Introduction to small RNA sequence analysis with applications in nematodes

Garry Wong, Ph.D.
RNA-seq Data Analysis Workshop
CSC – IT Center for Science - Espoo
9.1.2014

Today's Outline

- Introduction to *Caenorhabditis elegans and small noncoding RNAs*

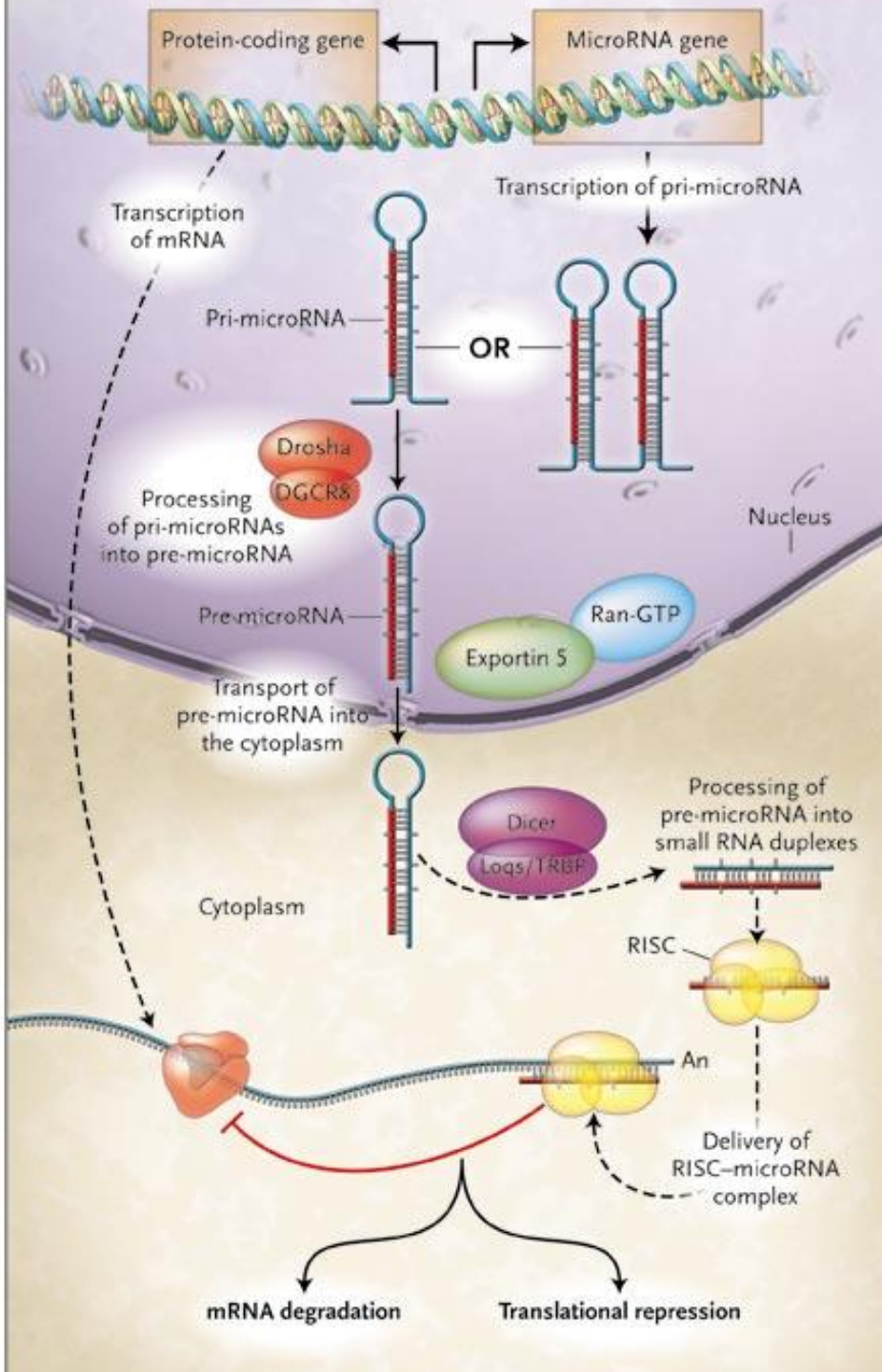
- *miRNA-seq analysis in Caenorhabditis elegans (differential analysis) and Panagrellus Redivivus (miRNA discovery)*

microRNAs (miRNAs)

MiRNAs are small , ~22nt, non-coding RNAs that act as posttranscriptional regulators of gene expression

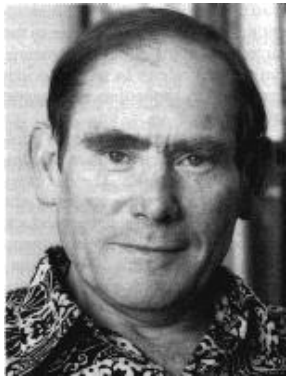
MiRNAs were discovered 1993 as small temporal RNAs that regulate developmental timing in *C. elegans (lin-4*).

