

Computational Genomics

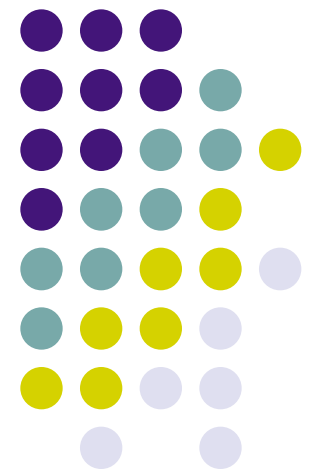
MSCBIO 2070

microRNAs:
deux ou trois choses que je sais d'elle

Takis Benos
Department of Computational & Systems Biology

February 29, 2016

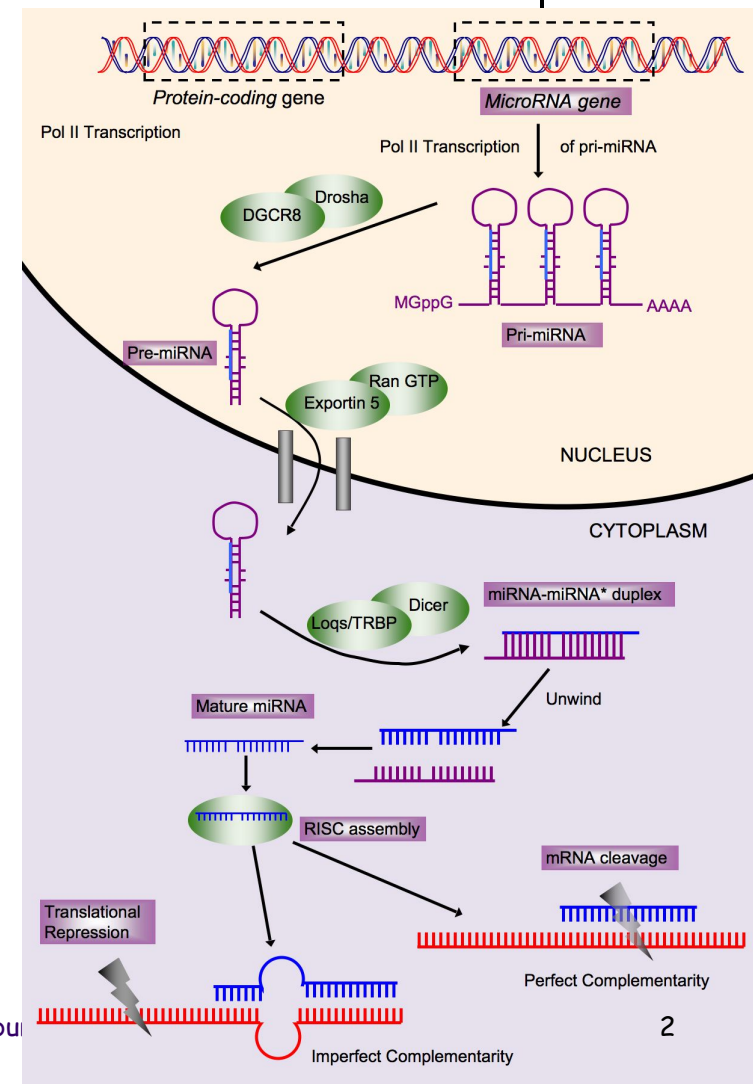
Reading: handouts & papers

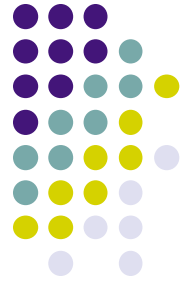


miRNA genes: a couple of things we know about them



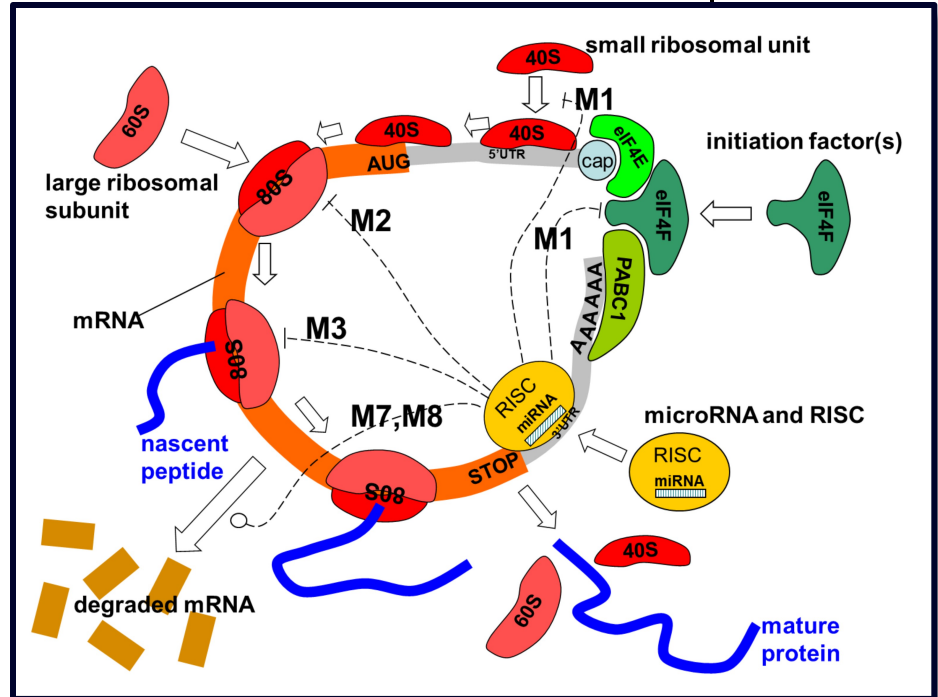
- Size
 - 60-80bp pre-miRNA
 - 20-24 nucleotides mature miRNA
- Role: translation regulation, cancer diagnosis
- Location: intergenic or intronic
- Regulation: pol II (mostly)
- They were discovered as part of RNAi gene silencing studies





microRNA mechanisms of action

- Translational inhibition in
 - Cap-40S initiation
 - 60S Ribosomal unit joining
 - Protein elongation
 - Premature ribosome drop off
 - Co-translational protein degradation
- mRNA and protein degradation
 - mRNA cleavage and decay
 - Protein degradation
 - Sequestration in P-bodies
- Transcriptional inhibition
 - (through chromatin reorganization)



What is RNA interference (RNAi) ?

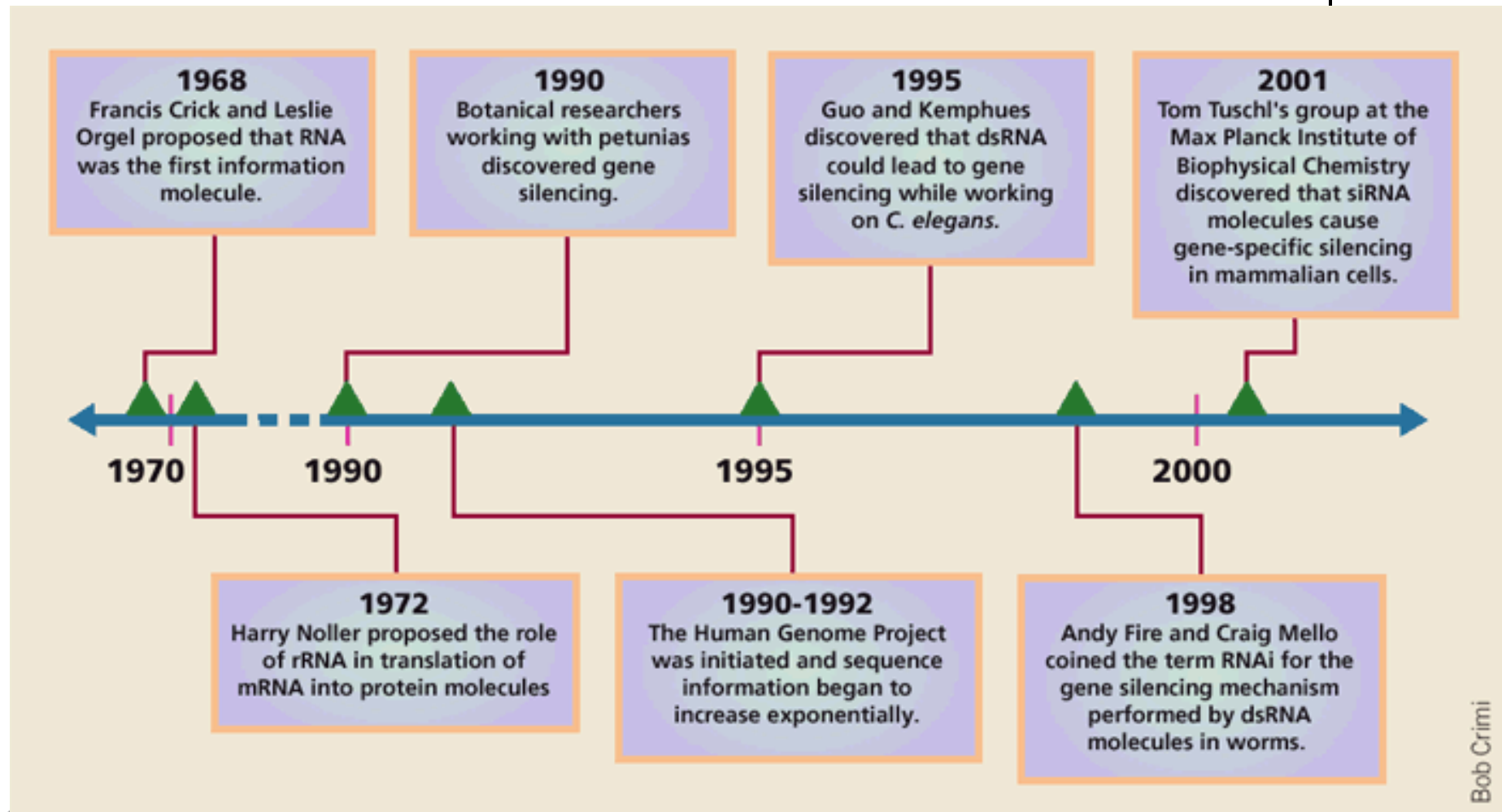


- RNAi is a cellular process by which the expression of genes is regulated at the mRNA level
- RNAi appeared under different names, until people realized it was the same process:
 - Co-suppression
 - Post-transcriptional gene silencing (PTGS)
 - Quelling





Timeline for RNAi Discoveries





From petunias to worms

- In the early 90's scientists tried to darken petunia's color by overexpressing the *chalcone synthetase* gene.

- The result:



Suppressed action of
chalcone synthetase

- In 1995, Guo and Kempthues used anti-sense RNA to *C. elegans* *par-1* gene to show they have cloned the correct gene.
 - Both sense and anti-sense *par-1* gene produced the same (mutant) phenotype. (Hmm! Hmmm! Hmmm!)
- Similar phenomena observed in fungus *N. crassa* and plant viruses
 - The phenomenon was shown to be post-transcriptional, but the mechanism remained unknown





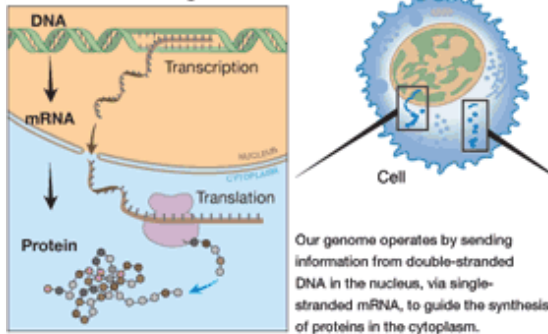
From petunias to worms (cntd)

- In 1998, Andy Fire and Craig Mello published something revolutionary.

RNA interference

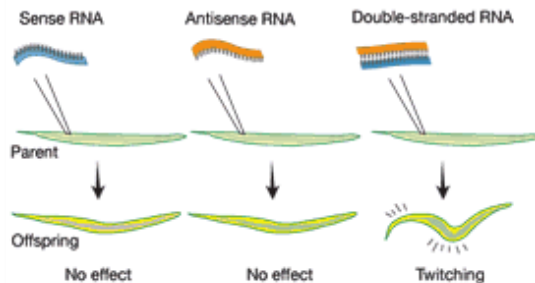
– gene silencing by double-stranded RNA

1. The central dogma



2. The experiment

RNA carrying the code for a muscle protein is injected into the worm *C. elegans*. Single-stranded RNA has no effect. But when double-stranded RNA is injected, the worm starts twitching in a similar way to worms carrying a defective gene for the muscle protein.



The Nobel Prize in Physiology or Medicine 2006
Andrew Z. Fire, Craig C. Mello

The Nobel Prize in Physiology or Medicine 2006

Nobel Prize Award Ceremony

Andrew Z. Fire

Craig C. Mello



Photo: L. Cicero

Andrew Z. Fire



Photo: J. Mottern

Craig C. Mello

The Nobel Prize in Physiology or Medicine 2006 was awarded jointly to Andrew Z. Fire and Craig C. Mello "for their discovery of RNA interference - gene silencing by double-stranded RNA"

Photos: Copyright © The Nobel Foundation



Plot thickens... the discovery of microRNAs (miRNAs)



lin-4 microRNA in worms (Ambros & Ruvkun '93)

- Non-coding, 22nt RNA
- lin-4 binds to sites in lin-14 3'UTR and negatively regulates lin-14 translation
- Not conserved outside the worm phyla
- **Yeah of course: Strange worms, right?**



Victor Ambros Ph.D.
University of Massachusetts
Medical School



Gary Ruvkun Ph.D.
Harvard Medical School
Mass. General Hospital

siRNAs/miRNAs in plants (Baulcombe '99)

Second miRNA - let-7 (Ruvkun '00)

- Non coding, 21nt RNA
- Highly Conserved
- Regulates lin-14 in same way as lin-4



David Baulcombe
University of Cambridge

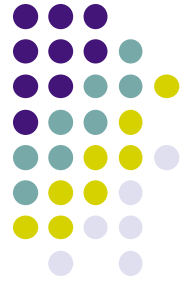


What is the difference between miRNA and siRNA?



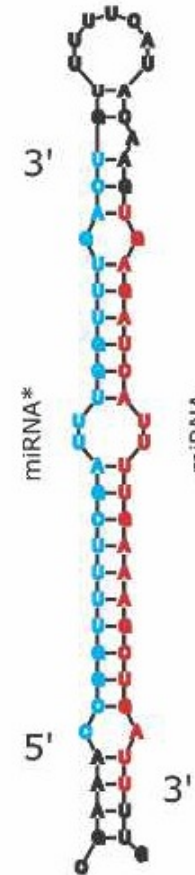
- Function of both species is regulation of gene expression
- Difference is in where they originate
- siRNA originates with dsRNA
- siRNA is most commonly a response to foreign RNA (usually viral) and is often 100% complementary to the target
- miRNA originates with ssRNA that forms a hairpin secondary structure
- miRNA regulates post-transcriptional gene expression and is often not 100% complementary to the target





microRNAs?

