Minireview

Towards a complete description of the microRNA complement of animal genomes

Julius Brennecke and Stephen M Cohen

Address: European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

Correspondence: Stephen M Cohen. E-mail: Stephen.Cohen@embl-heidelberg.de

Published: 21 August 2003

Genome ***Biology*** 2003, **4:**228

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/9/228

Abstract

Recent cloning and computational studies have sought to catalog all the microRNA genes encoded in animal genomes. Here, we highlight recent advances in identifying *Caenorhabditis elegans* and *Drosophila melanogaster* microRNAs.

One of the big surprises of the past few years has been the emergence of microRNAs (miRNAs) as a major new class of regulatory gene products. These small RNA molecules control gene expression by regulating the stability or translation of mRNAs. The growing awareness that there might be quite large numbers of miRNA genes has triggered the use of systematic approaches to define the complete set of miRNAs encoded in animal genomes. Here, we summarize these efforts, highlighting recent papers that estimate the miRNA complement of the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes.

microRNAs are thought to control gene expression by binding to complementary sites in target mRNAs from protein-coding genes. Much of our understanding of miRNA biogenesis and function comes from their similarity to small interfering RNAs (siRNAs), short RNAs of around 22 nucleotides that mediate RNA interference [1-5]. Like siRNAs, miRNAs are processed by endonucleolytic cleavage from larger double-stranded RNA precursor molecules. The resulting small single-stranded RNAs are incorporated into a multiprotein complex, termed RISC, in the case of siRNAs, or miRNP, in the case of miRNAs [2,6], although these may prove to be two names for the same thing. The small RNA provides sequence information that is used to guide the RNA-protein complex to its target RNA molecules [7,8]. The degree of complementarity between the small RNA and its target determines the fate of the bound mRNA [9,10]. Perfect pairing induces target RNA cleavage, as is the case for siRNAs and most plant miRNAs [11,12]. Imperfect pairing in the central part of the duplex presumably does not allow cleavage to occur and instead leads to a block in translation, as shown for the founding members of the miRNA family, *lin-4* and *let-7* of *C. elegans* [13-15]. Despite these similarities, miRNAs can be distinguished from siRNAs according to the following four criteria [16]. First, miRNAs are excised from endogenous transcripts that have the ability to locally form stable, primarily double-stranded, hairpin structures of around 70 nucleotides. Second, the hairpin structure is usually conserved in closely related species and often in distantly related species. Third, the hairpin is processed into one discrete mature miRNA. Finally, miRNAs regulate the expression of genes encoded at another locus, whereas siRNAs regulate the locus from which their sequence derives.

Evolutionary conservation of many miRNAs, even across phyla [17], suggests ancient and important roles for this class of regulator. The observation that they are found in multicellular plants and animals but not in unicellular eukaryotes led to speculation that miRNAs were essential in the evolution of multicellular organisms [18]. How many of these tiny regulators are hidden in animal genomes? Systematic miRNA-sequencing efforts and computational approaches are converging on an answer to this question.

Cloning endogenous RNAs that fall into the 18-25 nucleotide size range has proven to be a powerful way to identify

miRNAs [6,18-24]; more than 200 miRNA-coding genes have been identified in this way. Cloning efforts are limited by transcript abundance, and it can be expected that they might not find miRNAs expressed at low levels or in few cells, or miRNAs expressed only under particular conditions. Other difficulties include background from small RNAs that arise as degradation products of abundant cellular RNAs, which range from one quarter to over half of all clones in different studies. Where possible, affinity purification of the miRNP complex can enrich for miRNAs [6].