Presently miRNAs are found in a wide range of organisms

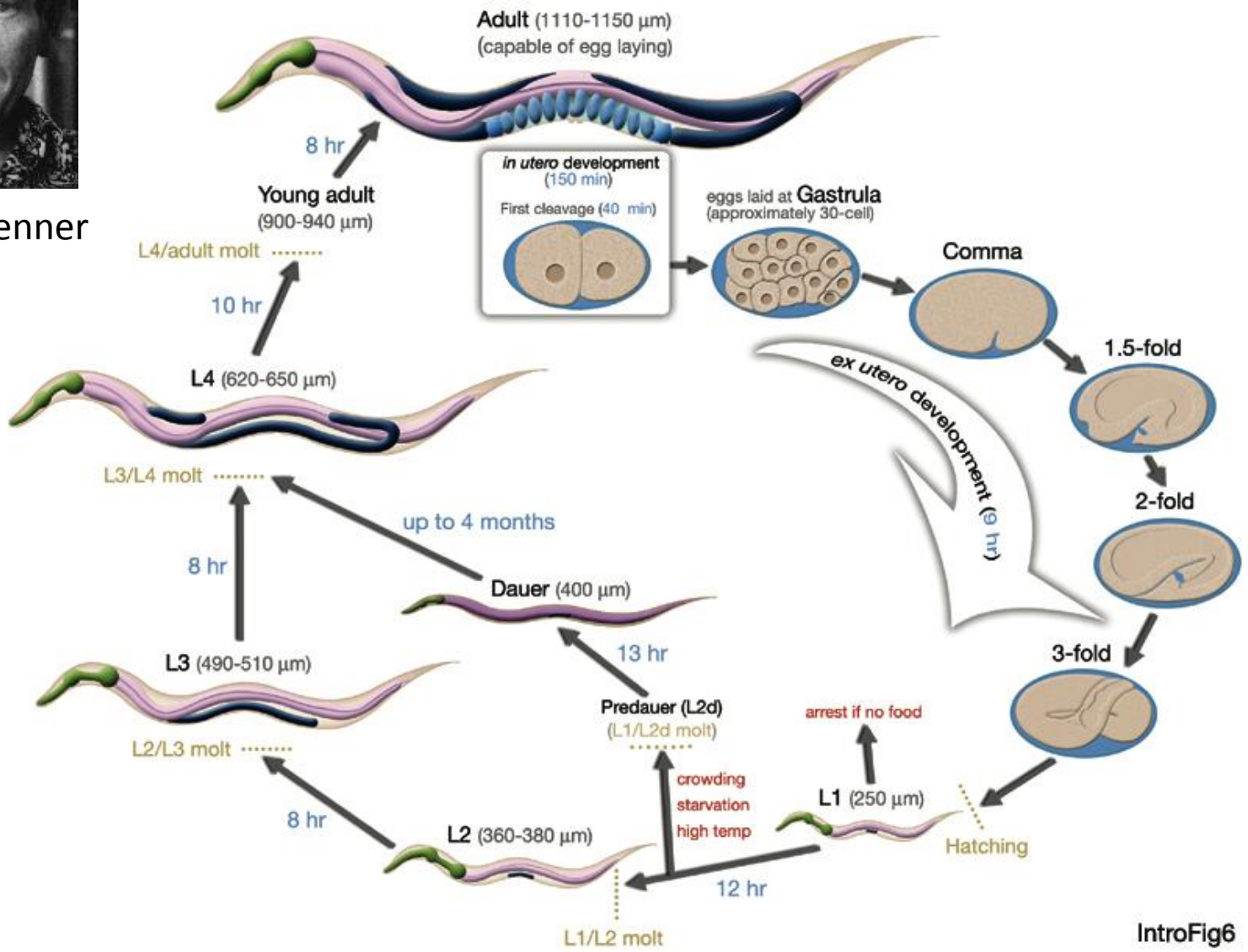


Chen, N Engl J med 2005

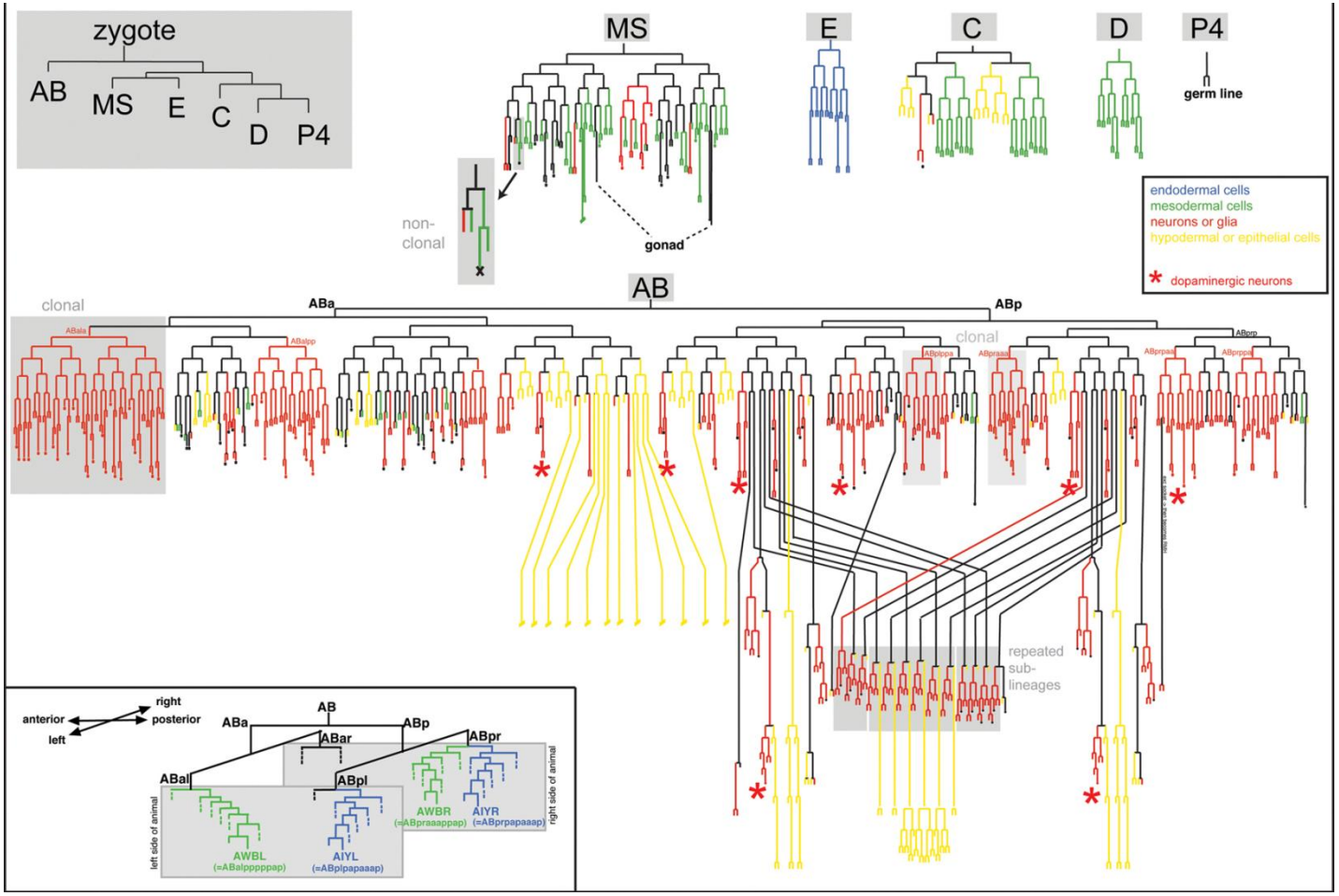
C. elegans Life Cycle



Sydney Brenner

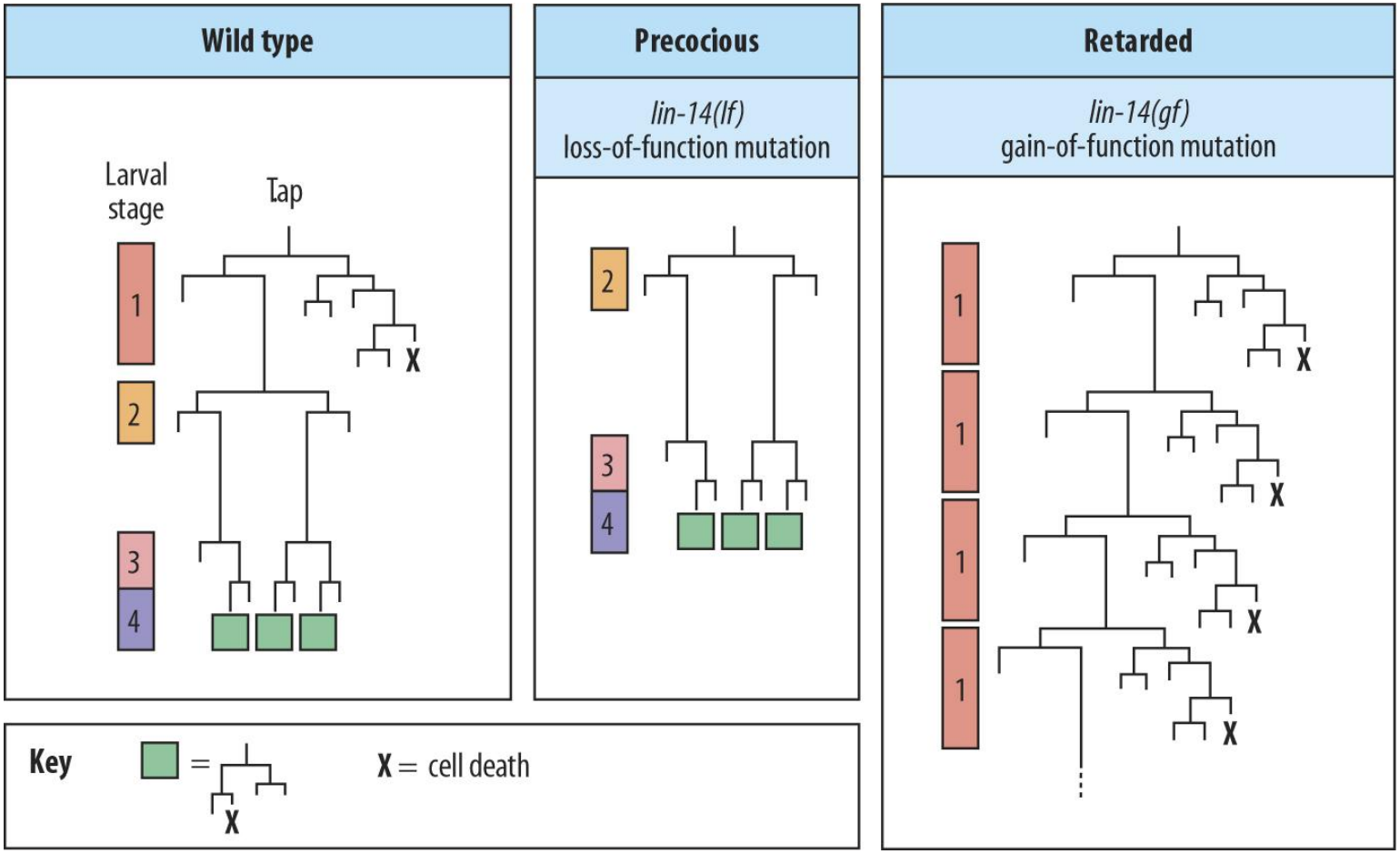


Neurogenesis in *C. elegans*: Cell lineages are invariant and known

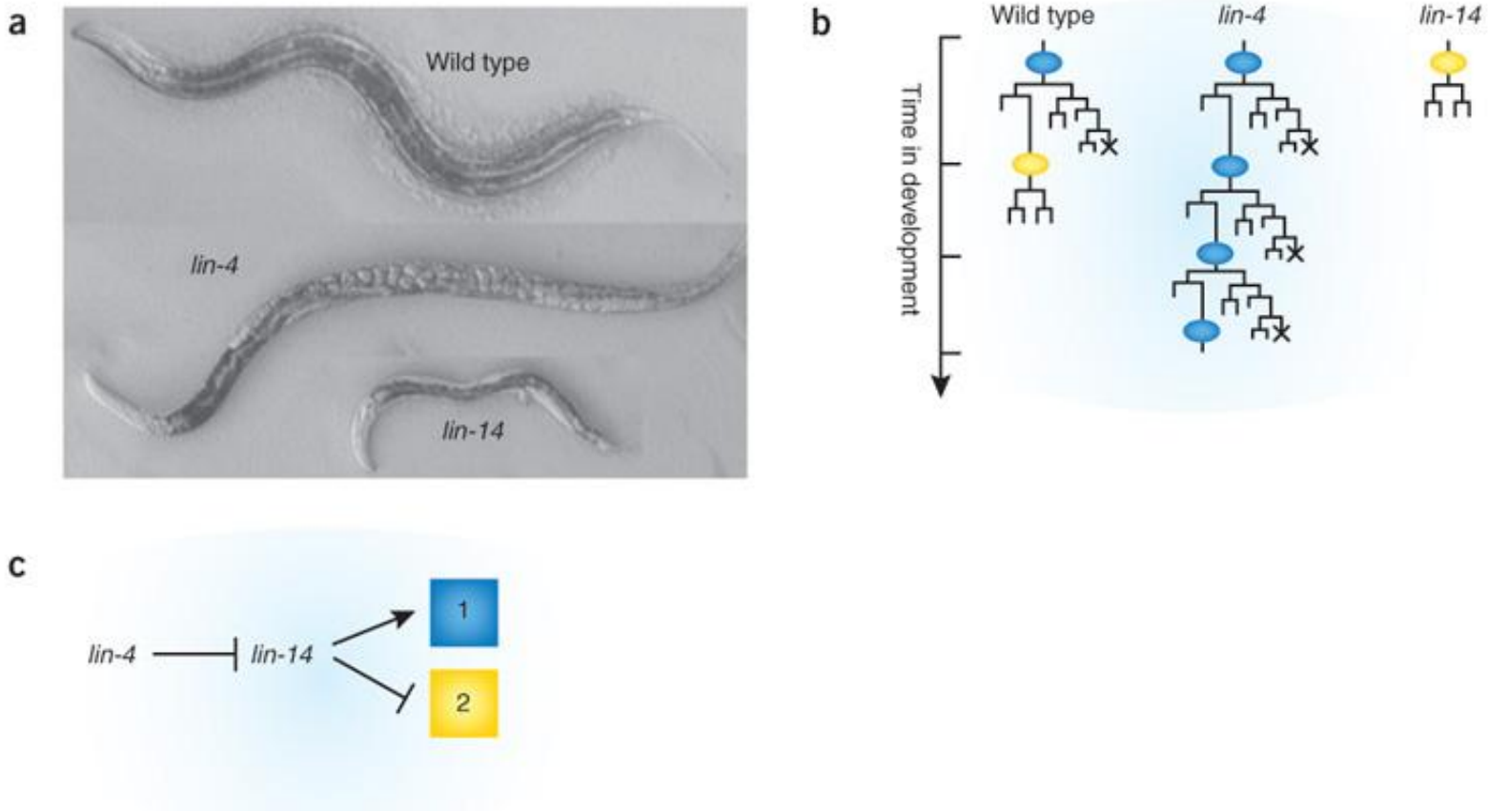


O. Hobert, Wormbook

Heterochronic/Developmental Timing mutants



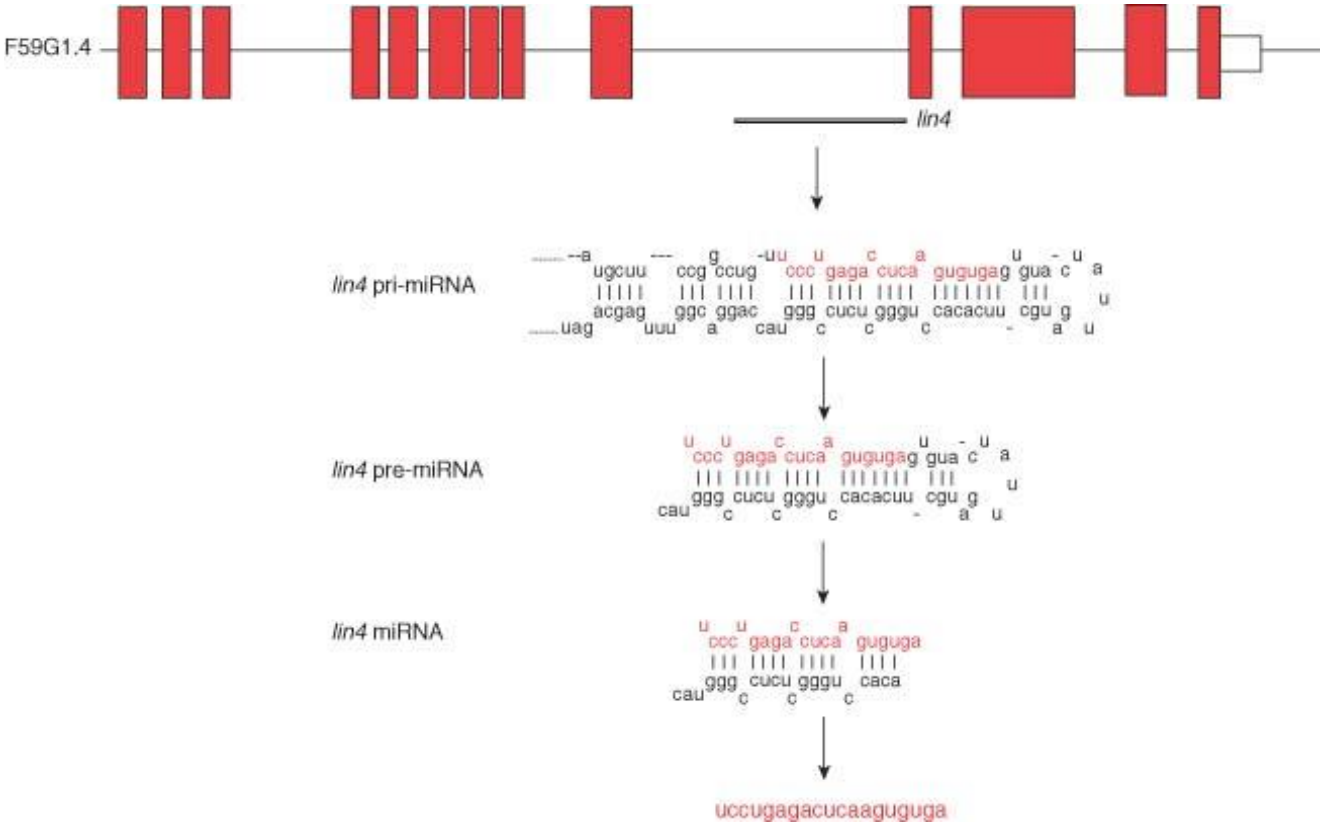
lin-4 animals lack adult structures and is epistatic to *lin-14*



