Predict miRNA target genes

Based on Cell 2005 120: 15-20

Bioinformatics. 2016 32(18): 2768–2775

Outline

- Review of miRNAs
- Three consecutive papers by one Bartel Group at MIT
  - Lewis et al. Cell. December 2003
  - Lewis et al. Cell. January 2005
  - Farh et al. Science. November 2005
  - One more paper from Kellis group at MIT 2005

- miRNA target identification at the NGS era

Introduction - miRNA

- Function: **silencing genes** through post-transcriptional regulation
- Single strand RNA (ssRNA);
- 19-25 nucleotides (~22 nucleotides);
- Hairpin-shaped;
- Endogenous
- Accounting for 1% (10%) of the genome (>200 (>2000) members per species);
- >1/3 of human genes are microRNA target;

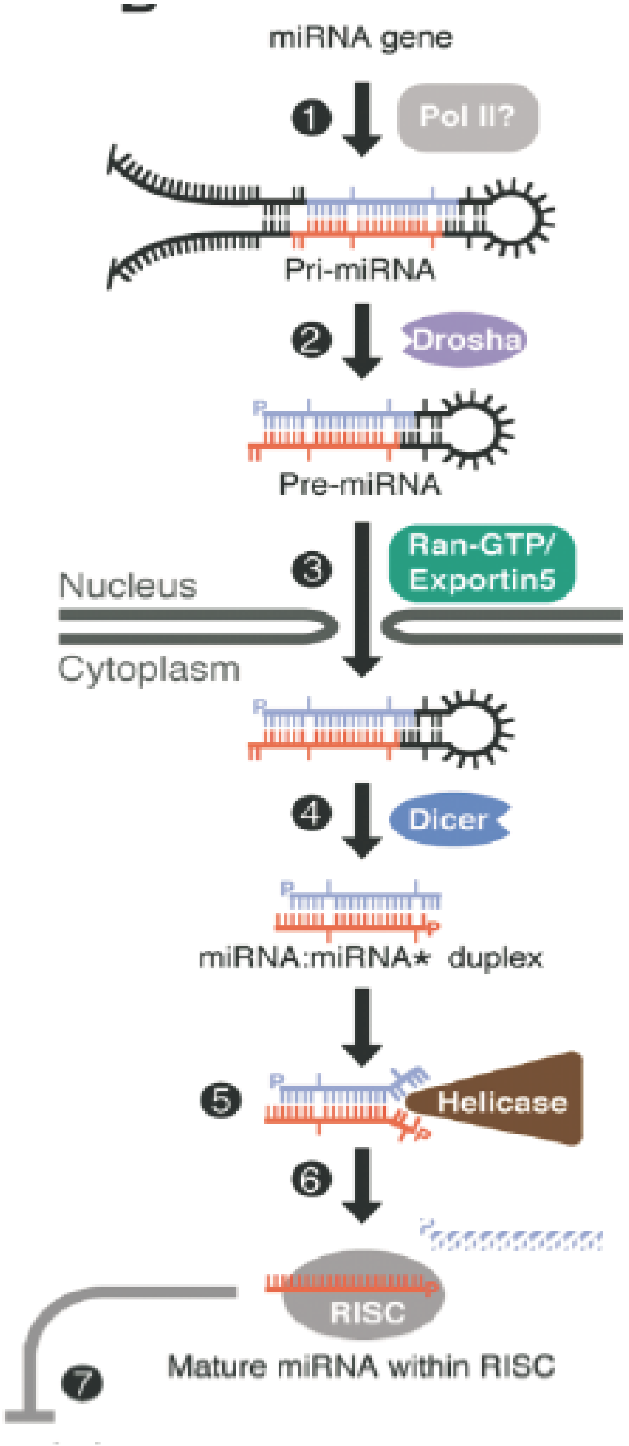
Nat Rev Mol Cell Biol. 2005 May;6(5):376-85.

Introduction - microRNA

- Combinatorial effects
  - Single microRNA can regulate many different mRNA
  - Single mRNA can be regulated cooperatively by several different microRNA
- microRNA has key roles in diverse regulatory pathways:
  - Control of developmental timing
  - Haematopoietic cell differentiation
  - Apoptosis
  - Cell proliferation
  - Organ development …

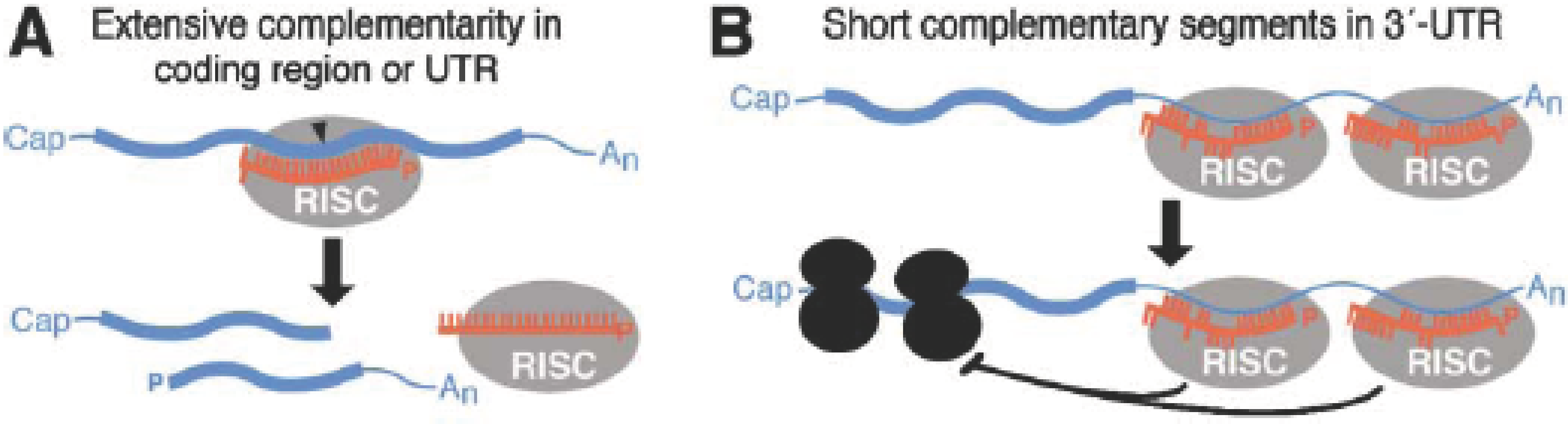
microRNA biogenesis

1. Transcription
2. Pri-microRNA
3. Pre-microRNA
4. Export into cytoplasm
5. Duplex
6. Mature microRNA
7. microRNA with RISC (RNA-Induced Silencing Complex)



Cell. 2004 Jan 23;116(2):281-97.

microRNA functions


A
Extensive complementarity in coding region or UTR
Cap
An
P
RISC
Cap
P
An
RISC
B
Short complementary segments in 3′-UTR
Cap
An
P
P
RISC
RISC
Cap
An
P
P
RISC
RISC

A. mRNA cleavage
B. Translational repression

Cell. 2004 Jan 23;116(2):281-97.

Questions to be answered?

- Identification of microRNA genes
- Identification of microRNA targets

Identification of microRNA genes

- Methods:
  - Prescreen size-fractionated RNA population (on gel)
  - Ligate 5’ and 3’ adapter molecules to both ends
  - Reverse transcription, amplification (PCR), Concatamerization, cloning, and sequencing.
- Total number of human microRNA genes
  - **200-250** (Lim et al. Science 2003) – estimation – representing ~1% human genome
  - **319** human microRNA (234 has been experimentally verified) in miRBase (release 7.1, October 2005)

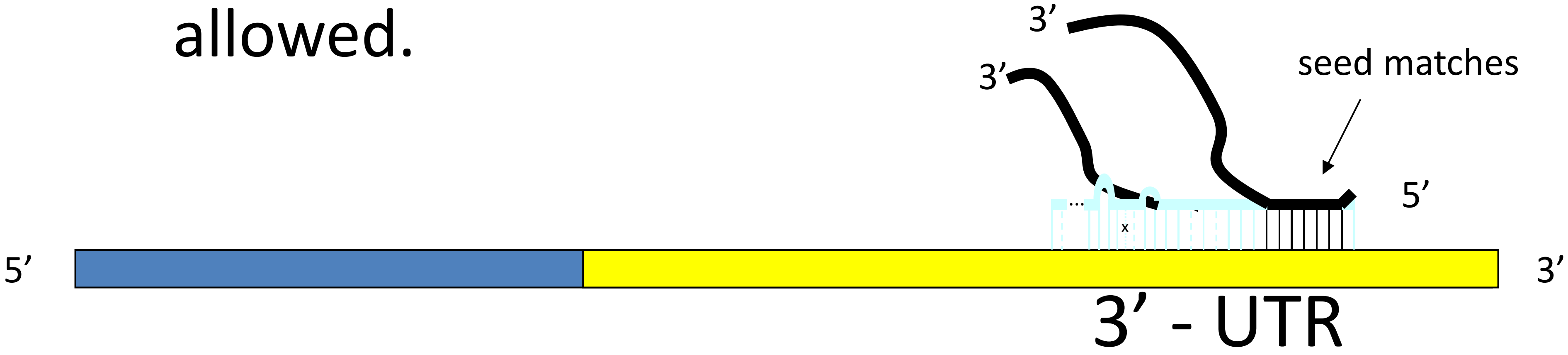
Identification of microRNA targets

- Goal: Identify mRNA targets that are regulated by a known microRNA
  - Each microRNA can regulate multiple genes
  - Each gene can be regulated by multiple microRNA

Summary of Lewis et al. (Cell 2003)

- Lewis et al. Cell 2003

1. Find perfect W-C match between 3’-UTR and base 2-8 of the microRNA
2. Extend the seed match as far as possible to each direction; stop at mismatch; G:U pairs are allowed.



3. Optimize basepairing of the remaining 3’ portion of the miRNA to the 35 bases of UTR immediate 5’ of seed using RNAfold.

