MicroRNAs: Genomics, Biogenesis, Mechanism, and Function

(D. Bartel Cell 2004)

The microRNAs of *Caenorhabditis elegans*

(Lim *et al*. Genes & Development 2003)

Vertebrate MicroRNA Genes

(Lim *et al*. Science 2003)

Jia Jian Liu

Eric Bishop

Steve Parker

September 22, 2004

Overview of miRNA

- Brief history of miRNA;
- miRNA genes and structure;
- miRNA transcription and maturation;
- siRNA;
- miRNA function, targets;

Brief history

- MicroRNAs (miRNAs) are endogenous ~22 nt RNAs that play important roles in regulating gene expression in animals, plants, and fungi.
- The first miRNAs, *lin-4, let-7*, were identified in C. elegans *(Lee R et al. 1993; Reihhart et al. 2000)* when they were called small temporal RNAs (stRNA);
- The *lin-4* and *let-7* stRNAs are now recognized as the founding members of an abundant class of tiny RNAs world, such as miRNA, siRNA, coRNA, ncRNA and so on *(Ruvkun G. 2001. Bartel DP, 2004. Herbert A. 2004).*

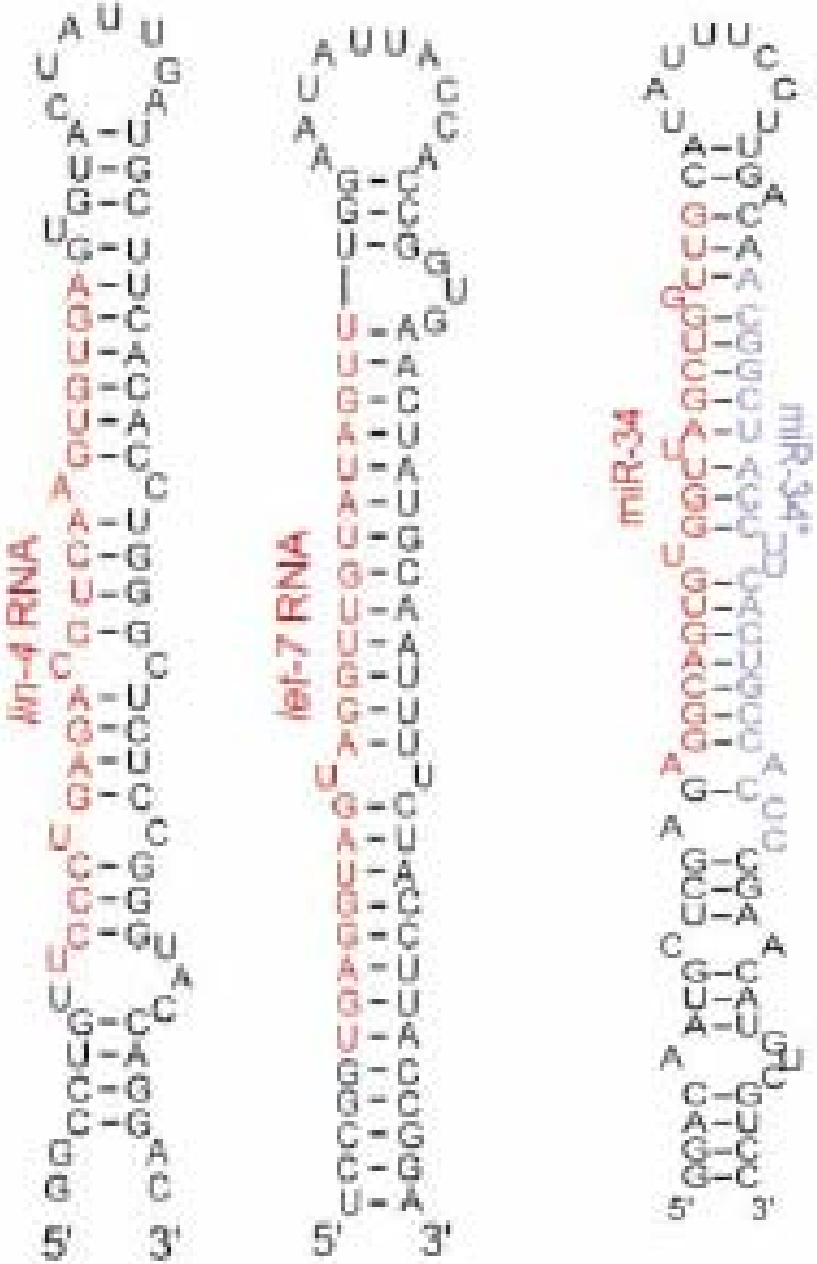
miRNA genes

- Most miRNA genes come from regions of the genome quite distant from previously annotated genes, implying they derive from different transcription units (TUs);
- The miRNAs within a genomic cluster are often related to each other (but not always);
- Not all of the cloned miRNAs are conserved even in very closely related animals, such as human/mouse, C. elegans/C. briggsae (see main paper for details);

miRNA expression/structure:

- Many miRNAs have intriguing expression patterns.
- It is tempting to speculate that the substantial expansion of miRNA genes/expression in plants and animals (and the apparent loss of miRNA in single celled eukaryotes such as yeast) is related to cell differentiation and developmental patterning (see the main paper).
- miRNA precursors Stem loop structure (thus computational methods searching hairpins).

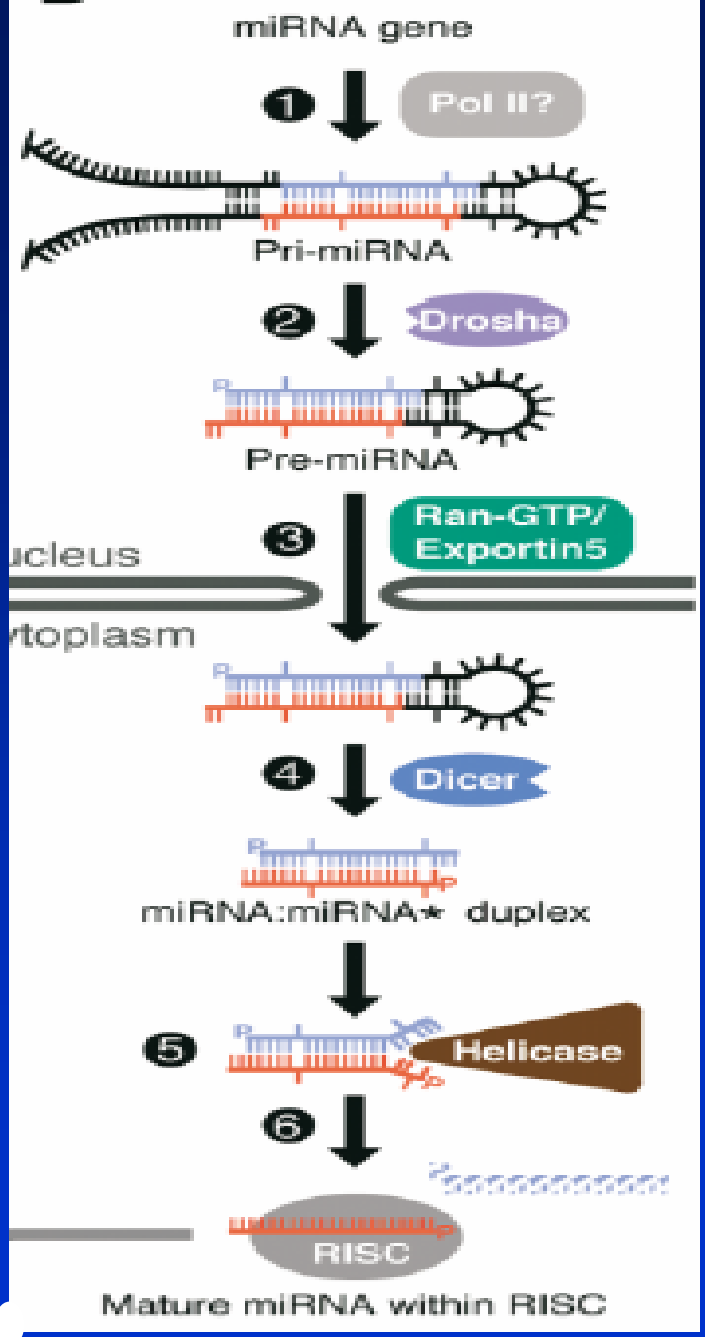
Predicted miRNA stem/loop structure. The precise sequences of the mature miRNAs(red) and miRNA* (blue) were defined by cloning.