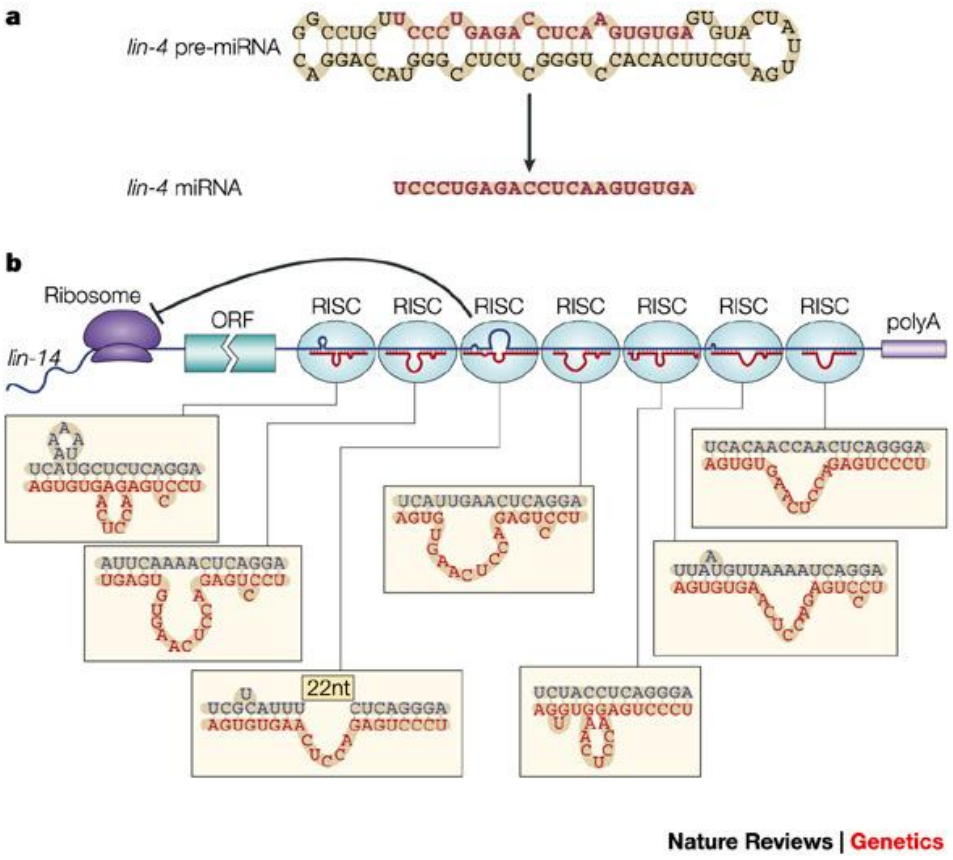
Victor Ambros
Nature Medicine 14, 1036 - 1040 (2008)

Victor Ambros

Discovery of miRNA



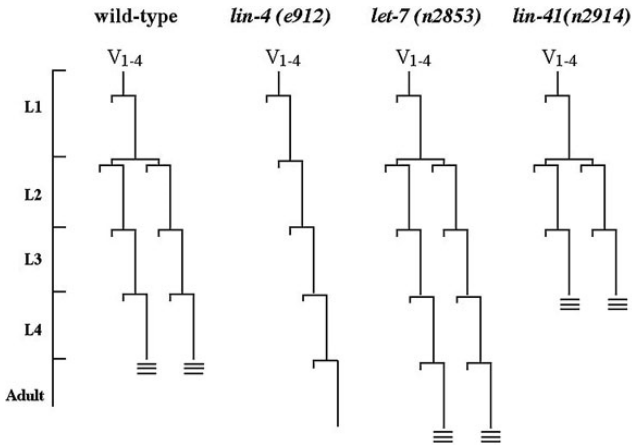
miRNAs target genes using 6-8nt seed sequences

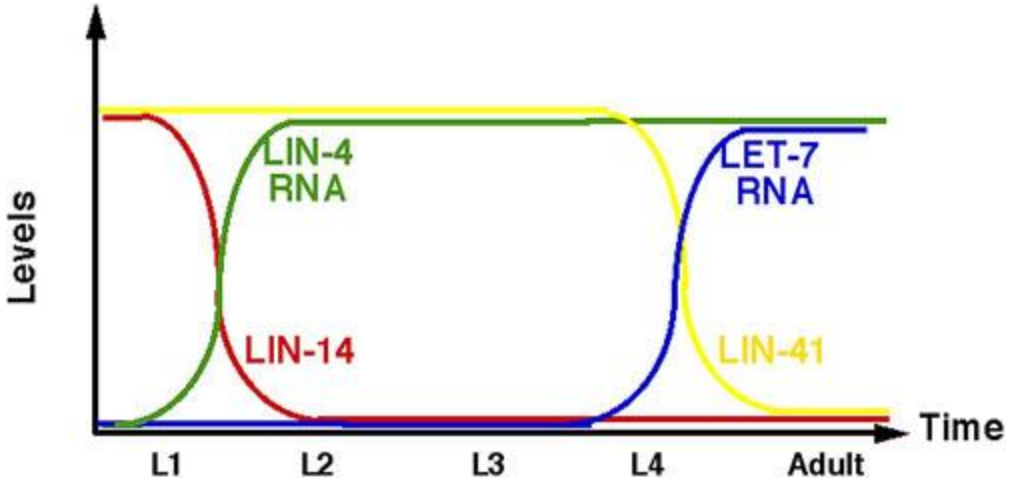


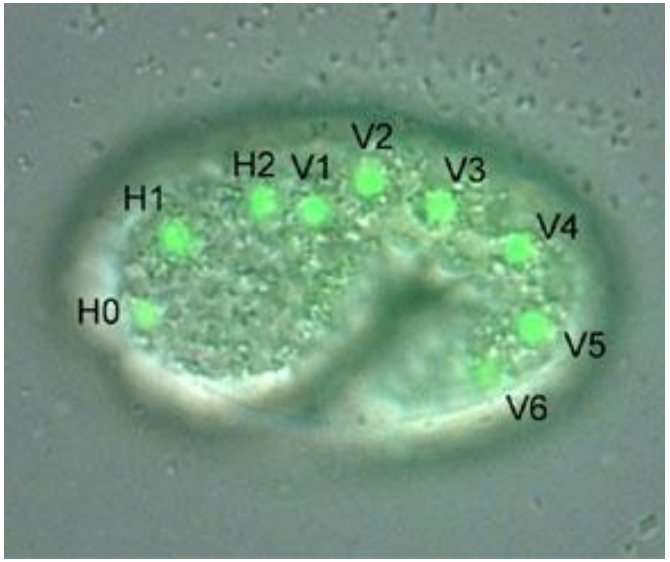
.

Nature Reviews Genetics 5, 522-531 (2004)

Multiple heterochronic genes identified by forward genetic screening



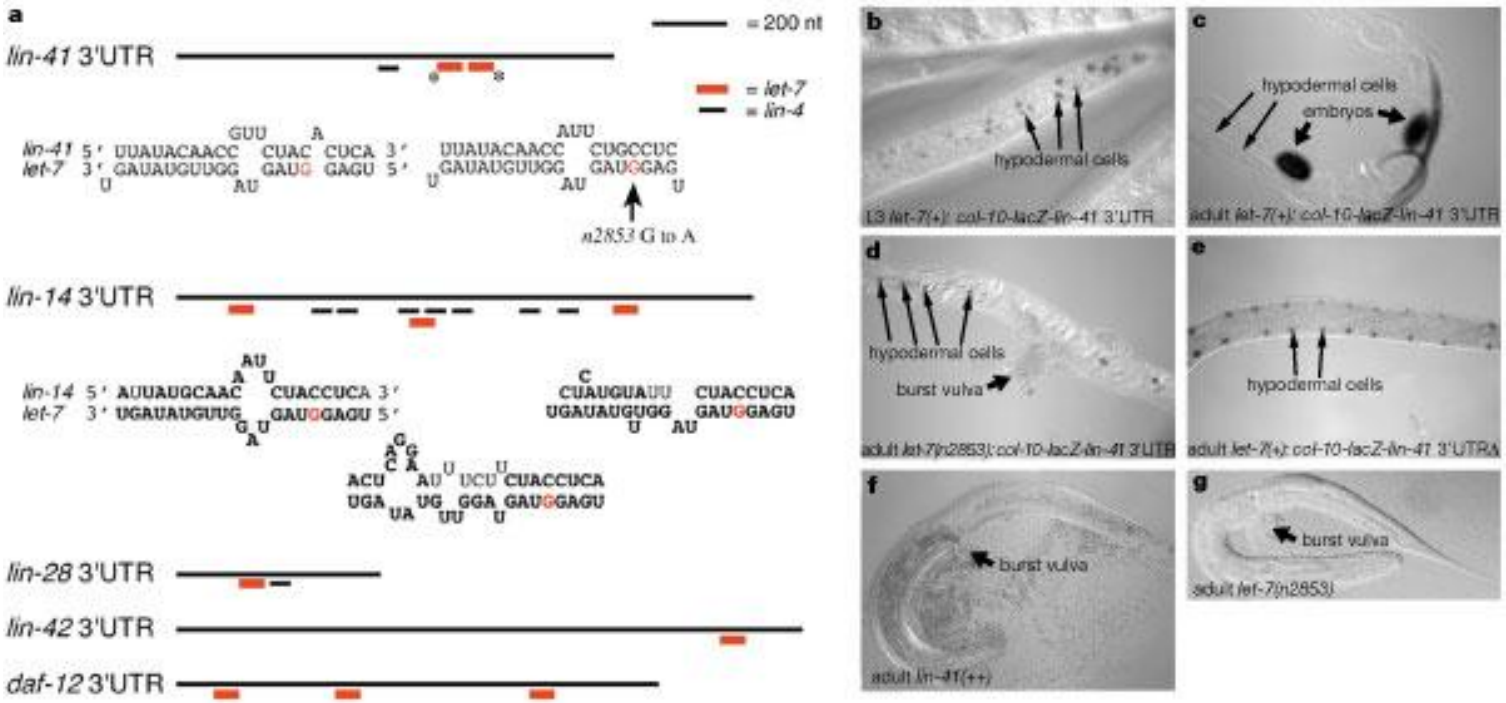




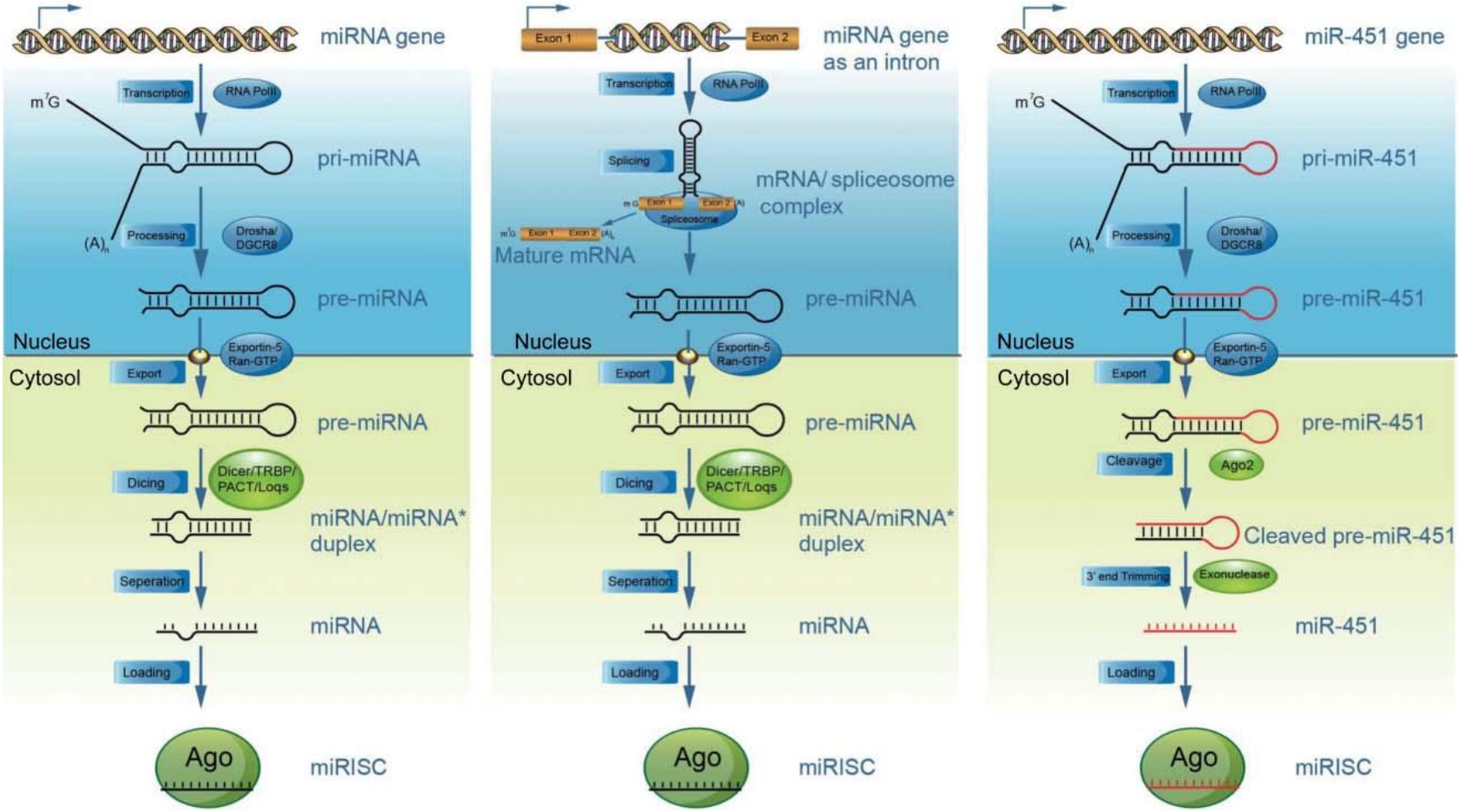
V. Ambros

Gary Ruvkun

let-7 targets *lin-41*



Three pathways for miRNA biogenesis



Huang and Lin, Wiley Interdisciplinary Reviews: Developmental Biology
Volume 1, Issue 1, pages 83–95, January/February 2012