As an alternative approach, four groups have used structural features of known miRNAs to develop computational strategies for searching nematode, fly and vertebrate genomes for miRNA genes [18,22,23,25,26]. Despite some slight differences, each study used a similar overall strategy consisting of three major steps: One, hairpin-like structures residing in intergenic or intronic sequences are identified; two, the identified hairpin-set is refined by applying a series of structural filters; and three, sequence conservation filters are applied between closely and distantly related species, or sequence similarity to already known miRNAs, to further refine the set. These searches were successful in identifying most cloned miRNAs and led to the identification of many new miRNAs. Expression of some newly predicted miRNAs was validated by northern blot analysis, and others too low in abundance to be detected in this way were validated using a more sensitive PCR-based assay [18,23]. The Bartel lab's *C. elegans* study [18] found no correlation between miRNA abundance and the success of the computational method in predicting them, emphasizing the utility of computational screens for identifying rare miRNAs.

One difficulty with the computational approaches that have been used is sorting out new miRNA genes from random sequences that can form plausible-looking hairpins. It is necessary to set thresholds that enrich for true positives while not including too many false positives. In each case, this was done by evaluating how well the methods predicted known miRNAs. The different studies [18,22,23,25,26] found most previously validated miRNAs, but only 50-75% were among the 'high-confidence' predictions; it was not possible to find rules that do not miss any of the known miRNAs. Some real miRNAs were also missed, for trivial reasons such as problems of genome annotation or incomplete genome data (conversely, some validated miRNAs were not found in closely related genomes). In most cases, however, the identification seems to have been hampered by our limited knowledge of the specific sequence or structural features in the short miRNA genes that distinguish them from background 'hits' in the genome.

The most powerful step in refining the initially large sets of candidates without losing many of the known examples seems to be evolutionary conservation. This is exemplified by the following comparison. Using the identical computational strategy, the sensitivity in identifying known miRNAs rose from 50% for worms (two genomes) to 75% for vertebrates (three genomes), despite the fact that the analyzed vertebrate genomes are between 4 and 30 times larger [18,25]. Third and fourth genome comparison was also successfully used in the *Drosophila* study, where mosquito and honeybee sequences were crucial in identifying new miRNAs among many potential candidates [26]. Sequencing additional vertebrate, worm and insect genomes is likely to be a powerful resource for improving computational prediction methods for miRNA genes. In addition, Lai *et al.* [26] found that sequence conservation was significantly higher in the miRNA-producing arm of the hairpin than in the opposite arm or the terminal loop, providing a powerful filter to reduce the number of false positives.

A closer look at the *Drosophila* studies [24,26] illustrates that a combination of experimental and computational approaches will be needed to identify the full set of miRNA genes. The total number of fly miRNAs validated by sequencing or by northern blot stands at 76. Of these, 61 were identified by sequencing and 60 were predicted in the top scoring set by computation, with 48 in common between the two sets. It is interesting to note that 3 of the 76 validated miRNAs were not found by either method, but were picked out because of their proximity to known miRNAs.

The combined results of the sequencing efforts and computational searches provide a basis for estimating the number of miRNA genes in invertebrate and vertebrate genomes. Upper limits have been derived by extrapolating from the fraction of previously known miRNAs in the high-confidence prediction sets. On this basis, the worm and fly genomes are estimated to contain 100-120 distinct miRNA genes each, of which 96 and 76 have been experimentally validated; 109 of the predicted 200-250 vertebrate miRNAs have been validated. Ruvkun and colleagues [23] estimated 140-300 miRNA genes in *C. elegans*, considerably more than the estimates in the two other studies [18,22]. Nevertheless, the various estimates for the number of miRNA genes are around 0.5-1% of the number of predicted protein-coding genes, underlining the potential importance of miRNAs as a class of regulatory gene products.

Despite several advantages, computational approaches only allow the identification of genes that resemble those in the training set. The miRNA sequencing projects identified two additional classes of Dicer-processed small RNAs. A large class of siRNAs (termed rasiRNAs) that derive from chromosomal repeats and others complementary to transposons have been identified in *Drosophila* [24]. Tuschl and colleagues suggest that rasiRNAs could play important roles in chromosomal maintenance and transposon silencing [24]. In addition, siRNAs complementary to more than 500 distinct protein-encoding genes were identified in *C. elegans*, suggesting that regulation of gene expression by RNAi is a

common feature of *C. elegans* development [22]. The Ambros lab also reports the identification of a new class of small noncoding RNAs that they call tncRNAs, which seem to be related to miRNAs except that they do not appear to be encoded by conserved hairpin-like precursors [22].