Bartel DP. 2004. Cell.116:281

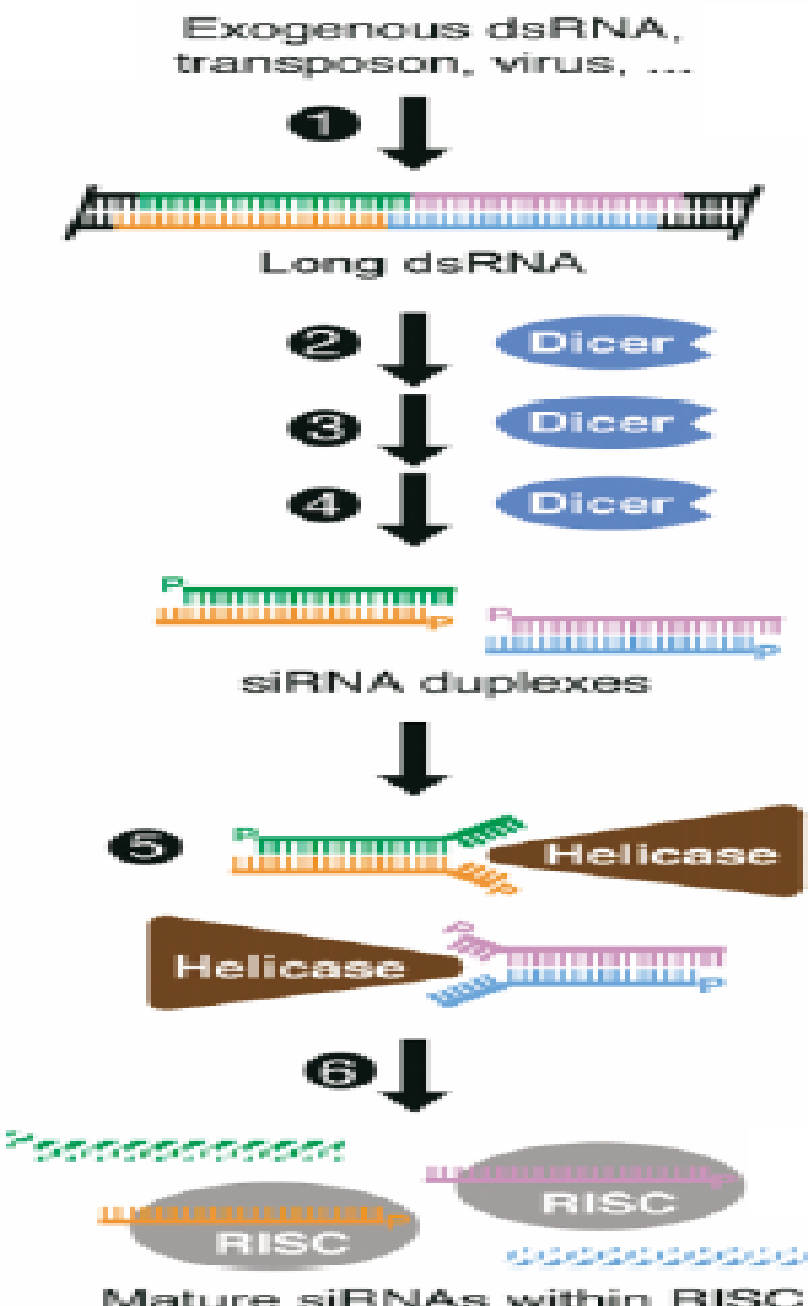
miRNA transcription and maturation

For Metazoan miRNA:
Nuclear gene to pri-miRNA(1); cleavage to miRNA precursor by Drosha RNaseIII(2); actively (5'-p, ~2nt 3'overhang) transported to cytoplasm by Ran-GTP/Exportin5 (3); loop cut by dicer(RNaseIII)(4); *duplex is generally short-lived, by Helicase to single strand RNA, forming RNA-Induced Silencing Complex, RISC/maturation (5-6).



Animal siRNA

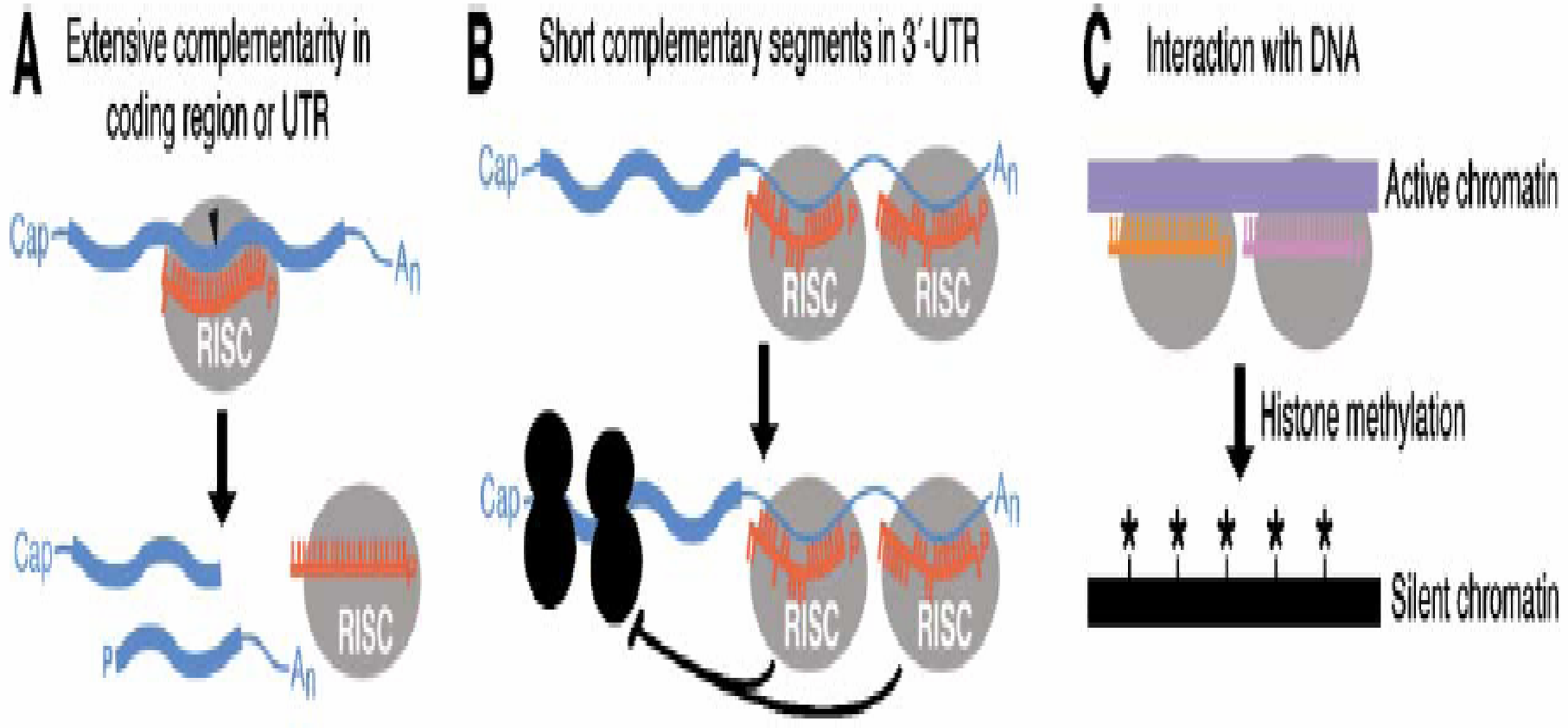
Dicer was first recognized for its role in generating small interfering RNAs (siRNAs), and was later shown play a role in miRNA maturation (the other end of miRNA maturation shown in previous fig). Difference of miRNA and siRNA: 1)miRNA derive from genome loci distinct from other recognized genes; whereas siRNA derive from mRNAs,transposons, viruses, or heterochromatic DNA (step1); 2)miRNAs precursors have hairpin structures, whereas siRNAs are processed from long bimolecular RNA duplex, generating more dif siRNAs; 3)miRNA sequences are nearly always conserved in related organism, whereas siRNA are rarely conserved; 4) miRNA/RISC hetero-silencing of loci unrelated to that from which it originated; whereas siRNA auto-silencing of the same/similar loci from which it originated **(gene knockdown);**