TABLE 1 | MAIN CLASSES AND FUNCTIONS OF MAMMALIAN NON-CODING RNAS

ncRNA*	No. of known transcripts†	Transcript lengths (nucleotides; nt)‡	Functions
Precursors to short RNAs			
miRNA	1,756	>1,000	Precursors to short (21–23 nt) regulatory RNAs
snoRNA	1,521	>100	Precursors to short (60–300 nt) RNAs that help to chemically modify other RNAs
snRNA	1,944	1,000	Precursors to short (150 nt) RNAs that assist in RNA splicing
piRNA	89	Unknown	Precursors to short (25–33 nt) RNAs that repress retrotransposition of repeat elements
tRNA	497	>100	Precursors to short (73–93 nt) transfer RNAs
Long ncRNAs			
Antisense ncRNA	5,446	100–>1,000	Mostly unknown, but some are involved in gene regulation through RNA interference
Enhancer ncRNA (eRNA)§	>2,000	>1,000	Unknown
Enhancer ncRNA (meRNA)‖	Not fully documented	As variable as the length of mRNAs	Unknown, but they resemble alternative gene transcripts
Intergenic ncRNA	6,742	10^2–10^5	Mostly unknown, but some are involved in gene regulation
Pseudogene ncRNA	680	10^2–10^4	Mostly unknown, but some are involved in regulation of miRNA
3′ UTR ncRNA	12	>100	Unknown

Kowalczyk et al, Nature 482, 310–311 (16 February 2012)

21U-RNA

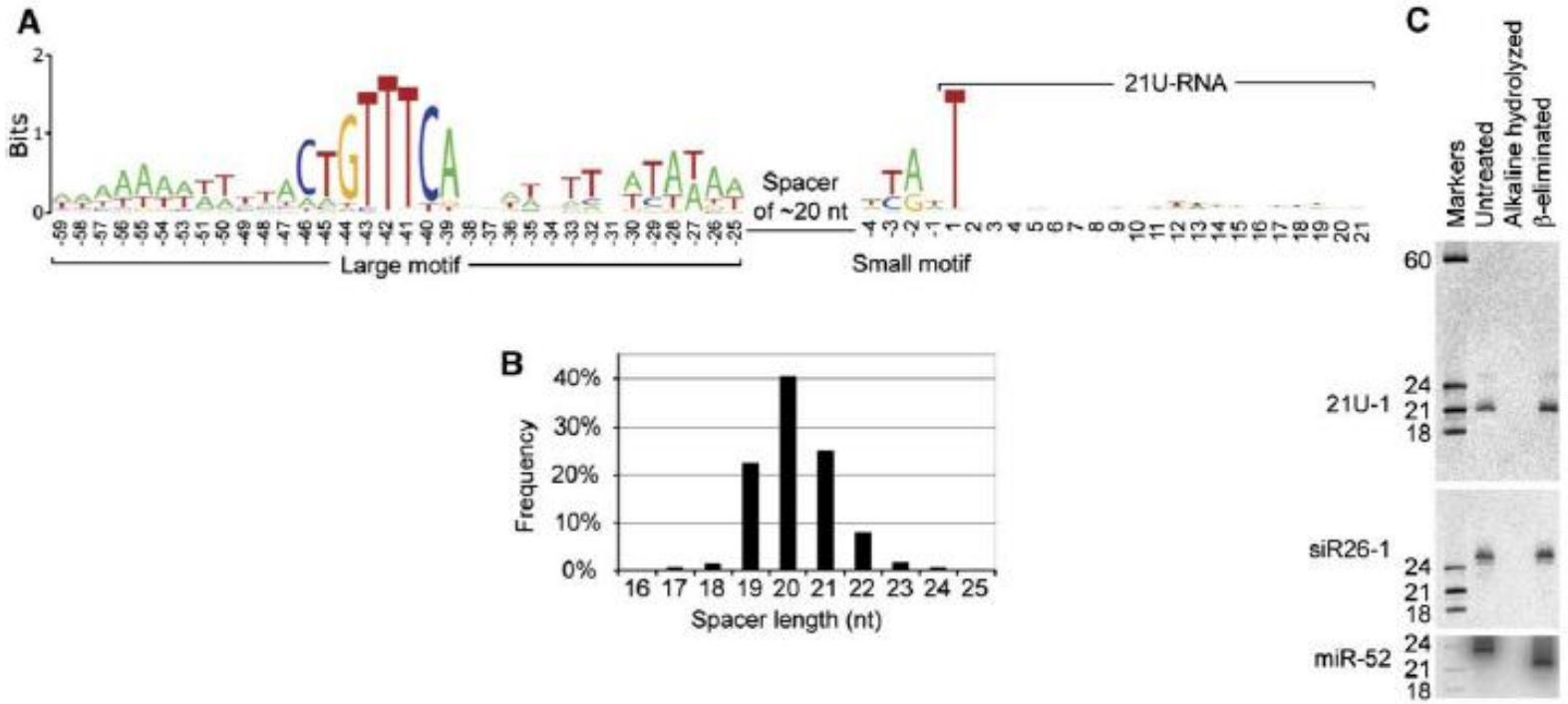


Figure 3. The 21U-RNA Sequence Motifs and Small RNA Chemical Reactivity

(A) The large and small motifs found upstream of 21U-RNA loci, depicted as a sequence motif (Crooks et al., 2004). The T at position 1 corresponds to the 5′ U of the 21U-RNA.

(B) The distribution of distances between the large and small motifs.

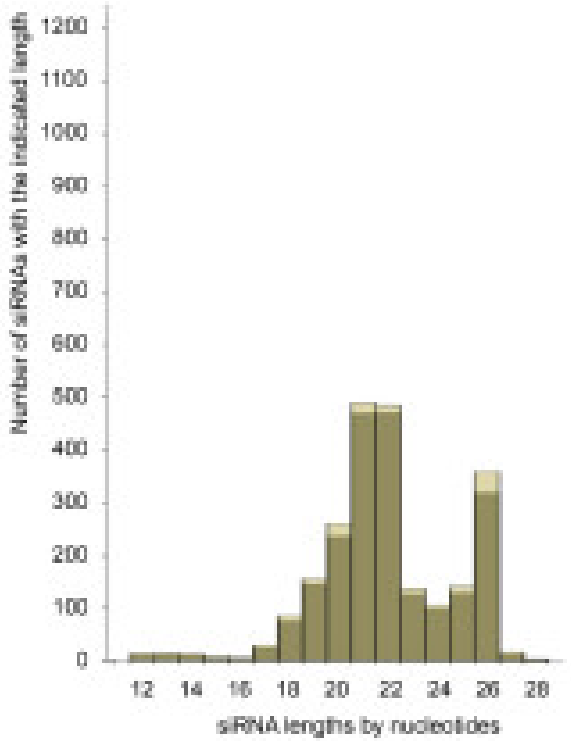
(C) Chemical reactivity of small RNAs. Total RNA (40 μg) was treated as indicated and analyzed by RNA blot, probing first for 21U-1, then stripping and reprobing for siR26-1, then miR-52.

Ruby et al., 2006

Endo siRNAs in *C. elegans* have specific sizes and structure

Figure 5.

siRNA length	Sequence logo	Number of siRNAs
12		16
13		20
14		15
15		10
16		11
17		50
18		111
19		225
20		408
21		784
22		1251
23		400
24		178
25		160
26		340
27		27
28		8
29		6



Asikainen et al., BMC Genomics, 2008

endo-siRNAs and their targets

BMC Genomics 2008, 9:270 http://www.biomedcentral.com/1471-2164/9/270

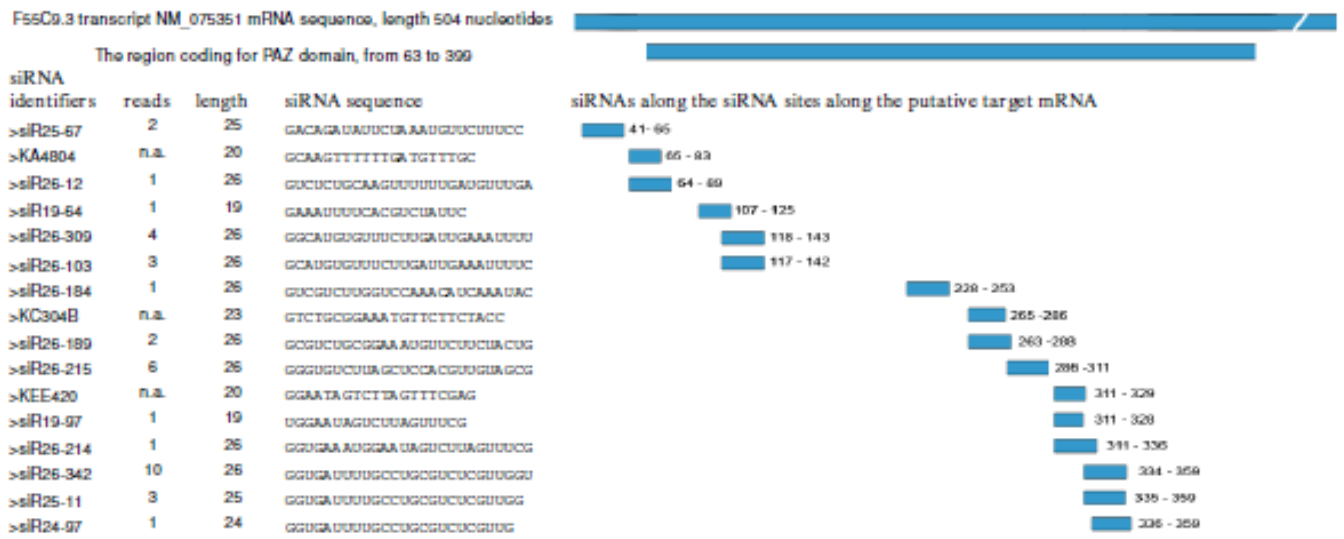


Figure 3
Length distribution and target sites of siRNAs complementary to Argonaute-related mRNA NM_075351 (F55C9.3). The siRNA identifier, number of reads, length, siRNA sequence, and position along the transcript is shown. The short bars indicate the position along the transcript and the numbers indicate the nucleotide position along the sequence. There are one to two nucleotide shorter versions of certain siRNAs that have been read multiple times. The number of reads was not available (n.a.) in some cases because those siRNAs were obtained from Lee et al. (2006).