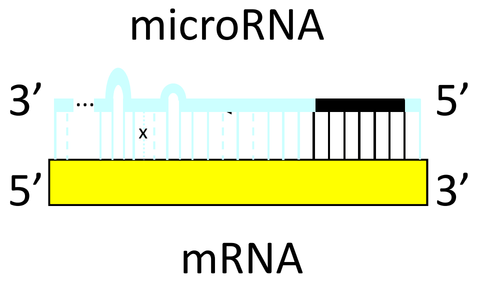
Summary of Lewis et al. (Cell 2003)

- Lewis et al. Cell 2003

4. Calculate the folding free energy of microRNA:target pair using RNAeval.

5. Assign a score Z to each 3’-UTR based on the free energy.

6. Rank the UTRs by Z score, and select the top ones.

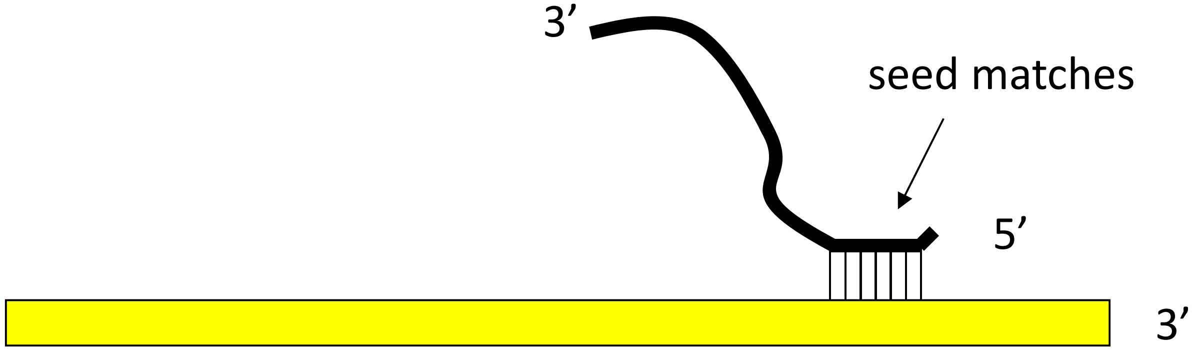


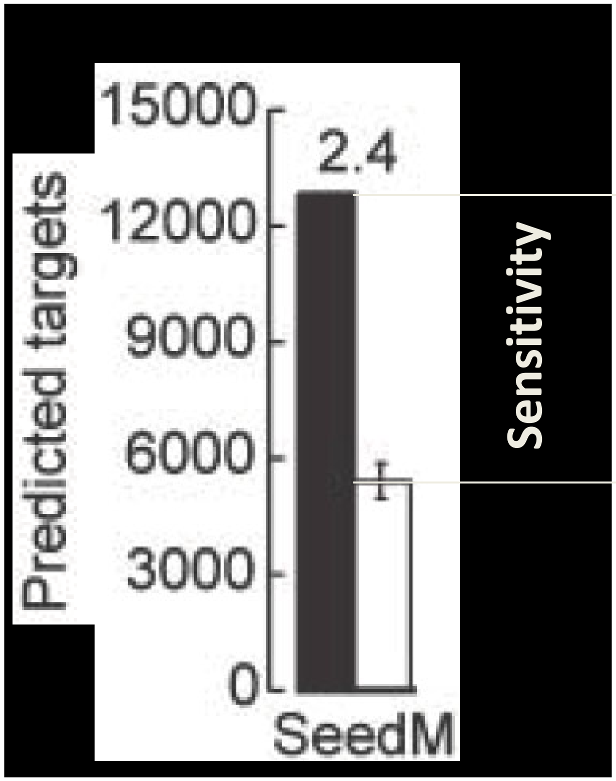
7. Repeat the process in multiple organisms such as human, mouse, rat and dog.

Goal – Lewis 2005

- Characterize features of microRNA binding sites
  - Looking at the number of predicted targets based on:
    1. 148 microRNA sequences (in miRBase) ⬜
    2. random sequences ⬛

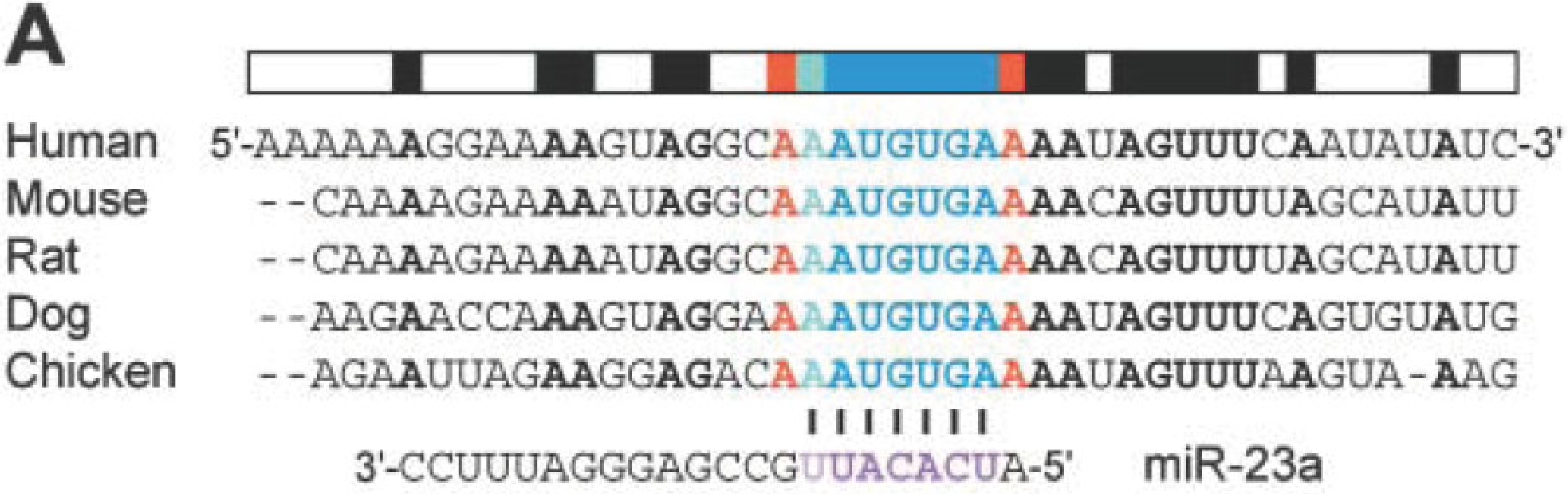
  Evaluation of the prediction:
    - Signal-to-noise ratio
    - Sensitivity





Simplified approach

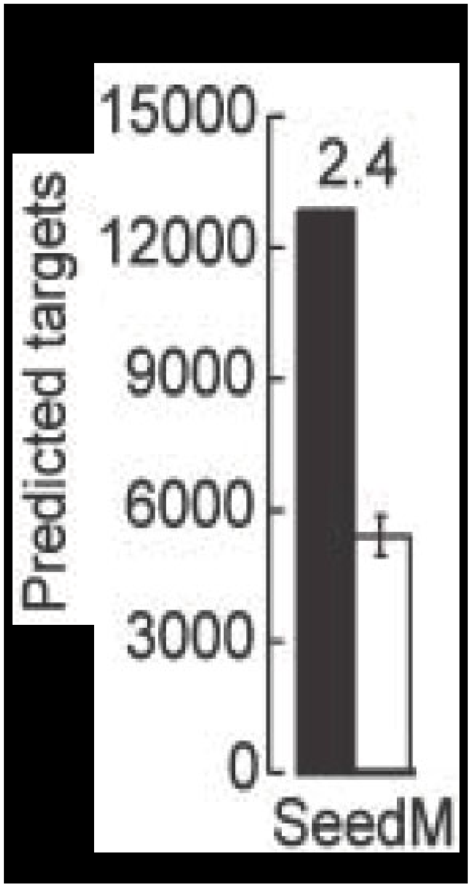
- Finding perfect (W-C) seed matches that are conserved in the UTR regions of whole-genome alignment.

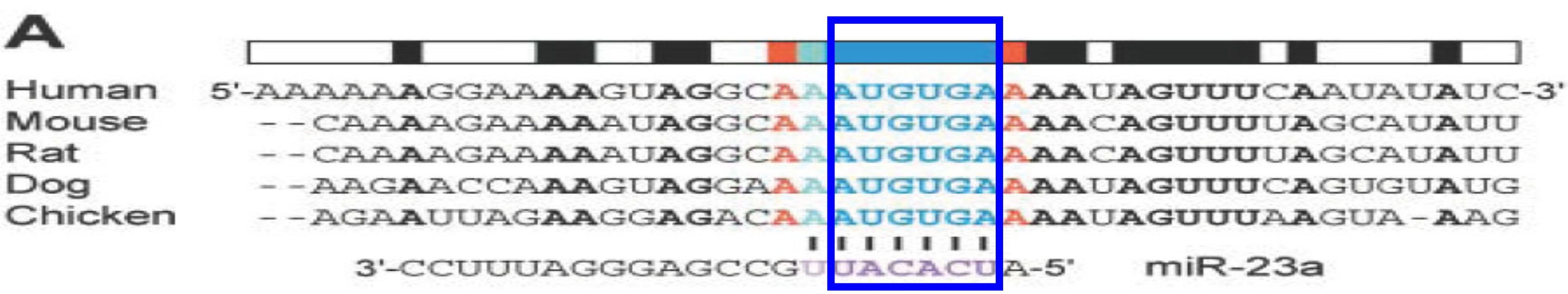

A
Human 5'-AAAAAAGGAAAAGUAGGCAAAUGUGAAAAUAGUUUCAAUAUAUC-3'
Mouse --CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
Rat --CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
Dog --AAGAACCAAAGUAGGAAAAUGUGAAAAUAGUUUCAGUGUAUG
Chicken --AGAAUUAGAAGGAGACAAAUGUGAAAAUAGUUUAAGUA-AAG
3'-CCUUUAGGGAGCCGUUACACUA-5'
miR-23a

on **10,938** orthologous genes

Results I – position 2-7 in microRNA

- 148 microRNA sequence
  - 14301 unique target sites in 3’ UTR
  - 12839 pairs of unique miRNA-targets
  - 3227 unique genes
- Random sequence (false positive):
  - 5817 pairs
- Signal
  - 8484 pairs
  - **2767** unique genes (25% of 10938 genes)

