02-716

Cross species analysis of genomics data

Computational Prediction of miRNAs and their targets

Outline

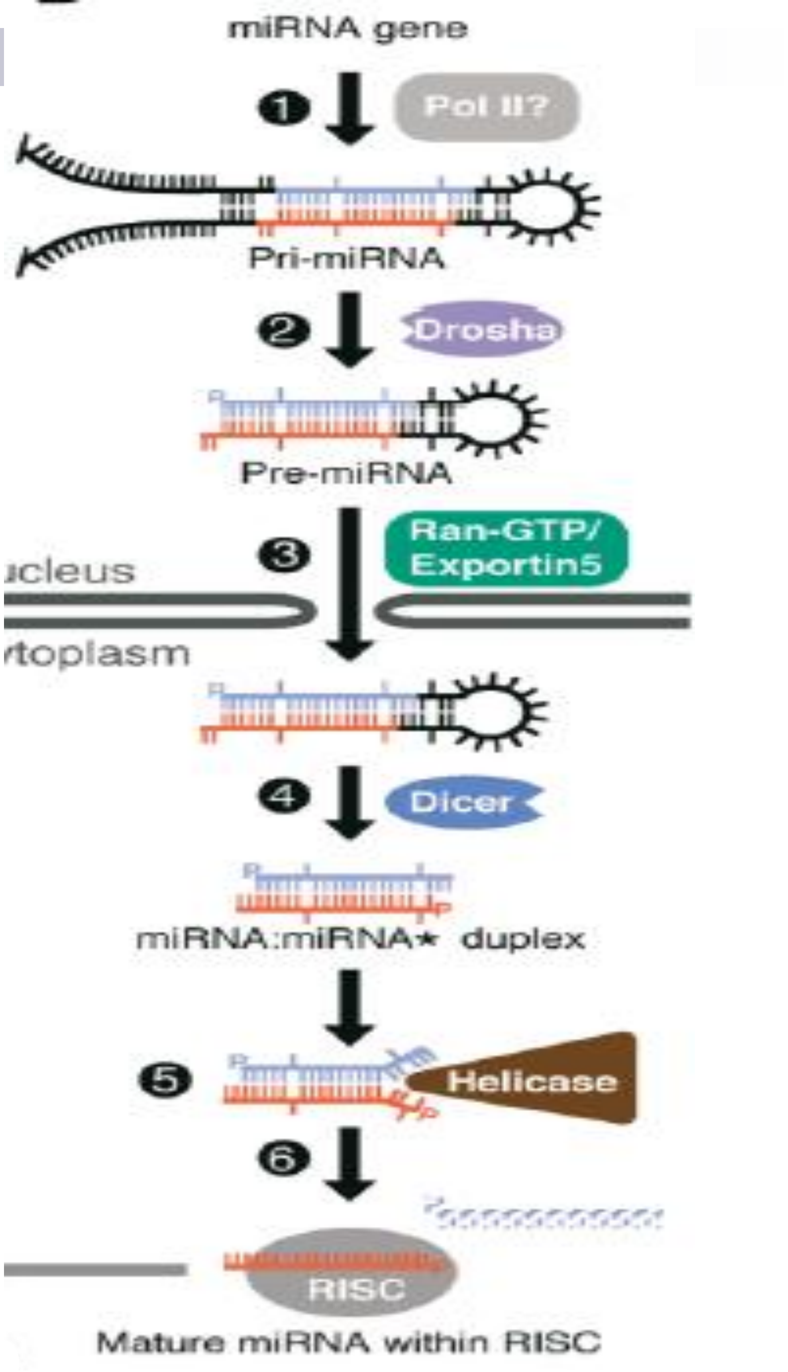
- Introduction
  - Brief history
  - miRNA Biogenesis
  - Why Computational Methods ?
- Computational Methods
  - Mature and precursor miRNA prediction
  - miRNA target gene prediction
- Conclusions

Brief history

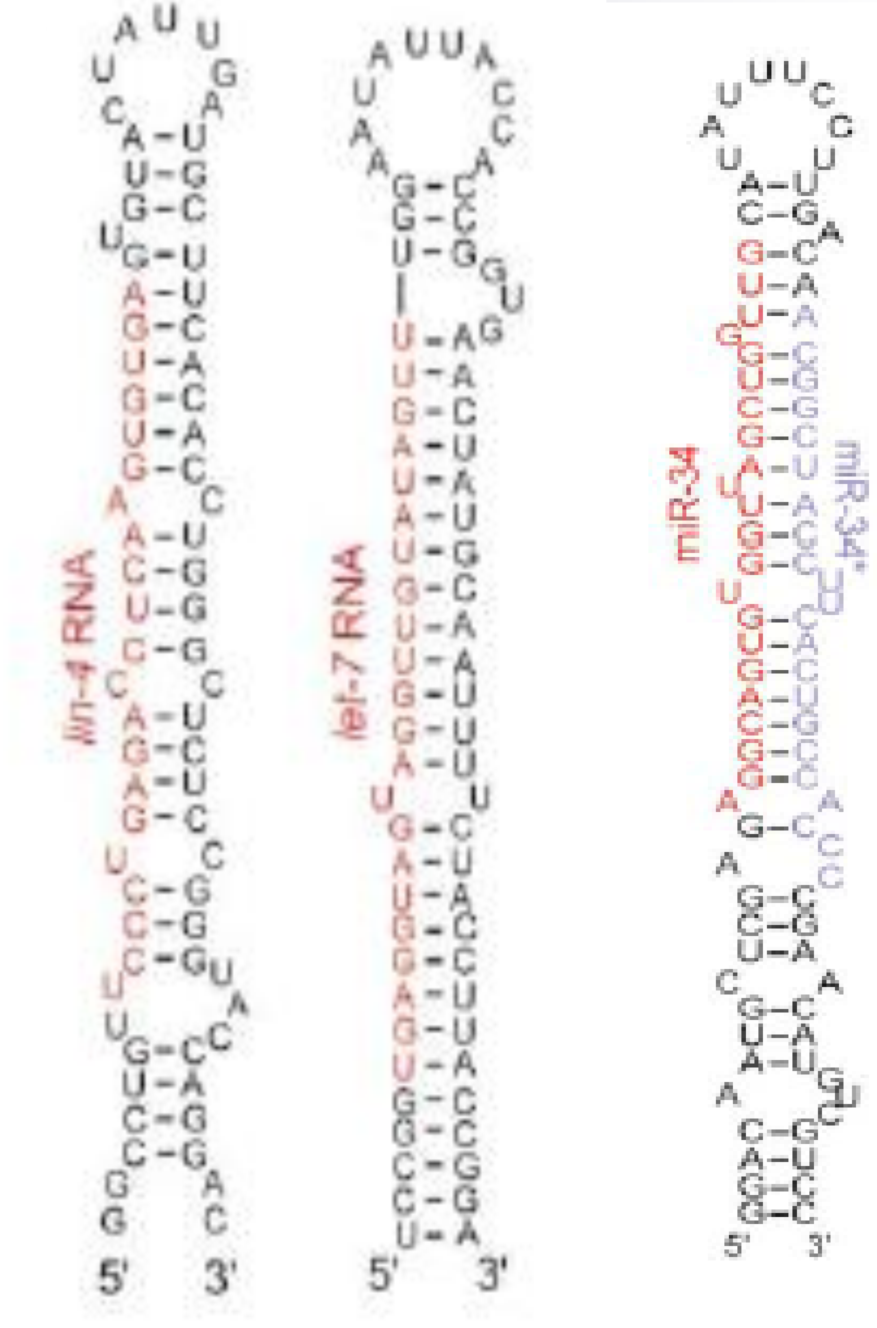
- MicroRNAs (miRNAs) are endogenous ~22 nt RNAs that play important roles in regulating gene expression in animals, plants, and fungi.
- The first miRNAs, lin-4, let-7, were identified in *C. elegans* (Lee R et al. 1993; Reihhart et al. 2000) when they were called small temporal RNAs (stRNA);
- The lin-4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

miRNA transcription and maturation

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



Predicted stem/loop secondary structure by RNAfold of known pre-miRNA. The sequence of the mature miRNAs(red) and miRNA* (blue).



Questions:

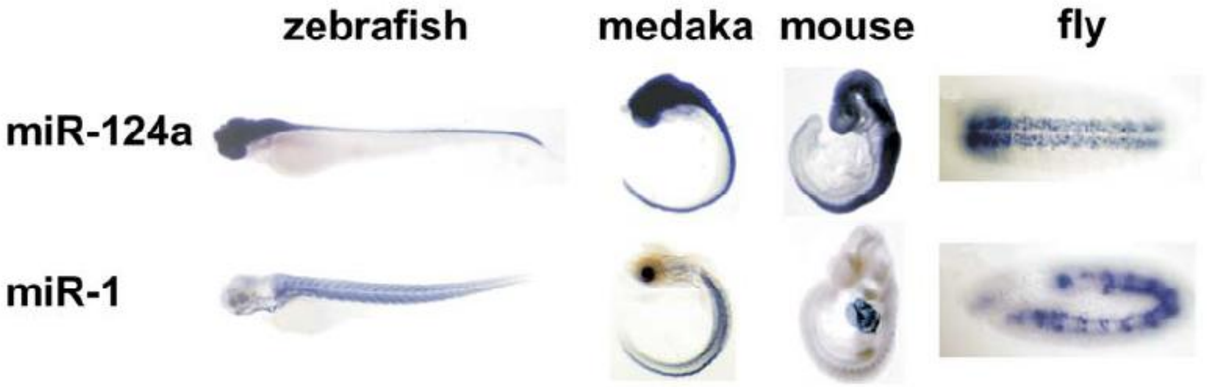
- How to find microRNA genes?