- RNA can fold like proteins: possess primary, secondary and tertiary structure
- Secondary hairpin structure crucial to processing of small RNAs



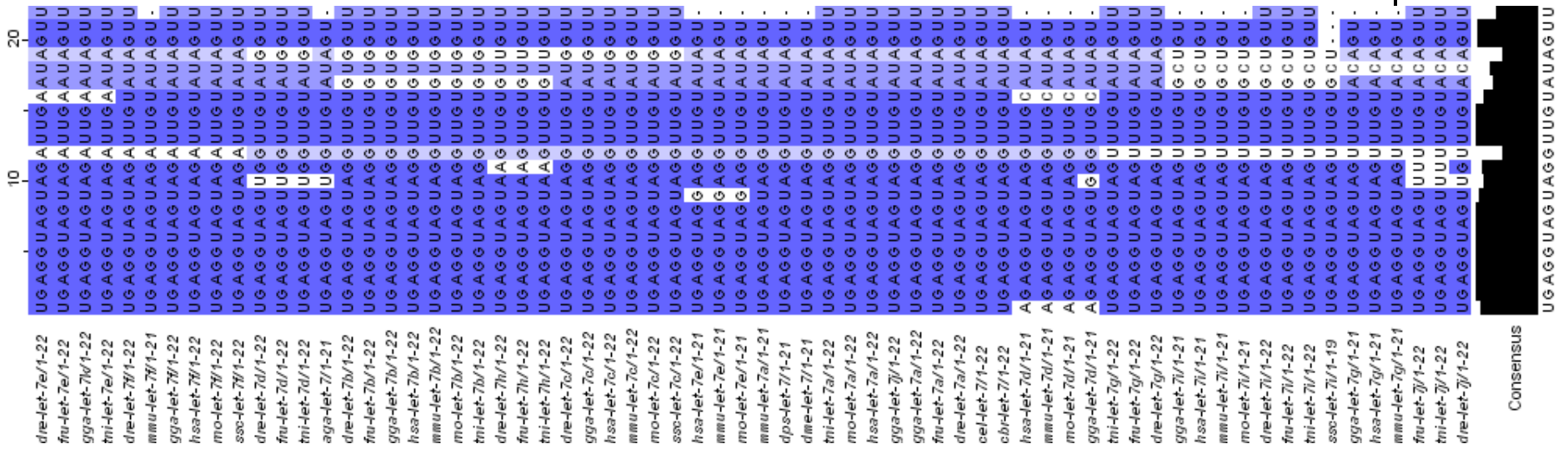
Stem-loop sequence MI0000070

Accession	MI0000070
ID	hsa-mir-16-1
Symbol	HGNC:MIR16-1
Description	Homo sapiens miR-16-1 stem-loop
Stem-loop	<pre> ag c - a c u gauu gucagc ugc u<u>agcagcac</u> gu <u>aaauugg</u> g uaa c caguug aug <u>agucgucgug</u> ca <u>uuugacc</u> c auu u ga a u a u u aaaa </pre> <p>Get sequence</p>
Comments	<p>Human miR-16 has been cloned by independent groups [1,2]. This precursor sequence maps to chromosome 13, and was nam reported 2 identical chromosome 13 loci, which appear to map to the same locus in subsequent genome assemblies. This gene region has been shown to be deleted in more than half of B cell chronic lymphocytic leukemias (CLL). Both miR-15a and miR-1 CLL cases [3]. A second putative mir-16 hairpin precursor is located on chromosome 3 (MI0000738).</p> <p><i>Coordinates (GRCh37)</i> 13: 50623109-50623197 [-]</p> <p><i>Overlapping transcripts</i> sense OTTHUMT00000044954; DLEU2-001; intron 3 OTTHUMT00000044957; DLEU2-004; intron 4 OTTHUMT00000044960; DLEU2-007; intron 4</p>

Two miRNAs from single precursor - 5p/3p, and * nomenclatures



Conservation across species of the second miRNA, let-7



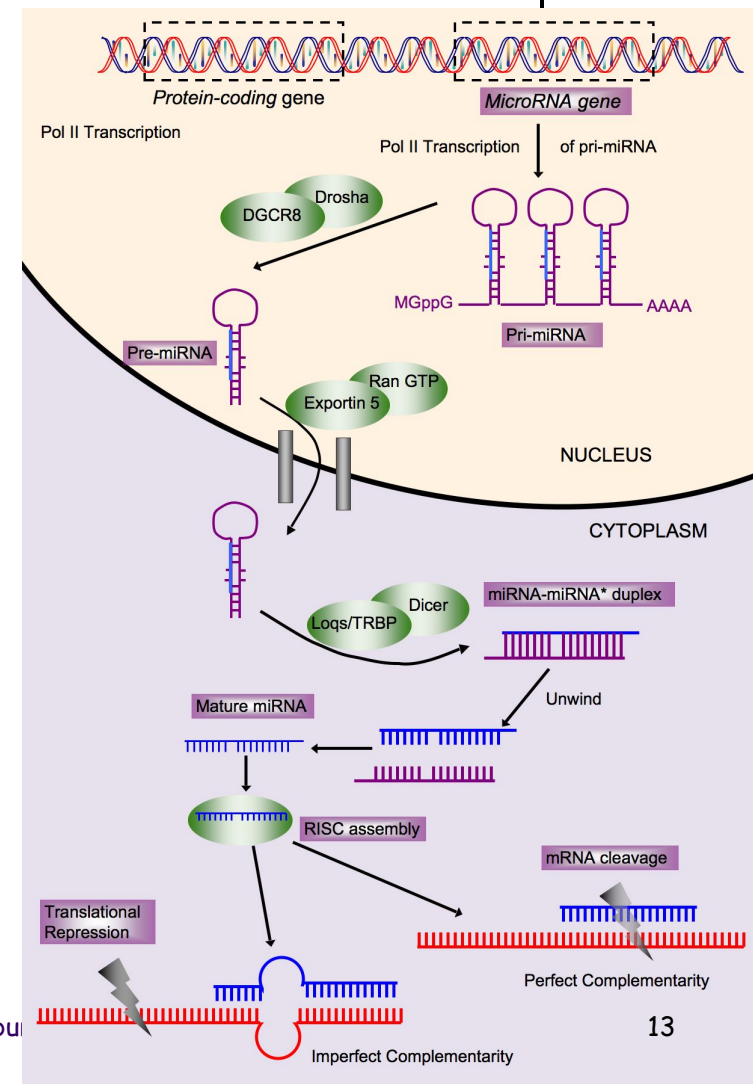
Observation that let-7 is highly conserved led to the "gold rush" of finding miRNAs, resulting in ~800 miRNAs in humans, and many other species including viruses.



miRNA genes: a couple of things we know about them

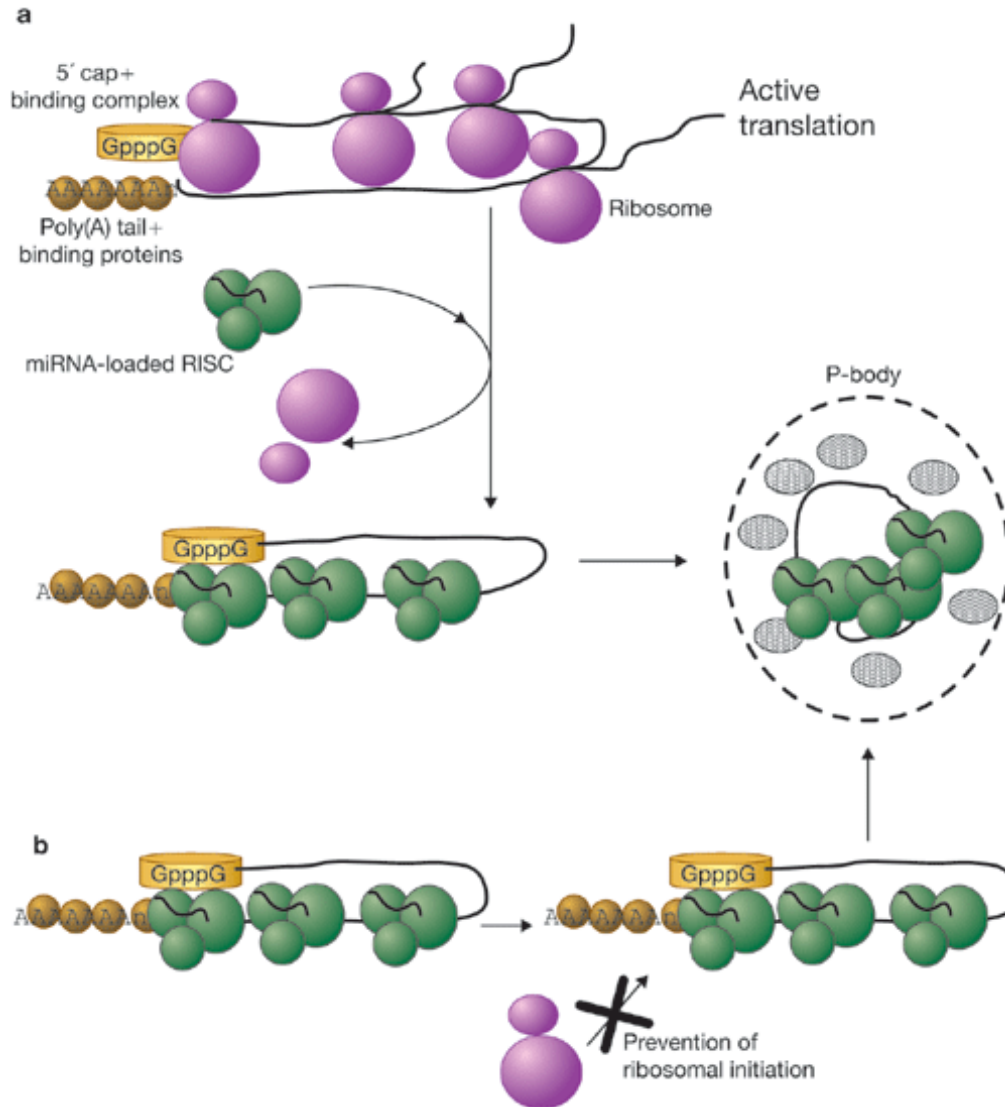


- Size
 - 60-80bp pre-miRNA
 - 20-24 nucleotides mature miRNA
- Role: translation regulation, cancer diagnosis
- Location: intergenic or intronic
- Regulation: pol II (mostly)

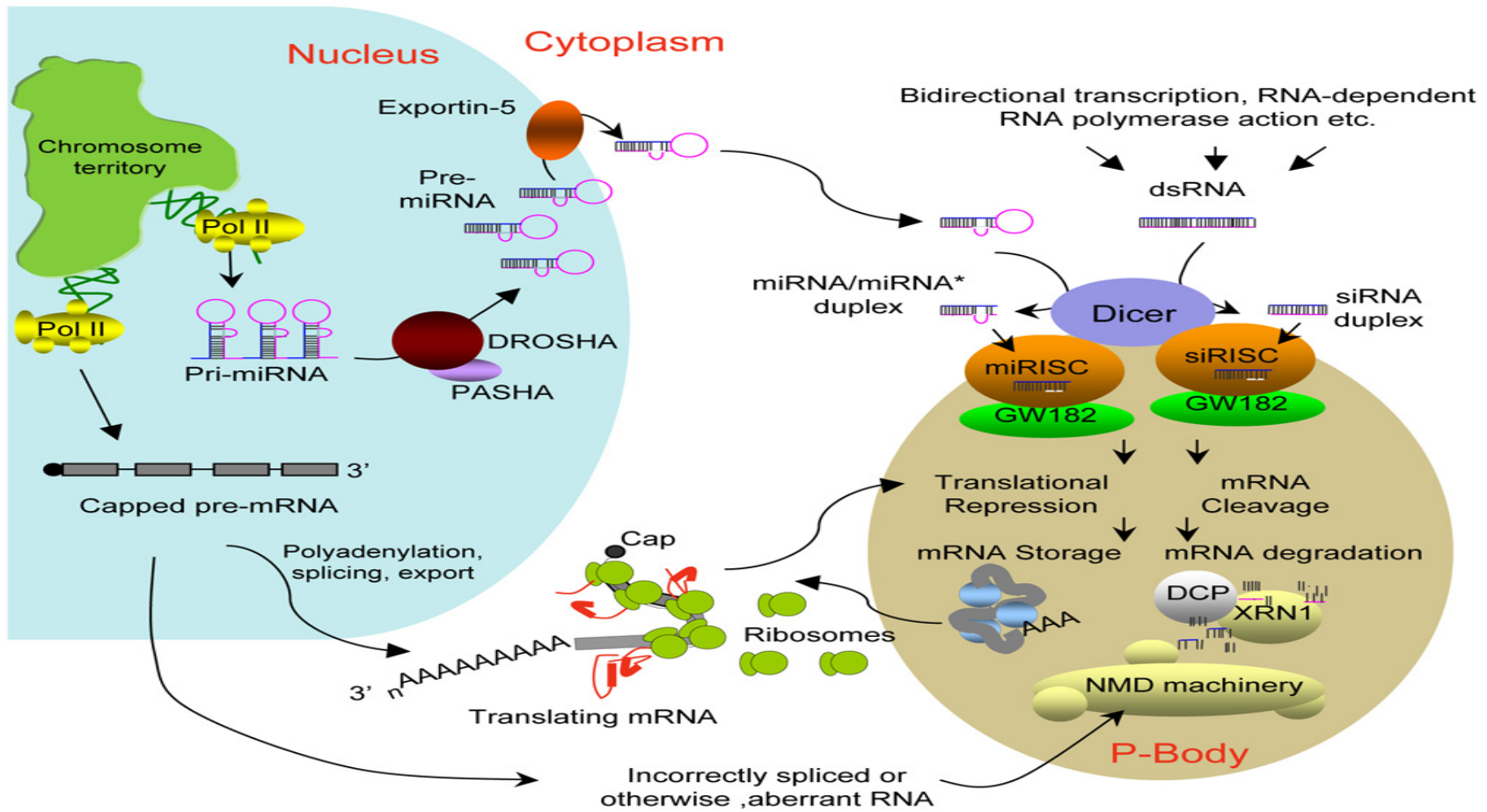


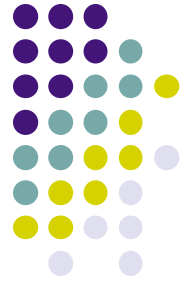


miRNA method of action



miRNA pathway





Summary of Players

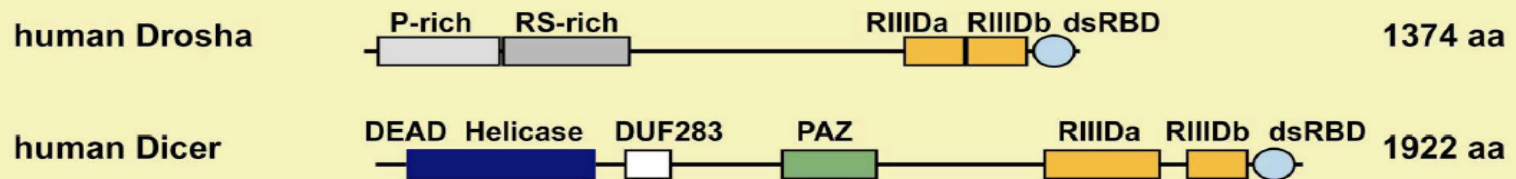
- **Drosha** and **Pasha** are part of the “Microprocessor” protein complex (~600-650kDa)
- Drosha and **Dicer** are RNase III enzymes
- Pasha is a dsRNA binding protein
- **Exportin 5** is a member of the karyopherin nucleocytoplasmic transport factors that requires Ran and GTP
- **Argonautes** are RNase H enzymes