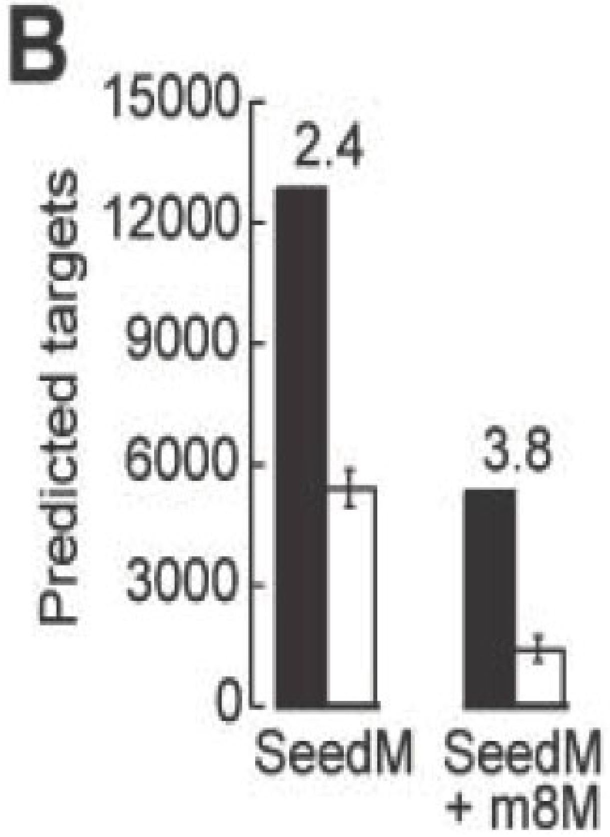


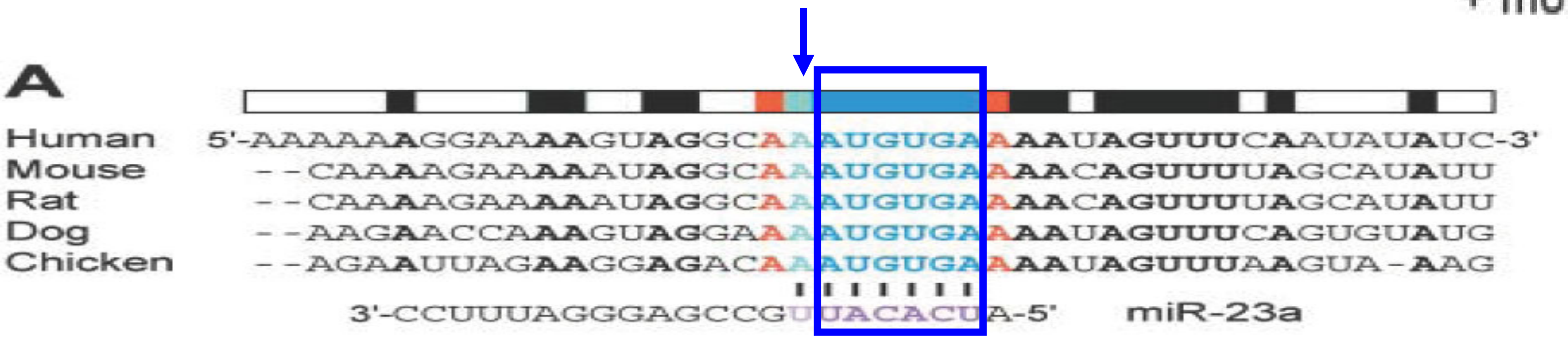

A
Human 5'-AAAAAAGGAAAAGUAGGCAAAUGUGAAAAUAGUUUCAAUAUAUC-3'
Mouse --CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
Rat --CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
Dog --AAGAACCAAAGUAGGAAAAUGUGAAAAUAGUUUCAGUGUAUG
Chicken --AGAAUUAGAAGGAGACAAAUGUGAAAAUAGUUUAAGUA-AAG
3'-CCUUUAGGGAGCCGUUACACUA-5'
miR-23a

Result II – position 8

- T8 is highly conserved, and likely to become an M8 with miRNA
- Sensitivity drops if require M8 – more than **3500** authentic target sites **lack** M8 matches

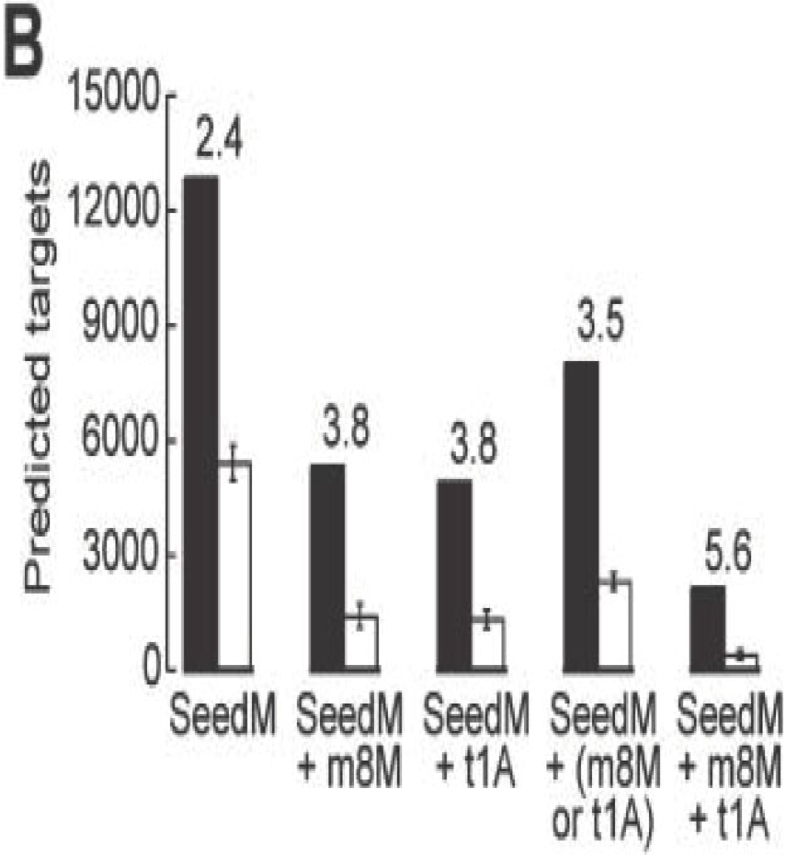
(M and T – matches and target)

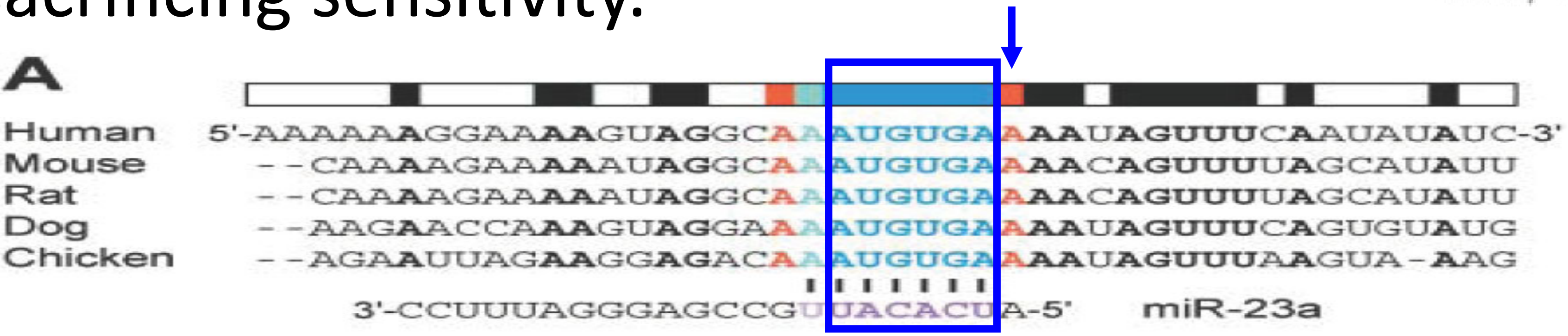



A
Human
Mouse
Rat
Dog
Chicken
5'-AAAAAAGGAAAAGUAGGCAAAUGUGAAAAUAGUUUCAAUAUAUC-3'
--CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
--CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
--AAGAACCAAAGUAGGAAAAUGUGAAAAUAGUUUCAGUGUAUG
--AGAAUUAGAAGGAGACAAAUGUGAAAAUAGUUUAAGUA-AAG
3'-CCUUUAGGGAGCCGUUACACUA-5'
miR-23a

Result III – position 1

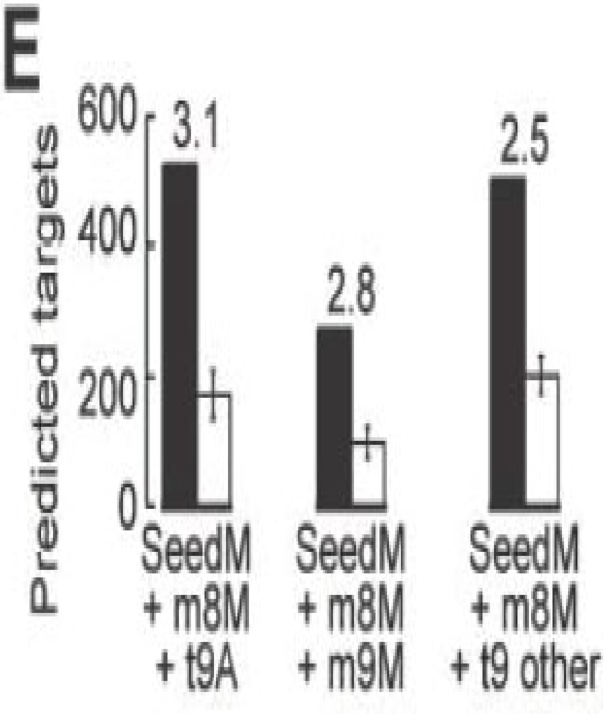
- T1 is often a conserved “A”, even if 1^{st} position of miRNA is not a “U”.
- SeedM+t1A reduced sensitivity to 51%.
- Requiring one of two anchors M8 or T1A increases signal-to-noise ratio, without sacrificing sensitivity.