- Given a microRNA gene, how to find its targets?

- Target-driven approach:
  - Xie et al. (2005) analyzed conserved motifs that are overrepresented in 3' UTRs of genes
  - Found out they are complementing the seed sequences of known microRNAs.
  - They predicted 120 new miRNA candidates in human.

How to find microRNA genes?

- Biological approach
  - Small-RNA-cloning to identify new small RNAs
- Most MicroRNA genes are tissue-specific



miR-124a is restricted to the brain and spinal cord in fish and mouse or to the ventral nerve cord in fly

miR-1 is restricted to the muscles and the hart in mous

Computational methods to identify miRNA genes: Why?

- Significant progress has been made in miRNA research since the report of the lin-4 RNA(1993). Hundreds of 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 expression
  - stability
  - tissue specificity
  - cloning procedure
- Thus, computational identification of miRNAs from genomic sequences provide a valuable complement to cloning.

Prediction of novel miRNA: Biological inference

- Biogenesis
  - miRNA
    - 20-to 24-nt RNAs derived from endogenous transcripts that form local hairpin structures.
    - Processing of pre-miRNA leads to single (sometimes 2) mature miRNA molecule
  - siRNA
    - Derived from extended dsRNA
    - Each dsRNA gives rise to numerous different siRNAs
- Evolutionary conservation
  - miRNA
    - Mature and pre-miRNA is usually evolutionary conserved
    - miRNA genomic loci are distinct from and often usually distant from those of other types of recognized genes. Sometimes reside in introns.
  - siRNA
    - Less sequence conservation
    - Correspond to sequences of known or predicted mRNAs, or heterochromatin.

Overview

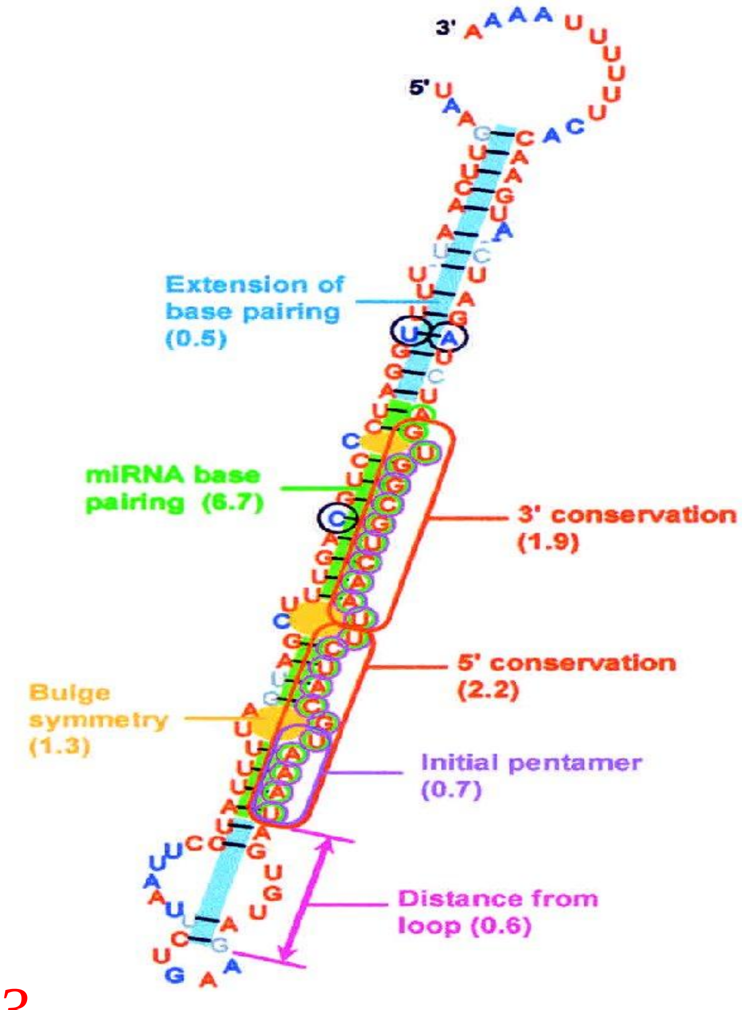
- Introduction
  - Brief history
  - MiRNA Biogenesis
  - Why Computational Methods ?
- Computational Methods
  - Mature and precursor miRNA prediction
  - miRNA target gene prediction
- Conclusions

Computational prediction of *C.elegans* miRNA genes

- Scanning for hairpin structures (RNAfold: free energy < -25kcal/mole) within sequences that were conserved between *C.elegans* and *C.briggsae* (WU-BLAST cut-off E < 1.8).
- 36,000 pairs of hairpins identified capturing 50/53 miRNAs previously reported to be conserved between the two species.
- 50 miRNAs were used as training set for the development of a program called “MiRscan”.
- MiRscan was then used to evaluate the 36,000 hairpins.

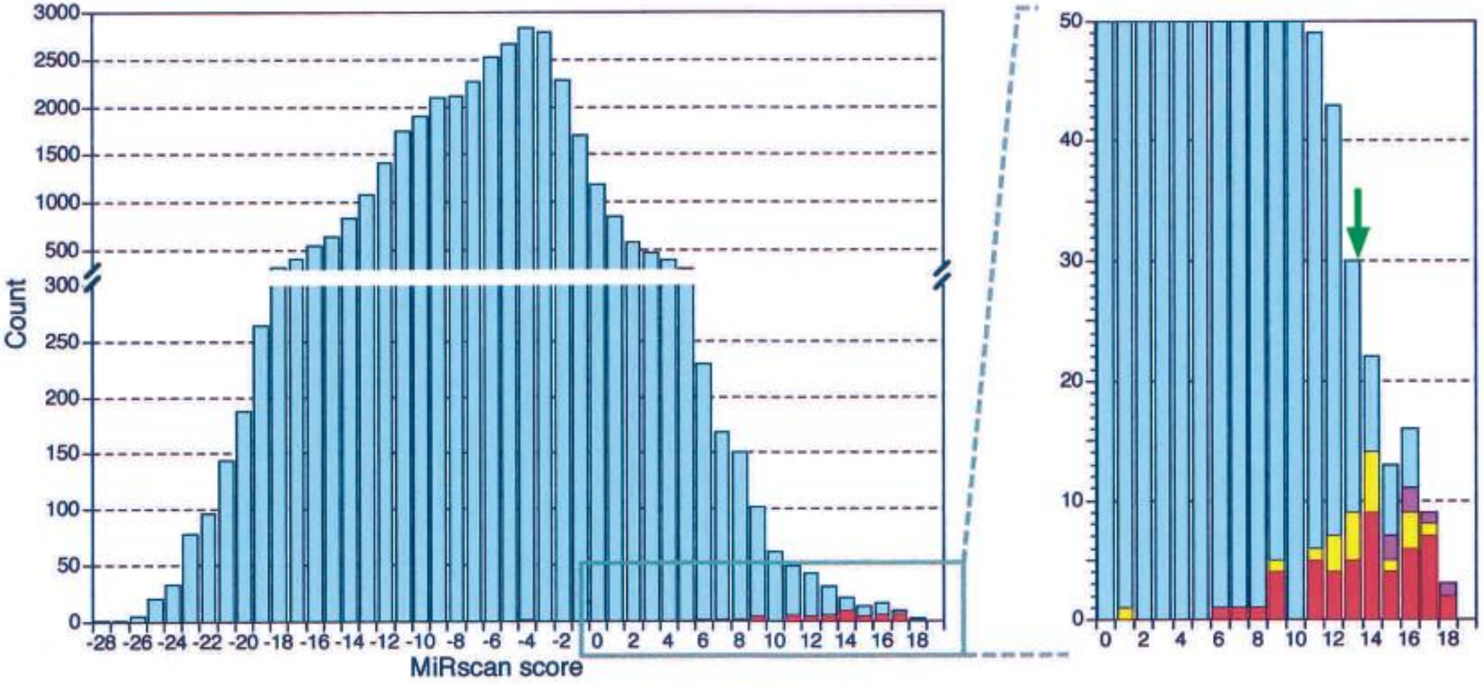
Features utilized by the Algorithm

- The MiRscan algorithm examines several features of the hairpin in a 21-nt window
- The total score for a miRNA candidate was computed by summing the score of each feature
- The score for each feature is computed by dividing the frequency of the given value in the training set to its overall frequency