Players



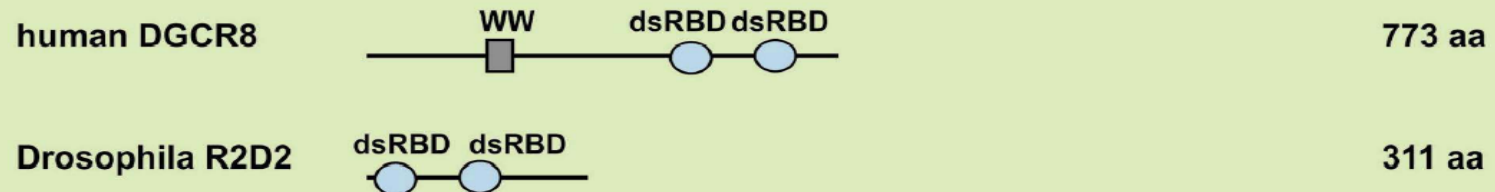
A. RNase III type proteins



B. Argonaute proteins



C. dsRNA-binding proteins



D. DEAD-box helicases





miRNA function: few examples

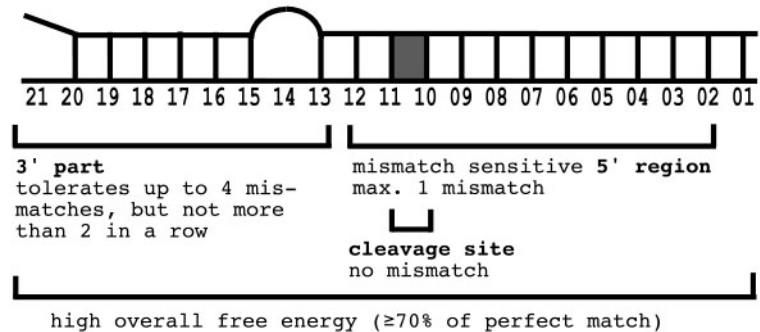
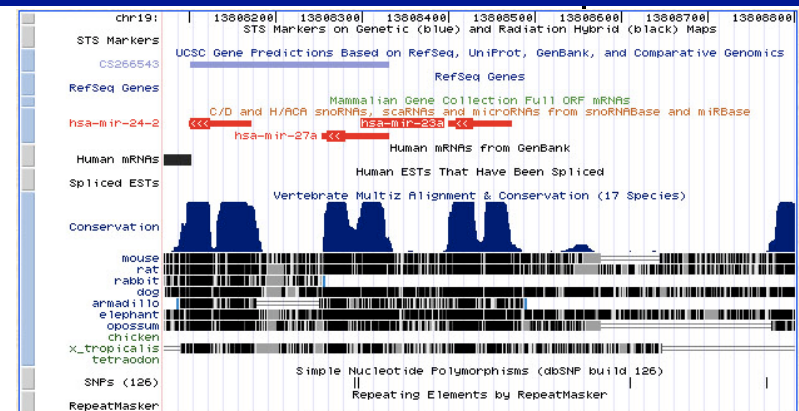
	miRNA	Target genes	Function
<i>C. elegans</i>	<i>lin-4</i>	<i>lin-14, lin-28</i>	Early Developmental timing
	<i>let-7</i>	<i>lin-41, hbl-1, daf-12,</i> ...	Late Developmental timing
	<i>lisy-6</i>	<i>cog1</i>	L/R neuronal symmetry
	<i>miR-273</i>	<i>die-1</i>	
<i>Drosophila</i>	<i>Bantam</i>	<i>hid</i>	Programmed cell death
Mouse	miR-196	<i>Hoxb8</i>	Developmental patterning
	miR-1	<i>Hand2</i>	Cardiomyocyte differentiation & proliferation



miRNA computational predictions



- miRNA gene prediction
 - miRNA features
 - Gene prediction methods
- miRNA target prediction
 - Physical characteristics
 - Target prediction methods



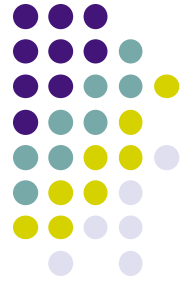
Computational methods to identify miRNA genes: Why?



- 800-1,000 human miRNA genes to date, thousands across species.
- However, experimental identification miRNAs is not easy:
 - low expression
 - stability
 - tissue specific
 - Expensive, and long cloning procedure
- Predicting miRNAs from genomic sequences provide a valuable alternative/support to cloning.

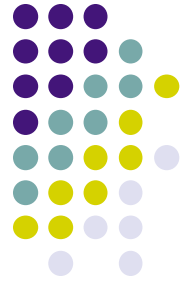


In the beginning, miRNA genes were identified...



- In the lab
 - Forward genetics: start from the mutant phenotype and look for the responsible gene
 - Very slow, inefficient (can only be applied to certain cases)
 - cDNA sequencing: size-fractionate RNA, clone, sequence
 - Slow, expensive
 - Deep sequencing of small RNAs (e.g., 454, Solexa)
 - Expensive, we do not know how many small RNA flavors exist
- *In silico* methods
 - Conservation-based
 - Clustering
 - SVMs

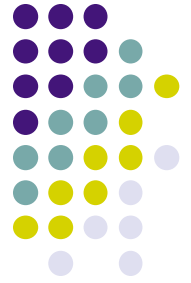




miRNA gene prediction

- Computational prediction
 - Structural features (e.g., hairpin length, thermodynamic stability, etc)
 - Sequence features (e.g., nucleotide content, location, etc)
 - Evolutionary conservation
- Methodologies
 - Neighbor stem loop searches (*identify closely located stem loops*)
 - Gene-finding (*identify conserved genomic regions, then run MFold*)
 - Homology search (*direct BLAST searches*)





miRNA gene prediction (cntd)

- Programs

- **miRseeker** (Lai *et al.* 2003): assesses folding patterns of RNA sequences conserved between two *Drosophila* species
- **MiRscan** (Lim *et al.* 2003): uses *RNAFold* to find hairpin structures in evolutionary conserved sequences (in worms)
- **Berezikov *et al.* (2005)**: uses *phylogenetic shadowing* together with other properties to identify miRNA genes
- **BayesmiRNAfind** (Yousef *et al.* (2006): uses *Naive Bayes* classifier with multi-species information
- **Kadri *et al.* (2009)**: uses *hierarchical HMM* with no evolutionary information



miRNA prediction - Initial methods



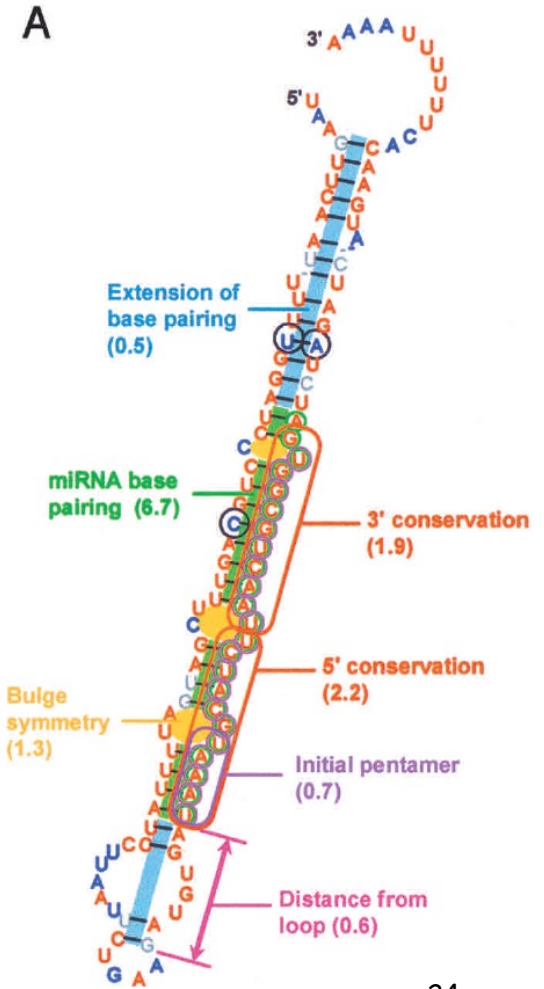
MiRscan -
find conserved hairpin structures in miRNAs

$$S_i(x_i) = \log_2 \frac{f_i(x_i)}{g_i(x_i)} ;$$

$$S = \sum_{i=1}^7 S_i(x_i)$$

f=freq. of feature in miRNA training data set (50 miRs);
g= Similar freq. in random (~36K random hairpins)

Lim et al, Genes and Development 2003

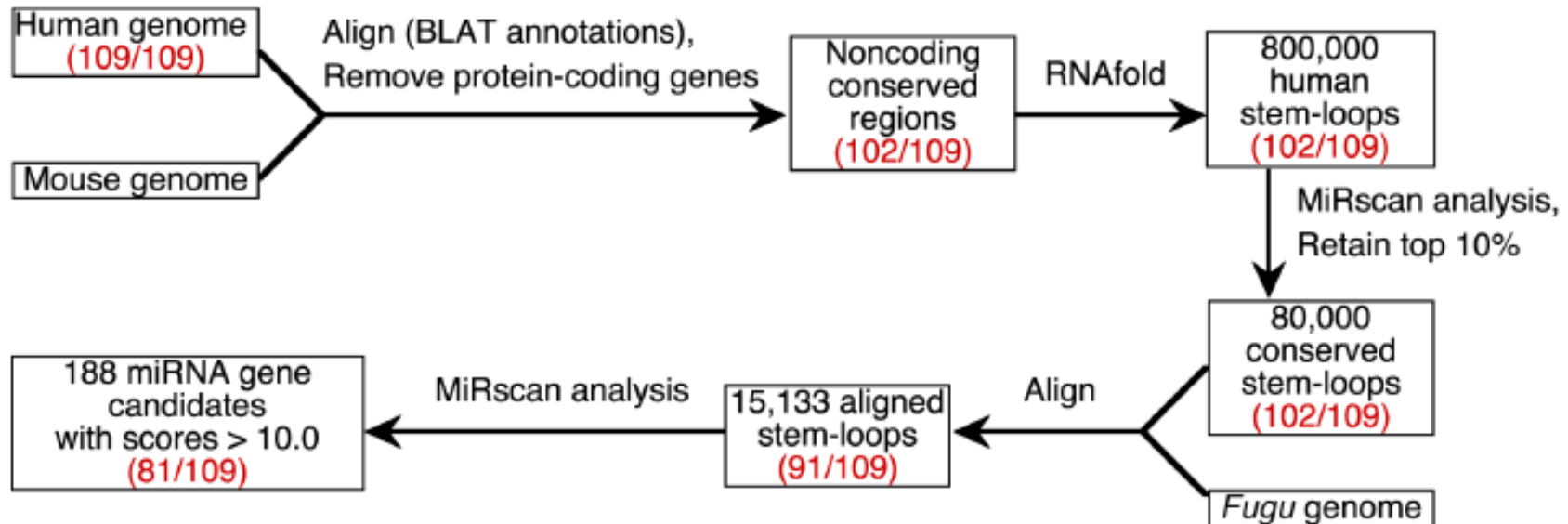




miRScan

MiRscan

- 81 of 109 (at that time) known miRNAs identified by *de novo* approach alone

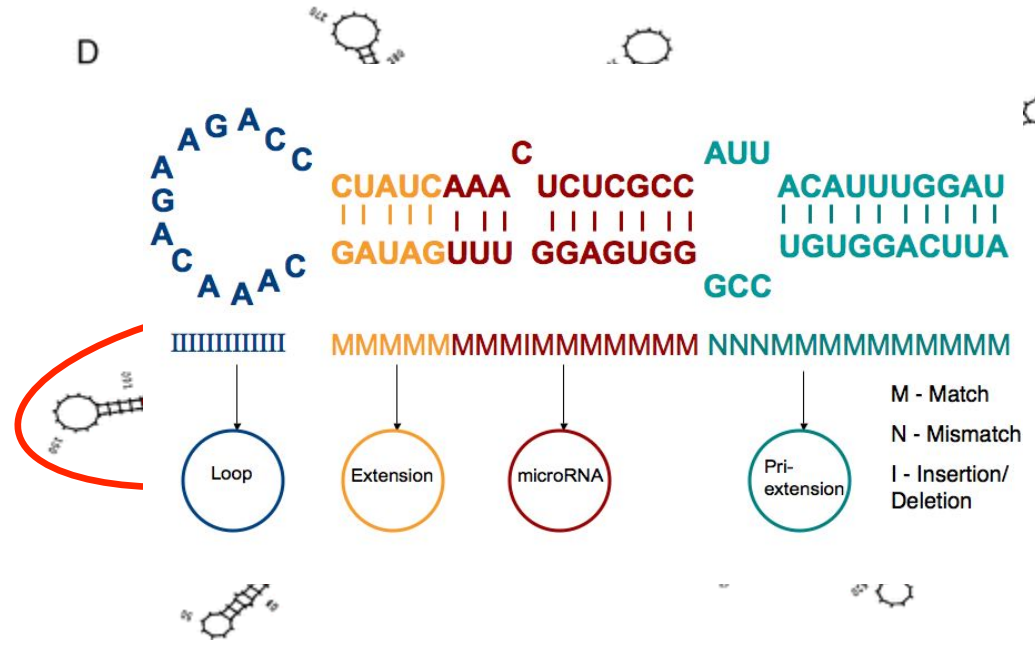
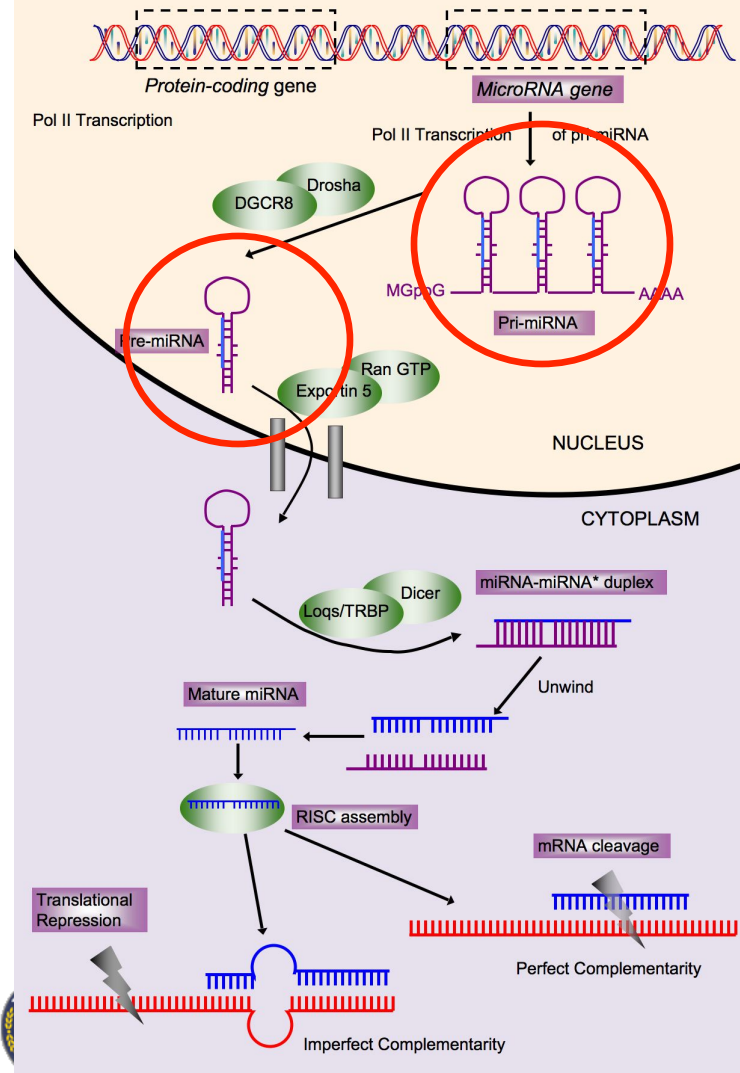