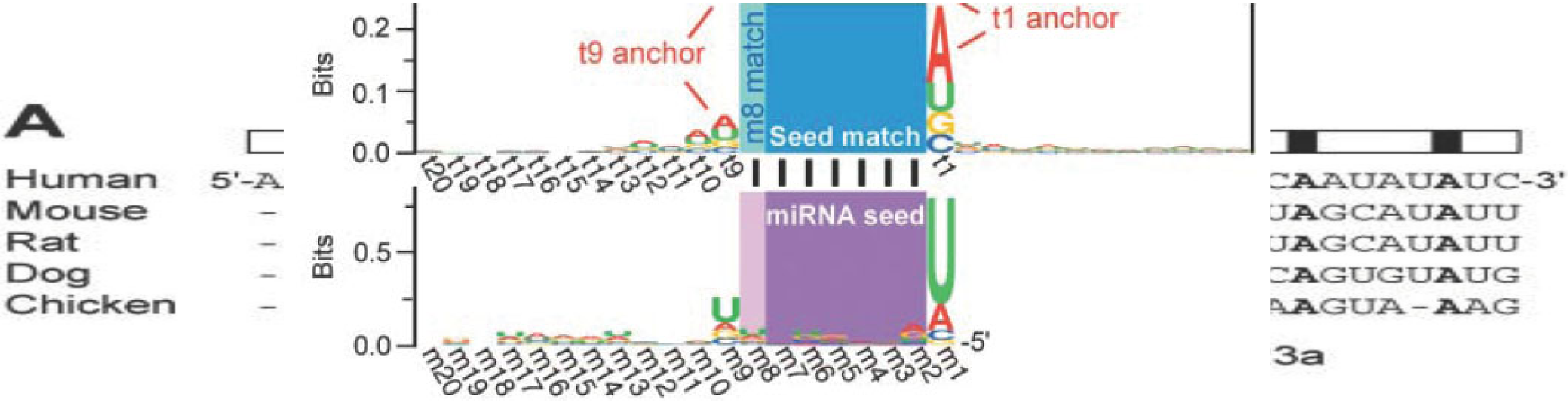



A
Human 5'-AAAAAAGGAAAAGUAGGCAAAUGUGAAAAUAGUUUCAAUAUAUC-3'
Mouse --CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
Rat --CAAAAGAAAAAUAGGCAAAUGUGAAAACAGUUUUAGCAUAUU
Dog --AAGAACCAAAGUAGGAAAAUGUGAAAAUAGUUUCAGUGUAUG
Chicken --AGAAUUAGAAGGAGACAAAUGUGAAAAUAGUUUAAGUA-AAG
3'-CCUUUAGGGAGCCGUUACACUA-5' miR-23a

Result IV – position 9

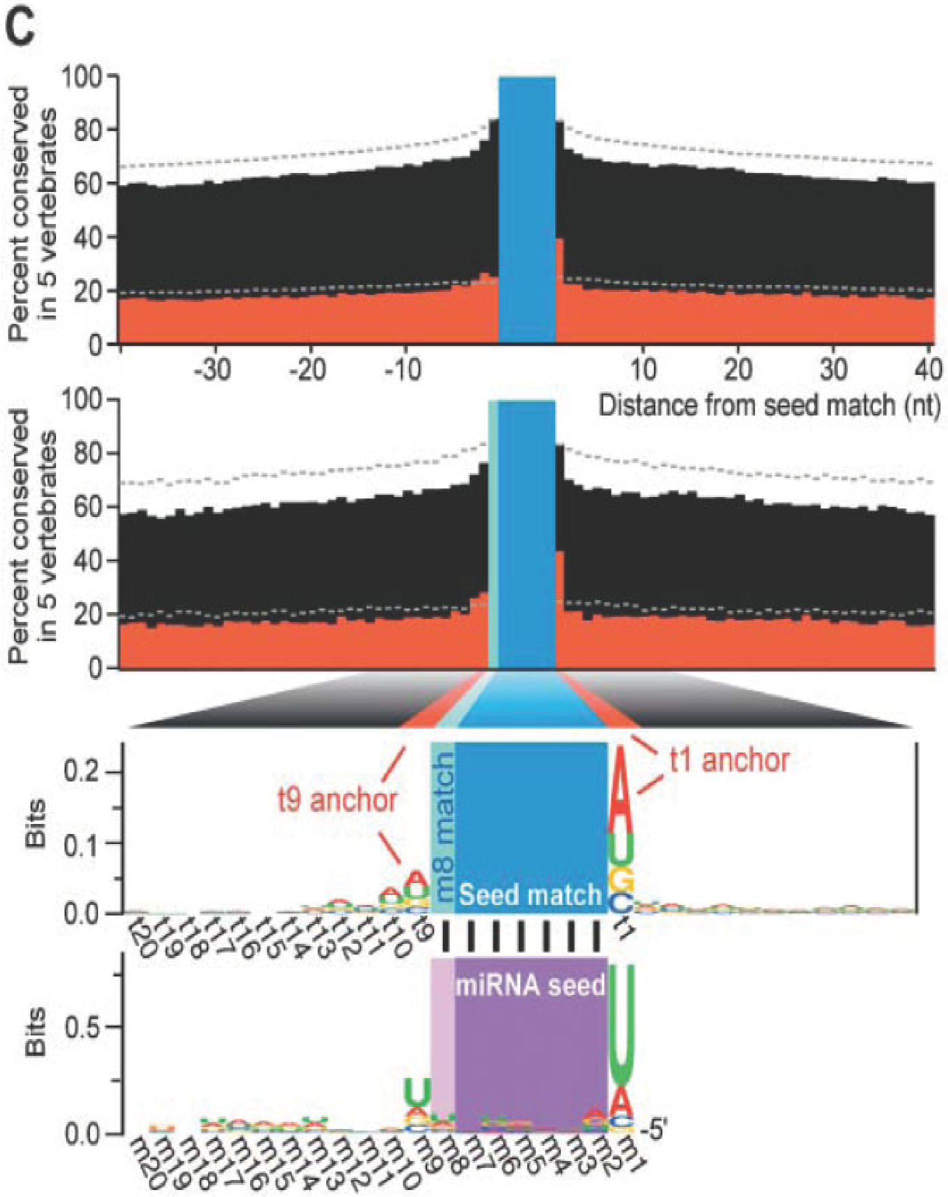
- Slightly conserved at T9 position;
- Enrichment of “A”
- Focusing on miRNA that do not have a “U” at position 9, over-abundance “A” at T9.





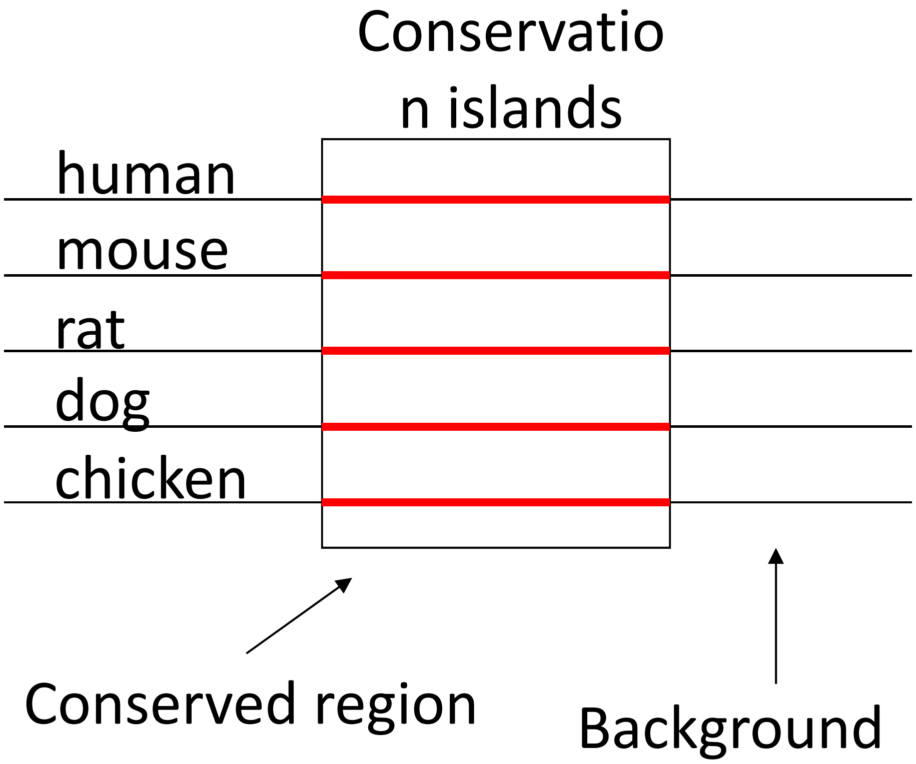
Results IV – Beyond seed matches

- Little conservation observed beyond seed match;
- Single conserved matches are sufficient to predict miRNA-target pairs.



Result V – Conservation islands

- **Including** seed matches that occur in the context of **more extensive conservation** improves signal-to-noise ratio.