miRNA Function

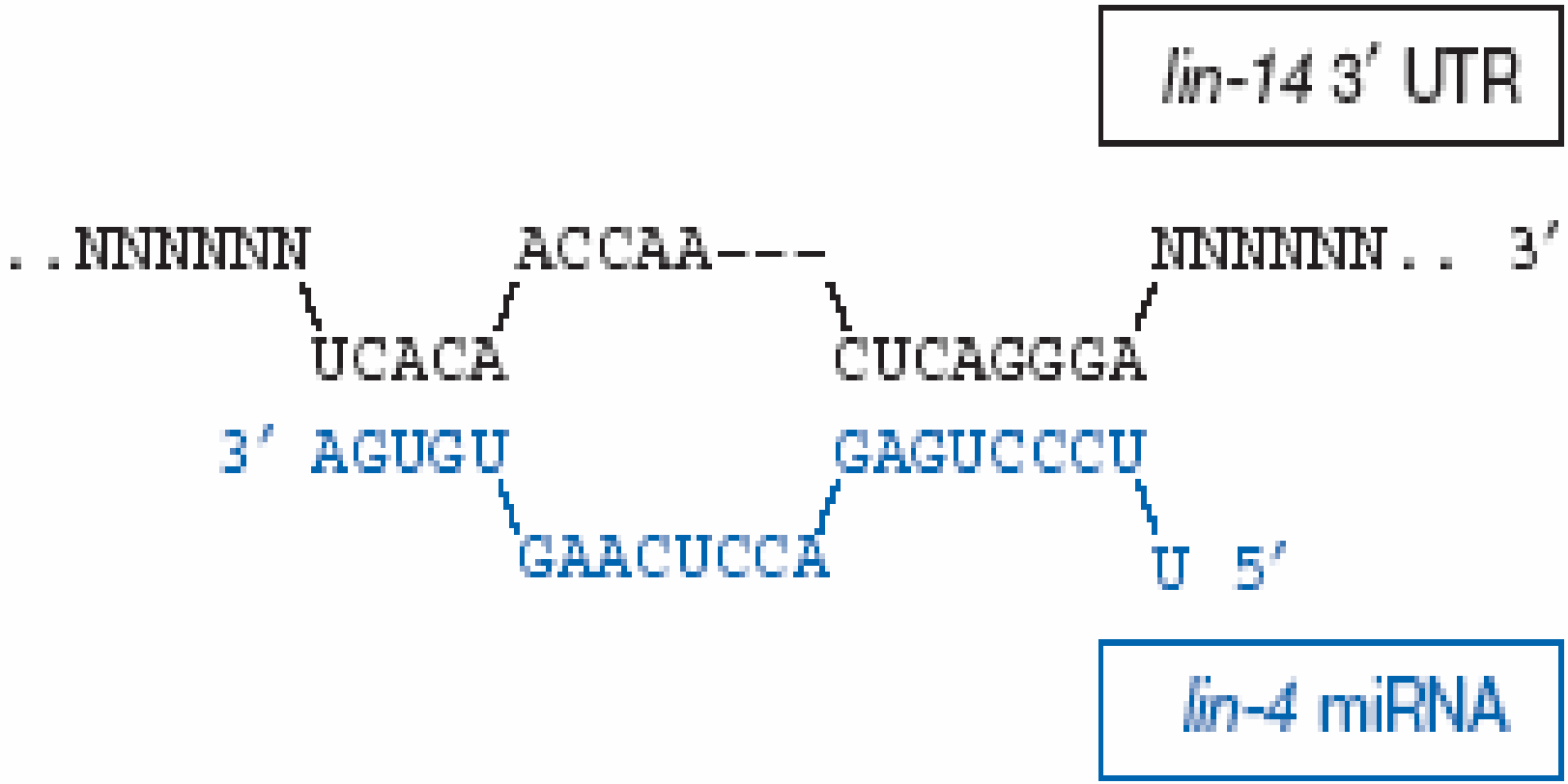
- miRNAs have important functions including control of cell proliferation, cell death, and fat metabolism; neuronal patterning; modulation of hematopoietic lineage differentiation, and control of leaf and flower development.

The actions of small silencing RNA



A, mRNA cleavage specified by a miRNA/siRNA; B, translational repression specified by miRNAs/siRNAs; C, transcriptional silencing, thought to be specified by heterochromatic siRNAs

miRNA Target



Ambros V. Nature. 2004.431:350

Computational program to identify miRNA genes

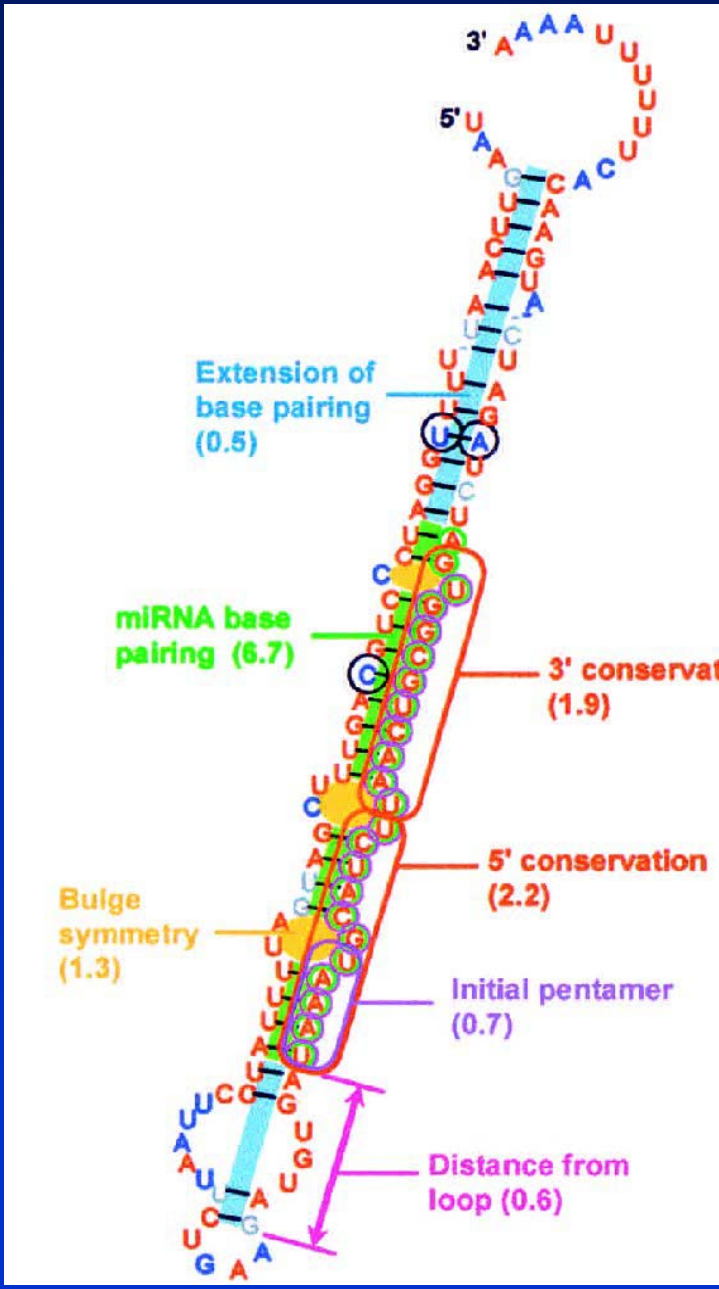
- Significant progress has been made in miRNA research since the report of the *lin-4* RNA(1993). About 300 miRNAs have been identified in different organisms to date.
- However, experimental identification miRNAs is still slow since some miRNAs are difficult to isolate by cloning due to low abundance /stability/ expression pattern/cloning procedure. Thus, computational identification of miRNAs from genomic sequences provide a valuable complement to cloning. Steve/Eric are going to talk more about this in the main paper…

Computational Prediction of miRNAs

- Lim et al. developed a tool called MiRscan to help identify new miRNA genes
  - This program looks at hairpin sequences conserved between species
  - The program was given a training set of known miRNAs in *C. elegans*
  - This data was then used to identify which conserved hairpin sequences were most similar to the training data.

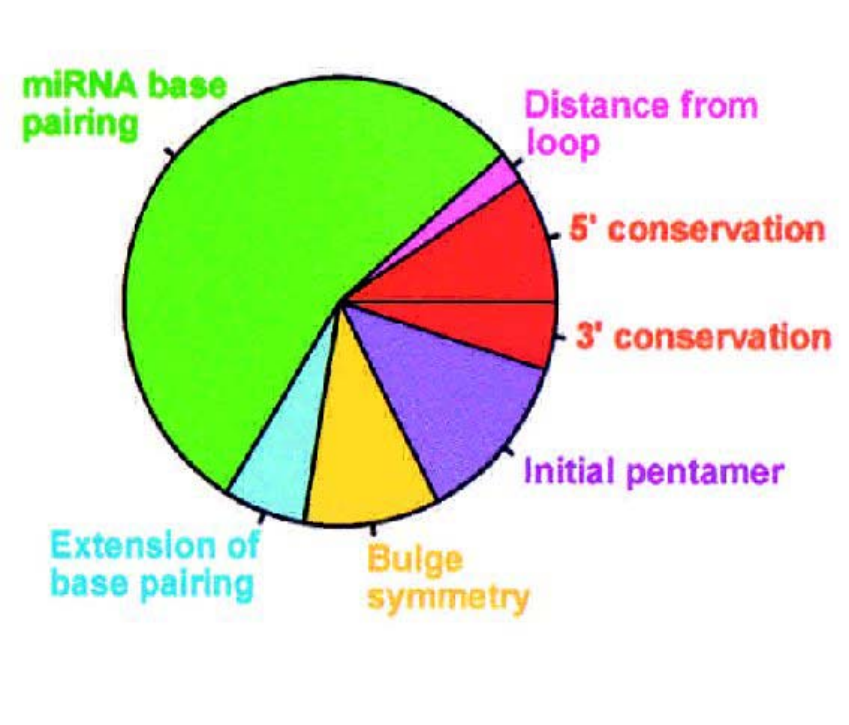
Algorithm

- The MiRscan algorithm examines several features of the hairpin
- The total score computed by summing the score of each feature
- The score for each feature is computing by dividing the frequency of the given value in the training set to its overall frequency