Note: Sequence conservation patterns in other related genomes are very powerful - Comparative genomics





miRNA biogenesis: stemloops

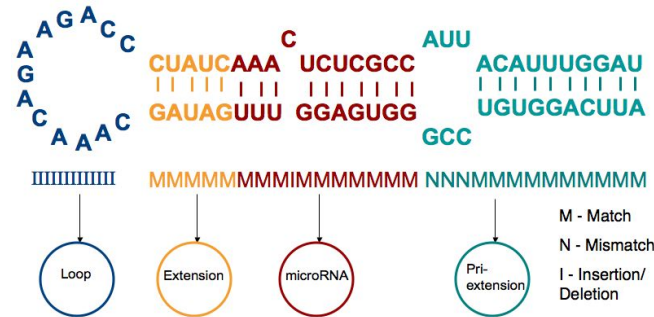


Millar AA, Waterhouse PM (2005) *Funct Integr Genomics* 5:129-135





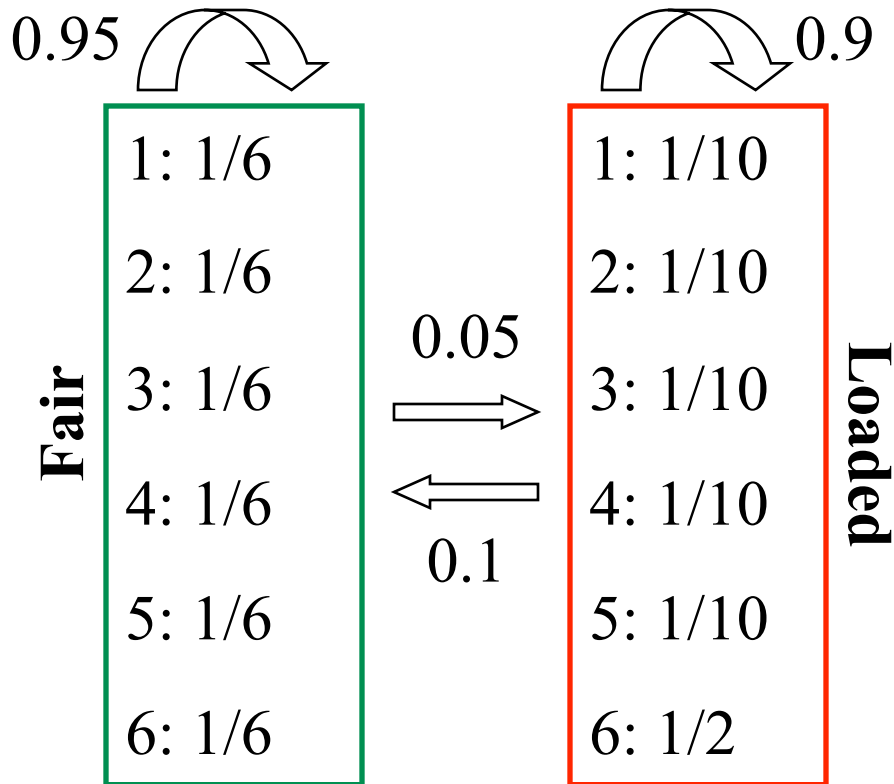
Stemloop characteristics (species)



	Hairpin (bases)	Loop (bases)	Extension (bp mostly)	miRNA (bp mostly)	Pri-miR ext (bp mostly)
Mean (SD)					
Vertebrates	86.7 (13.8)	7.3 (3.5)	5.0 (3.4)	22.0 (0.9)	12.6 (7.0)
Invertebrates	91.8 (13.1)	7.9 (3.9)	5.8 (4.5)	22.2 (1.3)	13.8 (5.9)
Plants	119.5 (43.2)	6.8 (3.7)	22.8 (18.5)	21.3 (1.0)	12.5 (9.9)
Min - Max					
Vertebrates	55 - 153	3 - 22	0 - 34	16 - 26	0 - 50
Invertebrates	54 - 215	3 - 30	0 - 55	18 - 28	0 - 32
Plants	57 - 337	3 - 35	0 - 102	16 - 24	0 - 78

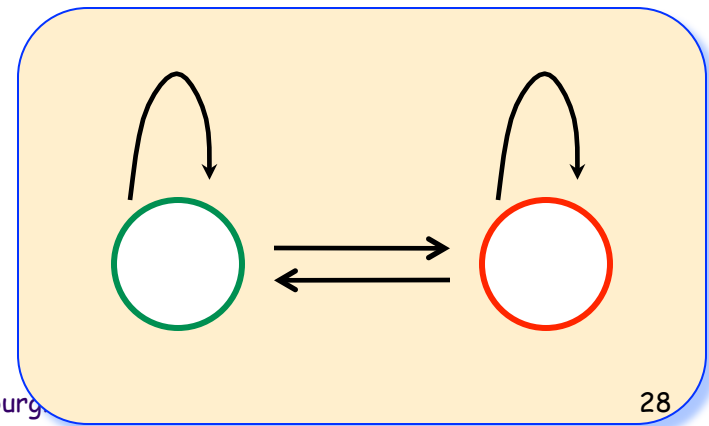


HMM example: the dishonest casino



Classification Problem
Given the model, parameters and a set of observations can we determine if they come from the *fair* or the *loaded* dice?

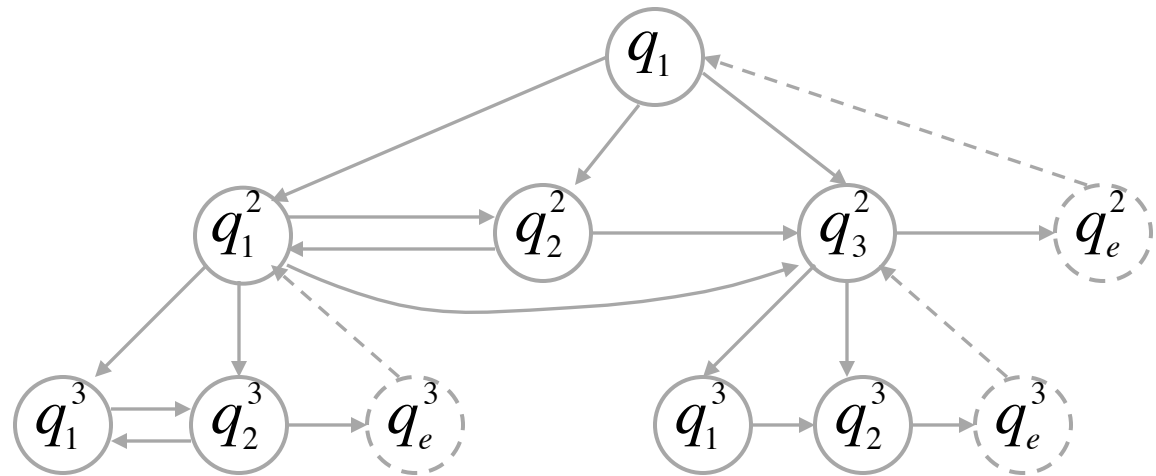
Q: what is hidden?





Hierarchical hidden Markov models

- Internal States
- Production States
- End States
- Parameter Set λ



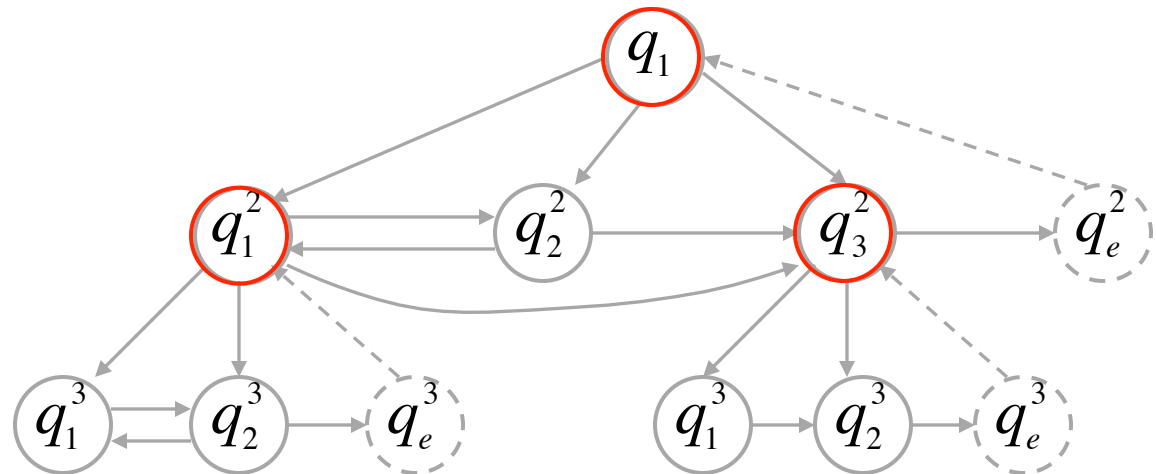
Fine et al., 1998; Machine Learning, 32, 41-62





Hierarchical hidden Markov models

- Internal States
- Production States
- End States
- Parameter Set λ



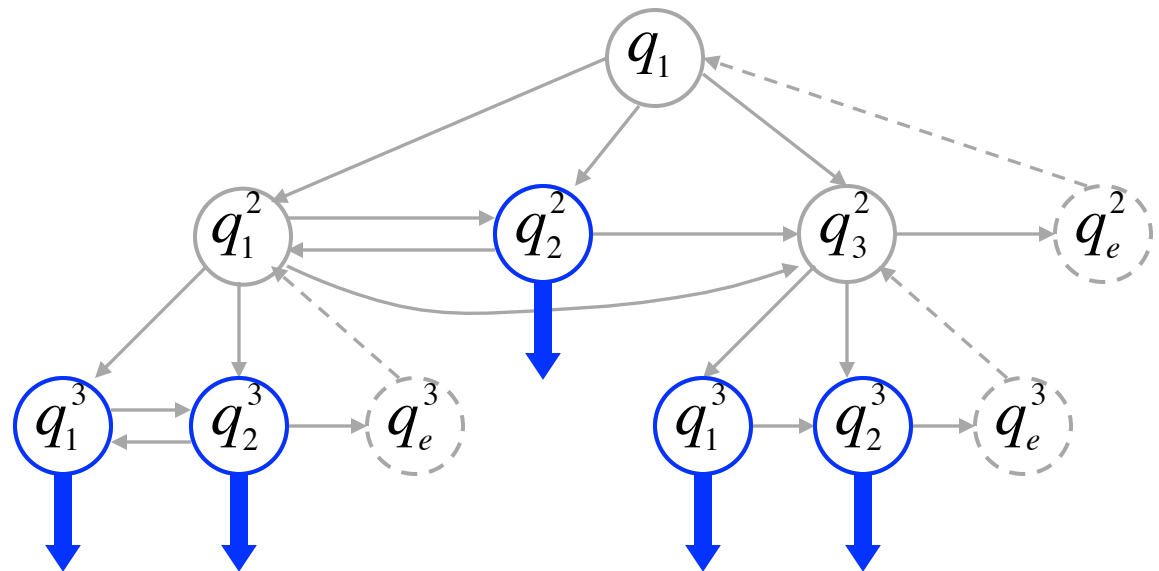
Fine et al., 1998; Machine Learning, 32, 41-62





Hierarchical hidden Markov models

- Internal States
- Production States
- End States
- Parameter Set λ



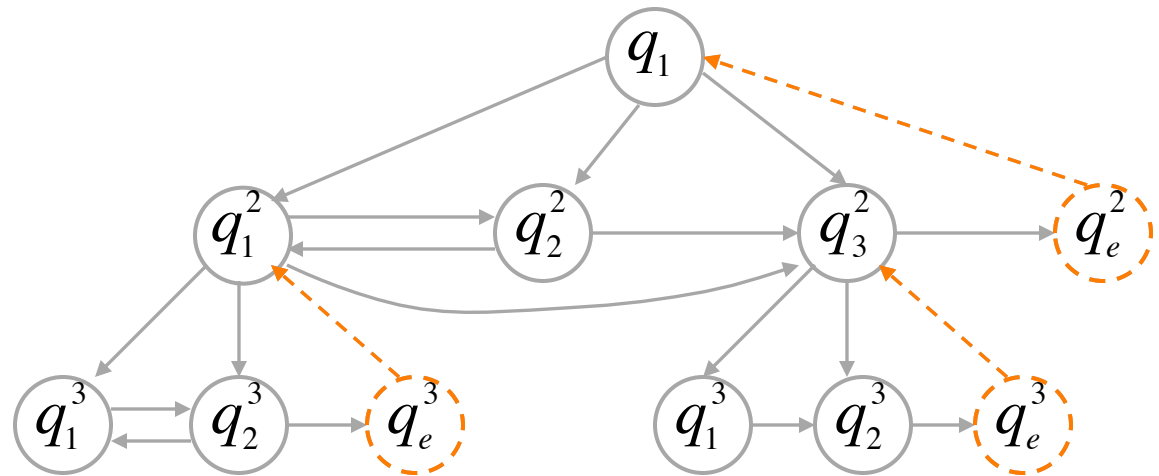
Fine et al., 1998; Machine Learning, 32, 41-62