Lim et al, Genes and Development 2003

MiRscan predicted results

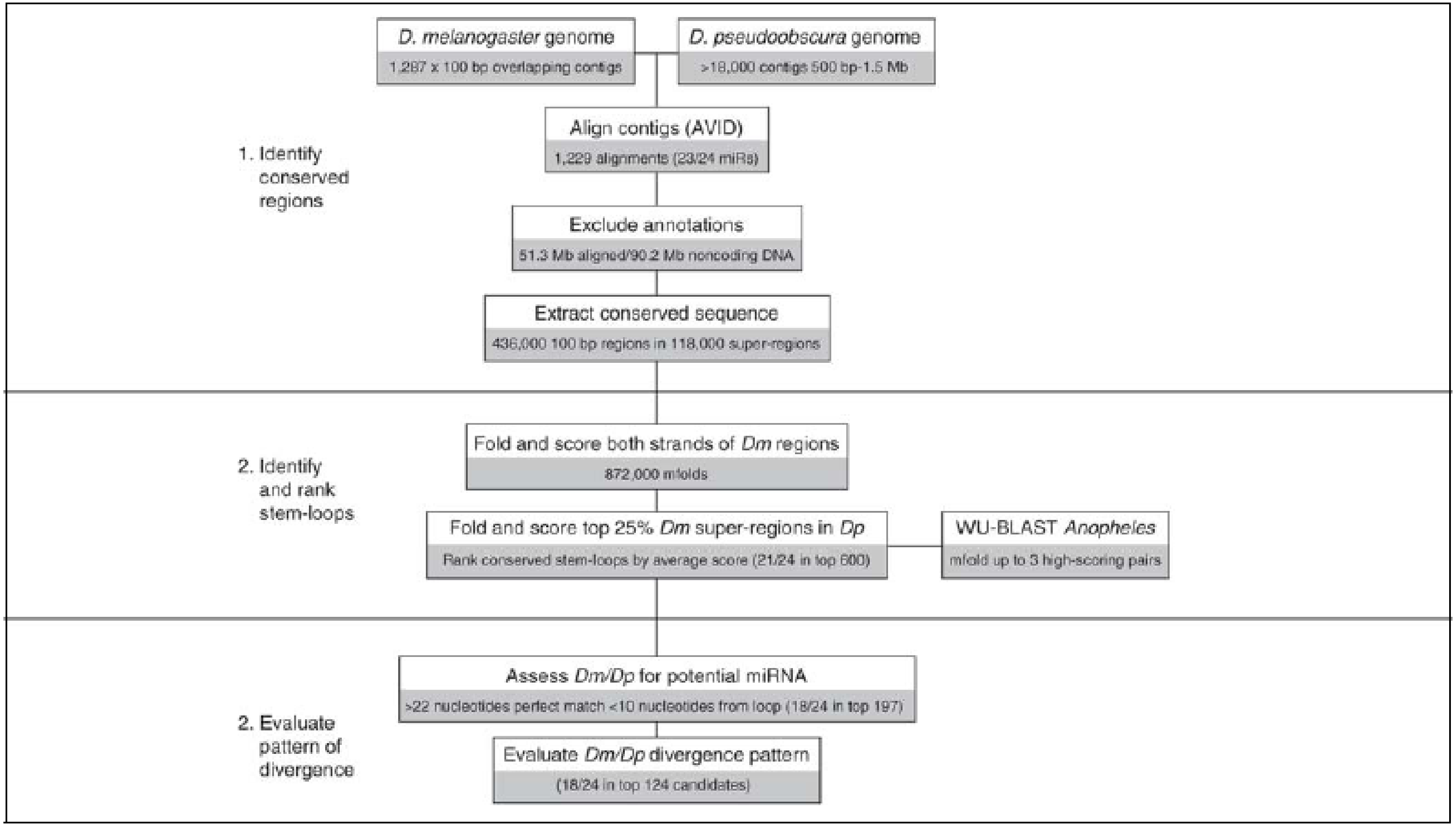


- Blue: distribution of MiRscan score of 35,697 sequences
- Red: training set
- Yellow and purple are verified by cloning or other evidence.

Computational Identification of *Drosophila* miRNA genes

- Two *Drosophila* species: *D.melanogaster* and *D.pseudoobscura* were used to establish conservation.
- 3-part computational pipeline called “miRseeker” to identify Drosophilid miRNA sequences
- Assessed algorithms efficiency by observing its ability to give high score to 24 known Drosophila miRNAs.

Overview of “miRseeker”



D. melanogaster genome

1,287 x 100 bp overlapping contigs

D. pseudoobscura genome

>18,000 contigs 500 bp-1.5 Mb

1. Identify conserved regions

Align contigs (AVID)

1,229 alignments (23/24 miRs)

Exclude annotations

51.3 Mb aligned/90.2 Mb noncoding DNA

Extract conserved sequence

436,000 100 bp regions in 118,000 super-regions

2. Identify and rank stem-loops

Fold and score both strands of *Dm* regions

872,000 mfolds

Fold and score top 25% *Dm* super-regions in *Dp*

Rank conserved stem-loops by average score (21/24 in top 600)

WU-BLAST *Anopheles*

mfold up to 3 high-scoring pairs

2. Evaluate pattern of divergence

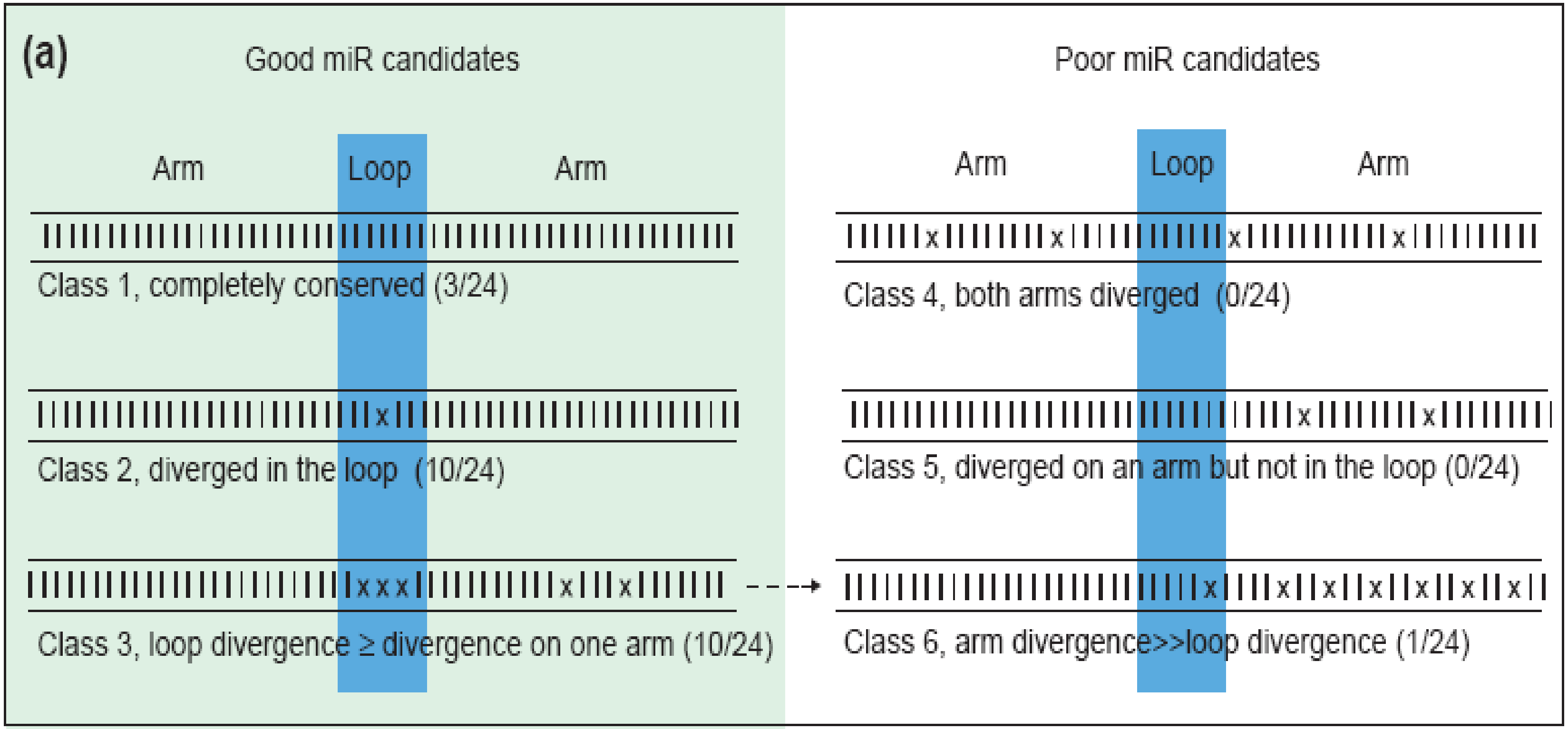
Assess *Dm/Dp* for potential miRNA

>22 nucleotides perfect match <10 nucleotides from loop (18/24 in top 197)

Evaluate *Dm/Dp* divergence pattern

(18/24 in top 124 candidates)

Step3: Patterns of nucleotide divergence


(a)
Good miR candidates
Poor miR candidates
Arm
Loop
Arm
Arm
Loop
Arm
Class 1, completely conserved (3/24)
Class 4, both arms diverged (0/24)
Class 2, diverged in the loop (10/24)
Class 5, diverged on an arm but not in the loop (0/24)
Class 3, loop divergence ≥ divergence on one arm (10/24)
Class 6, arm divergence>>loop divergence (1/24)