Conservation islands

human

mouse

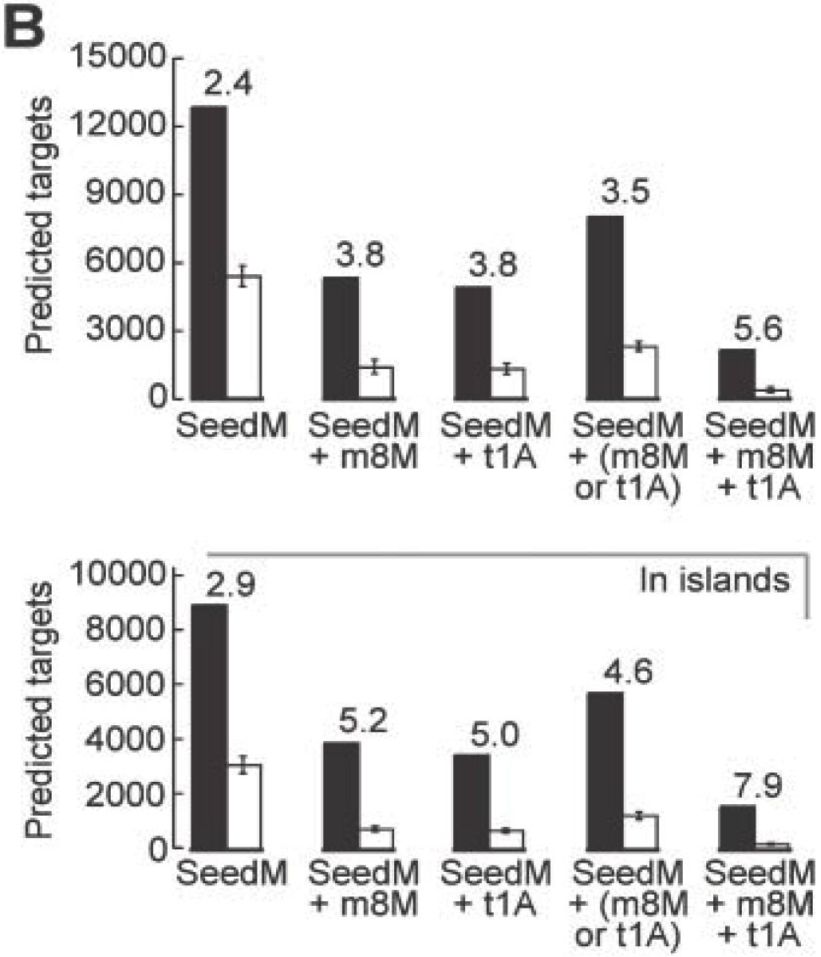
rat

dog

chicken

Conserved region

Background



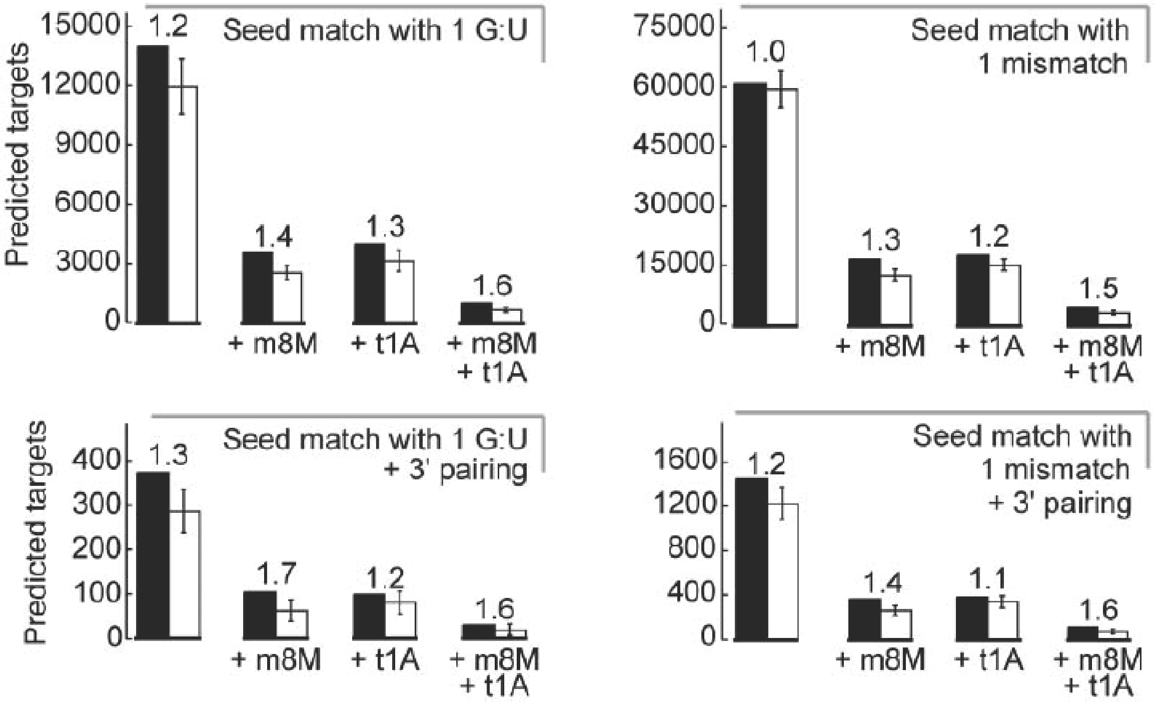
Results VI – Mammalian genome only

- Four genomes including human, rat, mouse, and dog;
- Sensitivity increases, 13,044 regulatory interactions above noise (comparing with 8,484 in five-genome analysis), including 5,300 unique genes (comparing with 2,767 unique genes in five-genome).

  (This is based on 17,850 orthologous genes)
- Average of 200 targets per microRNA

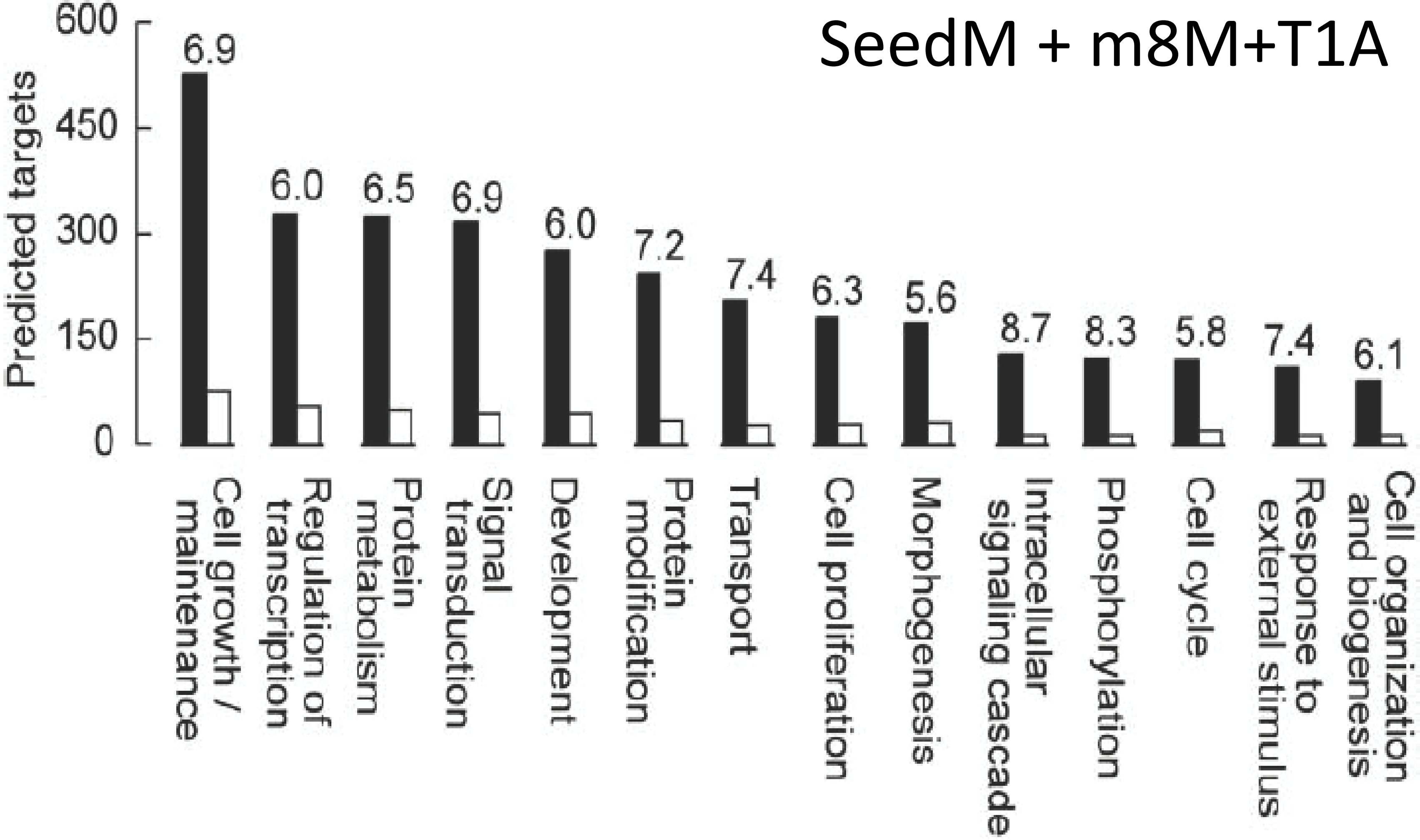
Results VII – Wobbles and mismatches

- Some microRNA-target pairs has wobbles and mismatches, such as *let-7-lin41* or *miR-196-HoxB8*
- Allowing wobbles and mismatches decreases signal-to-noise ratio dramatically



Result VIII – functions of microRNA targets

SeedM + m8M+T1A



Discussion – uniqueness of the method

- Requirement for perfect W-C seed pairing;
- Starts from whole-genome alignment;
- Focusing only on 8-nt segment that centers on the seed match;
- Careful design of the control sequences;

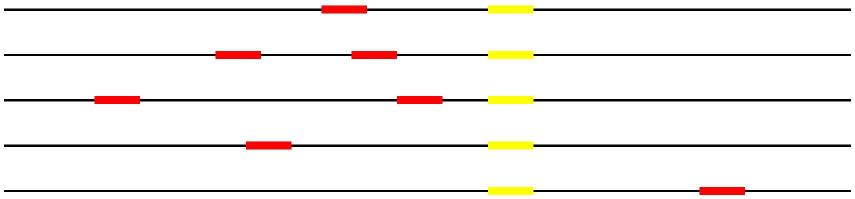
Conclusion

- Seed match (positions 2-7 of microRNA sequence) plus either of both M8 and T1A anchor determines microRNA-target interaction.

- “Biochemical specificity is augmented by additional determinants, such as mRNA structure, binding of accessory proteins, and/or the presence of nonconserved or imperfect seed matches at additional sites in the message.”

Question unanswered

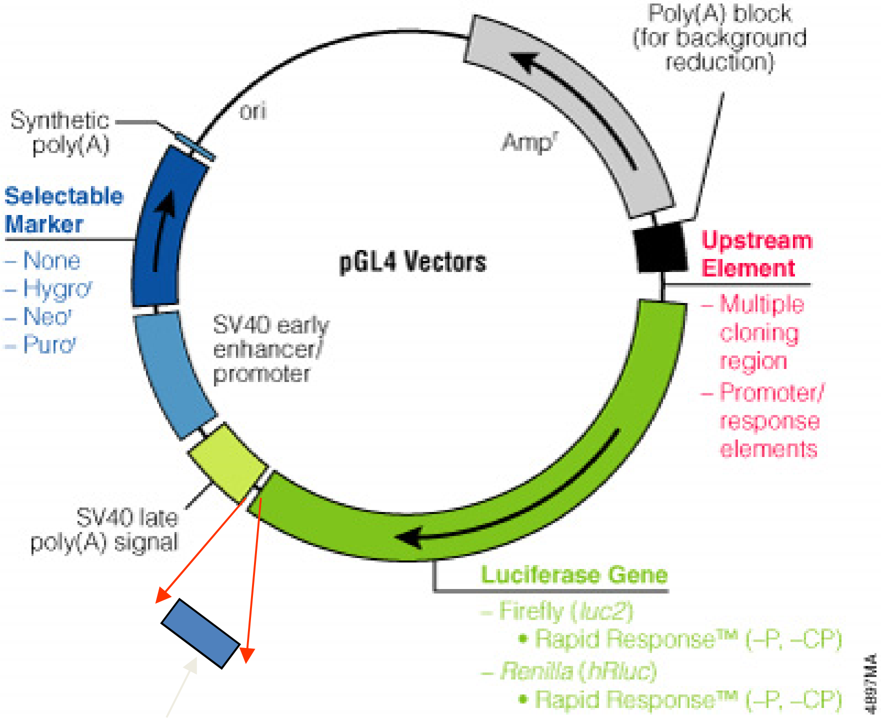
- Each mammalian microRNA have an average of ~200 conserved target sites. 1/10 of non-conserved 7-nt sites in the whole genome UTR
- Cells can not distinguish conserved or non-conserved sites
- Question: **Will the non-conserved sites be functional?**



Non-conserved binding sites

- Reporter assay: tests the luciferase activity from HeLa cells cotransfected with microRNA and reporter construct (wild-type or mutant UTRs).