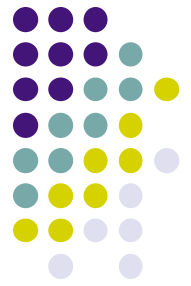
Hierarchical hidden Markov models

- Internal States
- Production States
- End States
- Parameter Set λ



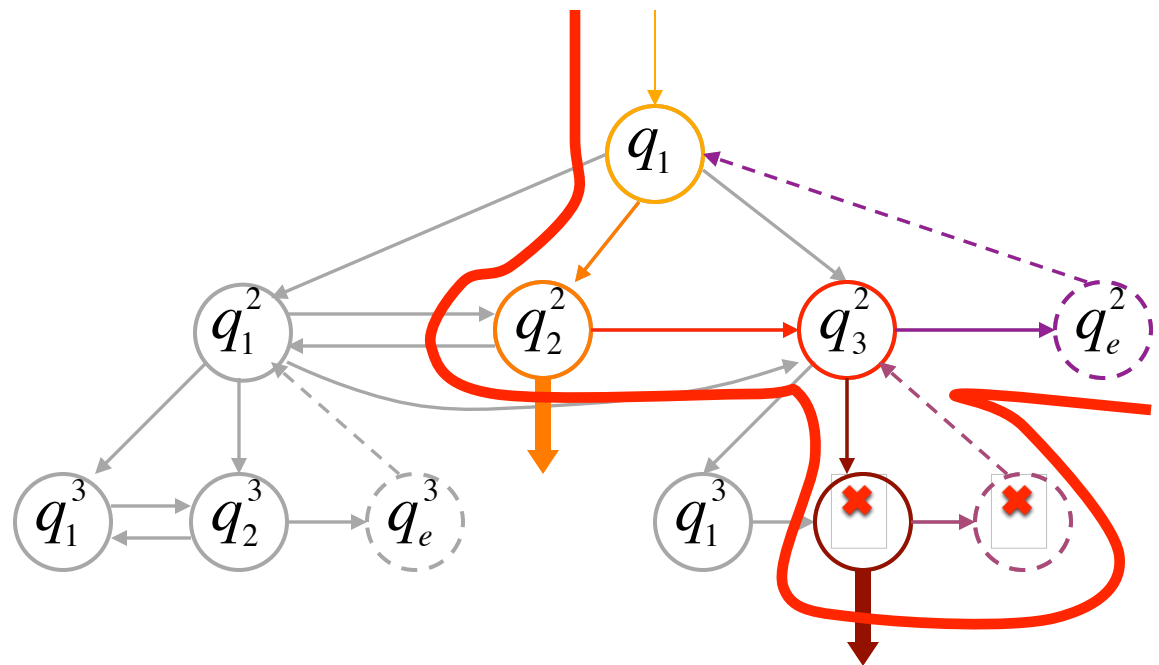
Fine et al., 1998; Machine Learning, 32, 41-62





Hierarchical hidden Markov models

- Internal States
- Production States
- End States
- Parameter Set λ



Fine et al., 1998; Machine Learning, 32, 41-62





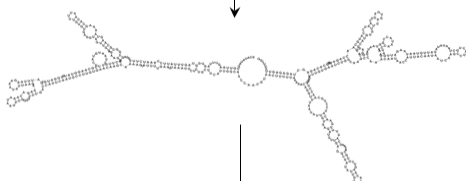
Datasets

- Positive examples
 - miRNA registry version 10.1 (December, 2007)

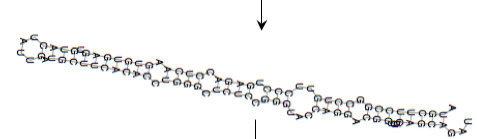
```

CCAUAUAGUGUUAAGUUUAGUGGUAGUUUUAUCCAAUCUUUGGAGCG
GCGCAAGUAAAAGCACAAUUUUGUGUGAGGUAGUUUGAAGUAGCUU
GUCUAAUUGUACUUGGCAAGGUAUGUGGUAUGGUAUUAAGAAUUAUAA
AAAGACAAUUAUUUGGGGUUGACCAUUCAGUCCUUGUGAACAAGCA
AAAGUUAUUUGAAUUAUUUCUGUUAUUGGCAACCAACGUAACUU
UUUAAAAACAGGUUUAUUAUUGUGAGAUUUUGAGGCAUUAUUGGACAAA
GUCUUUGCAAAUCAUUUUGAACAUCCUGCAACCGGUAUUUGUGAAUCU
GGCAAGCCAAACAUAACCAUUAACAUAUUAAGCCUACUGCUAAUUAU
  
```

Genome folding



Hairpin Extraction



Pre-processing

```

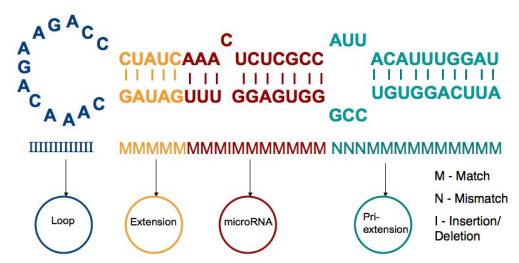
>mir-xyz
GCAAGCCCAUGAAGAAGAGUAAGAAUAGAGGAAAGGAGAAGGAUGAGAGACAGG
----CC----UUUCUUGUC-UUCUU--CUUCUUUCCCCUCUCUUC-CUCAUAUC
  
```

AND

```

miR-Mapping & Labeling
(For training only)

>mir-xyz
GCAAGCCCAUGAAGAAGAGUAAGAAUAGAGGAAAGGAGAAGGAUGAGAGACAGG
----CC----UUUCUUGUC-UUCUU--CUUCUUUCCCCUCUCUUC-CUCAUAUC
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
  
```



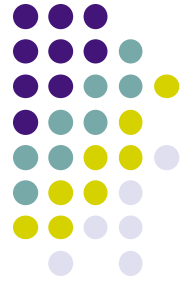


Datasets

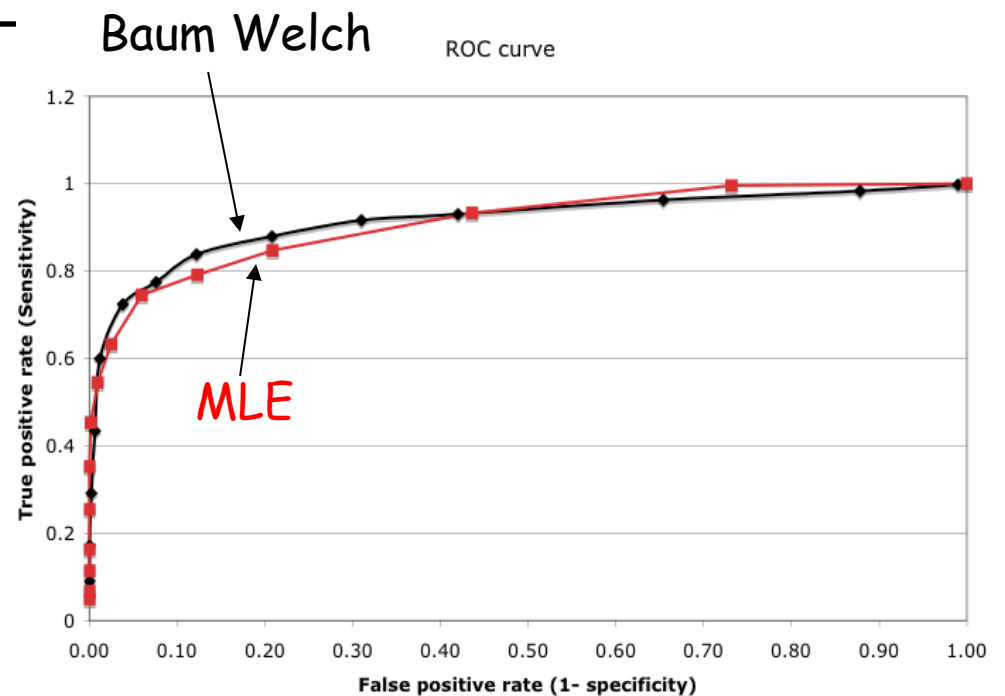
- Positive examples
 - miRNA registry version 10.1 (December, 2007)
- Negative examples
 - Folded hairpins derived from coding regions
- Alphabet
 - Match: $M = \{AU, GC, GU\}$
 - Mismatch: $N = \{AG, AC, CU, AA, CC, GG, UU\}$
 - Indel: $I = \{A-, C-, G-, U-\}$



Parameter estimation: Baum-Welch vs. MLE



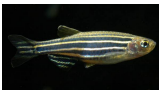

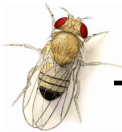




	Sn (SD)	FDR (SD)
Baum-Welch	0.84 (0.19)	0.12 (0.06)
MLE	0.74 (0.14)	0.16 (0.08)



Performance of HHMMiR across species (trained on human data)



Organism	Known hairpins	% predicted
  → <i>M. musculus</i>	422	74.7
<i>G. gallus</i>	147	89.1
 → <i>D. rerio</i>	334	88.3
 → <i>C. elegans</i>	131	85.5
 → <i>D. melanogaster</i>	143	93.0
 → <i>A. thaliana</i>	114	97.4
 → <i>O. sativa</i>	188	85.7
Total	1,479	85.1



Comparison of HHMMiR to tripletSVM



Test set	Known hairpins (at the time)	tripletSVM (%)	HHMMiR (%)
New human hairpins in registry at the time	39	92.3	97.4
<i>M. musculus</i>	36	94.4	88.9
<i>R. norvegicus</i>	25	80.0	84.0
<i>G. gallus</i>	13	84.6	100
<i>D. rerio</i>	6	66.7	100
<i>C. elegans</i>	110	86.4	90.9
<i>C. briggsae</i>	73	95.9	95.9
<i>D. melanogaster</i>	71	91.6	95.8
<i>D. pseudoobscura</i>	71	90.1	98.6
<i>A. thaliana</i>	75	92.0	97.3
<i>O. sativa</i>	96	94.8	86.5
Epstein Barr virus	5	100	80.0
TOTAL	620	91	93.2





To summarize...

Ab initio miRNA stemloop prediction

- The fundamental miRNA characteristics are similar between very diverse taxa (vertebrates, invertebrates, plants)
- **HHMMiR**: first HMM-based approach for classification of microRNA precursors
 - HHMiR classifies known miRNA genes from distant species with high accuracy
 - HHMMiR uses structural and sequence characteristics of distinct regions of the miRNA precursors



miRDeep: taking advantage of sequence read number



Nature Biotechnology **26**, 407 - 415 (2008)
doi:10.1038/nbt1394

Discovering microRNAs from deep sequencing data using miRDeep

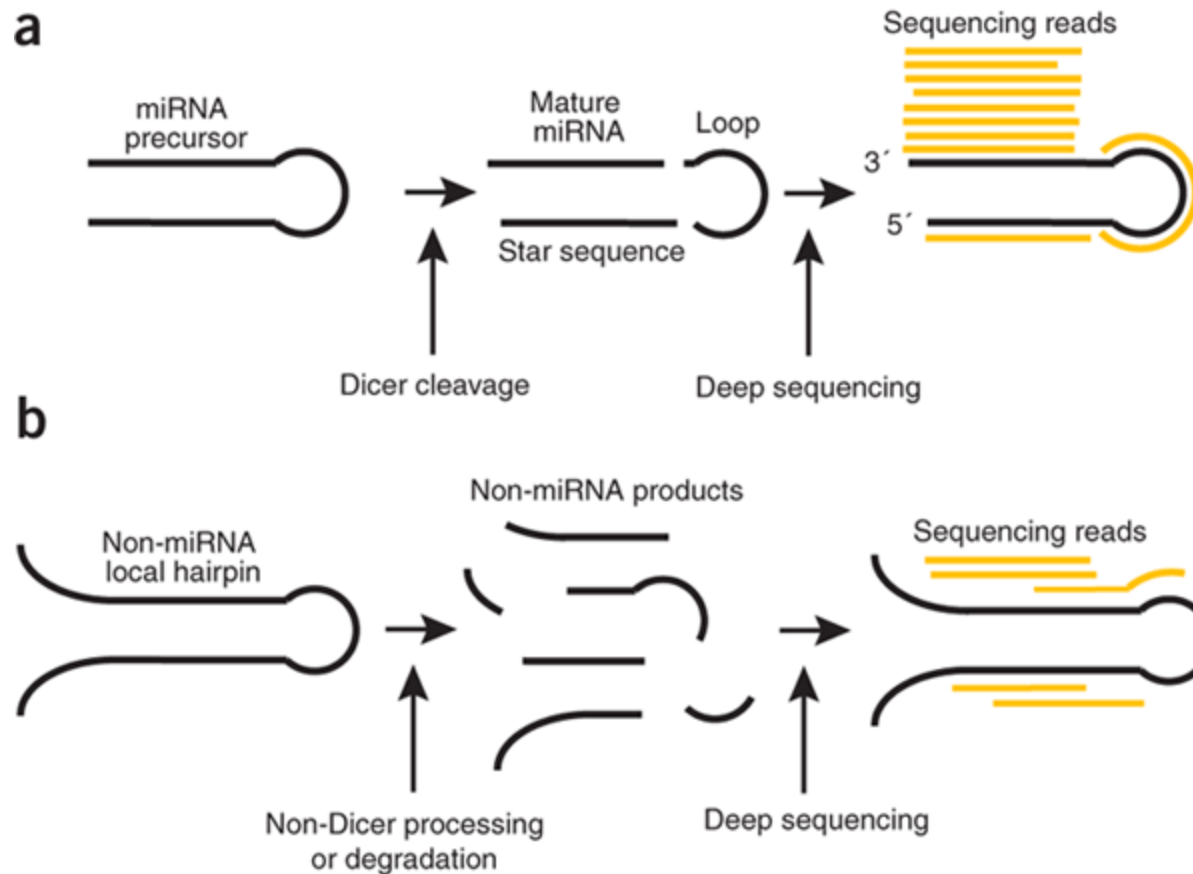
Marc R Friedländer¹, Wei Chen², Catherine Adamidi¹, Jonas Maaskola¹, Ralf Einspanier³, Signe Knespel¹ & Nikolaus Rajewsky¹

The capacity of highly parallel sequencing technologies to detect small RNAs at unprecedented depth suggests their value in systematically identifying microRNAs (miRNAs). However, the identification of miRNAs from the large pool of sequenced transcripts from a single deep sequencing run remains a major challenge. Here, we present an algorithm, miRDeep, which uses a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. We demonstrate its accuracy and robustness using published *Caenorhabditis elegans* data and data we generated by deep sequencing human and dog RNAs. miRDeep reports altogether ~230 previously unannotated miRNAs, of which four novel *C. elegans* miRNAs are validated by northern blot analysis.