Relative Importance of Hairpin Features

- Certain features were found to be more useful than others in distinguishing miRNAs

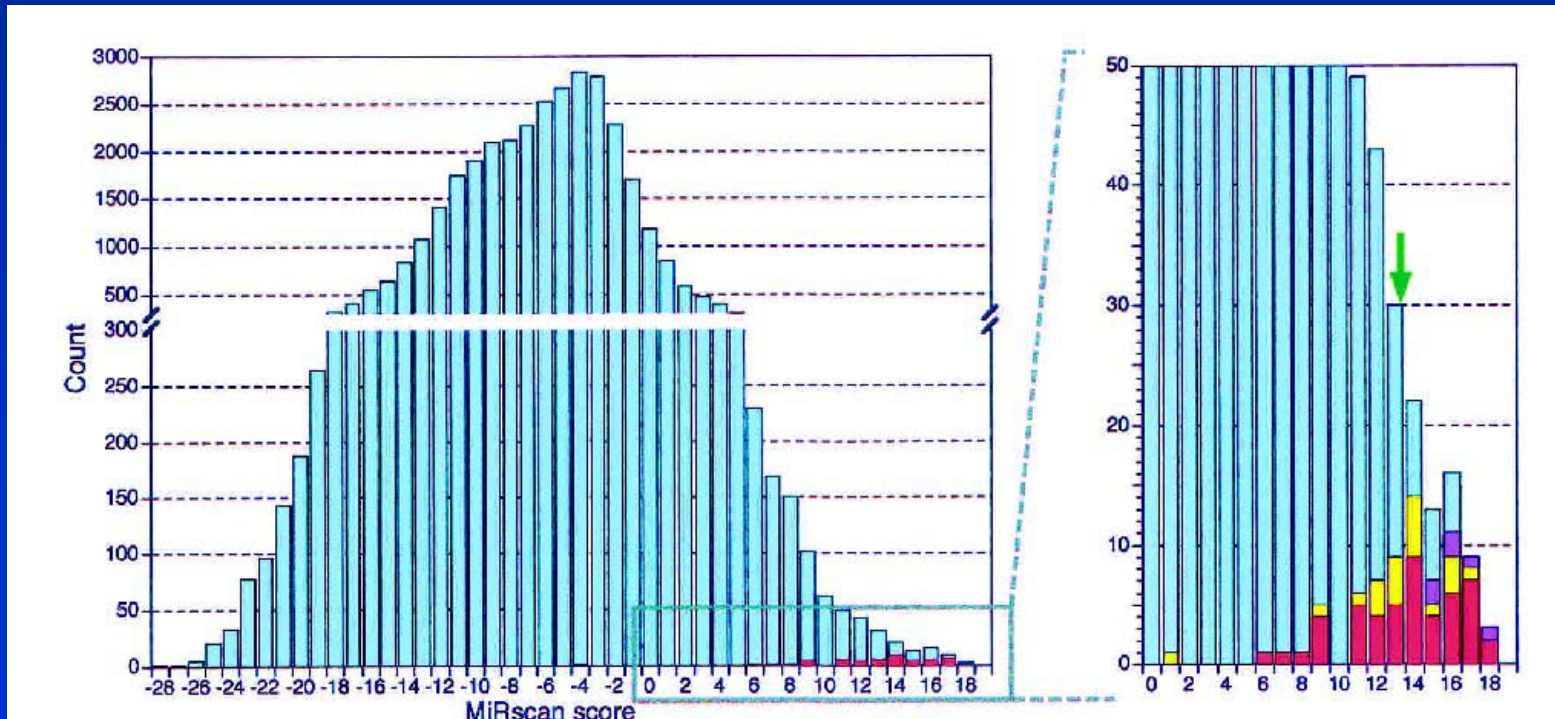


Testing the Algorithm

- In order to test their algorithm, Lim et al. ran MiRscan on the ~36,000 conserved hairpins in the *C. elegans* and *C. briggsae* genomes
  - The 50 known miRNA genes conserved between *C. elegans* and *C. briggsae* were used as a training set
  - 35 sequences received a MiRscan score greater than the mean score of the known genes
  - These sequences were given special attention in the experimental portion of this research

Testing the Algorithm (cont'd)

- A total of 58 miRNA genes are known in C. elegans, but the remaining 8 were not identified by MiRscan because they are not conserved in C. briggsae



Identification of New miRNA Genes

- Lim et al. scaled up their previous molecular cloning procedure to identify new miRNA genes
- Also, RNA was taken from worms in different stages of development, to obtain miRNA clones that might not have been expressed in mature worms
- 18-24 base RNA was purified, then ligated to 5' and 3' adapter sequences. RT PCR was done on these fragments, and the products were cloned and sequenced

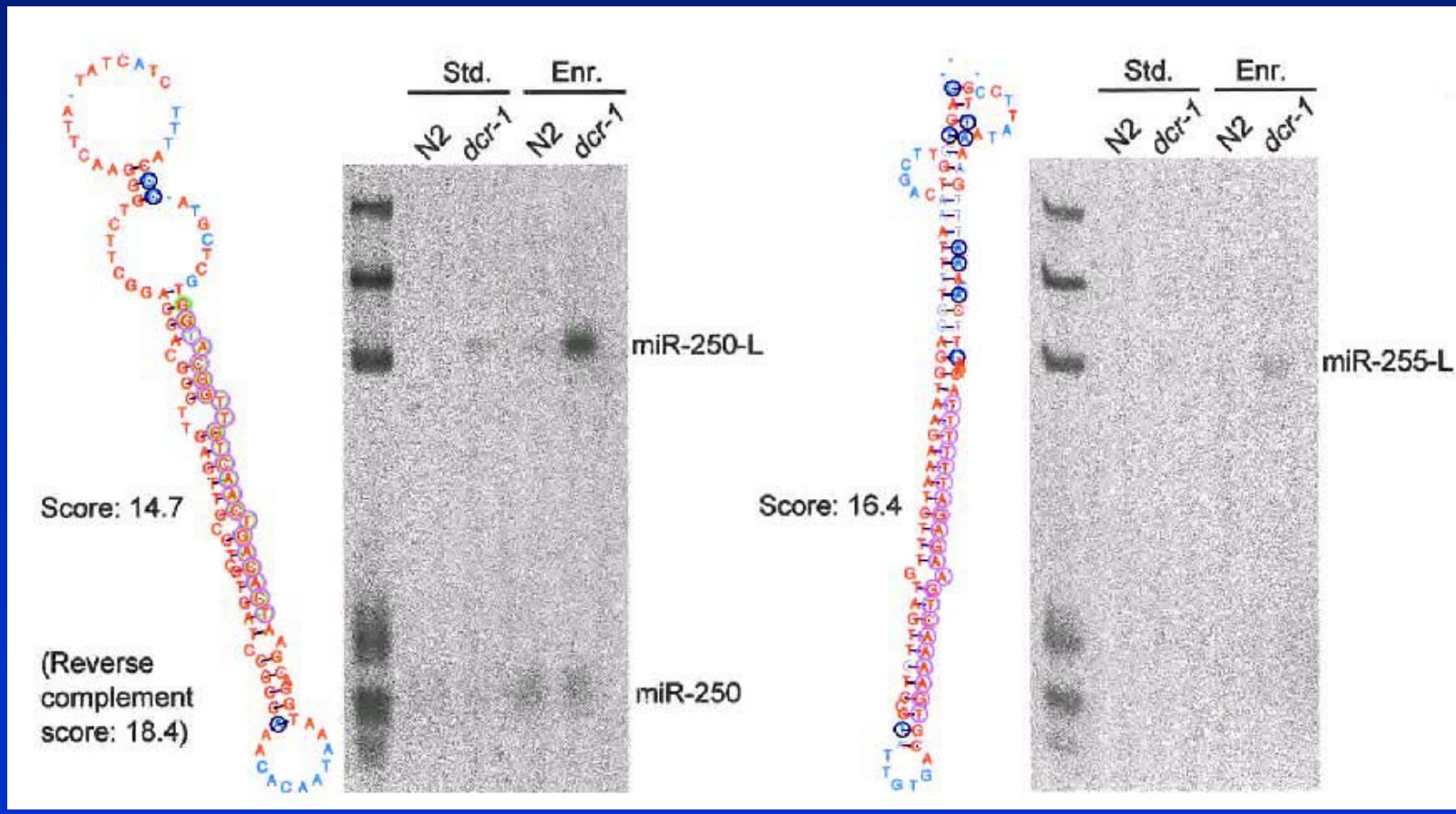
Identification of New miRNA Genes (cont'd)

- 3523 clones were identified as miRNA genes
  - Most of these were one of the 58 genes already identified
  - However, 404 of these corrospond to 23 new miRNA loci
  - 10 of the 23 newly identified genes were among the 35 top candidates identified by MiRscan

Northern Blots

- To validate the 25 genes predicted by MiRscan, but not cloned, northern blots were conducted
- To increase signal strength, RNA was enriched for small sequences
- Additionally, RNA from dicer mutants (dcr-1) was probed as well, to detect the precursor better
- Six of the 25 predicted genes were confirmed with this technique. However, signal strength tended to be weak, indicating low concentration in the sample.