Lai et al, Genome Biology 2003

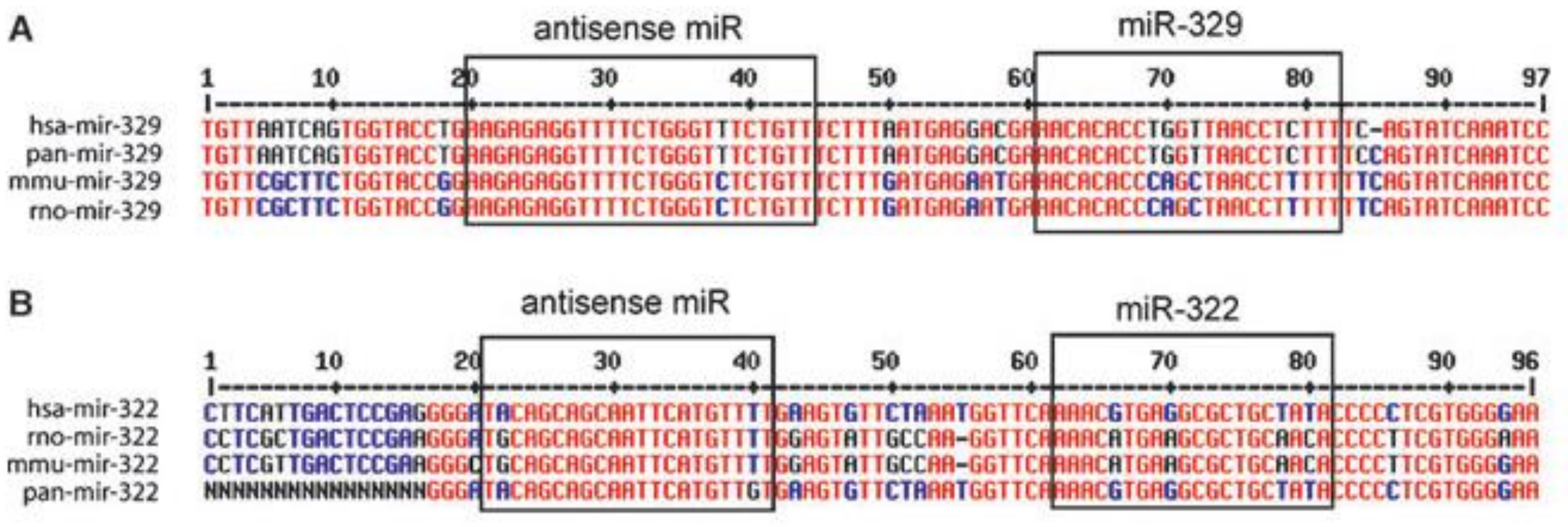
Results

Organism	**Program**	**Prediction accuracy**	**Experimental Verification**
C.elegans	MiRscan	50/58 known miRNAs fell in high scoring tail of the distribution.	35 hairpins had a score > 13,9 (median score of 58 known miRNAs). 16/35 were validated by cloning and northern blots
Drosophila	miRseeker	18/24 were in top 124 candidates	38 candidate genes selected for experimental validation. In 24/38 expression was observed by northern blot analysis

New human and mouse miRNA detected by homology

- Entire set of human and mouse pre- and mature miRNA from the miRNA registry was submitted to BLAT search engine against the human genome and then against the mouse genome.
- Sequences with high % identity were examined for hairpin structure using MFOLD, and 16-nt stretch base paring.

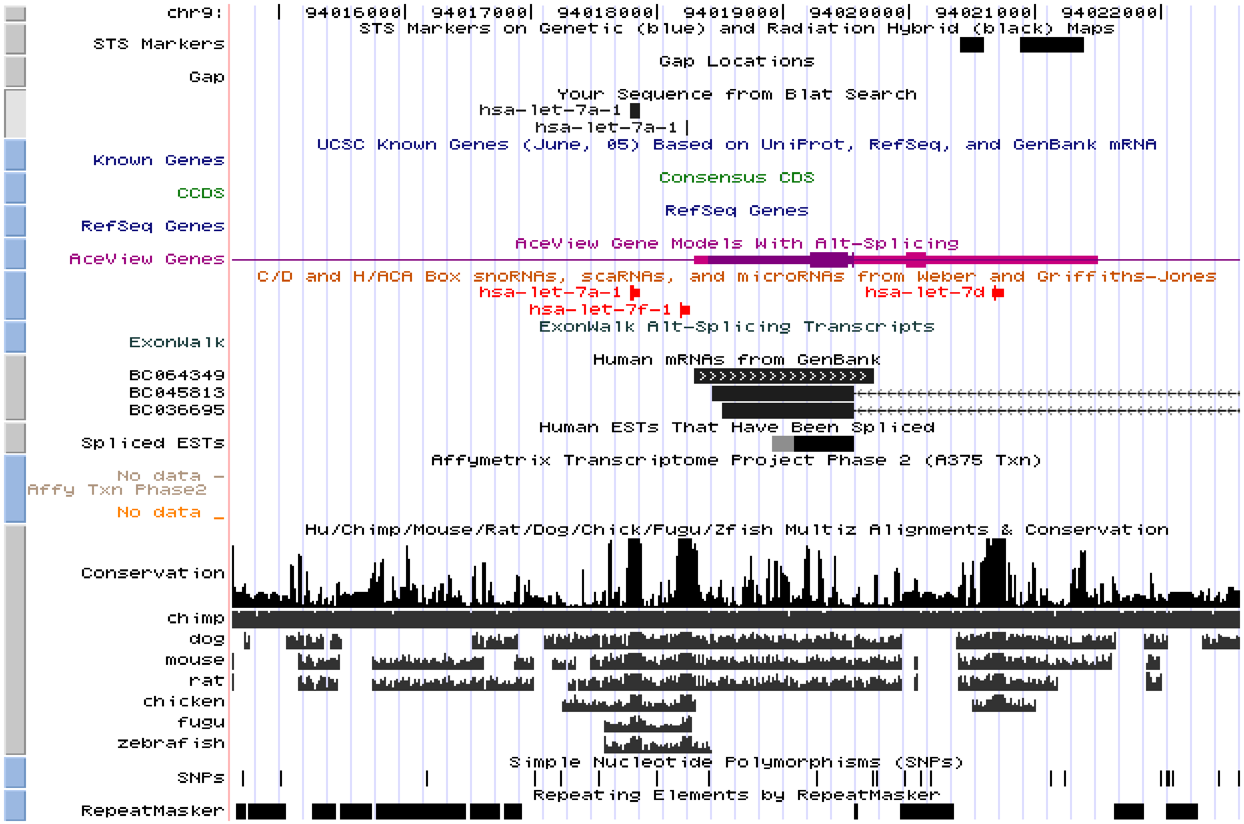
60 new potential miRNAs (15 for human and 45 for mouse)


A
antisense miR
miR-329
1
10
20
30
40
50
60
70
80
90
97
hsa-mir-329 TGTTAATCAGTGGTACCTGAAGAGAGGTTTTCTGGGTTTCTGTTTCTTTAATGAGGACGAAACACACCTGGTTAACCTCTTTTC-AGTATCAAATCC
pan-mir-329 TGTTAATCAGTGGTACCTGAAGAGAGGTTTTCTGGGTTTCTGTTTCTTTAATGAGGACGAAACACACCTGGTTAACCTCTTTTCCAGTATCAAATCC
mmu-mir-329 TGTTCGCTTCTGGTACCGGAAGAGAGGTTTTCTGGGTCTCTGTTTCTTTGATGAGAATGAAACACACCCAGCTAACCTTTTTTTCAGTATCAAATCC
rno-mir-329 TGTTCGCTTCTGGTACCGGAAGAGAGGTTTTCTGGGTCTCTGTTTCTTTGATGAGAATGAAACACACCCAGCTAACCTTTTTTTCAGTATCAAATCC
B
antisense miR
miR-322
1
10
20
30
40
50
60
70
80
90
96
hsa-mir-322 CTTCATTGACTCCGAGGGGATACAGCAGCAATTCATGTTTTGAAGTGTTCTAAATGGTTCAAAACGTGAGGCGCTGCTATACCCCCTCGTGGGGAA
rno-mir-322 CCTCGCTGACTCCGAAGGGATGCAGCAGCAATTCATGTTTTGGAGTATTGCCAA-GGTTCAAAACATGAAGCGCTGCAACACCCCTTCGTGGGAAA
mmu-mir-322 CCTCGTTGACTCCGAAGGGCTGCAGCAGCAATTCATGTTTTGGAGTATTGCCAA-GGTTCAAAACATGAAGCGCTGCAACACCCCTTCGTGGGGAA
pan-mir-322 NNNNNNNNNNNNNNNNGGGATACAGCAGCAATTCATGTTGTGAAGTGTTCTAAATGGTTCAAAACGTGAGGCGCTGCTATACCCCCTCGTGGGGAA

- Mature miRNA were either perfectly conserved or differed by only 1 nucleotide between human and mouse.

Weber, FEBS 2005

Human and mouse miRNAs reside in conserved regions of synteny


chr9:
94016000| 94017000| 94018000| 94019000| 94020000| 94021000| 94022000|
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Your Sequence from Blat Search
hsa-let-7a-1
hsa-let-7a-1
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
Known Genes
Consensus CDS
CCDS
RefSeq Genes
RefSeq Genes
AceView Gene Models With Alt-Splicing
AceView Genes
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from Weber and Griffiths-Jones
hsa-let-7a-1
hsa-let-7d
hsa-let-7f-1
ExonWalk Alt-Splicing Transcripts
ExonWalk
Human mRNAs from GenBank
BC064349
BC045813
BC036695
Human ESTs That Have Been Spliced
Spliced ESTs
Affymetrix Transcriptome Project Phase 2 (A375 Txn)
No data -
Affy Txn Phase2
No data _
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
Simple Nucleotide Polymorphisms (SNPs)
SNPs
Repeating Elements by RepeatMasker
RepeatMasker

- Mmu-mir-345 resides in AK0476268 *RefSeq* gene. Human orthologue was found upstream of C14orf69, the best BLAT hit for AK0476268.