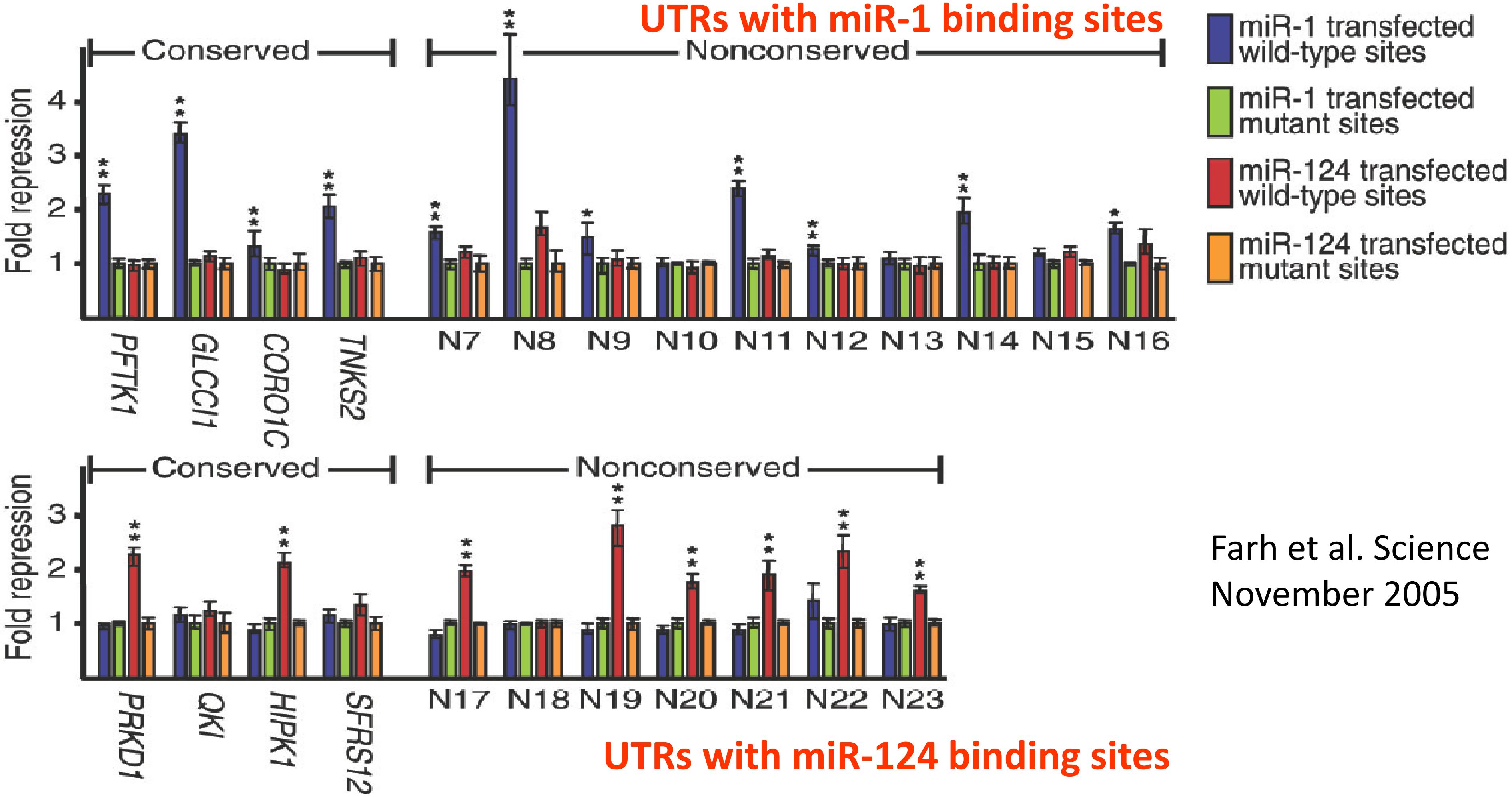
- Mutant UTRs were disrupted at three point substitutions in seed match.



Insert UTR here

Farh et al. Science November 2005

Non-conserved binding sites



UTRs with miR-1 binding sites

UTRs with miR-124 binding sites

Farh et al. Science
November 2005

Non-conserved sites accidentally reside have the potential to function when exposed to microRNA

Other conclusions

- Non-conserved sites are also functional;
- Seed match plus M8 or T1A is sufficient for microRNA-like regulation;
- “Additional recognition features, such as pairing to the remainder of the microRNA, accessible mRNA structure, or protein-binding sites, are usually **dispensable**, or occur so frequently that they impart little over specificity.”

miRNA target site prediction

- In plants, computational identification can be performed by simple blast search as miRNA:mRNA complementarity reaches 100%.
- Most animal miRNA are though to recognise their mRNA targets by partial complementarity.

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

Comparison of 3 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.

Summary of miRNA target prediction

- Differences in algorithm: one can state opinions about the strengths or weaknesses of each particular algorithm.
- 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.

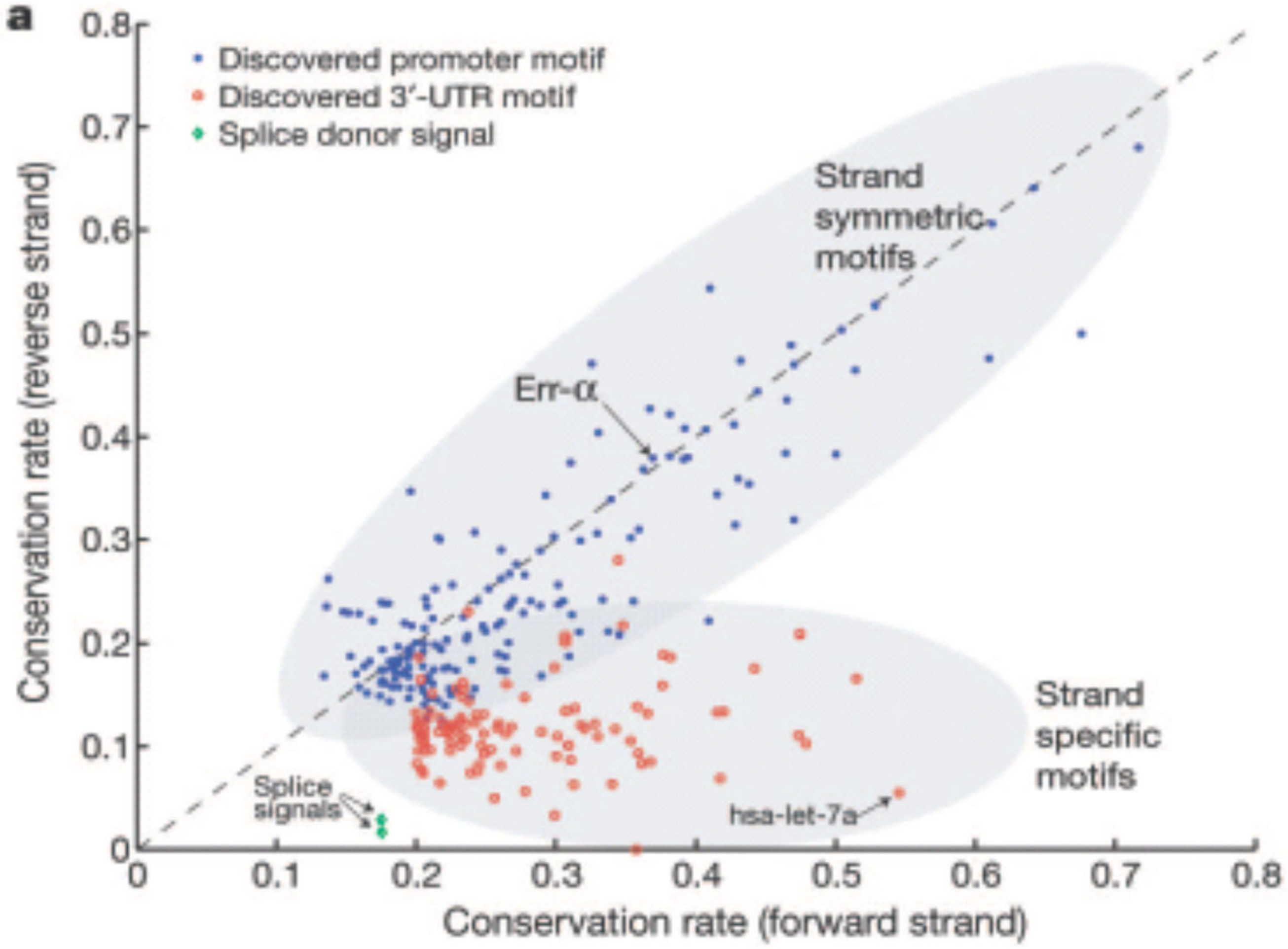
The second paper

- *Nature* **434**, 338-345 (17 March 2005) |

  Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals

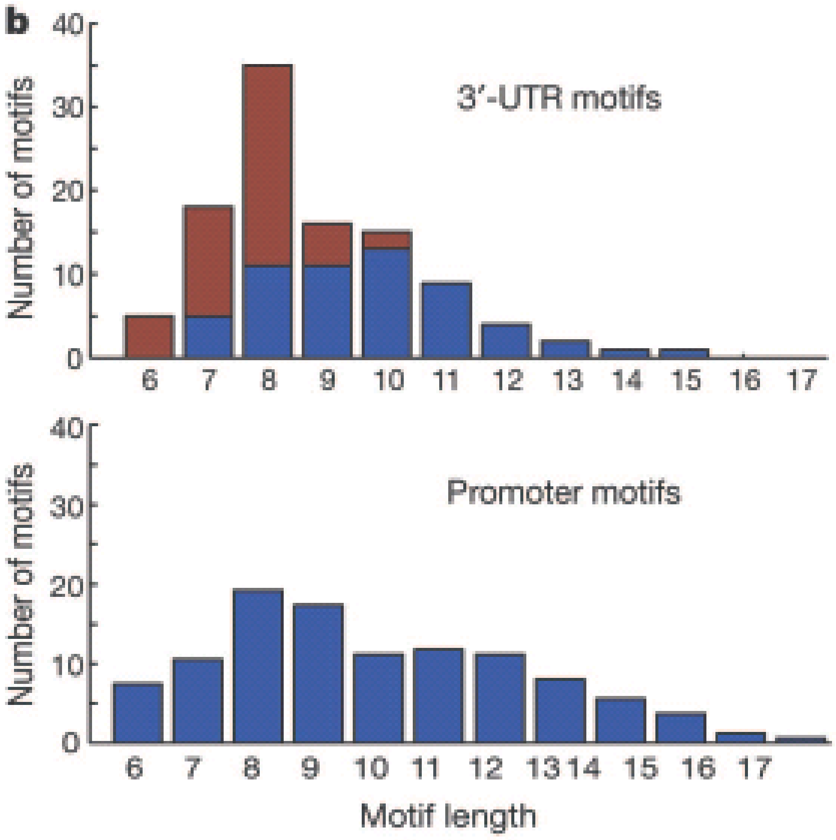
  Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R. Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S. Lander, Manolis Kellis