Example Northerns



PCR Assays

- In addition to the Northern blots, researchers used a PCR assay to investigate the presense of the 25 candidates not cloned
- Primers were designed for the 3’ and 5’ flanking regions of the candidates, and then the RNA library was probed for precursors
- Five of the six miRNA sequences identified by Northerns were found this way, but no others

Analysis of MiRscan Effectiveness

- Lim et al. conclude that their algorithm's success rate is 0.70 at a tolerance that detects ½ of known miRNA
  - 58 C. elegans miRNA genes were known initially
  - 16 of the 35 high-scoring candidates were confirmed experimentally
  - Half (29=58/2) of the known miRNA genes were given a score above the top 35 unknown candidates.
  - So, this success rate is computed: (29+16)/(29+35)

Evolutionary Conservation of miRNA sequences

- Lim et al. compared the identified miRNA sequences from *C. elegans* to the human genome, and found that over 1/3 of these genes had homologs in humans.

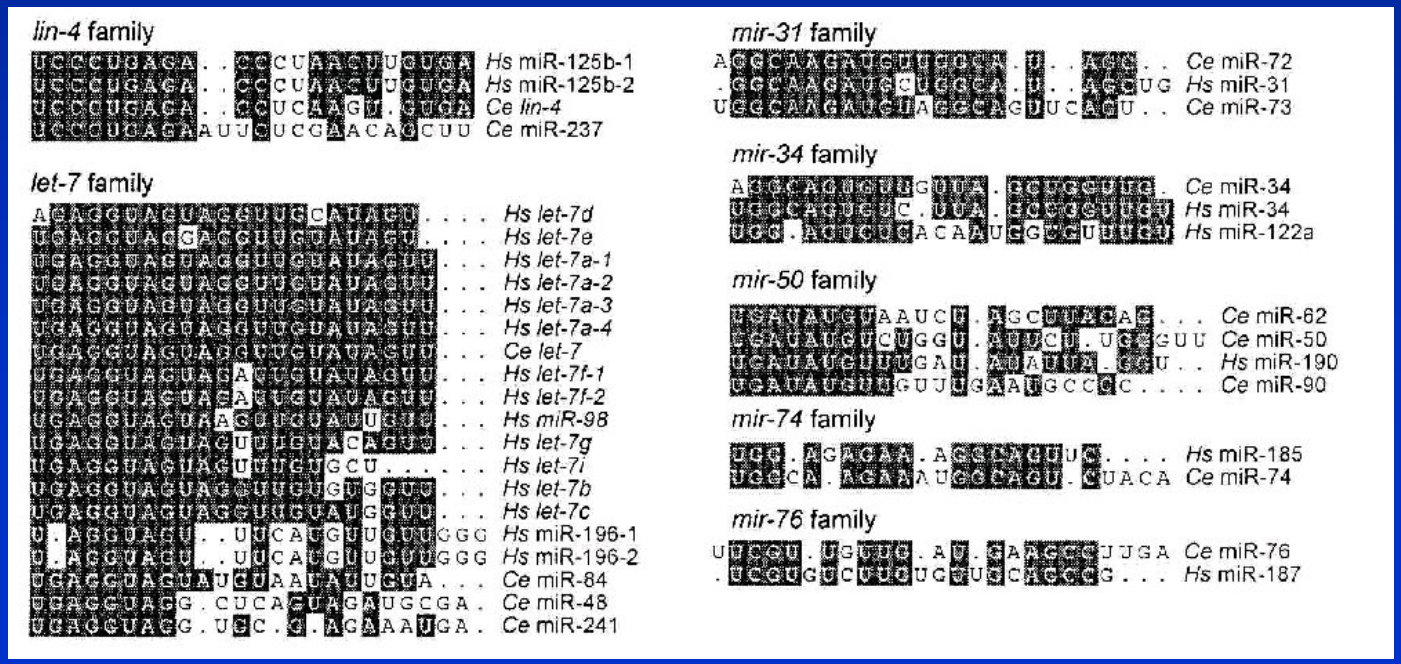
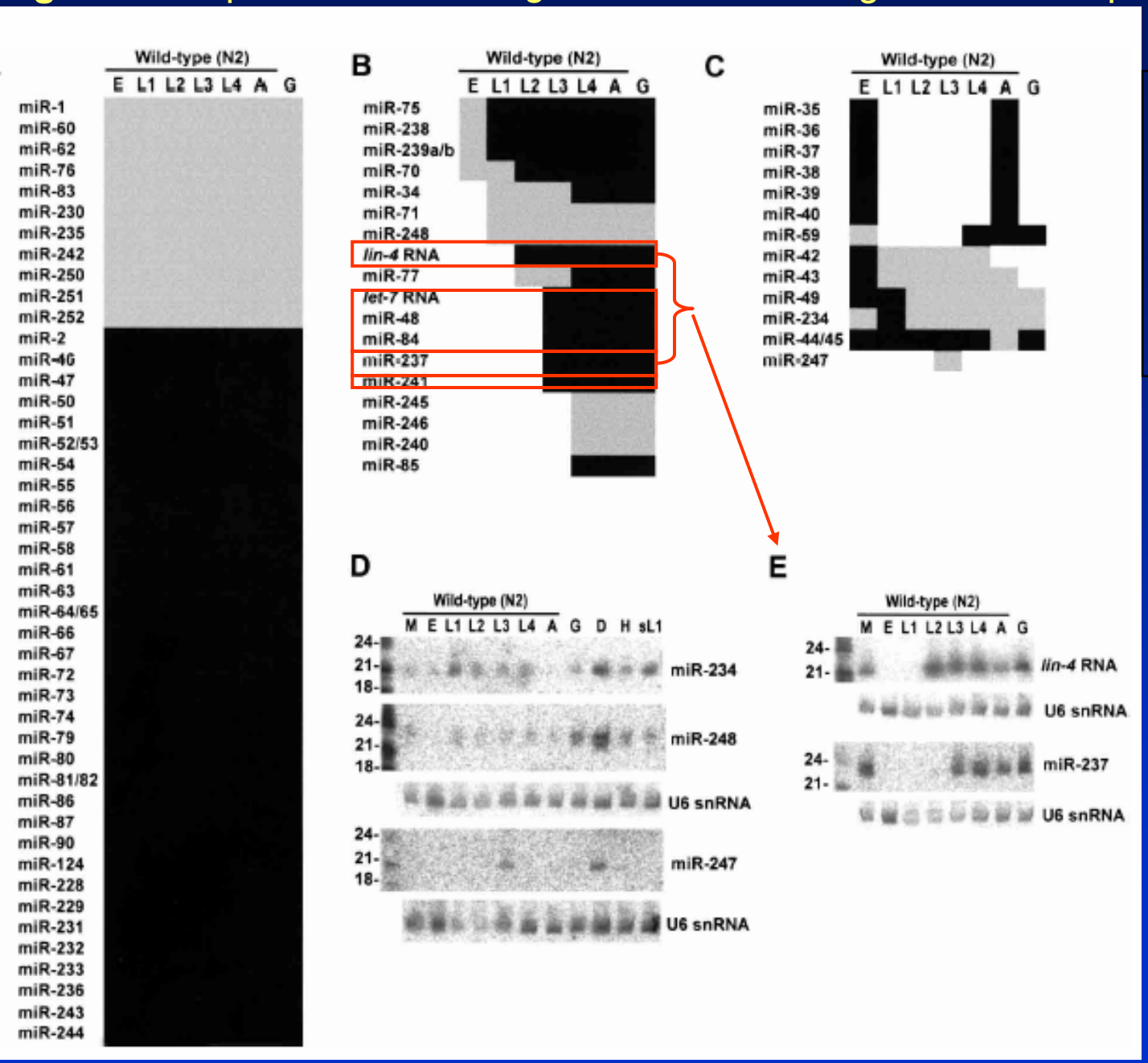


Figure 4. Expression of *C. elegans* miRNAs during larval development.



