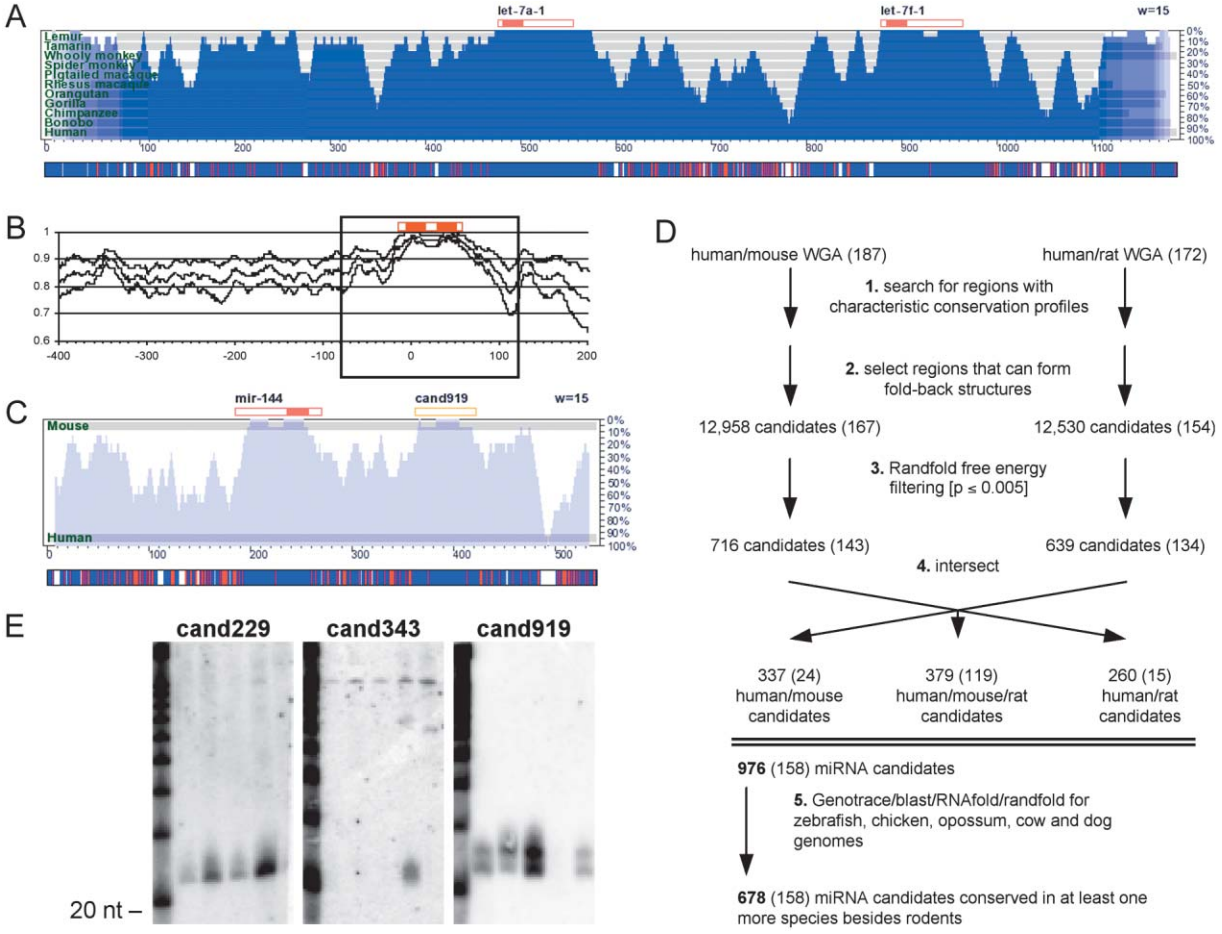

Figure 1. Prediction of Novel miRNA Genes Using Phylogenetic Shadowing Profiles

(A) Graphical VISTA-like representation of phylogenetic shadowing results in ten primate species for a genomic region harboring two known miRNA genes. Pre-miRNAs and mature miRNAs are represented as open and solid boxes on the top of the figure, respectively. For every position in an alignment, divergence is calculated in a 15-nt window centered on the position and plotted as a vertical blue line. Intensity of blue indicates sequence coverage depth and background horizontal gray rectangles show the coverage of individual monkey species. The bar below the alignment represents the nature of observed variations: blue—no variation, red—substitution, white—insertion/deletion.