Limitations of methods so far

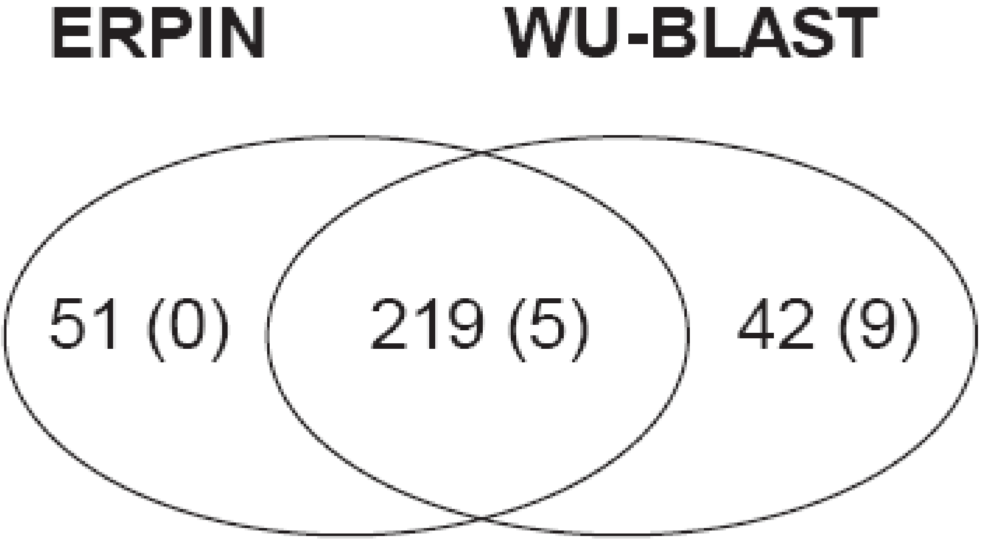
- Pipeline structure, use cut-offs and filtering/eliminating sequences as pipeline proceeds.
- Sequence alignment alone used to infer conservation (limited because areas of miRNA precursors are often not conserved)
- Limited to closely related species (i.e. *C.elegans, C.briggsae).*

Additional methods

Profile-based detection of mRNAs

- 593 sequences form miRNA registry (513 animal and 50 plant)
- CLUSTAL generated 18 most prominent miRNA clusters.
- Each cluster was used to deduce a consensus 2ry structure using ALIFOLD program.
- These training sets were then fed into ERPIN (profile scan algorithm - reads a sequence alignement and secondary structure)
- Scanned a 14.3 Gb database of 20 genomes.

Results: 270/553 top scoring ERPIN candidates previously un-identified



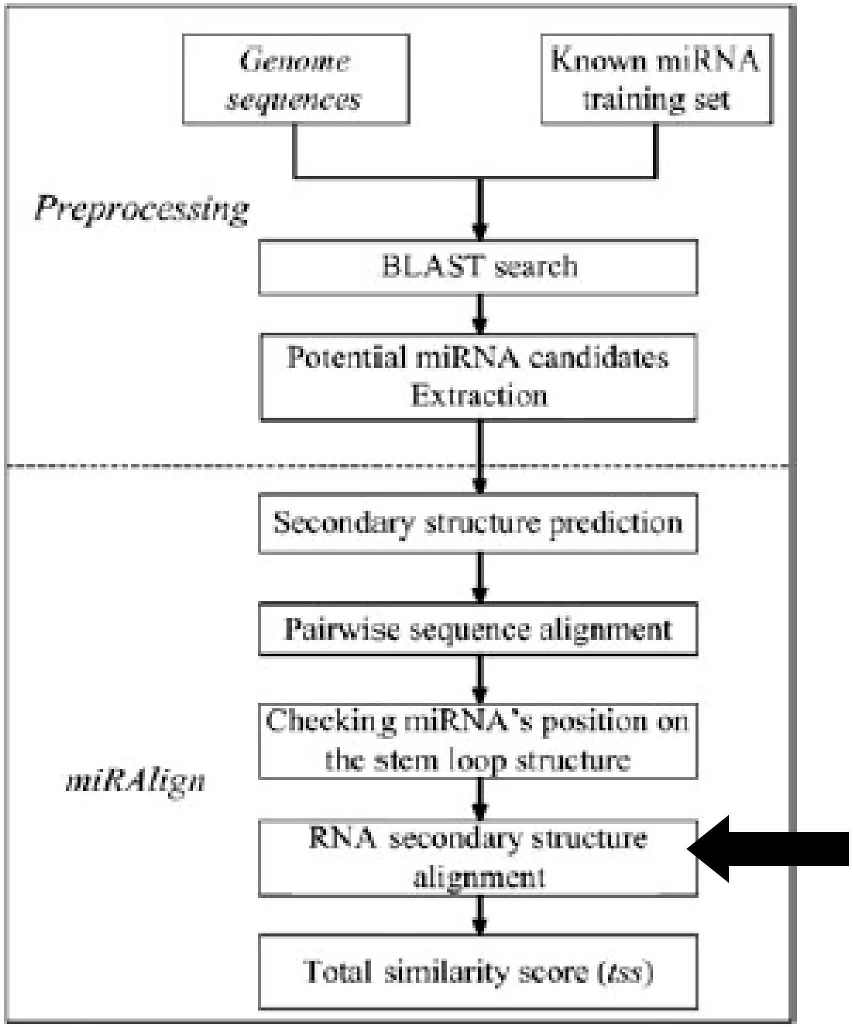
- *Adv:Takes into account 2ry structure conservation using Profiles.*
- *Disadv: Only applicable to miRNA families with sufficient known samples.*

Legendre et al, Bioinformatics 2005

Sequence and structure alignment - miRAlign

1. 1054 animal miRNA and their precursors (11040).
2. Train on all but C.briggsae miRNAs
3. Test programs ability to identify miRNAs in C.briggsae (79 known miRNAs).
4. Train on all but the C.briggsae and C.elegans
5. Repeat step (3) - Test programs ability to identify miRNAs in distantly related sequences.
6. Compare with other programs.

Overview of miRAlign


Genome sequences
Known miRNA training set
Preprocessing
BLAST search
Potential miRNA candidates Extraction
Secondary structure prediction
Pairwise sequence alignment
Checking miRNA's position on the stem loop structure
miRAlign
RNA secondary structure alignment
Total similarity score (tss)

RNAforeseter

Human miRNA prediction using Support Vector Machines

- DIANA-microH: Supervised analysis program based on SVM. (*Szafranski et al 2005*).
- Train on subset of human miRNAs present in RFAM and then test on the remaining.
- Negative sequences that appear to exhibit hairpin –like structure were also used derived from 3’UTRs.

Features used

First predicts 2ry structure and assessed the following:

1. Free Energy
2. Paired Bases
3. Loop Length
4. Arm Conservation

- DIANA-microH introduces two new features:

5. GC Content
6. Stem Linearity

Results

- 98.6% accuracy on test set: 43/45 true miRNAs correctly classified, 284/288 negative 3’UTR sequences correctly classified.
- Evaluation on chr 21:
  - 35 hairpins with outstandingly high score.
  - All four miRNA listed in RFAM on chr 21 where in the high scoring group.
- *Adv: Combines various biological features rather than follow a stringent pipeline. Sequence and structure conservation used.*
- *Disadv: Some feature may receive greater value than others (redundancy).*

Overview

- Introduction
  - Brief history
  - MiRNA Biogenesis
- Computational Methods
  - Mature and precursor miRNA prediction
  - miRNA target gene prediction
- Conclusions

How to find MicroRNA targets?

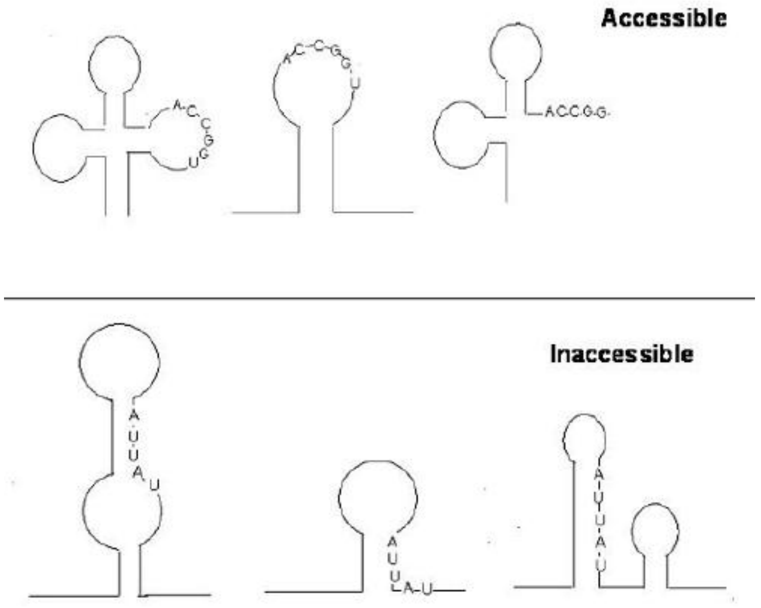
- MicroRNA targets are located in 3’ UTRs, and complementing mature microRNAs
  - For plants, the targets have a high degree of sequence complementarity.
  - … But for animals, this is not always the case.

  - They are short (~21 nt)
  - Allowing for G-U pairs, mismatches and gaps will likely lead to many false positives when using standard alignment algorithms.