Asikainen et al., 2008

3 distinct RNAi pathways

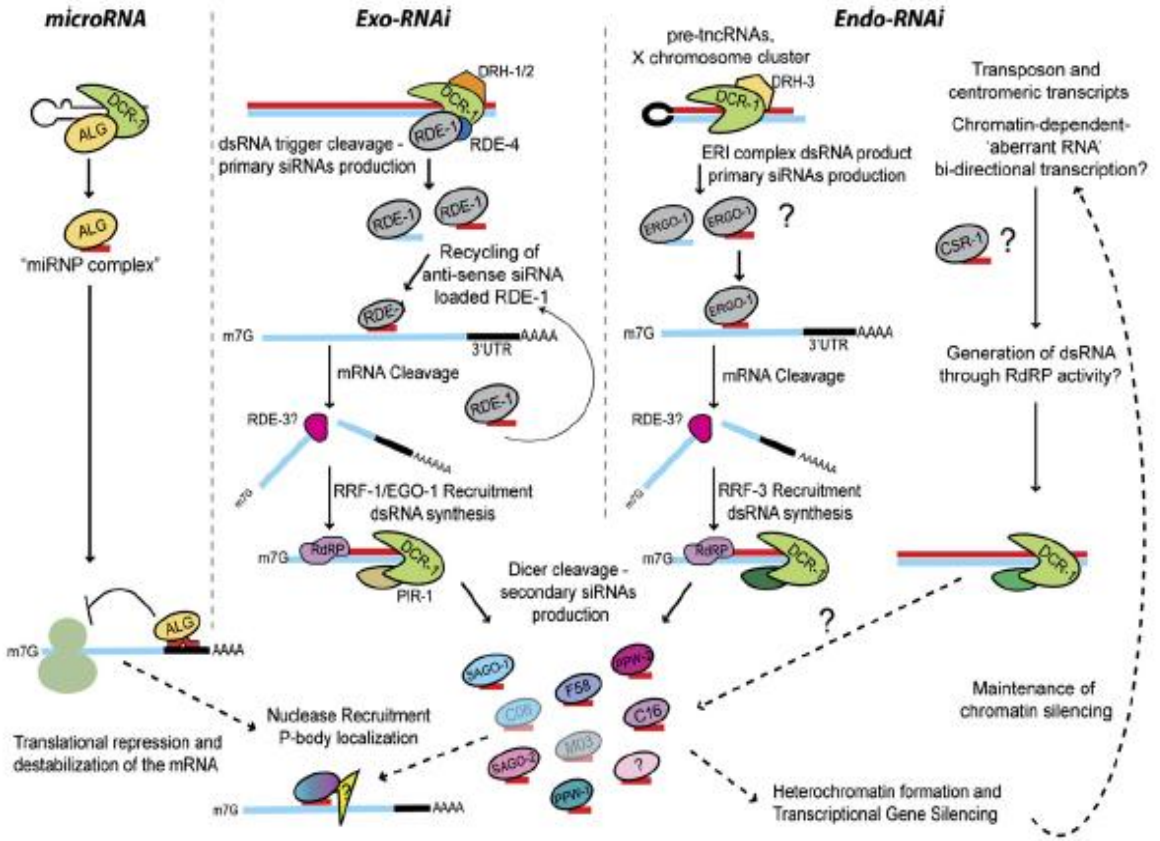


Figure 6. Model

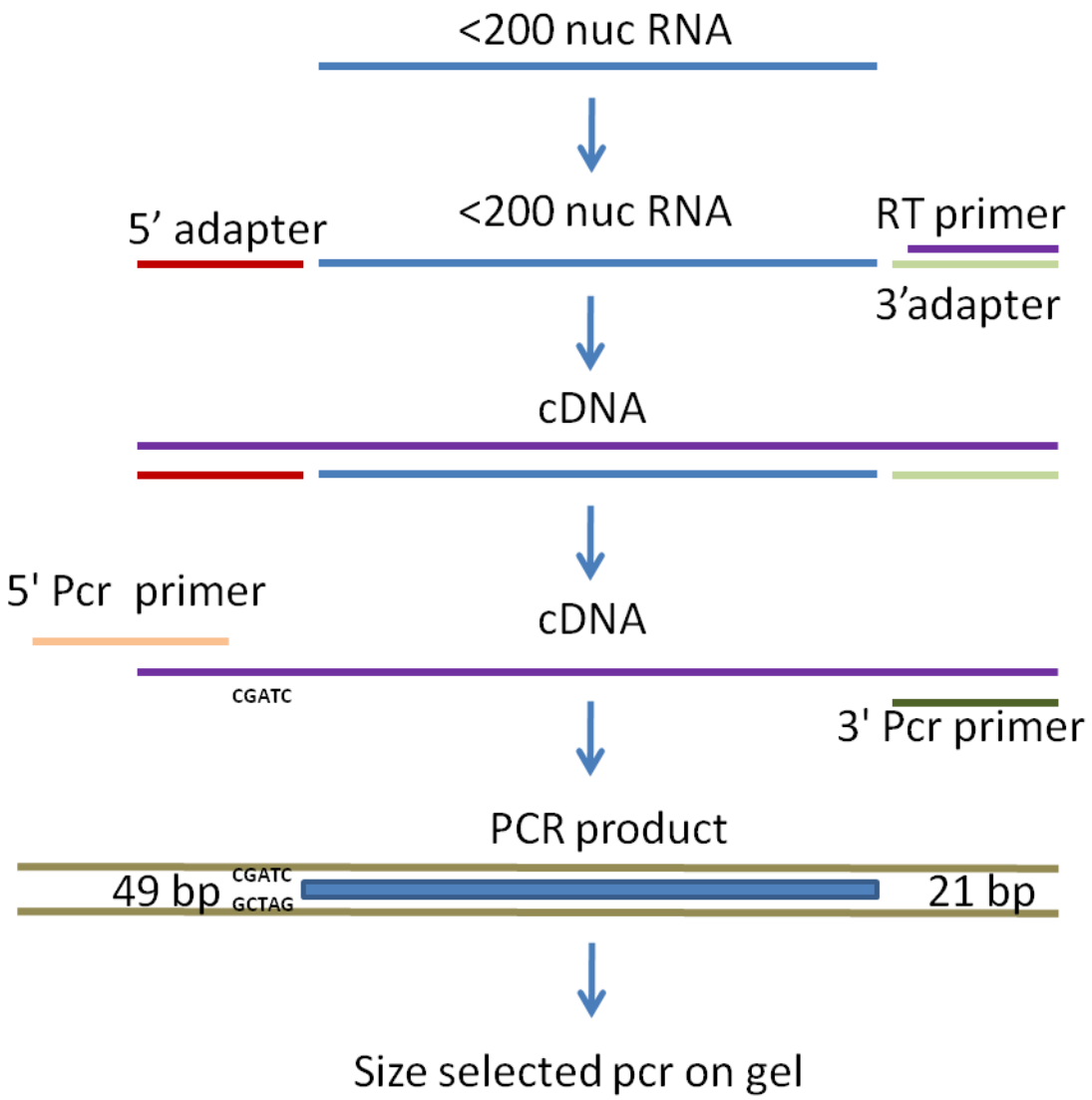
Schematic representations of RNAi-related pathways in *C. elegans*. Exo- and endo-RNAi pathways are proposed to involve sequential rounds of AGO action involving primary siRNA containing AGO complexes (gray ovals), and secondary siRNA containing AGO complexes (colored ovals). The miRNA pathway is proposed to involve a single AGO-mediated step. Distinct DCR-1 complexes are proposed to recognize the dsRNA substrates illustrated in the diagram. Evidence exists for several of these complexes, including the ALG, RDE-1, ERI, and PIR-1 containing DCR complexes (Tabara et al., 2002; Duchaine et al., 2006). After primary-siRNA-directed cleavage, a protein complex potentially containing RDE-3 (Chen et al., 2005, pink object) is proposed to mark the 3′ end of the 5′ cleavage product and to recruit RdRP. The question marks and dashed lines indicate speculative elements in the model.

Yigit et al., Cell 127, 747–757, November 17, 2006

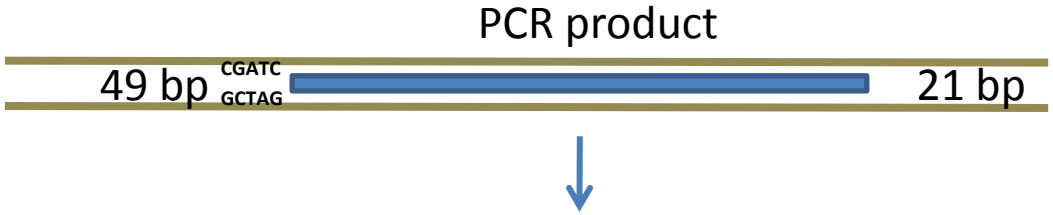
Summary of Part 1

- miRNAs represent a novel class of non-coding RNAs recently discovered serendipitously
- Many other classes of small noncoding RNAs exist
- Model organisms

Small RNA sample preparation



NEB small RNA sample prep

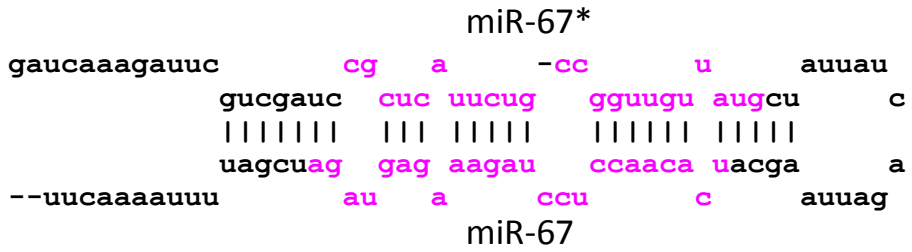


Size selected pcr products on gel

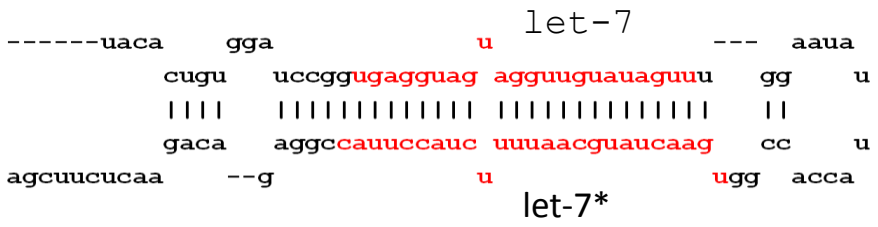
135-140 = tRNAs, pre miRNAs
100 bp = piwi RNAs
97 bp = 26G RNAs
91 bp = endo siRNAs, miRNAs
70 bp = primer dimers

Illumina Raw Data Output: FASTQ format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+HWUSI-EAS100R:6:73:941:1973#0/1
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```



cel-mir-67 hairpin precursor



cel-let-7 hairpin precursor

miRBase Release 18 Nov 2011

Nucleic Acids Research Advance Access published January 28, 2012

Nucleic Acids Research, 2012, 1–8
doi:10.1093/nar/gks043

Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis

Yue Li[1], Zhuo Zhang[2], Feng Liu[1], Wanwipa Vongsangnak[1], Qing Jing[2,3,*] and Bairong Shen[1,4,*]

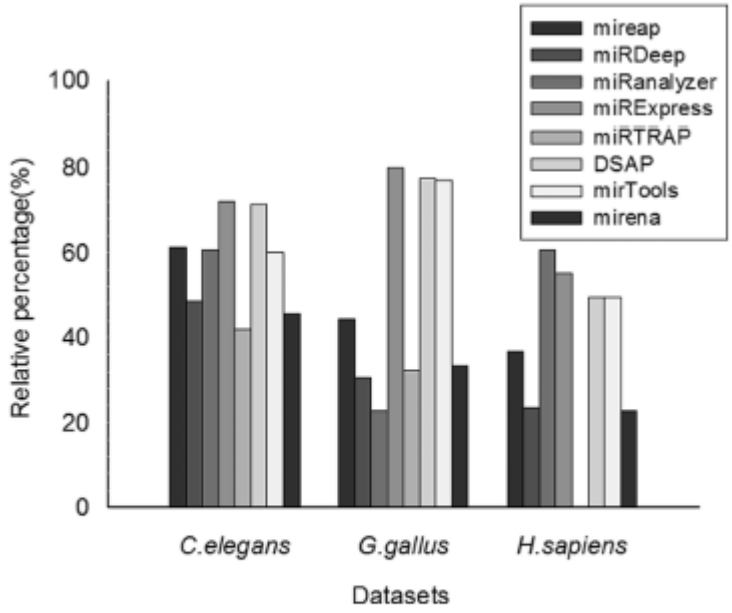


Figure 3. Comparison of the sensitivity of various software tools when predicting known miRNAs. Programs reported different numbers of miRNAs when run with their default or recommended settings using the same data sets. The percentage of predicted miRNAs in miRBase using different data sets is shown.