(B) Cumulative conservation profile of microRNA gene regions. Phylogenetic shadowing data for 38 "left-arm" and 26 "right-arm" solitary miRNA genes are used for calculations. Conservation profiles of individual miRNAs were transformed to relative coordinates with zero corresponding to the first nucleotide of mature miRNA in case of left-arm miRNAs and to the nucleotide pairing to the last base of mature miRNA in case of right-arm miRNAs, and an averaged conservation level (middle line) with 95% confidence intervals (upper and lower lines) was calculated for every position.

(C) VISTA-like plot for human/mouse miRNA comparison. A novel candidate miRNA (cand919) was discovered in a cluster containing a known miRNA (mir-144) by visual inspection of the conservation profile between mouse and human.

(D) Outline of the algorithm for identification of miRNA candidate genes. The number of known miRNAs found at different stages of analysis is shown in brackets.

(E) Northern blot detection of candidate miRNAs. Every blot contains five lanes: decade marker (starting with 20 nt), 8.5, 12.5, and 16.5 dpc mouse embryos, mouse ES cells and brain.

file. Additional stringent filtering for the ability of candidate regions to fold into a thermodynamically favorable stable hairpin, as calculated by Randfold software (Bonnet et al., 2004), resulted in the identification of 976 candidate miRNAs (Supplemental Dataset 2), containing 83% of all known human miRNAs (158 out of 189, based on miRNA registry v.3.1).

Screening for orthologs in additional vertebrate genomes (zebrafish, chicken, opossum, cow, and dog) revealed that 678 candidates are conserved in at least one other species besides rodents (Supplemental Table S2). Notably, in zebrafish we found 130 known miRNAs and only 27 new candidates, whereas in other categories the number of novel regions substantially exceeds the number of known miRNAs. We think that this may be due to the bias in the set of known miRNAs, many of which were identified by computational screens for conserved human/fugu regions followed by experimental verification in zebrafish (Lim et al., 2003). A substantial part of our predictions consists of miRNAs unique to mammals. Although the risk for false positives in this set may be higher, both the genomic distribution and the extent of supportive data for expression are comparable for the mammalian-specific subset and the set

of candidates that are also conserved in at least one nonmammalian species (Supplemental Table S3). Even though the degree of genome coverage varies for the species used in the comparisons, these data suggest that there are a significant number of lineage-specific miRNAs and indicate that both rapidly and slowly evolving miRNAs exist (let-7 being a typical example of a slow evolver).

Fourteen novel candidates share homology with known miRNAs and an additional 60 share homology with at least one other candidate (Supplemental Table S4), making up novel subfamilies. In addition to the established clustering behavior of miRNAs (Bartel, 2004; Rodriguez et al., 2004), the ratio between the number of miRNA genes in inter- and intragenic regions is similar (1.4) for both known and novel miRNAs (Supplemental Table S5). Although a fair proportion of candidates are predicted on the strand opposite to annotated transcripts, the disproportionate presence of miRNA genes in introns is intriguing and may reflect expression mechanisms by co-transcription with the host gene and processing of spliced introns. One hundred and seventy-one of the predicted novel miRNAs reside in genomic regions that are annotated as exons. In experimental approaches, such candidates are often discarded as potential cloning artifacts, but there is no experimental evidence that these regions cannot be processed into mature miRNAs. On the contrary, work by Cullen and coworkers (Cai et al., 2004) demonstrated that a transcript harboring simultaneously a miRNA and an ORF is efficiently used for both miRNA and protein production. About 25% (44) of the exonic candidates reside in noncoding parts and although 127 candidates overlap with annotated protein coding sequences, 75 are predicted on the opposite strand.

Support for the expression of candidate miRNAs is provided through various sources. Three candidates are present in the FANTOM2 database of expressed sequences and 11 candidates reside in gene clusters containing one or more known miRNAs. These miRNAs are likely to be coexpressed from the same primary transcript (Bartel, 2004; Rodriguez et al., 2004). Systematic human transcriptome analysis using high-density oligonucleotide tiling arrays (Kapranov et al., 2002) is in progress, and we found that the genomic regions encoding 64 known and 214 novel miRNAs have now been covered. From this set, 13 known (20%) and 72 novel (34%) miRNAs are expressed in the SK-N-AS cell line, for which data are publicly available (Supplemental Table S6). Although poly(A)$^+$ RNA was used for these experiments and properties of miRNA-containing transcripts remain largely to be elucidated, both intergenic and intronic miRNAs were detected. Various lines of research support the finding that at least some miRNAs are processed from poly-adenylated RNA (Cai et al., 2004; Lee et al., 2004).