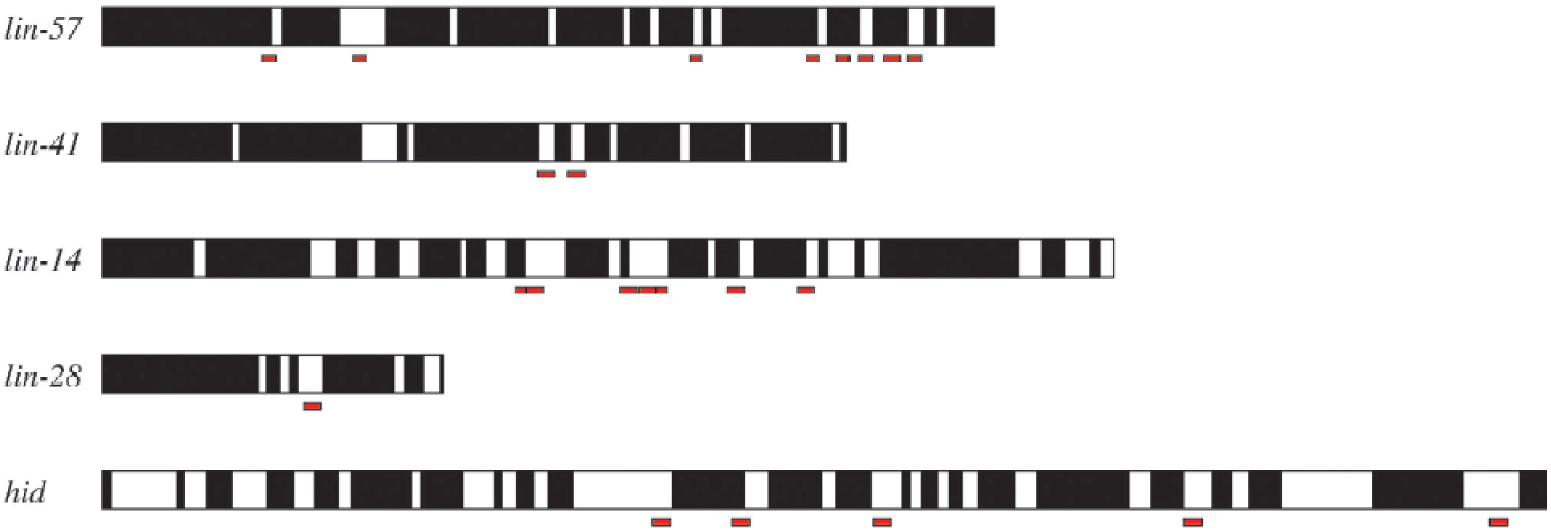
  - How to remove the false positives?

Improving prediction accuracy

- Incorporating mRNA UTR structure to predict microRNA targets (Robins et al. 2005)
  - □ Make sure the predicted target is “accessible”.
  - □ Not forming base pairing itself.



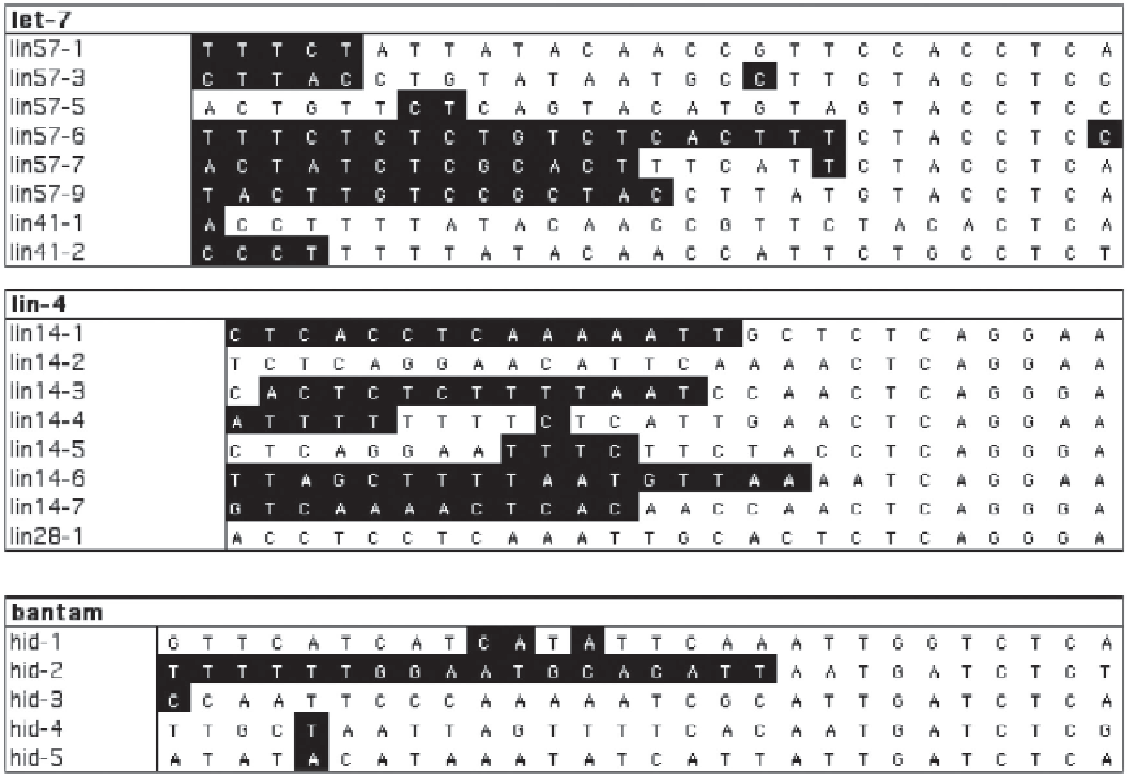
Other properties of microRNA targets

- MicroRNA targets are often conserved across species. (Stark et al. 2003)
- 

For lins, comparison is between *C. elegans* and *C. briggsae*.
For hid, comparison is between *D. melanogaster* and *D. pseudoobscura*.

Other properties of microRNA targets

- Sequence conservations of target sites
  - □ Better complementarity to the 5’ ends of the miRNAs.



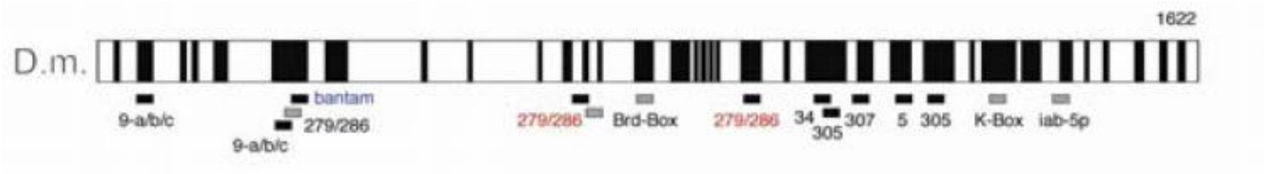
let-7	
lin57-1	T T T C T A T T A T A C A A C C G T T C C A C C T C A
lin57-3	C T T A C C T G T A T A A T G C C T T C T A C C T C C
lin57-5	A C T G T T C T C A G T A C A T G T A G T A C C T C C
lin57-6	T T T C T C T C T G T C T C A C T T T C T A C C T C C
lin57-7	A C T A T C T C G C A C T T T C A T T C T A C C T C A
lin57-9	T A C T T G T C C G C T A C C T T A T G T A C C T C A
lin41-1	A C C T T T T A T A C A A C C G T T C T A C A C T C A
lin41-2	C C C T T T T T A T A C A A C C A T T C T G C C T C T

lin-4	
lin14-1	C T C A C C T C A A A A A T T G C T C T C A G G A A
lin14-2	T C T C A G G A A C A T T C A A A A C T C A G G A A
lin14-3	C A C T C T C T T T T A A T C C A A C T C A G G G A
lin14-4	A T T T T T T T T C T C A T T G A A C T C A G G A A
lin14-5	C T C A G G A A T T T C T T C T A C C T C A G G G A
lin14-6	T T A G C T T T T A A T G T T A A A A T C A G G A A
lin14-7	G T C A A A A C T C A C A A C C A A C T C A G G G A
lin28-1	A C C T C C T C A A A T T G C A C T C T C A G G G A