M = mixed stage N2
E = embryos
L1-4 = larval stages
A = adults
G = *glp-4(bn2)* adults
D = N2 dauers
H = *him-8(e1489)* mixe
sL1 = starved L1

Figure 5. Quantitative analysis of miRNA expression.

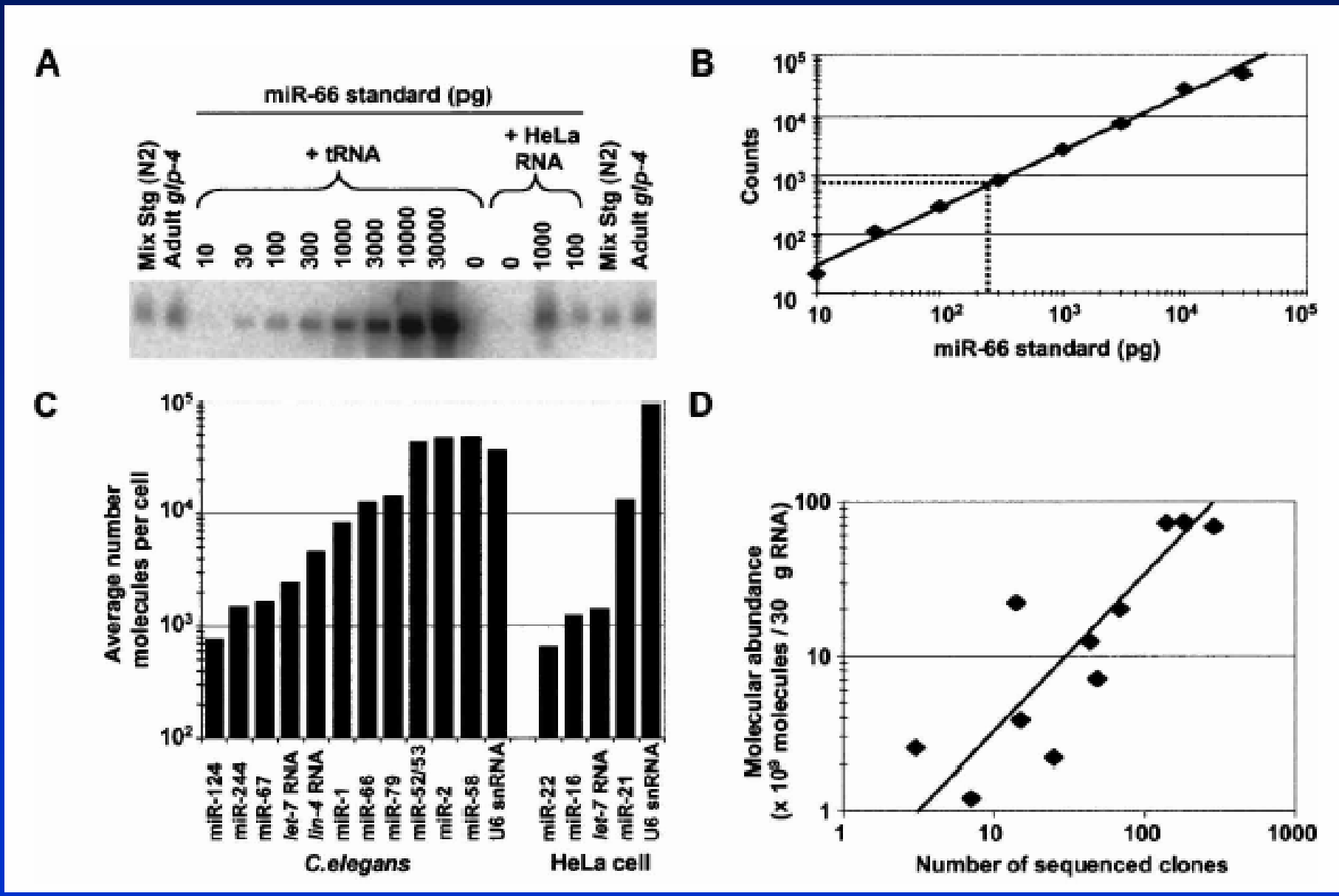
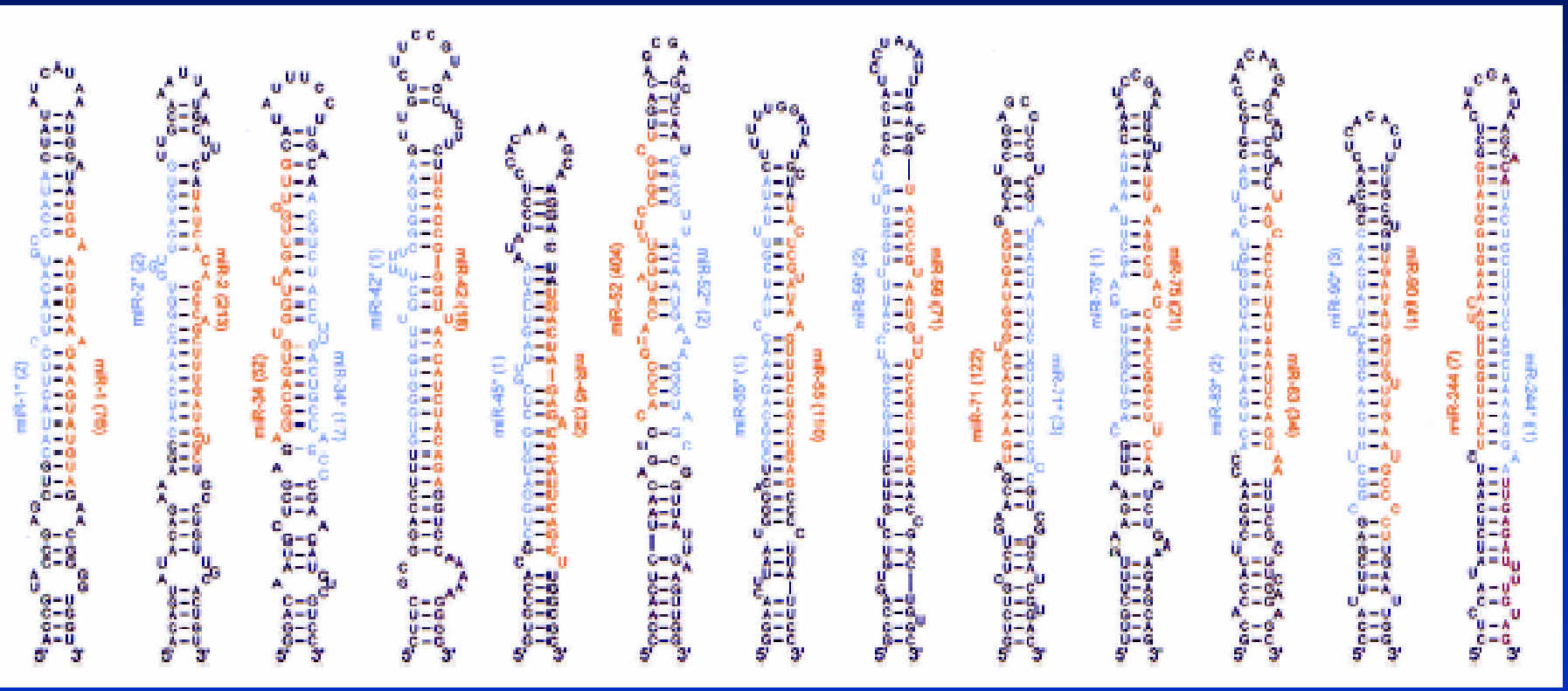


Figure 6. miRNA (red) and miRNA* (blue) sequences within the context of their predicted fold-back precursors.