To provide experimental support for our predictions, we have performed Northern blotting experiments for 69 candidates (Figure 1E and Supplemental Figure S3), confirming the expression of 16 mature miRNAs (23%). Interestingly, 11 of them show strong bands, potentially corresponding to precursor miRNAs (between 60 and 100 nt), which is also seen for 19 other candidates for which no mature miRNA could be detected by Northern blots (Supplemental Figure S3C). This may at least partially be explained by picking the wrong hairpin arm for probe design. Indeed, for two of these candidates additional evidence for expression is provided by others (Poy et al., 2004). Although these verification rates are lower than previously published rates using cloning- and PCR-based approaches (38 out of 93; Lim et al., 2003), they may be an underrepresentation as a result of a bias in the set of already known miRNAs for highly expressed and thus most easily detectable miRNAs, the sensitivity of the detection method, and spatio-temporal limitations of the RNA samples used. Indeed, several miRNAs can only be detected in one out of the five samples tested (Figure 1E). Of those miRNAs also conserved in zebrafish (comparable to Lim et al., 2003), expression of 6 out of the 12 tested candidates was confirmed (Supplemental Figure S3B).

After our predictions were made, several papers appeared providing experimental evidence for 40 novel miRNA genes. Half of these miRNAs were present in our predictions, further supporting the value of this data set. For example, Stoffel and coworkers (Poy et al., 2004) cloned 67 different miRNAs from a pancreatic β cell line. Although cloning approaches are sometimes thought to be close to saturation, this effort yielded 11 novel miRNAs, indicating that tissue-specific and developmental stage-specific expression may affect the results of experimental approaches. However, 9 out of these 11 novel miRNAs were present in our computationally derived set of candidates.

Taken together, phylogenetic shadowing of miRNAs in primate species revealed a characteristic conservation profile that can be utilized to efficiently detect the vast majority (83%) of known miRNAs and predict an extensive set of novel miRNAs based on genome-wide human-mouse-rat comparisons. Although final numbers will depend on extensive detection experiments for mature miRNA expression, a reasonable estimate can be made based on our results. First, extrapolation of our Northern blot experiments suggests that about 190 novel miRNAs in our set could be confirmed by Northern blotting (23% of the 818 newly predicted candidates). Secondly, additional support for the expression of a nonredundant set of 160 novel candidates can be found in transcript (3) and microarray-based expression (72) databases, their presence in miRNA gene clusters (74), and literature that appeared after starting the predictions (20). Finally, systematic transcriptome analysis is publicly available only for part of the human chromosome and a single cell line tested, and extrapolation to complete genome coverage suggests that 278 of the novel miRNAs are expressed in this cell line. Conservative interpretation of these data supports the presence of 200 to 300 novel miRNA genes in our data set, more than doubling current estimates (Lim et al., 2003). Taking into account that there may be classes of miRNAs that have a highly restricted temporal and spatial expression pattern and that significant numbers of taxon- and species-specific candidates exist that are missed by computational comparative methods like ours, this number may still be an underestimate. It may even be possible that as many as 1,000 miRNAs exist in a vertebrate genome. Based on the previously estimated 250 miRNAs, 10% of all protein-coding transcripts were thought to

be regulated by miRNAs (John et al., 2004). Our results indicate that this fraction may be substantially larger.

Eugene Berezikov, Victor Guryev, José van de Belt, Erno Wienholds, Ronald H.A. Plasterk,* and Edwin Cuppen
Hubrecht Laboratory
Uppsalalaan 8
3584 CT, Utrecht
The Netherlands

*Correspondence: plasterk@niob.knaw.nl

Selected Reading

Ambros, V. (2004). Nature *431*, 350–355.

Bartel, D.P. (2004). Cell *116*, 281–297.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Science *299*, 1391–1394.

Bonnet, E., Wuyts, J., Rouze, P., and Van De, P.E.Y. (2004). Bioinformatics *20*, 2911–2917.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). RNA *10*, 1957–1966.

Griffiths-Jones, S. (2004). Nucleic Acids Res. *32 Database issue*, D109–D111.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004). PLoS Biol 2(11): e363 DOI: 10.1371/journal.pbio.0020363.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. (2002). Science *296*, 916–919.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V.N. (2003). Nature *425*, 415–419.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). EMBO J. *23*, 4051–4060.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003). Science *299*, 1540.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. (2000). Bioinformatics *16*, 1046–1047.

Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P., and Burge, C.B. (2004). RNA *10*, 1309–1322.

Ovcharenko, I., Boffelli, D., and Loots, G.G. (2004). Genome Res. *14*, 1191–1198.

Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P.E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P., and Stoffel, M. (2004). Nature *432*, 226–230.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. (2004). Genome Res. *14*, 1902–1910.

Zeng, Y., and Cullen, B.R. (2003). RNA *9*, 112–123.