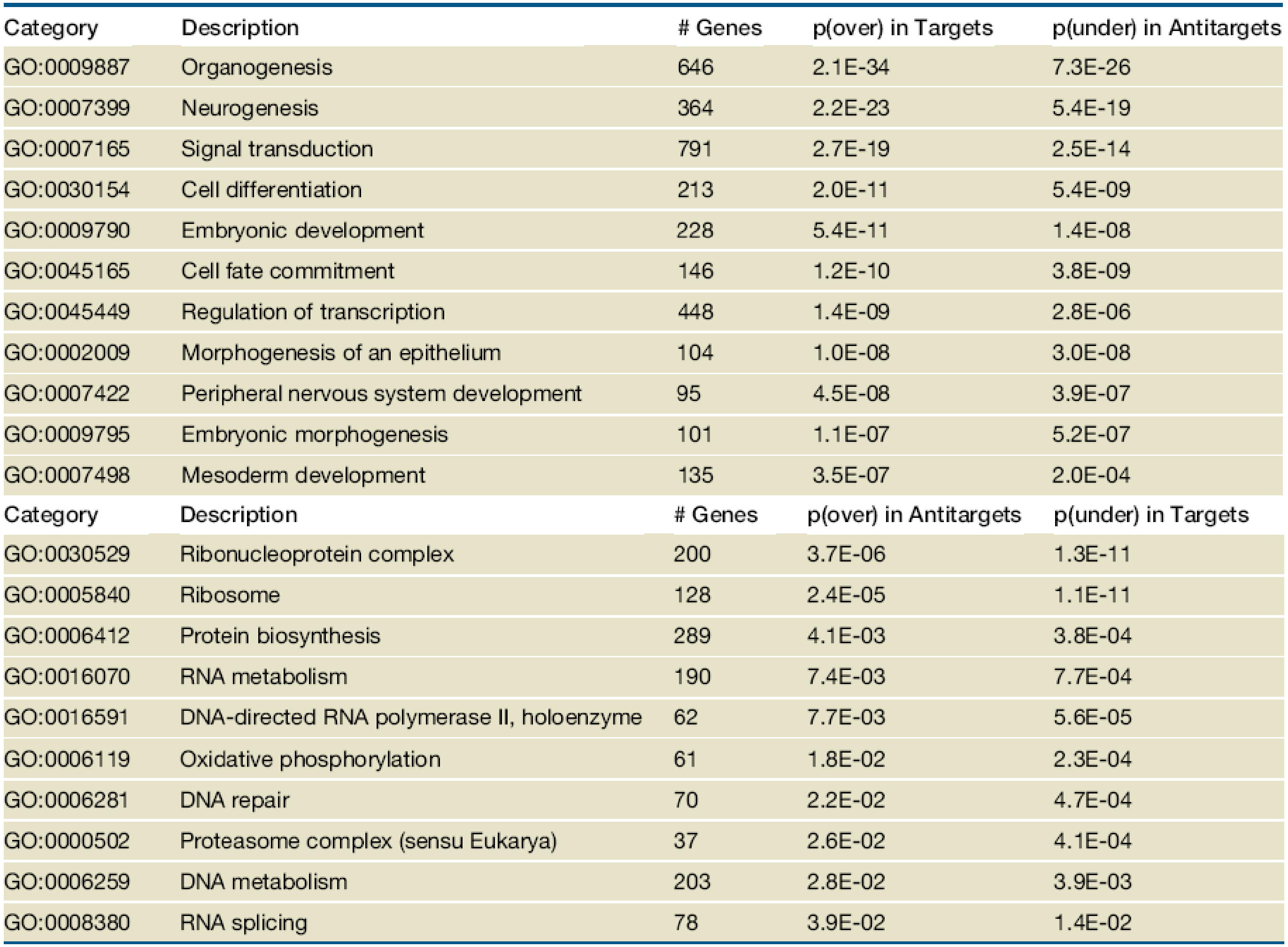
bantam	
hid-1	G T T C A T C A T C A T A T T C A A A T T G G T C T C A
hid-2	T T T T T T G G A A T G C A C A T T A A T G A T C T C T
hid-3	C C A A T T C C C A A A A A T C G C A T T G A T C T C A
hid-4	T T G C T A A T T A G T T T T C A C A A T G A T C T C G
hid-5	A T A T A C A T A A A T A T C A T T A T T G A T C T C A

Other properties of microRNA targets

- Clusters of microRNA targets
  - Extensive cooccurrence of the sites for different microRNAs in target 3’ UTRs.



Presence and absence of target sites correlate with gene function

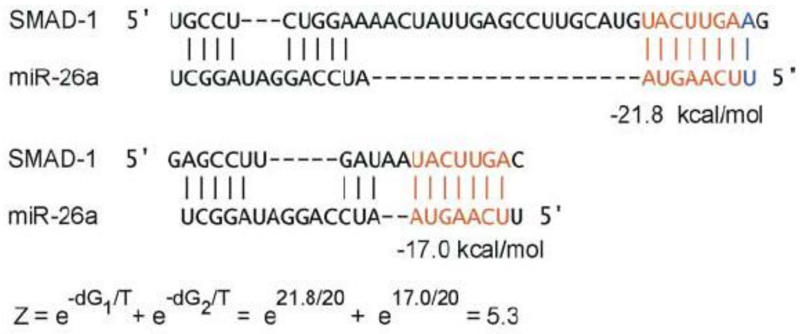


Category	Description	# Genes	p(over) in Targets	p(under) in Antitargets
GO:0009887	Organogenesis	646	2.1E-34	7.3E-26
GO:0007399	Neurogenesis	364	2.2E-23	5.4E-19
GO:0007165	Signal transduction	791	2.7E-19	2.5E-14
GO:0030154	Cell differentiation	213	2.0E-11	5.4E-09
GO:0009790	Embryonic development	228	5.4E-11	1.4E-08
GO:0045165	Cell fate commitment	146	1.2E-10	3.8E-09
GO:0045449	Regulation of transcription	448	1.4E-09	2.8E-06
GO:0002009	Morphogenesis of an epithelium	104	1.0E-08	3.0E-08
GO:0007422	Peripheral nervous system development	95	4.5E-08	3.9E-07
GO:0009795	Embryonic morphogenesis	101	1.1E-07	5.2E-07
GO:0007498	Mesoderm development	135	3.5E-07	2.0E-04
Category	Description	# Genes	p(over) in Antitargets	p(under) in Targets
GO:0030529	Ribonucleoprotein complex	200	3.7E-06	1.3E-11
GO:0005840	Ribosome	128	2.4E-05	1.1E-11
GO:0006412	Protein biosynthesis	289	4.1E-03	3.8E-04
GO:0016070	RNA metabolism	190	7.4E-03	7.7E-04
GO:0016591	DNA-directed RNA polymerase II, holoenzyme	62	7.7E-03	5.6E-05
GO:0006119	Oxidative phosphorylation	61	1.8E-02	2.3E-04
GO:0006281	DNA repair	70	2.2E-02	4.7E-04
GO:0000502	Proteasome complex (sensu Eukarya)	37	2.6E-02	4.1E-04
GO:0006259	DNA metabolism	203	2.8E-02	3.9E-03
GO:0008380	RNA splicing	78	3.9E-02	1.4E-02

Targetscan
(Lewis et al. Cell 2003)

Given a microRNA that is conserved in multiple species and a set of *orthologous* 3’ UTR sequences:

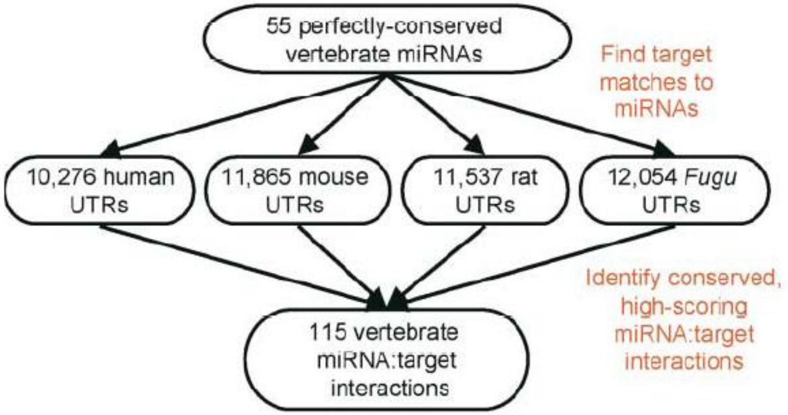
1. Use 7 nt segment of the miRNA as the ‘microRNA seed’ to find the perfect complementary motifs in the UTR regions.
2. Extend each seed to find the best energy match
3. Assign a Z score.
4. Rank (Ri) for each species.
5. Repeat above process for all species.
6. Keep those genes for which Zi > Z_c and Ri < R_c.



```
SMAD-1   5' UGCCU---CUGGAAAACUAUUGAGCCUUGCAUGUACUUGAAG
               ||||    |||||                      ||||||||
miR-26a     UCGGAUAGGACCUA-------------------AUGAACUU 5'
                                            -21.8 kcal/mol

SMAD-1   5' GAGCCUU-----GAUAAUACUUGAC
             |||||      |||  |||||||
miR-26a      UCGGAUAGGACCUA--AUGAACUU 5'
                        -17.0 kcal/mol
```

$$Z = e^{-dG_1/T} + e^{-dG_2/T} = e^{21.8/20} + e^{17.0/20} = 5.3$$



55 perfectly-conserved vertebrate miRNAs

Find target matches to miRNAs

10,276 human UTRs

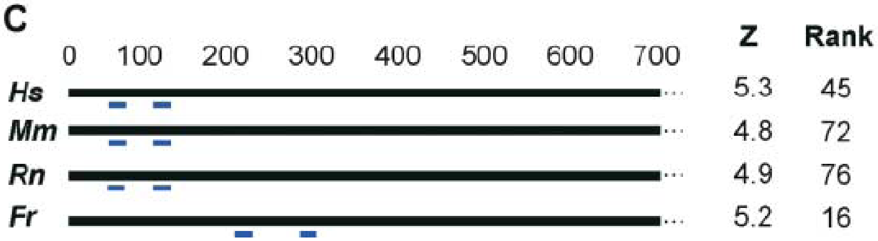
11,865 mouse UTRs

11,537 rat UTRs

12,054 *Fugu* UTRs

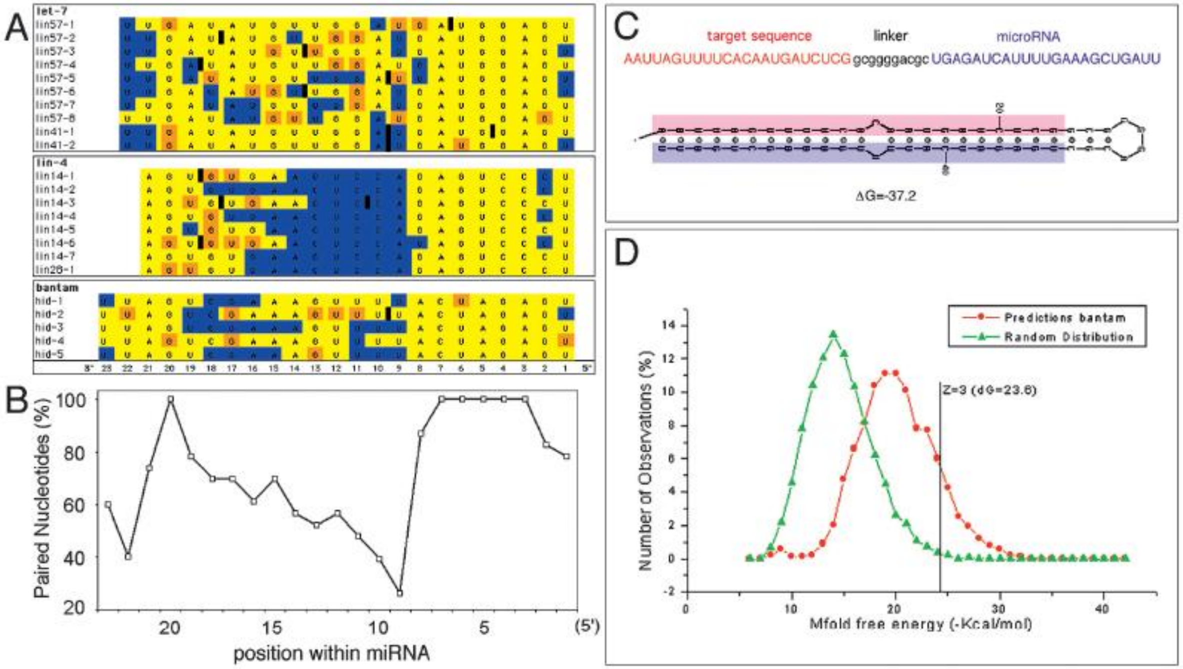
Identify conserved, high-scoring miRNA:target interactions