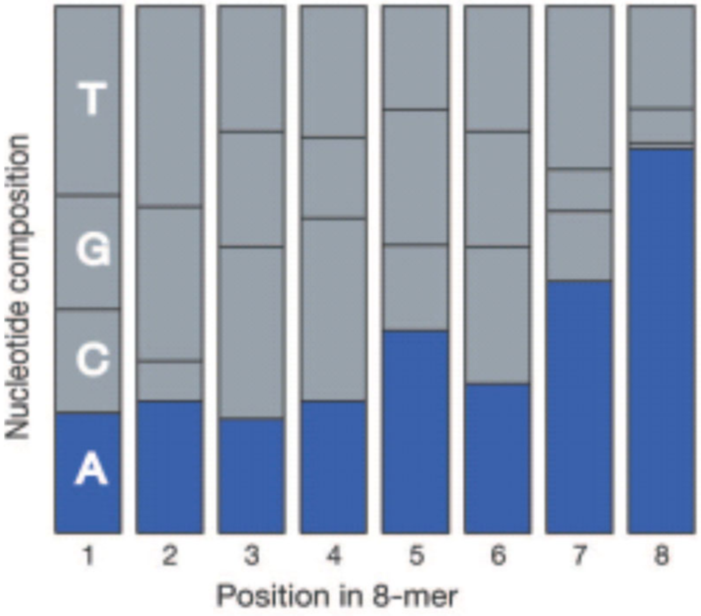
Property1: strand specificity



Xie, X. et al., Nature, 2005

Property2 : bias towards 8-mers

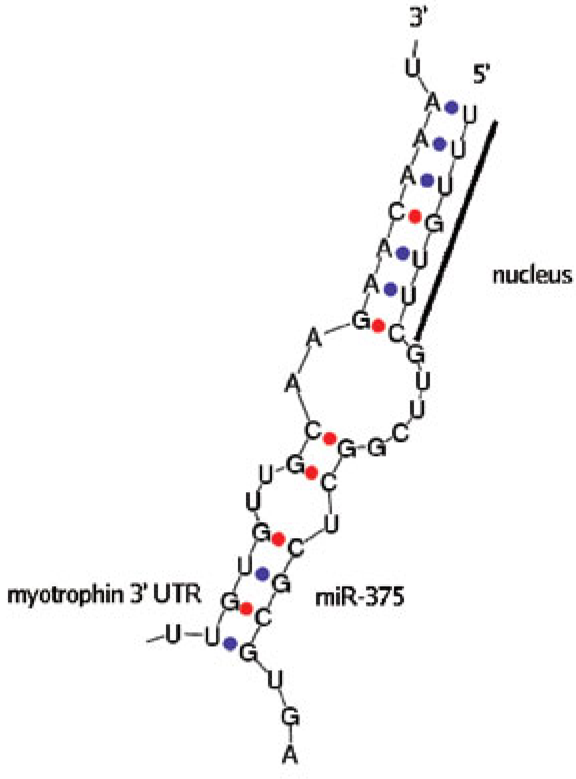




Xie, X. et al., Nature, 2005

Digression: miRNA

- Single stranded RNA
- transcribed from DNA but not translated into protein
- Many mature miRNA start with U followed by a 7-base “seed” complementary to a site in the 3’ UTR of target mRNAs.
- Thus many are 8 mers



microRNA that regulates insulin secretion by an NYU study published in Nature.

Inference

- Thus we can infer many of the conserved 8-mer motifs act as binding sites for miRNA
- Leads to discovery of 52% existing miRNA genes
- Leads to discovery of 129 new miRNA genes

miRNA target gene prediction in the NGS era

Know important features

- seed match, the exact sequence matching between the positions 2–7 of an miRNA and a segment of 6 nucleotides (nt) long in target mRNAs
- Accessibility, how likely a region in an mRNA sequence is ‘open’ or accessible for an miRNA to bind
- folding energy
- Conservation
- AU content

Tools based on these features

- miRanda (Enright *et al.* , 2004): seed match, conservation and free energy for target site prediction. http://www.microrna.org/microrna/getDownloads.do

- TargetScan (Friedman *et al.* , 2009 ; Grimson *et al.* , 2007): seed match, pairing of mRNAs with 3′ of miRNAs, local AU content, etc. http://www.targetscan.org/vert_71/

- PicTar (*Nature Genetics* **37**, 495 - 500 (2005)): seed match, conservation, etc. http://www.pictar.org/

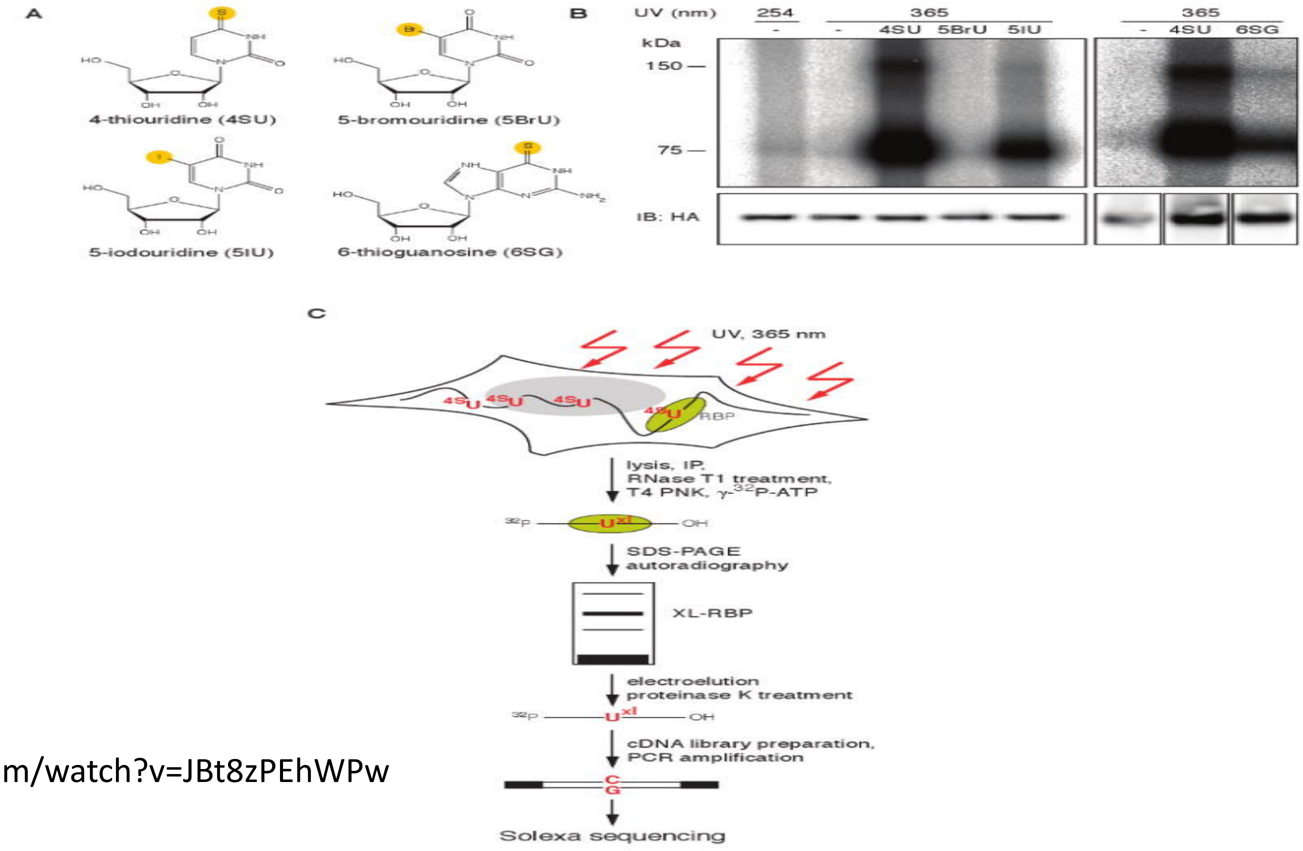
- miRWalk, (PLoS One 13(10), 2018): http://mirwalk.umm.uni-heidelberg.de/