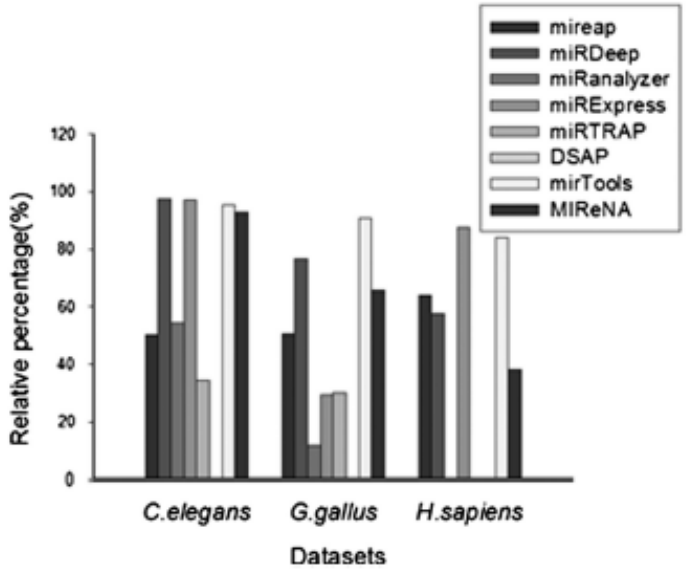


Figure 5. Comparison of the accuracy of various software tools when predicting known miRNAs. Programs reported different numbers of miRNAs when run with their default or recommended settings using the same data sets. The percentage of predicted miRNAs in miRBase is compared with the total number of predicted miRNAs with different data sets.

Methods

- miRAnalyzer - web-based, but data needs to reformatted to readcount or multifasta format, slow (Hackenberg et al., 2009)
- miRDeep2 – limited control, easy to implement, standard tool (Friedländer et al., 2012)
- Manual – complete control over the process, requires skill and effort to implement (Srinivasan et al., 2013)

READ-COUNT format:
ACTCTCGATCTATTT 57882
TCTCACGTGCGGTAAGC 23815
GTGATTGCATATCAT 432

MULTIFASTA format:
>gene1 57882
ACTCTCGATCTATTT
>gene2 23815
TCTCACGTGCGGTAAGC
>gene3 432
GTGATTGCATATCAT

Queing & Execution

Analysis completed

You can bookmark this page

Download all results in plain text here

Parameters

Species:	cel	Assembly:	ce6
Input:	mehg_ctrl.txt...	Mismatches (known):	1
Mismatches (libraries):	1	Mismatches (genome):	1
Score threshold:	0.9	Min. positives:	3
Type	Full analysis	Solid	no

Brief Summary

unique reads:	443720	read count:	18450502
filtered unique reads:	20056	filtered read count:	101762
No. known microRNA	274	No. known microRNA*	---
No. microRNA* (not miRBase)	---	No. new microRNAs	115
unique reads (after known)	395939	read count (after known):	1
unique reads (after lib)	393256	read count (after lib):	1
unique reads matched	277749	read count matched:	1
unique reads not-matched	115507	read count not-matched:	1477974

Mapping to known microRNA (miRBase 19)

Library/ Parameters	mature	ambiguous mature	mature-star	ambiguous mature-star	unobs. mature-star	ambiguos unobs. mature-star	hairpin	ambiguous hairpin
No. microRNA	274	7	0	0	0	0	121	3
fraction (number) of known microRNAs	74.7% (367)	---	0.0% (57)	---	---	---	54.3% (223)	---
unique reads	26882	70	0	0	0	0	768	7
fraction of unique reads	6.3%	0.017%	0.000%	0.000%	0.000%	0.000%	0.181%	0.002%
read count	2371520	488	0	0	0	0	2639	9
fraction of read count	12.9%	0.003%	0.000%	0.000%	0.000%	0.000%	0.014%	0.000%
links to detail pages	details	details	no results	no results	no results	no results	details	details

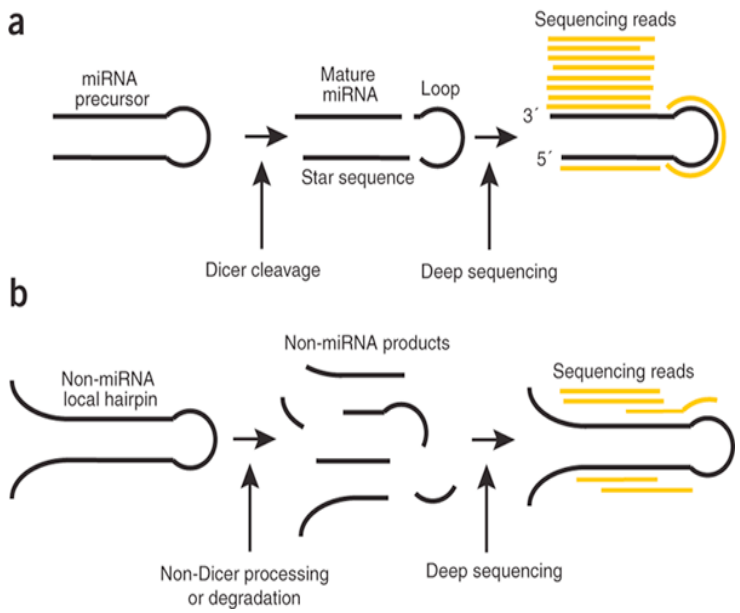
Alignment to other transcibed entities

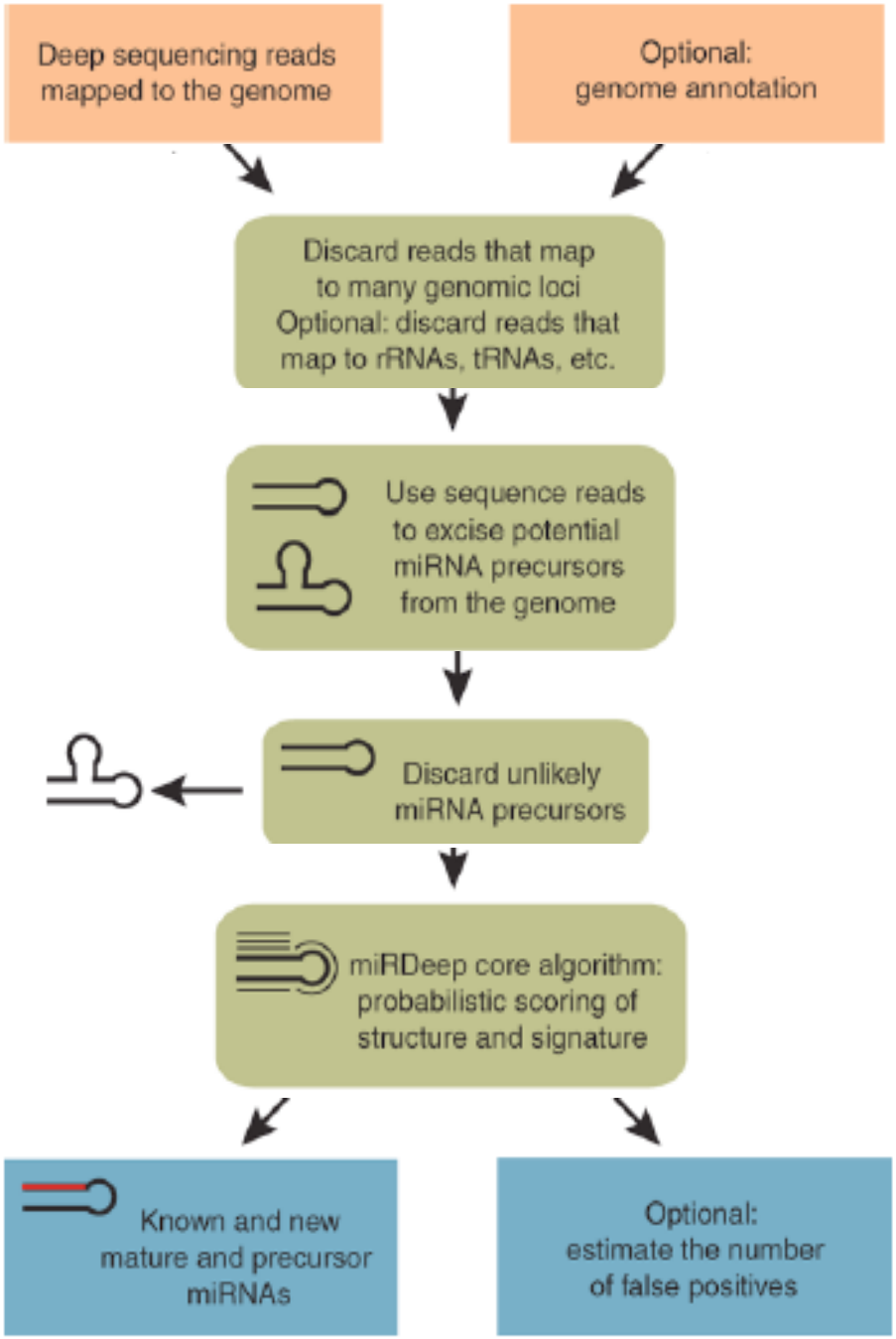
Library/ Parameters	RefSeq_genes	Rfam
number of unique reads	58333	1196
fraction of unique reads	13.77%	0.28%
number of reads	1	3711
fraction of reads	67.55%	0.02%
Links	details	details

Predicted candidate microRNAs

No. of read clusters:	173783			
No. of checked candidates:	46844			
No. new microRNAs :	115	Unique reads (read count):	629 (4440)	details
No. new microRNAs (trans filtered):	109	Unique reads (read count):	623 (4434)	details

miRDeep2



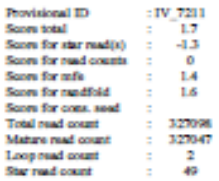


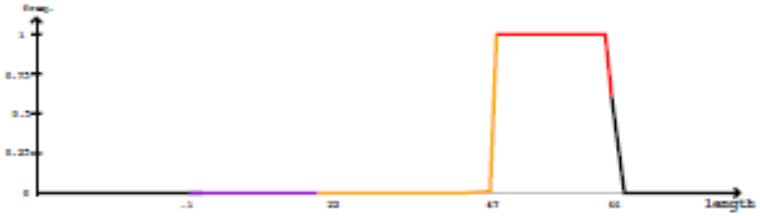
Friedländer et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *NAR*, **September 12, 2011**

miRDeep2 output

data set	total read count	mature read count	loop read count	star read count	mature miRBase miRNA	consensus mature sequence	consensus star sequence
mehg_ctrl	38150	38130	0	20	cel-let-7-5p_MIMAT0000001_Caenorhabditis_elegans_let-7-5p	gguaguagguuguauaguu	ugcaauuuucuaccuuacc
mehg_ctrl	2616	2614	0	2	cel-lin-4-5p_MIMAT0000002_Caenorhabditis_elegans_lin-4-5p	ccugagaccucaaguguga	accugggcucuccggguacca
mehg_ctrl	62	62	0	0	cel-lsy-6_MIMAT0000749_Caenorhabditis_elegans_lsy-6	uguaugagacgcauuucga	caaaugcgucuaguaucaaa
mehg_ctrl	114605	114591	0	14	cel-miR-1-5p_MIMAT0020301_Caenorhabditis_elegans_miR-1-5p	aauguaaagaaguaugua	cuuccuuacaugcccaua
mehg_ctrl	15	15	0	0	cel-miR-1018_MIMAT0005031_Caenorhabditis_elegans_miR-1018	gagaucauuggacuuacag	guaaguucaugauuucuc
mehg_ctrl	208	208	0	0	cel-miR-1022-5p_MIMAT0005523_Caenorhabditis_elegans_miR-1022-5p	ucauuguuaggacgccauc	ugauaguccaaugaugauc
mehg_ctrl	94	93	0	1	cel-miR-124-5p_MIMAT0015111_Caenorhabditis_elegans_miR-124-5p	aggcacgcggugaaugcca	cacccuagugacuuuagu
mehg_ctrl	29	29	0	0	cel-miR-1817_MIMAT0006584_Caenorhabditis_elegans_miR-1817	ccaaugucuucucuaucaug	ugauagugaaauauuugcu
mehg_ctrl	8	8	0	0	cel-miR-1819-5p_MIMAT0020358_Caenorhabditis_elegans_miR-1819-5p	aaugauugagcuugauggau	caaucaugcucaaaacauucg
mehg_ctrl	1011	940	0	71	cel-miR-1820-5p_MIMAT0006587_Caenorhabditis_elegans_miR-1820-5p	ugauuguuuuucgaugauguuc	uuguaaacaaucaaagaa
mehg_ctrl	30	30	0	0	cel-miR-1821_MIMAT0006588_Caenorhabditis_elegans_miR-1821	ggucuuauaguuagguaga	ugcccaacuugcagacuuu