- 3’ heterogeneity for some miRNA*s and most miRNAs
- **No** 5’ heterogeneity for miRNA*s; **very rare** (only one clone per species) for miRNAs

Conclusions

Upper bound of 120 miRNAs in *C. elegans*

- 64 loci have scores > the median for the 58 previously reported miRNAs
- 4 false positives (15 ambiguous)
- 2 X (64 – 4) = 120

Fig. 2b

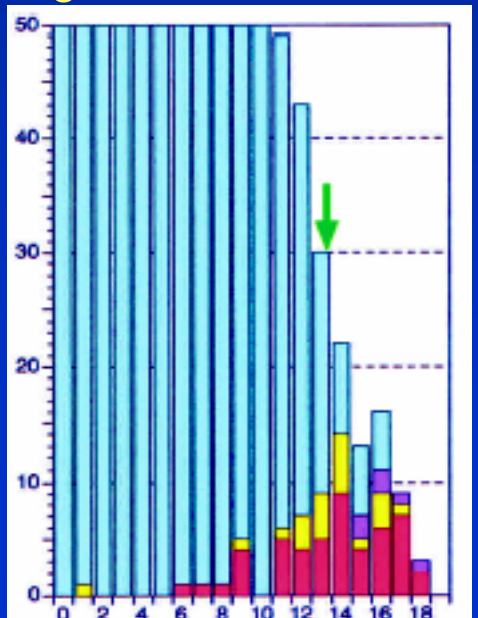
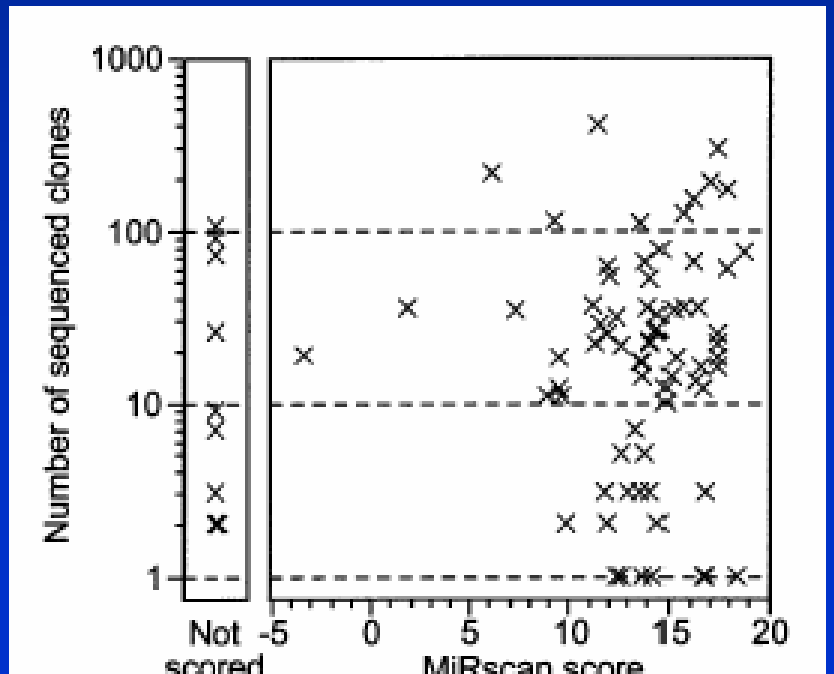


Figure 7



BREVIA

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 outside 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*).

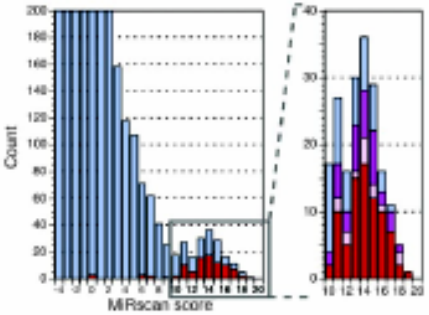


Fig. 1. Computational identification of vertebrate miRNA genes (*6*). The histogram represents the distribution of MiRscan scores for 15,133 human/*Fugu* consensus structures. Of the 109 reference-set loci, 91 were retained among these aligned segments (red), indicating that at least 80% of the human miRNAs are conserved in fish. The distribution peaks at the score of −4, with a count of 1188, 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.

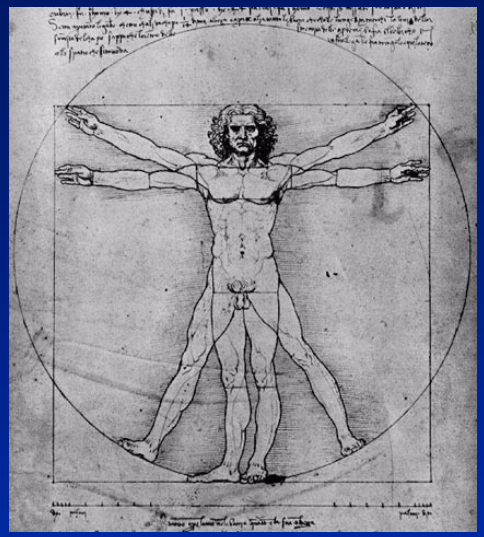
References and Notes

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, Curr. Biol. **12**, 689 (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. Wellbrock and H. Stiz 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 (CBB and DPB) and a grant from the David H. Koch Cancer Research Fund (DPB).

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

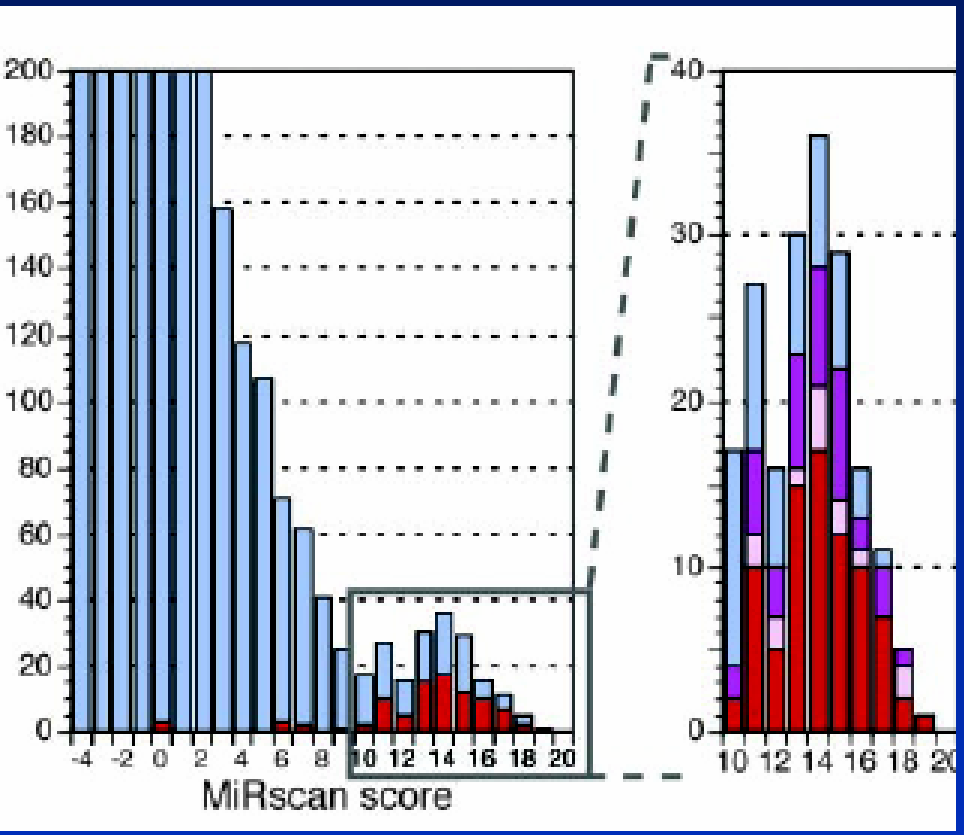
[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA.

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



~15,000 conserved human stem loops

Fig. 1. Computational identification of vertebrate miRNA genes.



- MiRscan identified 188 human loci
- 81 (red) of 109 known human miRNAs
- 14 (pink) paralogs of known miRNAs
- 38 (purple) found in zebrafish library
- 55 experimentally unverified

Upper bound of 255 miRNAs in human

81/109 = 0.74 sensitivity
188/.74 = 255 total

Considerations

- Pilot experiment detected no miRNA in *S. pombe*
- No evidence for Dicer (or Dicer-like proteins) in *S. cerevisae*
- Some miRNAs are known to regulate *C. elegans* development (likely plants too)

- **miRNA gene expansion may be linked to novel developmental patterning; evolution of multicellular body plans**

Limitations

Assay

Detection is limited to evolutionarily conserved miRNAs.

≠

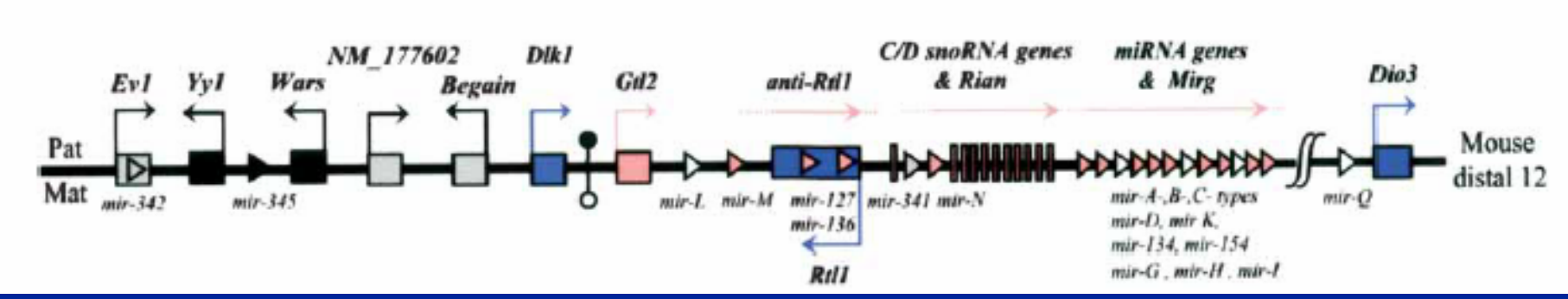
Species specific miRNAs not detected.