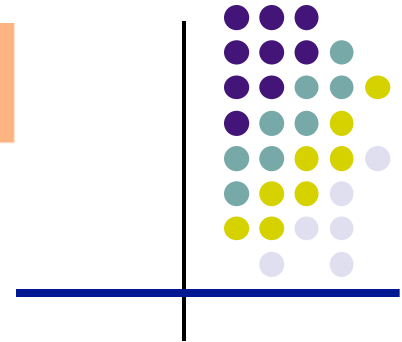
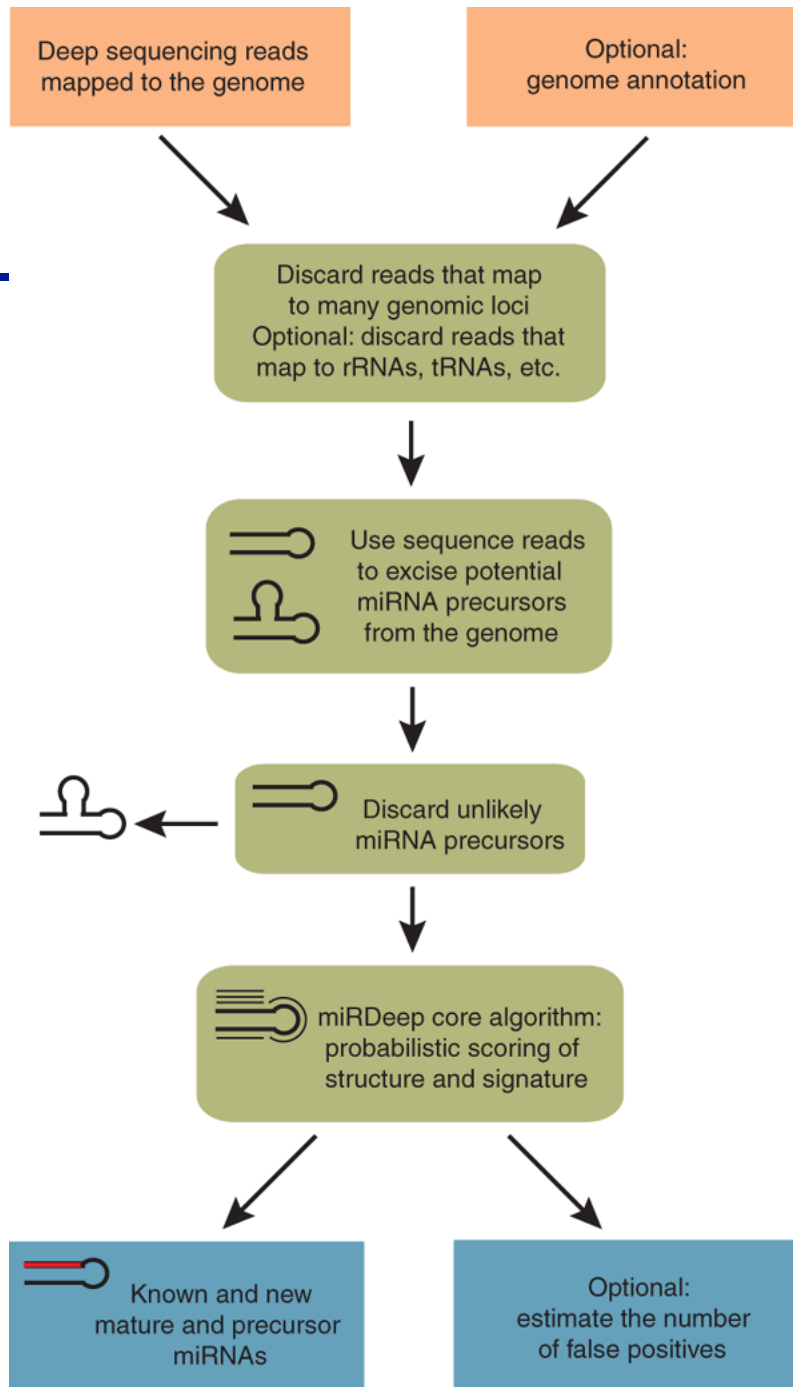




miRDeep: the idea



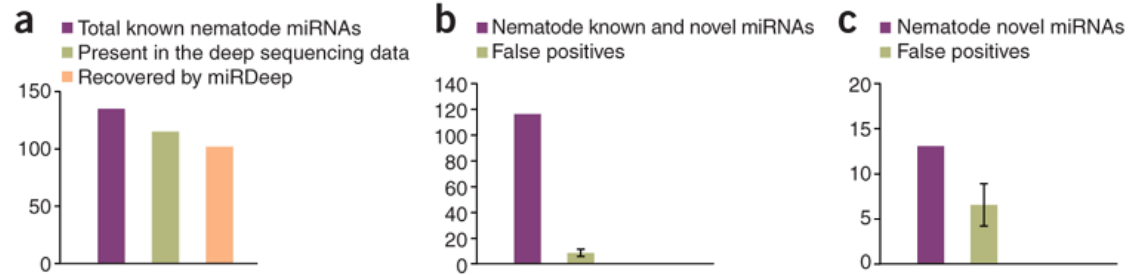
miRDeep: the pipeline



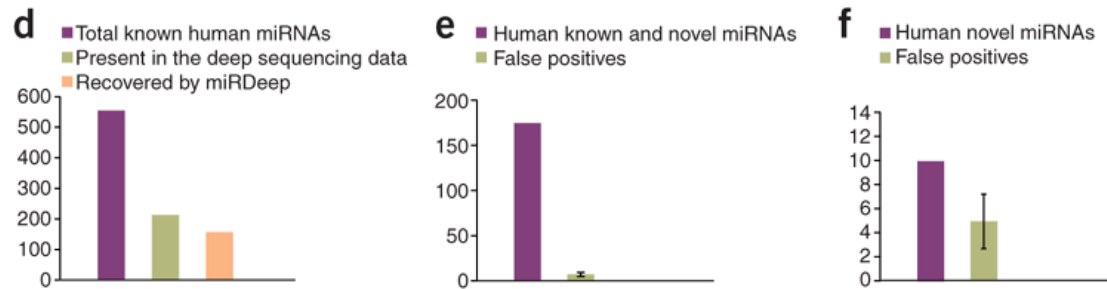


miRDeep: some results

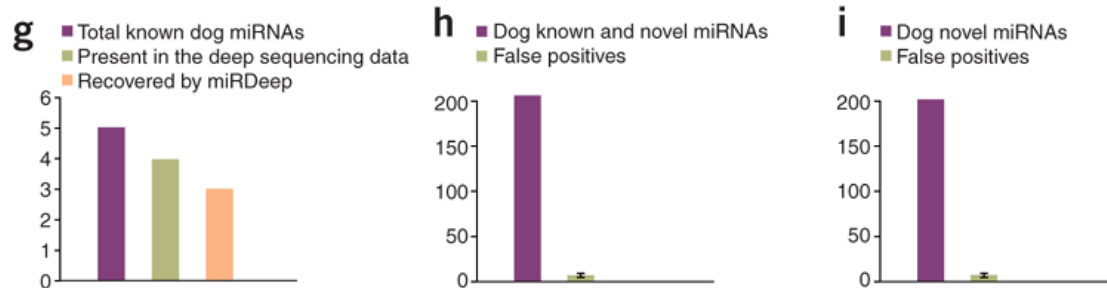
Nematode



Human

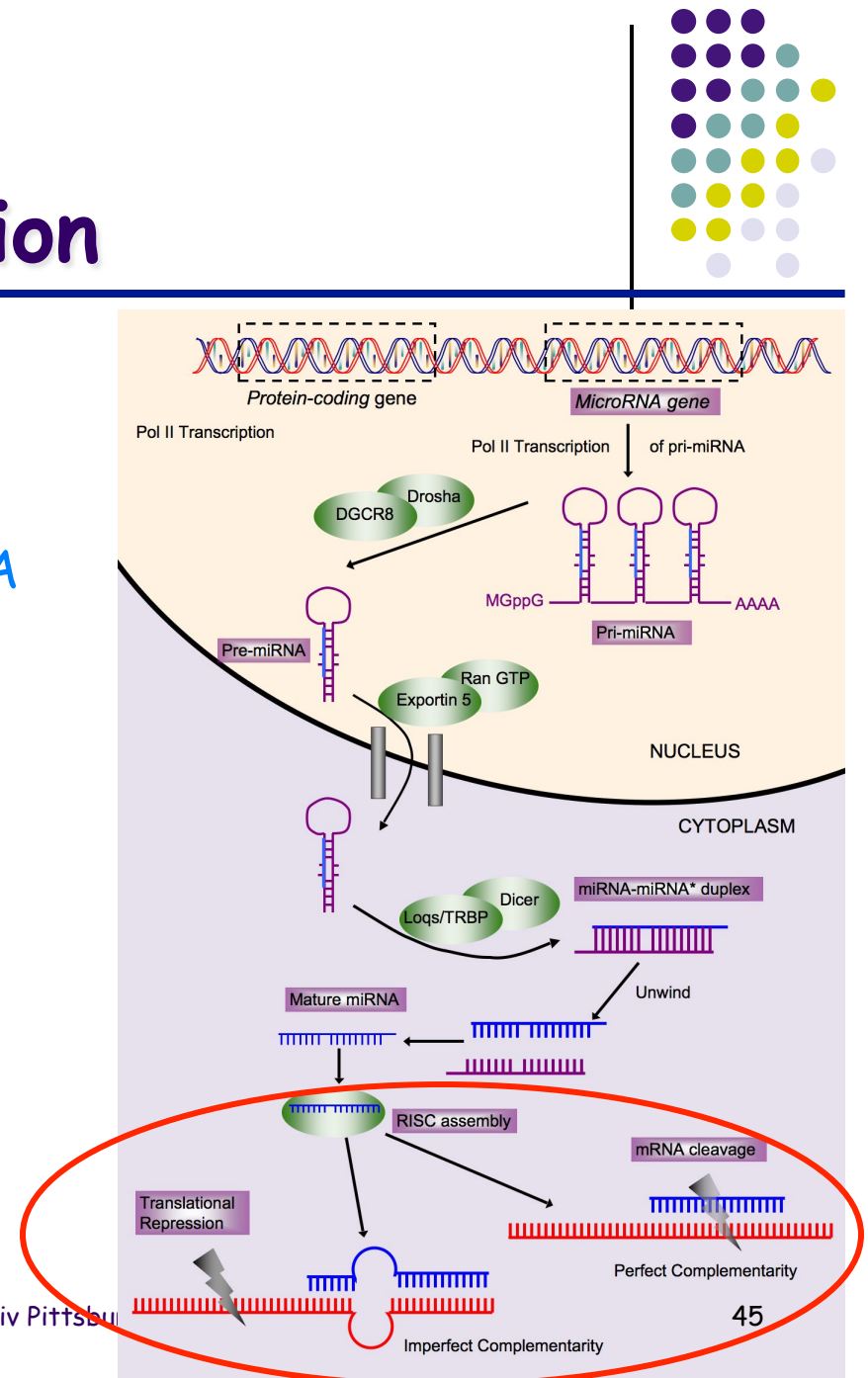


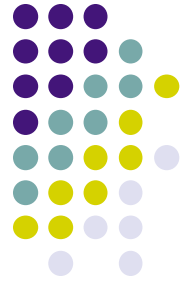
Dog



miRNA target prediction

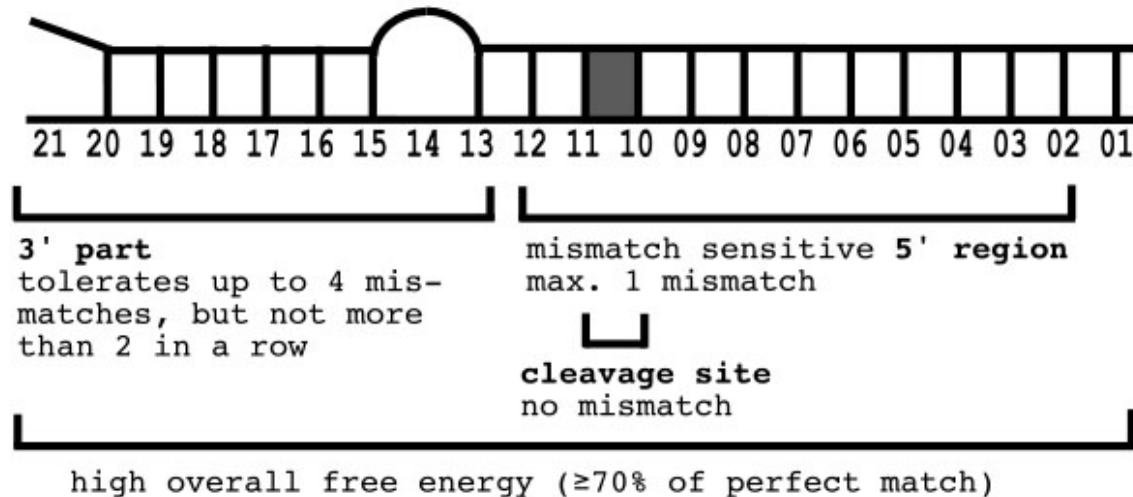
- Size
 - 60-80bp pre-miRNA
 - 20-24 nucleotides mature miRNA
- Role: translation regulation, cancer diagnosis
- Location: intergenic or intronic
- Regulation: pol II (mostly)

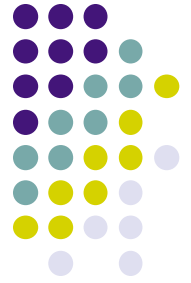




miRNA target prediction

- Physical characteristics
 - 5' end “seed” conservation (6-8 nt long)
 - Compensatory 3' end (to increase miRNA stability/efficiency)
 - Multiple target sites: are they important to have?
 - Structure of the target sequence



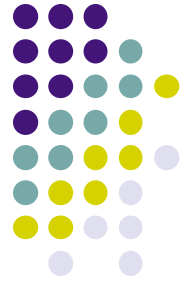


miRNA target prediction (cntd)

- Programs

- **Stark *et al.* (2003)**: detecting base complementarity on the 5' -end 8 nt seed w/ evolutionary conservation → *MFold* to calculate stability
- **RNAHybrid (Rehmsmeier *et al.* 2004)**: new RNA folding algorithm; uses only 6 nt at the 5' -end seed (nts 2-7)
- **TargetScan (Lewis *et al.* 2003, 2005)**: uses only 7 nt at the 5' -end seed → *RNAFold* to calculate binding energy
- **DIANA-MicroT (Kyriakidou *et al.* 2004)**: focuses on single target sites; seeks targets w/ central “bulge” and 3' complement





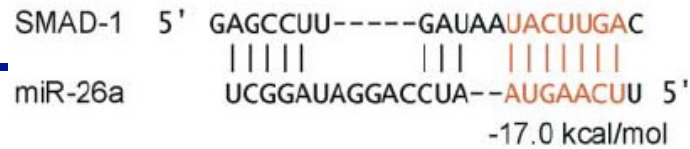
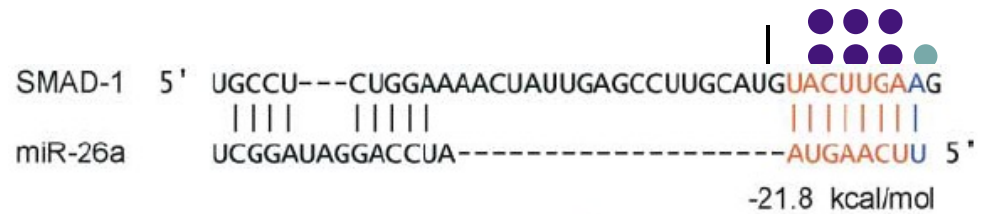
miRNA target prediction (cntd)

- Programs (cntd)
 - **miRanda** (Enright *et al.* 2005): uses weight matrices to emphasize 5' -end binding → *RNAFold* to calculate binding energy
 - **Xie *et al.*** (2005): whole genome conservation scan identified a large class of 8 nt motifs (not a formal miRNA finder)
 - **rna22** (Miranda *et al.* 2006): seeks overrepresented motifs in 3' UTR of the genes → *Vienna package* to calculate binding energy

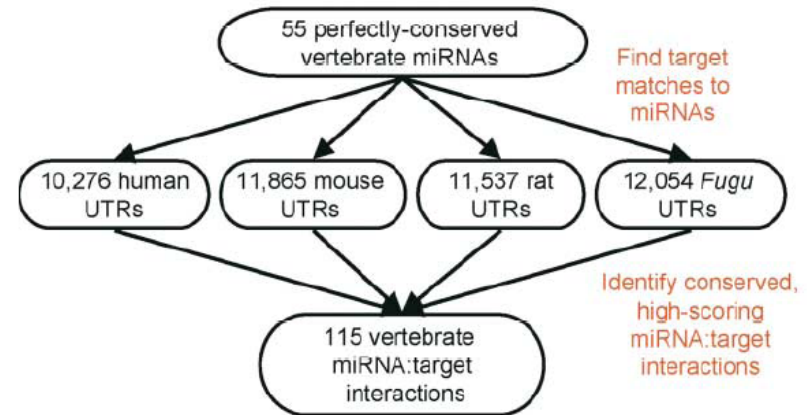


TargetScan - Initial methods

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 seeds to find the best energy
3. Assign a score, Z .
4. Give a rank (R_i) according to that species.
5. Repeat above process.
6. Keep those genes for which $Z_i > Z_c$ and $R_i < R_c$.