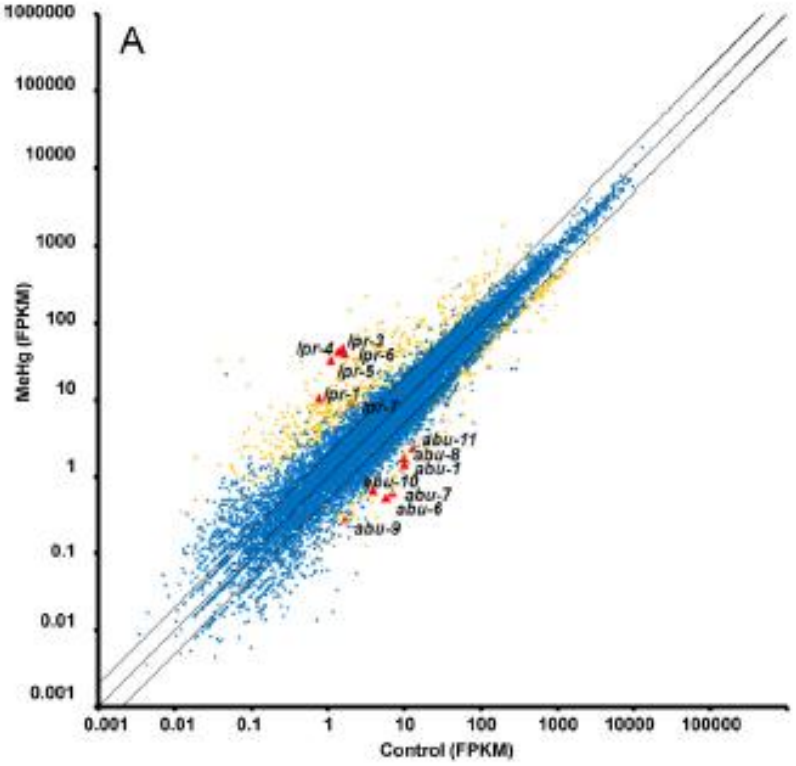
Linked output from mirDeep2

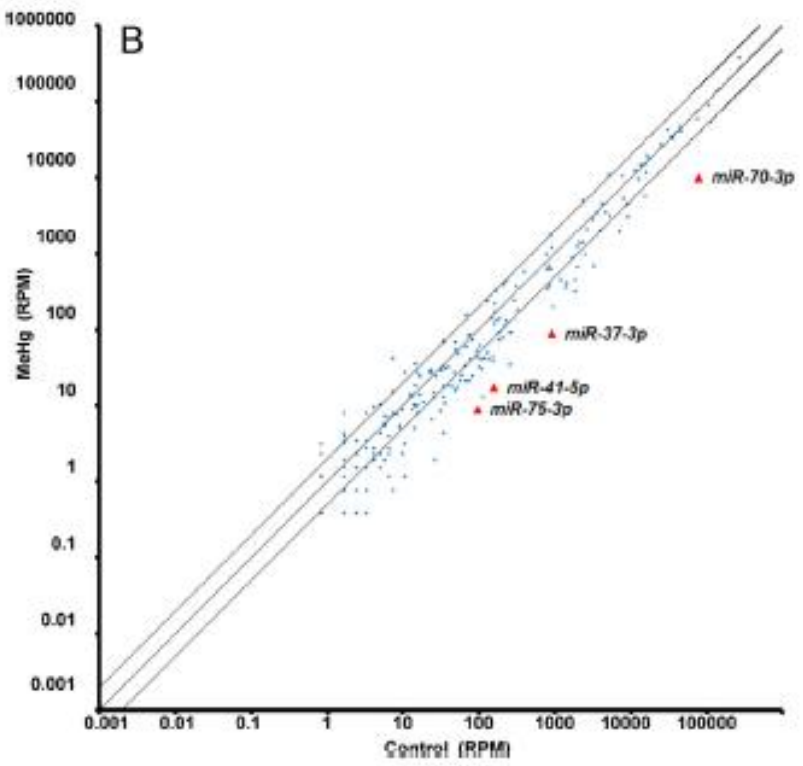




mRNA and miRNA differential expression in MeHg treated *C. elegans*

NEED FOR DATA INTEGRATION

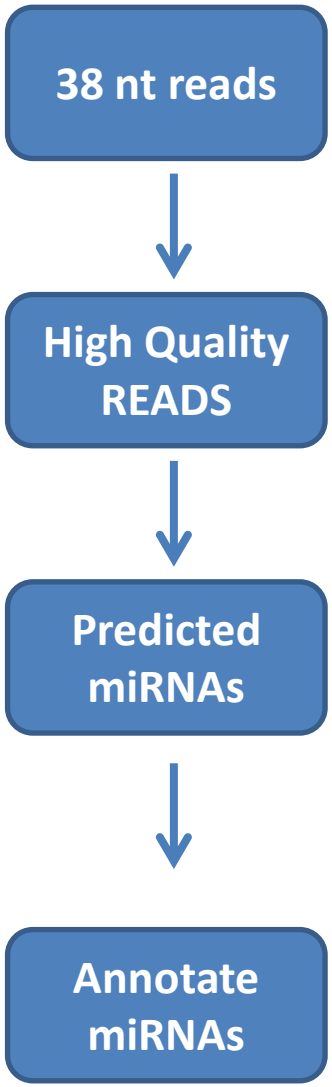




Rudgalvyte et al., 2013

miRNA discovery in Panagrellus redivivus

miR-seq Bioinformatics Workflow



Trim adapters

Filter/Quality Step

1. Remove E. Coli sequences
2. Remove tRNA sequences
3. Remove primer-dimers
4. Stratify for length

Analysis Step

5. Map to Genome or Contigs
6. Identify Hairpins
7. Calculate Reads per Hairpin
8. Identify miRNA*, miRNA clusters

Annotation Step

9. Determine MFE Structure
10. Identify orthologs in miRBase

Conserved sequence
- seed
- mature miR
- hairpin precursor

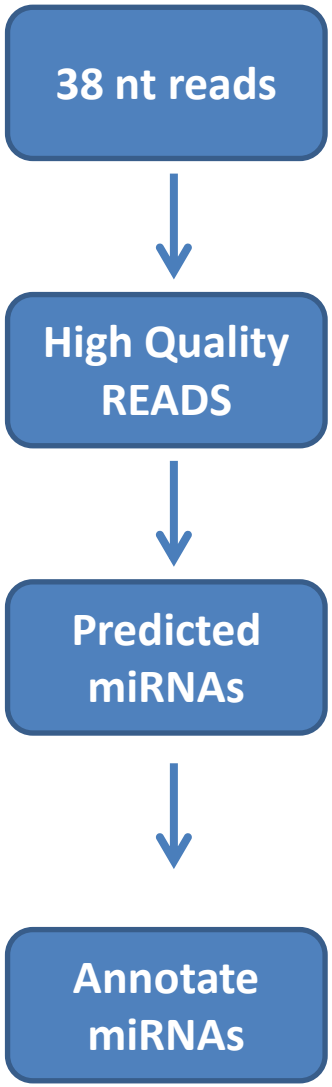
miR-seq Bioinformatics Workflow

Raw data: 24161842 reads

Cleaned data:
23981346 reads (99.3%)
10-28 nt reads: 749677
29-38 nt reads: 23231669

# Reads Mapped (no mismtches):	
1 location	659159 (88%)
2-10 locations	14505 (2%)
>10 locations	450

Predicted miRNAs: **224**
MFE form: hairpin
>10 reads
63 (25.4%) conserved
(6 worms, fruit fly and human)



Trim adapters

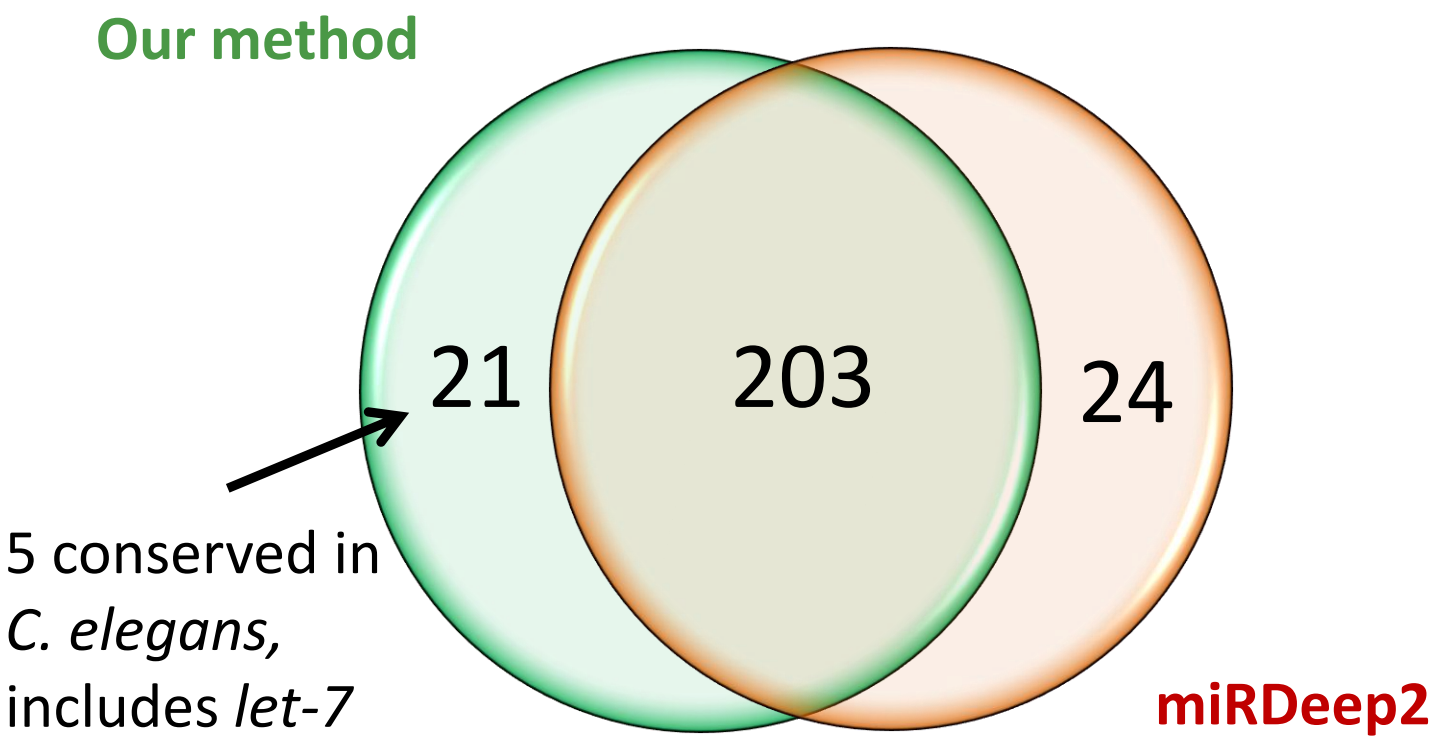
Filter/Quality Step
1. Remove E. Coli sequences
2. Remove tRNA sequences
3. Remove primer-dimers
4. Stratify for length

Analysis Step
5. Map to Genome or Contigs
6. Identify Hairpins
7. Calculate Reads per Hairpin
8. Identify miRNA*, miRNA clusters

Annotation Step
9. Determine MFE Structure
10. Identify orthologs in miRBase

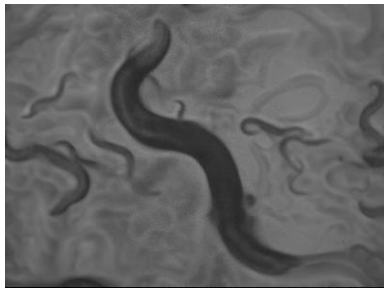
Conserved sequence
- seed
- mature miR
- hairpin precursor