In retrospect, it is perhaps surprising that the large class of genes that encode small RNAs could have gone almost unnoticed for so many years. Elegant new cloning strategies and computational methods have brought us to the point where we can now say that most genes that fit the current definition of miRNAs have been identified (although some surprises might still come). The next big challenge will be to find out what all these miRNAs do. Genetics will help, but it seems likely that a combination of new experimental and computational approaches will provide the solution to this problem as well.

References

1. Zamore PD, Tuschl T, Sharp PA, Bartel DP: **RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals.** *Cell* 2000, **101:**25-33.
2. Hammond SM, Bernstein E, Beach D, Hannon GJ: **An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells.** *Nature* 2000, **404:**293-296.
3. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T: **Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells.** *Nature* 2001, **411:**494-498.
4. Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD: **A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA.** *Science* 2001, **293:**834-838.
5. Bernstein E, Caudy AA, Hammond SM, Hannon GJ: **Role for a bidentate ribonuclease in the initiation step of RNA interference.** *Nature* 2001, **409:**363-366.
6. Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, Dreyfuss G: **miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs.** *Genes Dev* 2002, **16:**720-728.
7. Martinez J, Patkaniowska A, Urlaub H, Luhrmann R, Tuschl T: **Single-stranded antisense siRNAs guide target RNA cleavage in RNAi.** *Cell* 2002, **110:**563-574.
8. Schwarz DS, Hutvagner G, Haley B, Zamore PD: **Evidence that siRNAs function as guides, not primers, in the *Drosophila* and human RNAi pathways.** *Mol Cell* 2002, **10:**537-548.
9. Doench JG, Petersen CP, Sharp PA: **siRNAs can function as miRNAs.** *Genes Dev* 2003, **17:**438-442.
10. Hutvagner G, Zamore PD: **A microRNA in a multiple-turnover RNAi enzyme complex.** *Science* 2002, **297:**2056-2060.
11. Llave C, Xie Z, Kasschau KD, Carrington JC: **Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA.** *Science* 2002, **297:**2053-2056.
12. Tang G, Reinhart BJ, Bartel DP, Zamore PD: **A biochemical framework for RNA silencing in plants.** *Genes Dev* 2003, **17:**49-63.
13. Lee RC, Feinbaum RL, Ambros V: **The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*.** *Cell* 1993, **75:**843-854.
14. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*.** *Cell* 1993, **75:**855-862.
15. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*.** *Nature* 2000, **403:**901-906.
16. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, *et al.*: **A uniform system for microRNA annotation.** *RNA* 2003, **9:**277-279.
17. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P, *et al.*: **Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA.** *Nature* 2000, **408:**86-89.
18. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of *Caenorhabditis elegans*.** *Genes Dev* 2003, **17:**991-1008.
19. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*.** *Science* 2001, **294:**858-862.
20. Lee RC, Ambros V: **An extensive class of small RNAs in *Caenorhabditis elegans*.** *Science* 2001, **294:**862-864.
21. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs.** *Science* 2001, **294:**853-858.
22. Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D: **MicroRNAs and other tiny endogenous RNAs in *C. elegans*.** *Curr Biol* 2003, **13:**807-818.
23. Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J: **Computational and experimental identification of *C. elegans* microRNAs.** *Mol Cell* 2003, **11:**1253-1263.
24. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T: **The small RNA profile during *Drosophila melanogaster* development.** *Dev Cell* 2003, **5:**337-350.
25. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP: **Vertebrate microRNA genes.** *Science* 2003, **299:**1540.
26. Lai EC, Tomancak P, Williams RW, Rubin GM: **Computational identification of *Drosophila* microRNA genes.** *Genome Biol* 2003, **4:**R42.