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



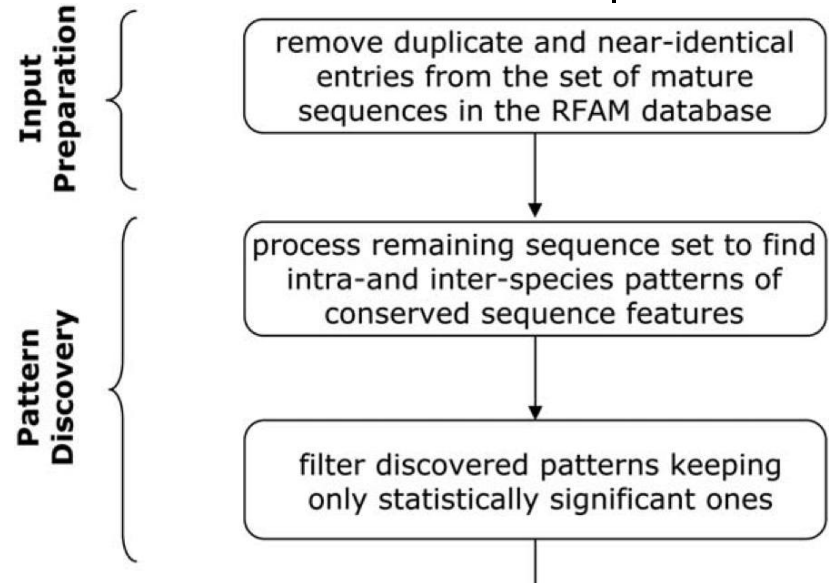
C	0	100	200	300	400	500	600	700	Z	Rank
Hs	[Progressive bars showing increasing Z score]								5.3	45
Mm	[Progressive bars showing increasing Z score]								4.8	72
Rn	[Progressive bars showing increasing Z score]								4.9	76
Fr	[Progressive bars showing increasing Z score]								5.2	16



rna22: a different strategy



- **Start:** 644 mature miRNA sequences (2004 version of RFAM)
- **End:** 354 sequences with $\leq 90\%$ identity (*training set*)
- Pattern identification: *Teiresias* (on the training set)
- Significance: compare to a 2nd order Markov from the genome
- E.g.: *[AT][CG].TTTTTT[CG]G..[AT]*



rna22 (cntd)

- **Target islands:** “hot spots” with ≥ 30 statistically significant mature miRNA patterns
- **Results:** *rna22* identifies correctly 17/21 “new” full-length sites

Identification of target islands

generate the reverse complement of statistically significant patterns and locate their instances in the target 5'UTRs

identify “target islands” supported by a minimum number of pattern hits

Assignment of microRNAs to target islands

pair-up each target island with each candidate microRNA

identify & report microRNA/target-island partners whose interaction satisfies user-specified thresholds

experimentally evaluate selected microRNA/target-island interactions



rna22 (cntd)





rna22: results (cntd)

Table 2. *Rna22*'s Estimates of the Number of MicroRNA Precursors for the Worm, Fruit Fly, Mouse, and Human Genomes

Genome	Number of MicroRNA Precursors Contained in the Used Training Set	Number of MicroRNA Precursors that Are in the Training Set and Can Be Detected by <i>ma22</i>	Total Number of MicroRNA Precursors Detected by <i>ma22</i> Including Already Known Ones ≤ -25 Kcal/mol (≤ -18 Kcal/mol)	Estimated Error when Predicting MicroRNA Precursors ≤ -25 Kcal/mol (≤ -18 Kcal/mol)
<i>C. elegans</i>	106	78 (73.6%)	359 (745)	$\leq 1\%$ ($\leq 2\%$)
<i>D. melanogaster</i>	78	62 (79.5%)	654 (1,236)	$\leq 1\%$ ($\leq 2\%$)
<i>M. musculus</i>	202	165 (81.7%)	>25,000 (>44,000)	$\leq 1\%$ ($\leq 2\%$)
<i>H. sapiens</i>	176	154 (87.5%)	>25,000 (>55,000)	$\leq 1\%$ ($\leq 2\%$)

Results are reported for two folding energy cutoffs: -25 Kcal/mol and -18 Kcal/mol.





rna22: evaluation

- **Advantages**

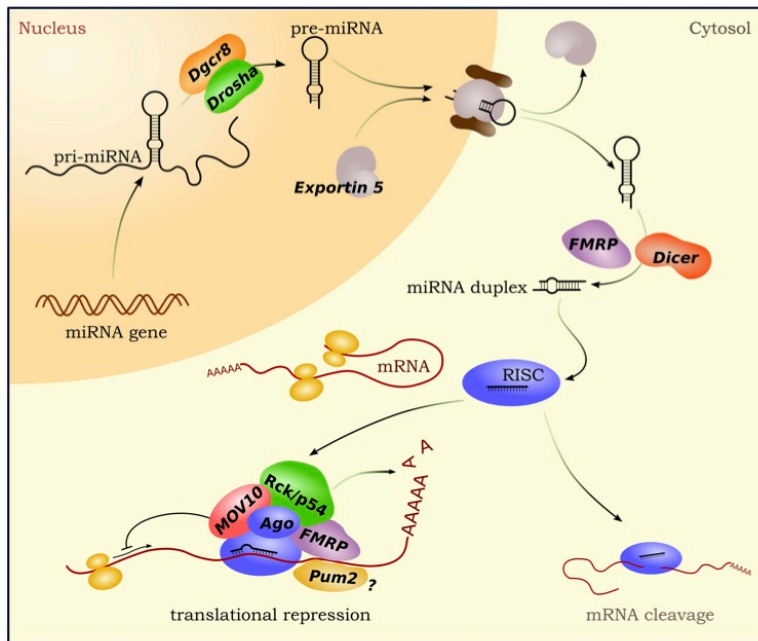
- Predicts miRNA target genes w/o knowledge of the miRNA gene
- No need for evolutionary conservation
- Performs better when miRNA genes have multiple targets in the same mRNA

- **Disadvantages**

- No consideration of the miRNA constrains *per se* (e.g., 5' "seed")
- May miss target genes with one or few target sequences in their 3' UTR
- Number of false positives cannot be estimated
- Heuristics



Predicting miRNA targets: not so easy...

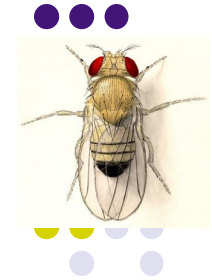


the-schratt-lab.de

Tool	Features
miRanda	Sequence binding, thermodynamics-based miRNA-mRNA duplex prediction and comparative sequence analysis
PITA	Also thermodynamics-based method; it considers the mRNA secondary structure in determining the miRNA-target accessibility.
TargetScan	Thermodynamics-based miRNA-mRNA duplex prediction and comparative sequence analysis. <u>Focus on seed region.</u>
mirSVR ^{***}	Based on regression method for predicting likelihood of target mRNA down-regulation from sequence and structure features in microRNA/mRNA predicted target sites.



The Drosophila AGO IP dataset



Immunoprecipitation of Ago1 miRNPs selects for a distinct class of microRNA targets

Xin Hong^{a,1}, Molly Hammell^{b,1}, Victor Ambros^{b,2}, and Stephen M. Cohen^{a,2}

^aTemasek Life Sciences Laboratory and Department of Biological Sciences, National University of Singapore, 1 Research Link, Singapore 117604; and ^bProgram in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605

Contributed by Victor Ambros, July 22, 2009 (sent for review June 26, 2009)

microRNAs comprise a few percent of animal genes and have been recognized as important regulators of a diverse range of biological processes. Understanding the biological functions of miRNAs requires effective means to identify their targets. Combined efforts from computational prediction, miRNA over-expression or depletion, and biochemical purification have identified thousands of potential miRNA-target pairs in cells and organisms. Complementarity to the miRNA seed sequence appears to be a common principle in target recognition. Other features, including miRNA-target duplex stability, binding site accessibility, and local UTR structure might affect target recognition. Yet computational approaches using such contextual features have yielded largely nonoverlapping results and experimental assessment of their impact has been limited. Here, we compare two large sets of miRNA targets: targets identified using an improved Ago1 immunoprecipitation method and targets identified among transcripts up-regulated after Ago1 depletion. We found surprisingly limited overlap between these sets. The two sets showed enrichment for target sites with different molecular, structural and functional properties. Intriguingly, we found a strong correlation between UTR length and other contextual features that distinguish the two groups. This finding was extended to all predicted microRNA targets. Distinct repression mechanisms could have evolved to regulate targets with different contextual features. This study reveals a complex relationship among different features in miRNA-target recognition and poses a new challenge for computational prediction.

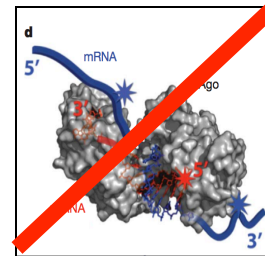
Argonaute | gene regulation | RISC complex

miRNAs (e.g., ref. 16). Whole proteome analyses have shown that miRNA induced changes in protein expression correlate with changes in mRNA level, in trend if not in magnitude (17, 18). Yet, there are well-documented instances of miRNA-mediated regulation at the protein level that do not involve changes in mRNA level (14, 17, 18). Therefore, methods to identify targets by miRNA-induced changes in expression profile can only tell part of the story. This highlights the need for alternative means to identify miRNA targets.

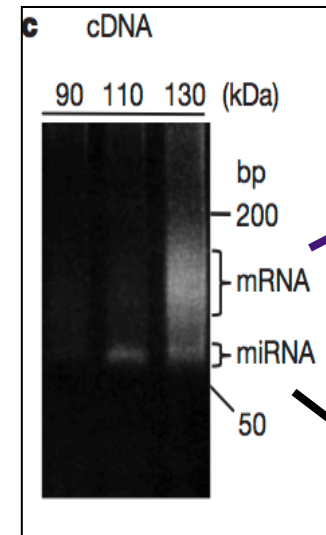
One such alternative involves identification of microRNA targets by virtue of their physical association with miRNA-containing ribonucleoprotein complexes (19–24). In ref. 19, we reported a method based on Ago1 immunoprecipitation (IP) that proved to be effective. Eleven new targets were identified for miR-1, including some that had not been predicted. Although the specificity was high, with all new targets experimentally validated, the method had limited sensitivity, identifying ~1/10th of the expected number of targets. Here, we present an improved Ago1 IP protocol, which permits identification of hundreds of potential miRNA targets, and compare the contextual features of targets identified by IP to the targets destabilized at the mRNA level upon Ago1 depletion.

Results

In an effort to improve the sensitivity of miRNA IP, with minimal loss of specificity, we tested a variety of antibody concentrations, incubation times and wash conditions (Fig. S1). Sensitivity was assessed by quantitative PCR to monitor miRNA levels (over a broad range of abundance: miR-184 comprises 17% of S2 cell miRNA; miR-305, 1.5%; miR-7, 0.1%; miR-92b,



*Image from Wook Chi et al 2009



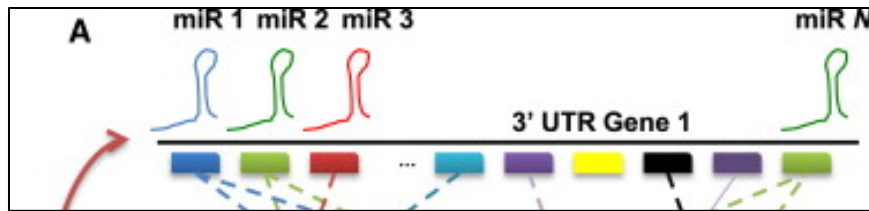
Drosophila S2 AGO IP

- 38 miRNAs expressed + IPed
- 6,285 mRNAs expressed
 - 1,091 AGO-bound
 - 5,194 not AGO-bound

	IP bound	IP not bound
Up-reg. after AGO1 depletion	SET I 142 mRNAs	SET III 287 mRNAs
Not up-reg. after AGO1 depletion	SET II 949 mRNAs	SET IV 4907 mRNAs
	1,091 mRNAs total	5,194 mRNAs total