203 miRNAs were predicted with both methods



203 + 21 + 24 = 248 miRNAs

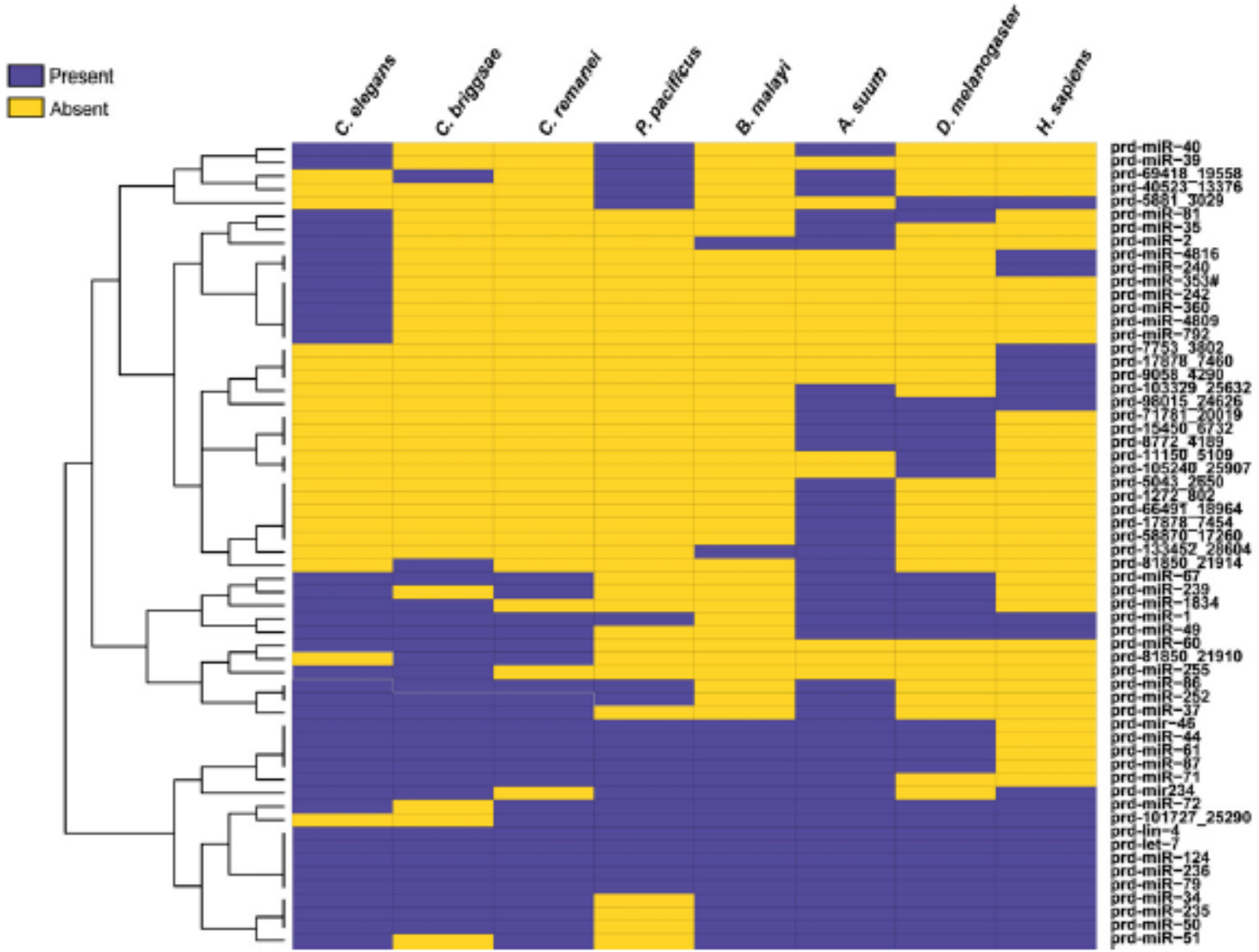
C. elegans vs. *P. redivivus* miRNAs



	C. elegans	*P. redivivus*
miRNAs	223	248
miRNA*	153	108
miRNA clusters	30	9

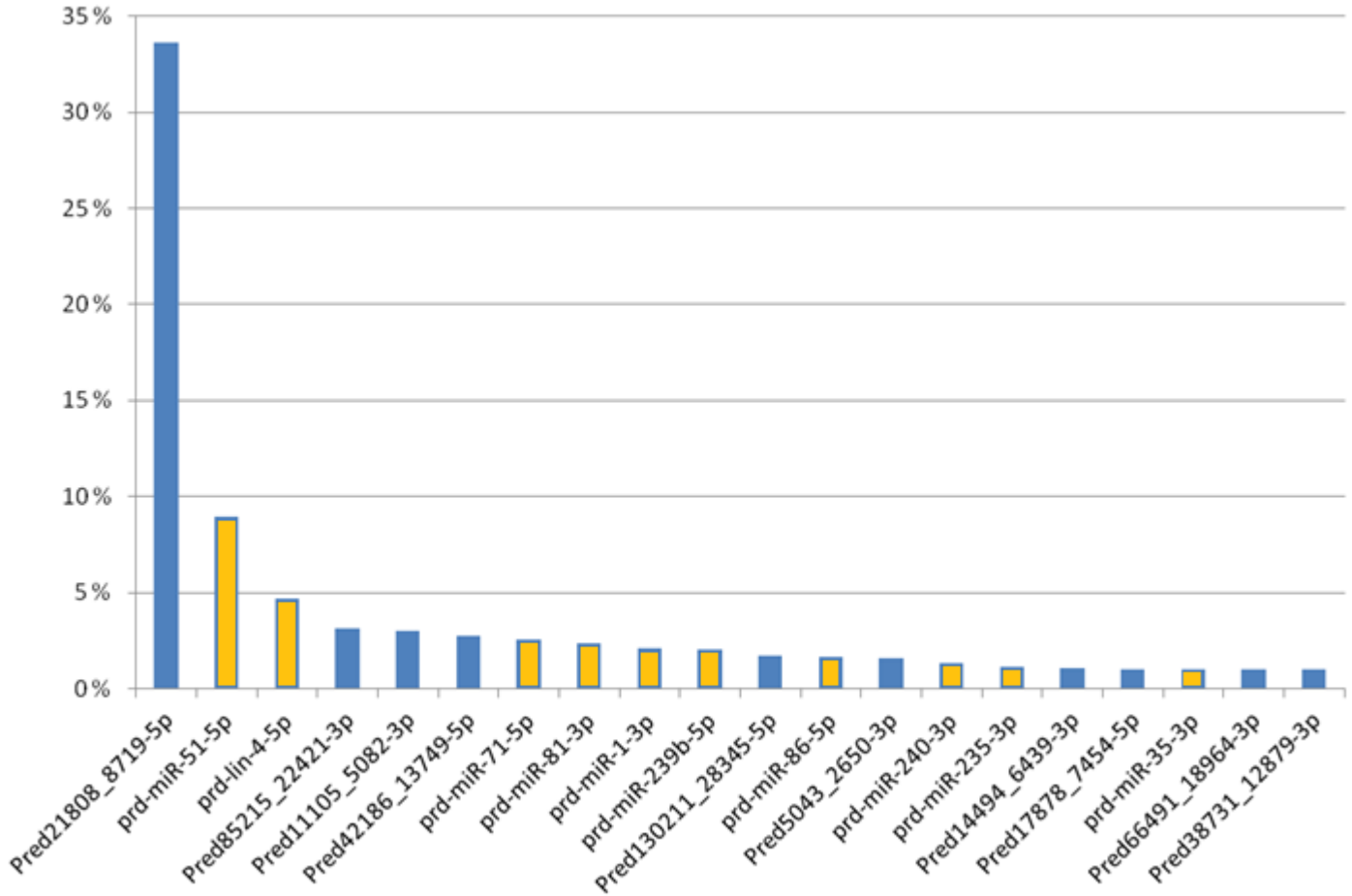
Shrinivasan et al., 2013

Clustering Heat Map of Orthology Between P. *redivivus* and other species



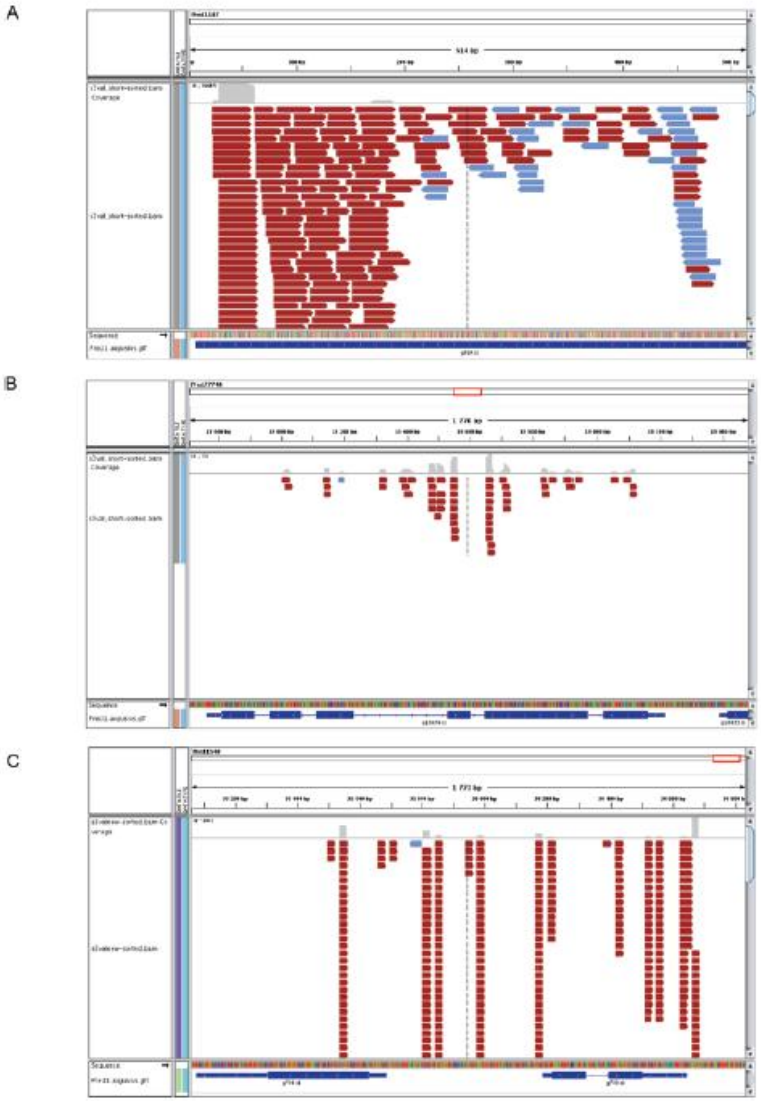
Srinivasan et al., 2013

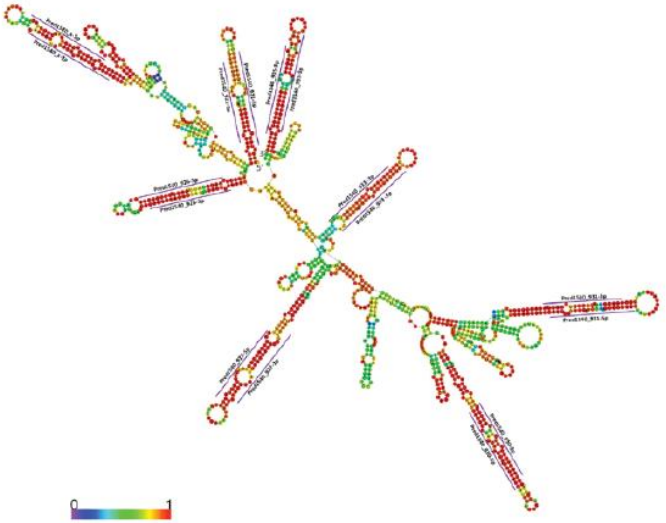
Top 20 miRNAs comprise 78% of all miRNA reads



Srinivasan et al., 2013

Small RNA classes in *P. redivivus*





Srinivasan et al., 2013

Summary

- miRNAs represent a class of an expanding family of small noncoding RNAs
- Small RNA-seq analysis tools can be used annotate known and unknown sequences from RNA seq data – KNOW WHAT TO LOOK FOR !
- Comparative genomics shows some well know miRNAs, and also some well conserved miRNAs with unknown function

Acknowledgements

Wong Laboratory
University of Eastern Finland
Liisa Heikkinen, M.Sc.
Juhani Peltonen, M.Sc.
Vuokko Aarnio, M.Sc.
Suvi Asikainen, Ph.D.
Merja Lakso, Ph.D.

Sternberg Laboratory
California Institute of Technology (Caltech)
Jagan Srinivasan, Ph.D.
Adler Dillman , Ph.D.
Ali Mortazavi, Ph.D.
Igor Antoshechkin, Ph.D.
Prof. Paul Sternberg

UNIVERSITY OF EASTERN FINLAND
ACADEMY OF FINLAND
norden
NordForsk
BF Biocenter Finland

F2 Mutagenesis Screen