115 vertebrate miRNA:target interactions

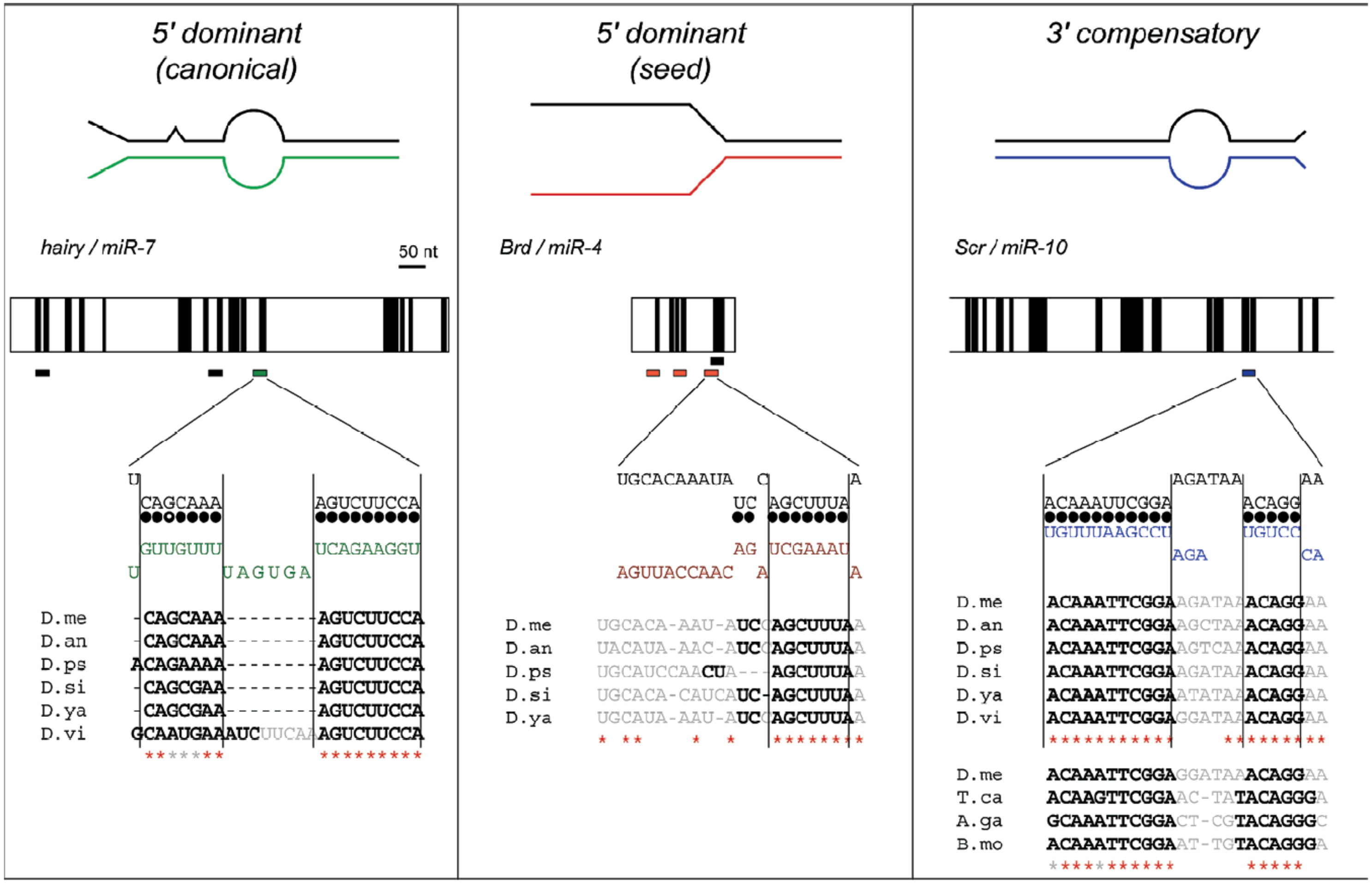


Profile based target search
(Stark et al. Plos Biology 2003

1. Build profiles for each microRNA family (using HMMer) for first 8 residues, allowing for G:U mismatches.
2. Only search conserved 3' UTRs (in two fly genomes) using the profiles.
3. Sequence matches are extended to miRNA length + 5nt.
4. Compute the energy using the Mfold and provide the z-scores.



Three Classes of microRNA Target Sites (Brennecke et al. Plos Biology 2005)


5' dominant
(canonical)
5' dominant
(seed)
3' compensatory
hairy / miR-7
50 nt
Brd / miR-4
Scr / miR-10
U
CAGCAAA
AGUCUUCCA
GUUGUUU
UCAGAAGGU
U
UAGUGA
D.me -CAGCAAA---------AGUCUUCCA
D.an -CAGCAAA---------AGUCUUCCA
D.ps ACAGAAAA---------AGUCUUCCA
D.si -CAGCGAA---------AGUCUUCCA
D.ya -CAGCGAA---------AGUCUUCCA
D.vi GCAAUGAAAUCUUCAAAGUCUUCCA
UGCACAAAUA C A
UC AGCUUUA
AG UCGAAAU
AGUUACCAAC A A
D.me UGCACA-AAU-AUCCAGCUUUAA
D.an UACAUA-AAC-AUCCAGCUUUAA
D.ps UGCAUCCAACUA---AGCUUUAA
D.si UGCACA-CAUCAUC-AGCUUUAA
D.ya UGCAUA-AAU-AUCCAGCUUUAA
AGATAA AA
ACAAAUUCGGA ACAGG
UGUUUAAGCCU UGUCC
AGA CA
D.me ACAAATTCGGAAGATAAACAGGAA
D.an ACAAATTCGGAAGCTAAACAGGAA
D.ps ACAAATTCGGAAGTCAAACAGGAA
D.si ACAAATTCGGAAGATAAACAGGAA
D.ya ACAAATTCGGAATATAAACAGGAA
D.vi ACAAATTCGGAGGATAAACAGGAA
D.me ACAAATTCGGAGGATAAACAGGAA
T.ca ACAAGTTCGGAAC-TATACAGGGA
A.ga GCAAATTCGGACT-CGTACAGGGC
B.mo ACAAATTCGGAAT-TGTACAGGGA

Comparison of miRNA gene target prediction programs

Common set of rules:

1. Complementarity i.e. 5’end of miRNAs has more bases complementary to its target than the 3’end.
2. Free energy calculations i.e. G:U wobbles are less common in the 5’end of the miRNA:mRNA duplex
3. Evolutionary arguments i.e. targets site that are conserved across mammalian genomes.
4. Cooperativity of binding: many miRNAs can bind to one gene.

Results and differences

	3'UTR datasets	miRNA used	Cooperativity of binding	Statistical assessment (shuffling miRNA sequences)	Validation experiments	algorithm	Gene targets
TargetScan	14,300 Ensemble Conserved h/m/r	79	multiple target sites by same miRNA on a target gene	50% false positives	Direct validation by reporter constructs in cell line	7-nt seed sequence comp	400 conserved mammalian targets 107 conserved in Fugu
DIANA-microT	13,000 Ensemble Conserved m/h	94	Single sites	50% false positives	Direct validation by reporter constructs in cell line	Uses experimental evidence to extrapolate rules	5031 human targets. 222 conserved in mouse.
miRanda	29,785 Ensemble Conserved h/m/r	218	High score to multiple hits on same gene, even by multiple miRNA	50% false positives	Some agreement with exp detected target sites	ten 5' nt more important than ten 3' nt	4467 targets 240 conserved in both mammals and fugu

Summary of miRNA target prediction

- Each of the three methods, falls substantially short of capturing the full detail of physical, temporal, and spatial requirements of biologically significant miRNA–mRNA interaction.
- As such, the target lists remain largely unproven, but useful hypotheses.

Conclusions

- Computational methods can provide a useful complement to cloning, speed, cost.
- Candidates have to be verified experimentally.
    - Doubts about the validity of experimental evidence,
    - very little in vivo validation in which native levels of specific miRNAs are shown to interact with identified native mRNA targets.
    - What are the observable phenotypic consequences under normal physiological conditions.
- More biological inference. (*e.g. Argonautes facilitate miRNA:RISC complex*).
- Computational time and power have to be taken into consideration (use of clusters, parallelization)