Fermi-Dirac binding model improves miRNA target prediction efficiency

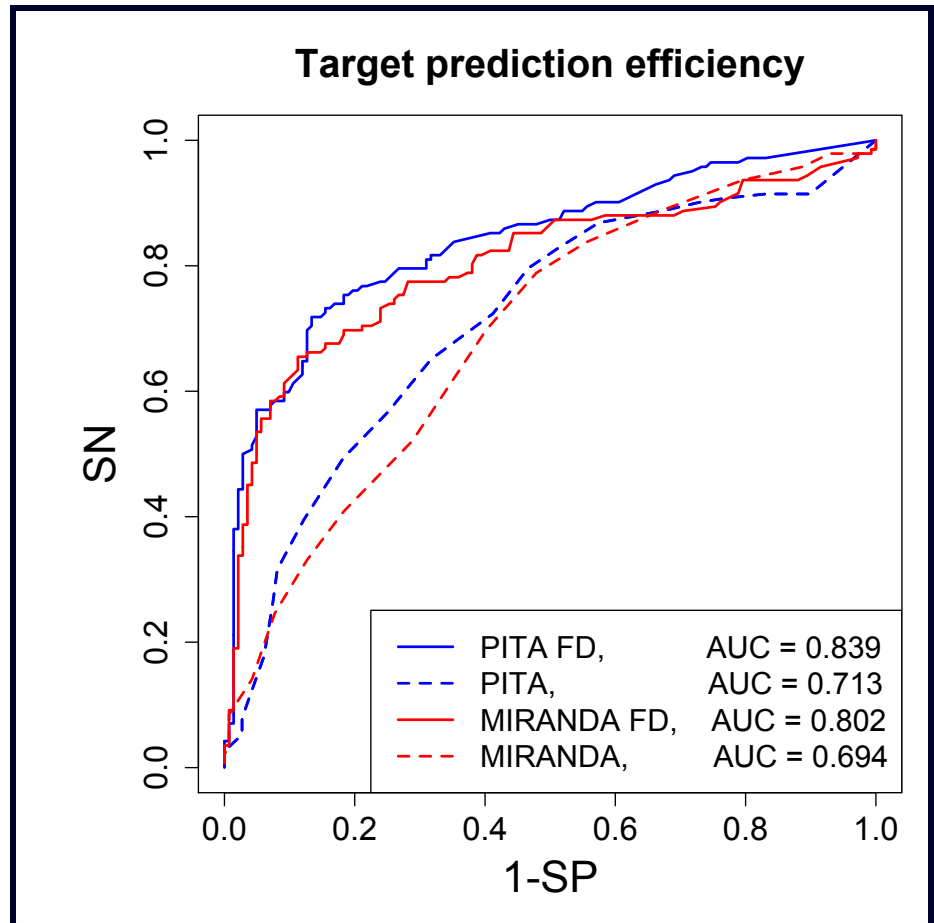


Naïve target combination

$$SN_k = \max_{i,j} (score_{ij})$$

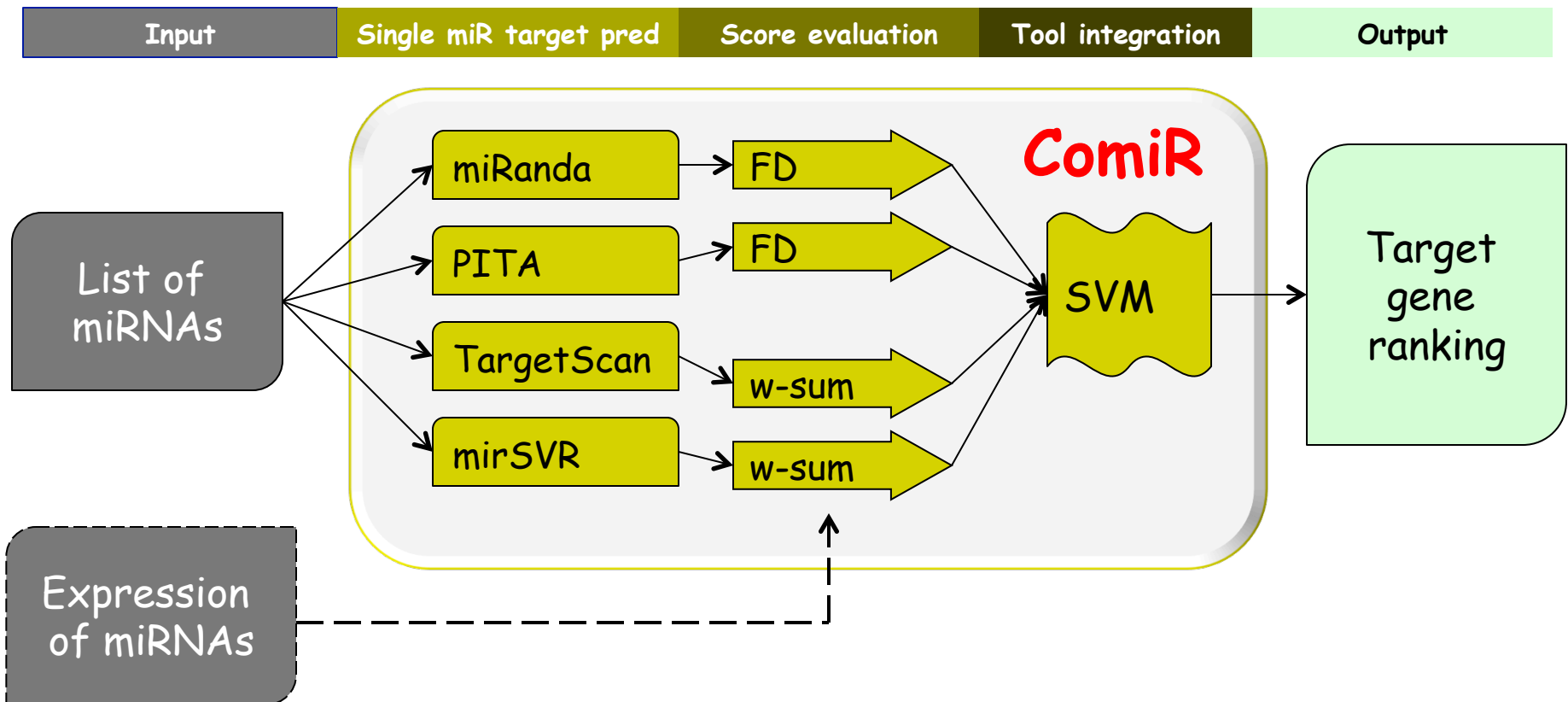
FD target combination

$$S_k = \sum_i^{miRs} \sum_{j=1}^{N_{ij}} \frac{1}{1 + e^{(S_{ijk} - \mu_i)/RT}}$$





ComiR: combinatorial miRNA targeting

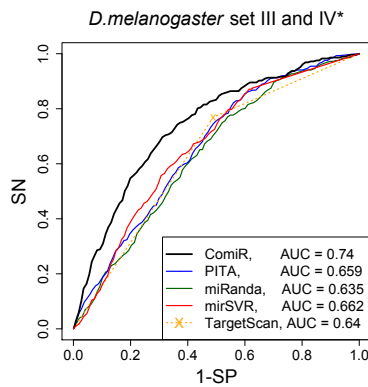
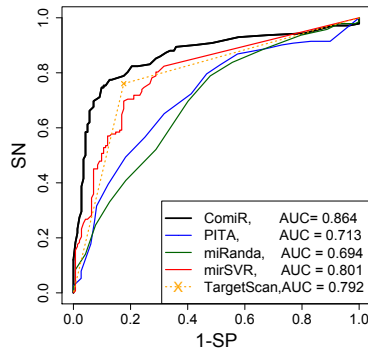
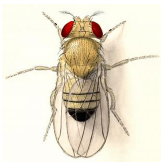


ComiR: results on various high-throughput datasets

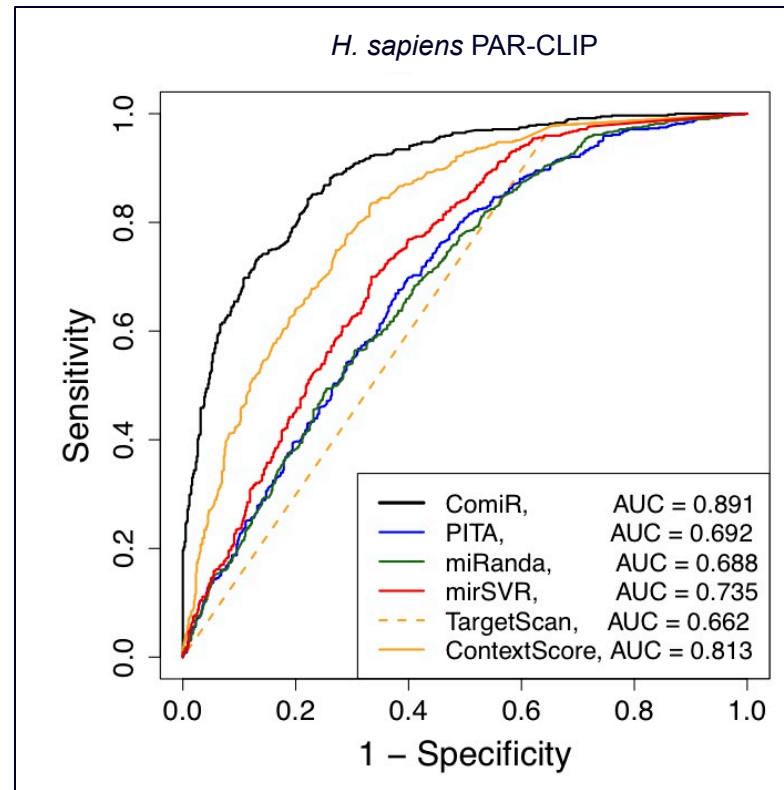


Claudia Coronello

D. melanogaster self-test



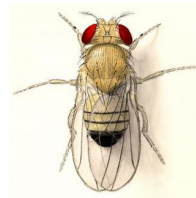
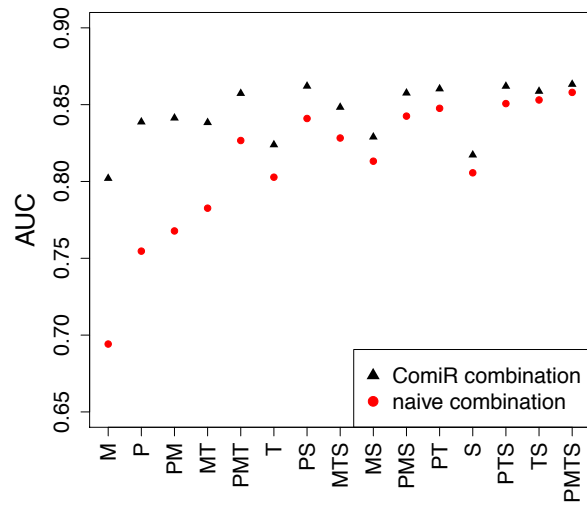
H. sapiens PAR-CLIP



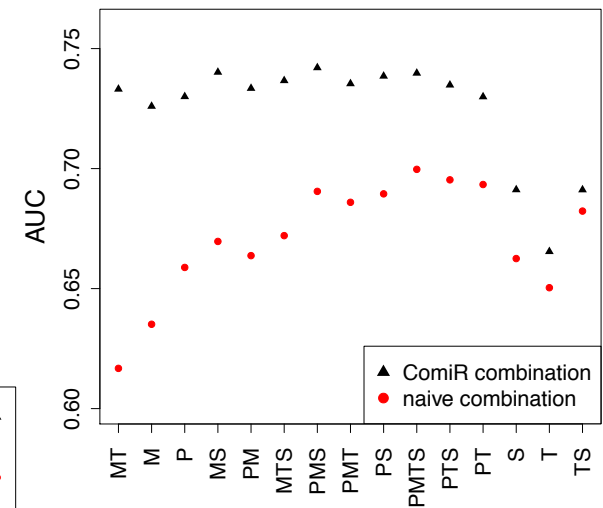


Why ComiR performs better?

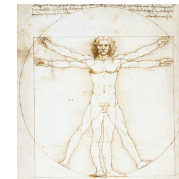
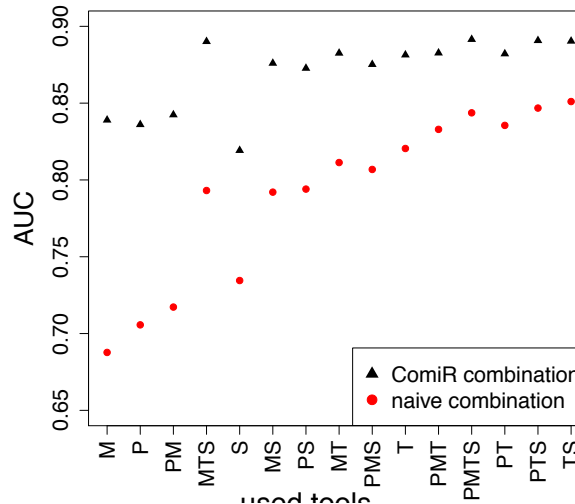
D. melanogaster self-test



D. melanogaster set III and IV*



H. sapiens PAR-CLIP





Acknowledgements

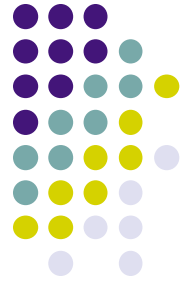
Some of the slides used in this lecture are adapted or modified slides from lectures of:

- Sarah Aerni, Universitaet Wien
- Bino John, Dow AgroSciences
- Brian Reinert, University of New Mexico

Theory and examples from the following:

- S.R. Eddy, “[How do RNA folding algorithms work?](#)”, *Nature Biotechnol*, 2004, **22**:1457-1458
- R. Durbin, S. Eddy, A. Krogh, G. Mitchison, “[Biological Sequence Analysis](#)”, 1998, Cambridge University Press





Acknowledgements

HHMMiR

Sabah Kadri, PhD (now at University of Chicago)



In collaboration with:

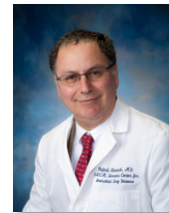
Veronica Hinman, PhD (Biology, CMU)



ComiR Claudia Coronello, PhD
(now at Ri.Med, Italy)



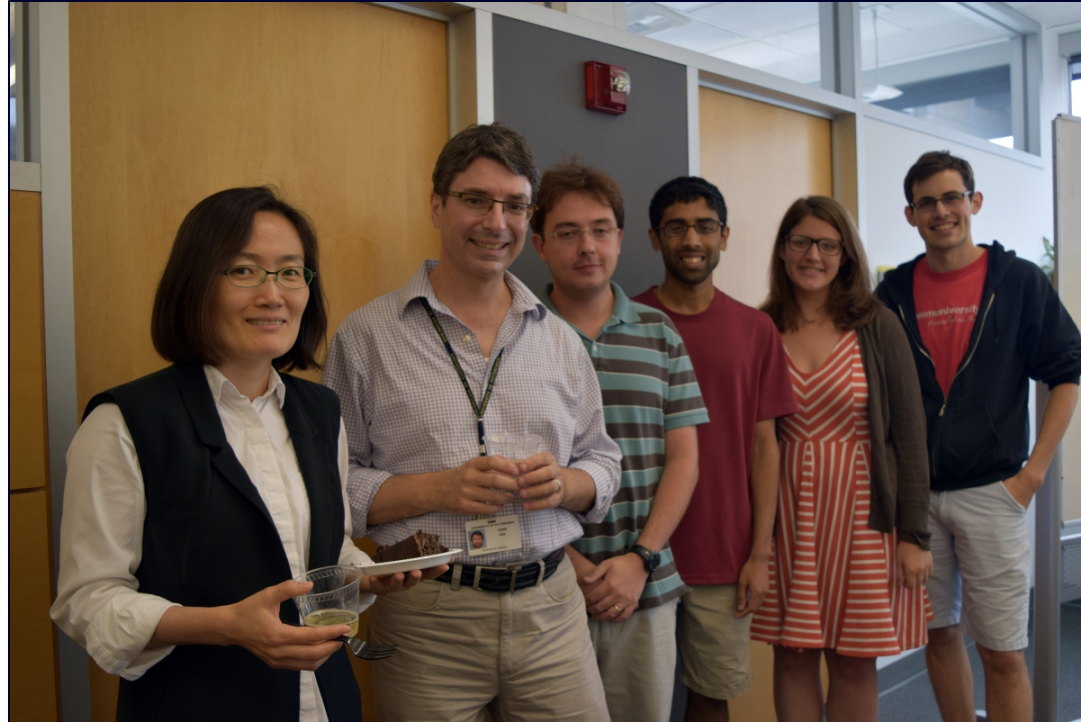
Naftali Kaminski, MD, Pulmonary, UPitt
Steffi Oesterreich, PhD, Magee-Womens
Gary Stormo, PhD, Washington Univ St Louis



Naftali Kaminski, MD, Pulmonary, UPitt



Thank you!



Electronic contacts:

benos@pitt.edu

<http://www.benoslab.pitt.edu>