Conclusion/Future

- Estimation of total number of miRNA genes.
- miRNA expansion mediated body plan evolution.

A Large Imprinted microRNA Gene Cluster at the Mouse Dlk1-Gtl2 Domain

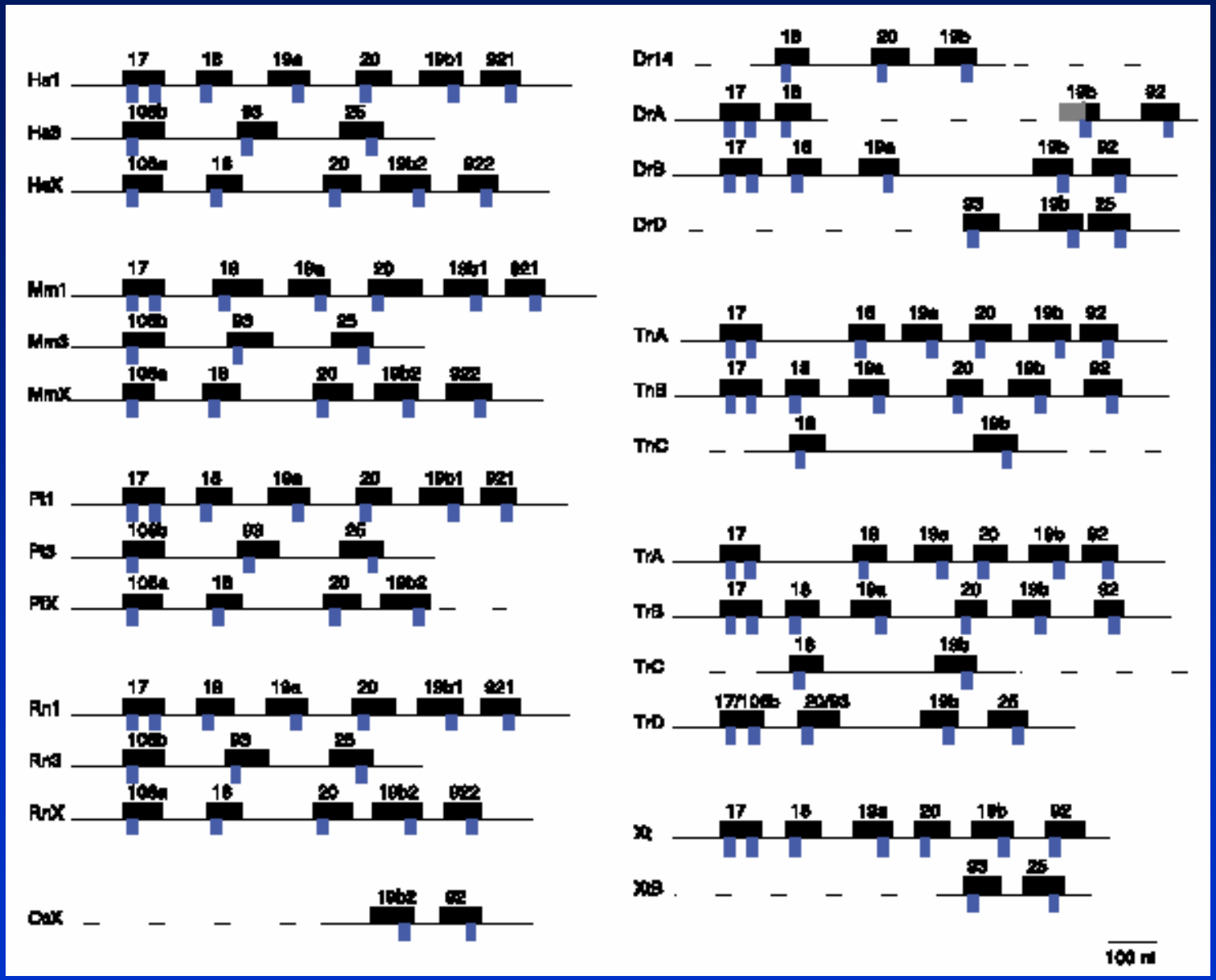
(Seitz *et al*. Genome Research. 2004)



Based on conservation criteria between human and fish *F. rubripes*, a recent study argued that the number of miRNAs in human genome should not exceed ~250, with ~40 remaining to be determined (Lim et al. 2003). Our study clearly shows that for mammalian systems, this number may be an underestimate."

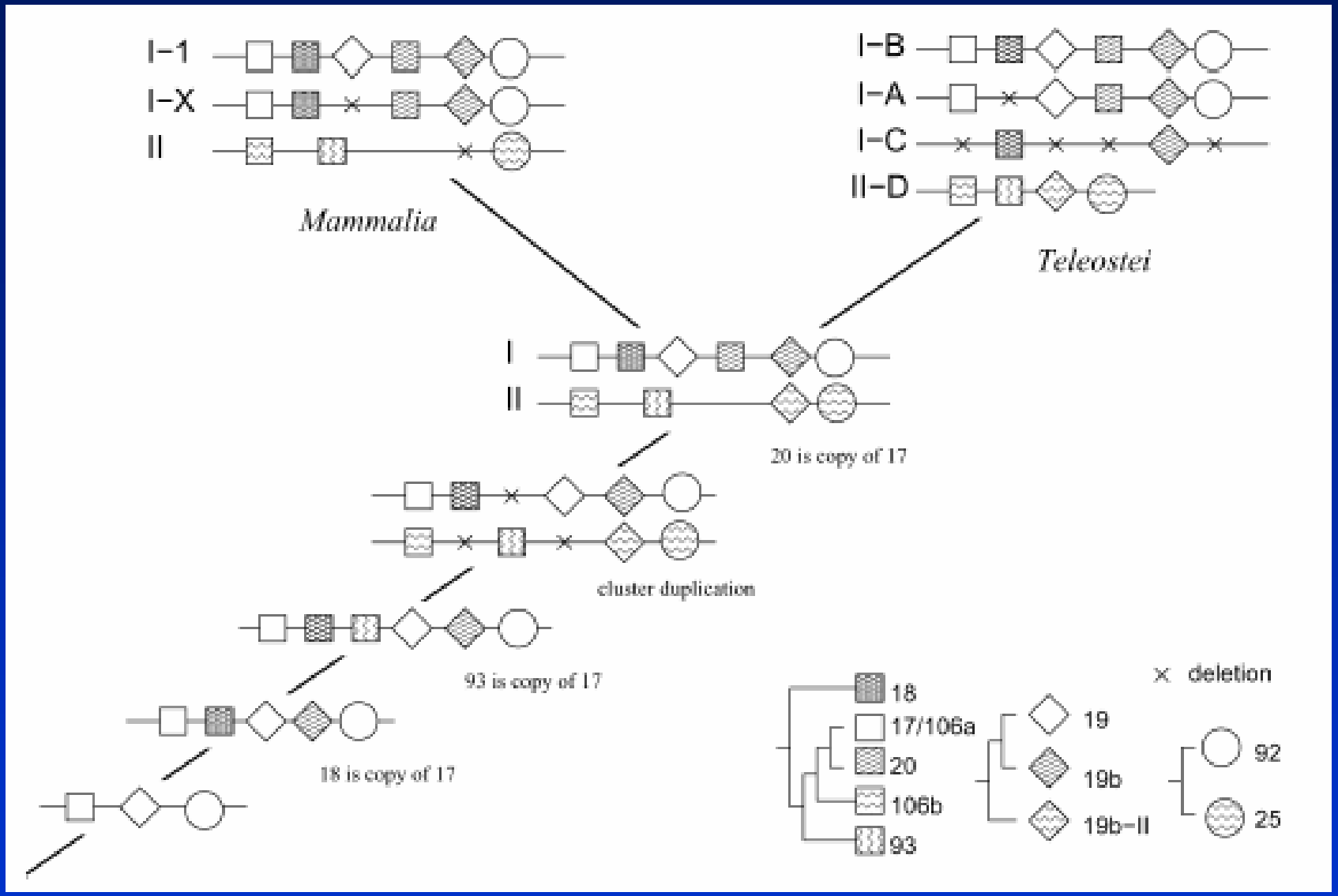
"Reinforcing this notion, **poorly conserved** embryonic stem-cell–specific miRNA genes, also organized in a tandem array, have been recently described and proposed to play a key role in the **regulation of early mammalian development** (Houbaviy et al. 2003)."

mir17 clusters from several species



(Tanzer and Stadler. Molecular Evolution of a MicroRNA Cluster. JMB 2004)

Evolution of the mir17 family



(Tanzer and Stadler. Molecular Evolution of a MicroRNA Cluster. JMB 2004)