-

Drawbacks in existing tools

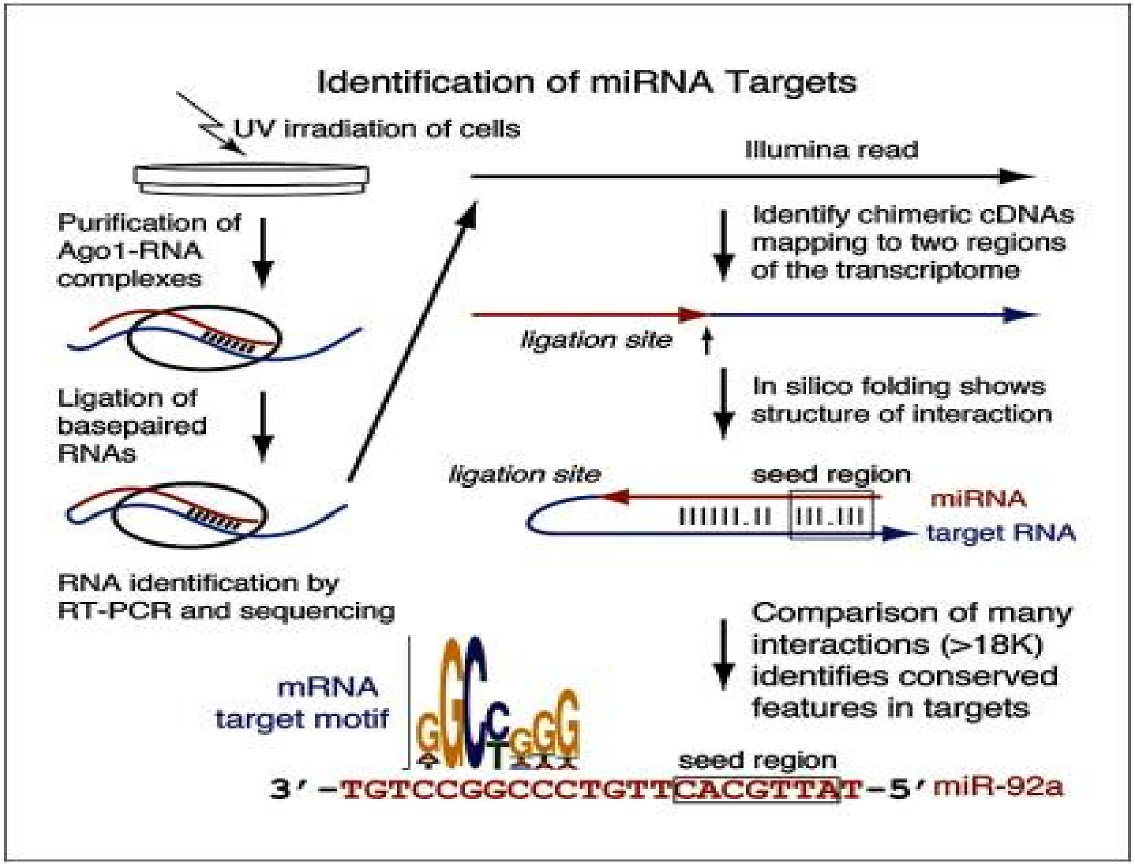
- matching seed is not always sufficient for a functional miRNA–mRNA interaction (Brennecke *et al.* , 2005 ; Didiano and Hobert, 2006)
- Seed matching is also not necessary: non-canonical pairings that allow G:U wobbles and even mismatches can be functional (Brennecke *et al.* , 2005 ; Didiano and Hobert, 2006).
- Recent photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) and crosslinking ligation and sequencing of hybrids (CLASH) experiments (Hafner *et al.* , 2010, Helwak *et al.* , 2013) have further shown that seed match, including canonical and non-canonical seed-matching, is not required for certain miRNA–mRNA interactions.

PAR-CLIP experiment



https://www.youtube.com/watch?v=JBt8zPEhWPw

CLASH experiment



Identification of miRNA Targets

UV irradiation of cells

Purification of Ago1-RNA complexes

Ligation of basepaired RNAs

RNA identification by RT-PCR and sequencing

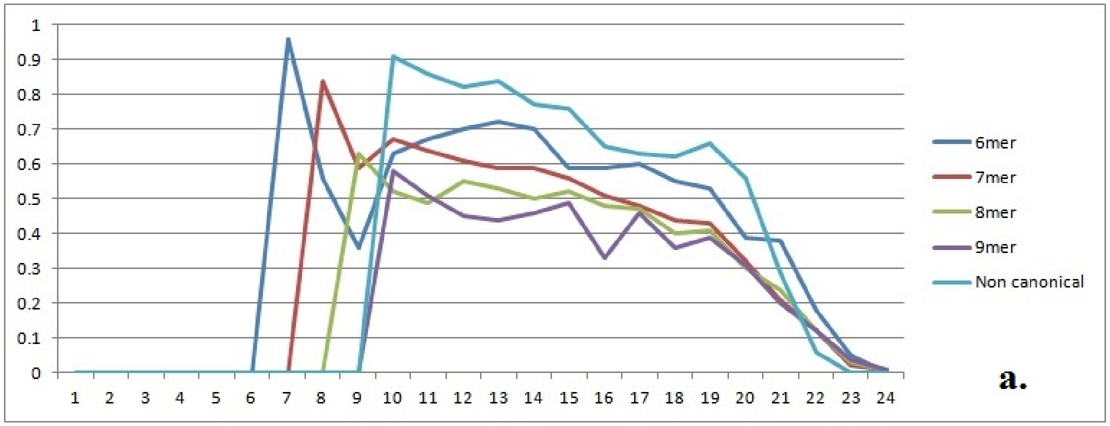
Illumina read

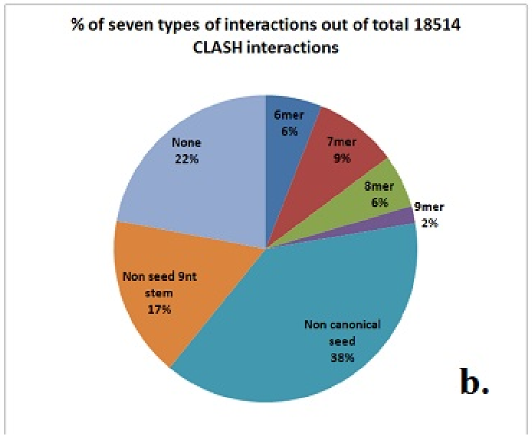
Identify chimeric cDNAs mapping to two regions of the transcriptome

In silico folding shows structure of interaction

Comparison of many interactions (>18K) identifies conserved features in targets

Summary of CLASH data





TarPmir: A new approach based on conventional and new features

- http://hulab.ucf.edu/research/projects/miRNA/TarPmiR/ .

Training datasets

- Positive: CLASH data, 18 514 miRNA target sites of 399 miRNAs from CLASH experiments (Helwak *et al.* , 2013).
- Negative: 18 514 corresponding negative or ‘false’ target sites in a manner similar to a previous study (Li *et al.* , 2014).

Select negative sites

- A positive site and its corresponding negative site are on the same mRNA;
- The positive and its corresponding negative site has similar CG dinucleotide frequency;
- The positive and its corresponding negative site has similar number of the nucleotide G;
- A negative site does not overlap with any positive site; and
- With multiple candidate negative sites in an mRNA, select the one with the lowest folding energy.

Testing data

- CLASH data with cross-validation
- Two PAR-CLIP datasets (17 310 CCRs was from Hafner *et al.* (2010) ; 44 497 CCRs was obtained from Kishore *et al.* (2011).
- HITS-CLIP dataset from the mouse cortex cell (Chi *et al.* , 2009). This dataset provided an Argo–miRNA–mRNA ternary interaction map related to 20 miRNA families, 2953 mRNAs and 11 080 miRNA–mRNA interactions.
- 421 086 POSITIVE TarBase 7.0 miRNA–mRNA interactions in human. We chose the top 100 and 50 miRNAs, which had the largest number of interactions in TabBase 7.0, for further analyses. The top 100 and 50 miRNAs in TarBase 7.0 accounted for 100 608 (23.9%) and 60 818 (14.4%) of human TarBase 7.0 interactions, respectively. There were 9869 and 9823 mRNAs associated with these 100 and 50 top miRNAs, respectively. We ran TarPmiR and other tools with the 100 or 50 miRNAs and the corresponding mRNAs they interacted as input to predict miRNA target sites.

Potential features considered (1)

- (i) folding energy;
- (ii) seed match;
- (iii) accessibility;
- (iv) AU content;
- (v) stem conservation;
- (vi) flanking conservation;
- (vii) difference between stem and flanking conservation;

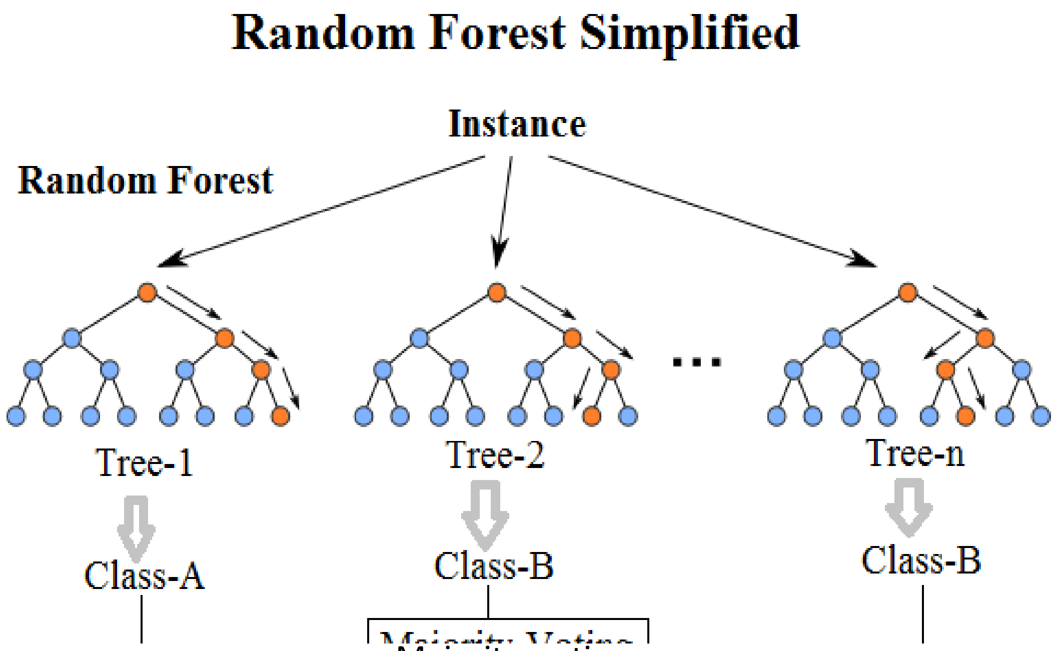
Potential features considered (2)

- (viii) m/e motif;
- (ix) the total number of paired positions;
- (x) the length of the target mRNA region;
- (xi) the length of the largest consecutive pairs;
- (xii) the position of the largest consecutive pairs relative to the miRNA 5′;
- (xiii) the length of the largest consecutive pairs allowing 2 mismatches;
- (xiv) the position of the largest consecutive pairs allowing 2 mismatches;
- (xv) the number of paired positions at the miRNA 3′ end, where 3′ miRNA end was defined as the last 7 positions of the miRNA;
- (xvi) the total number of paired positions in the seed region and the miRNA 3′ end; (xvii) the difference between the number of paired positions in the seed region and that in the miRNA 3′ end
- (xviii) exon preference (Ding *et al.* , 2015).

Four methods for feature selection

- step-wise logistic regression (Ralston and Wilf, 1960)
- least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996)
- randomized logistic regression (Meinshausen and Bühlmann, 2010)
- random forests (Svetnik *et al.* , 2003).

Random forests



Majority voting

Features used

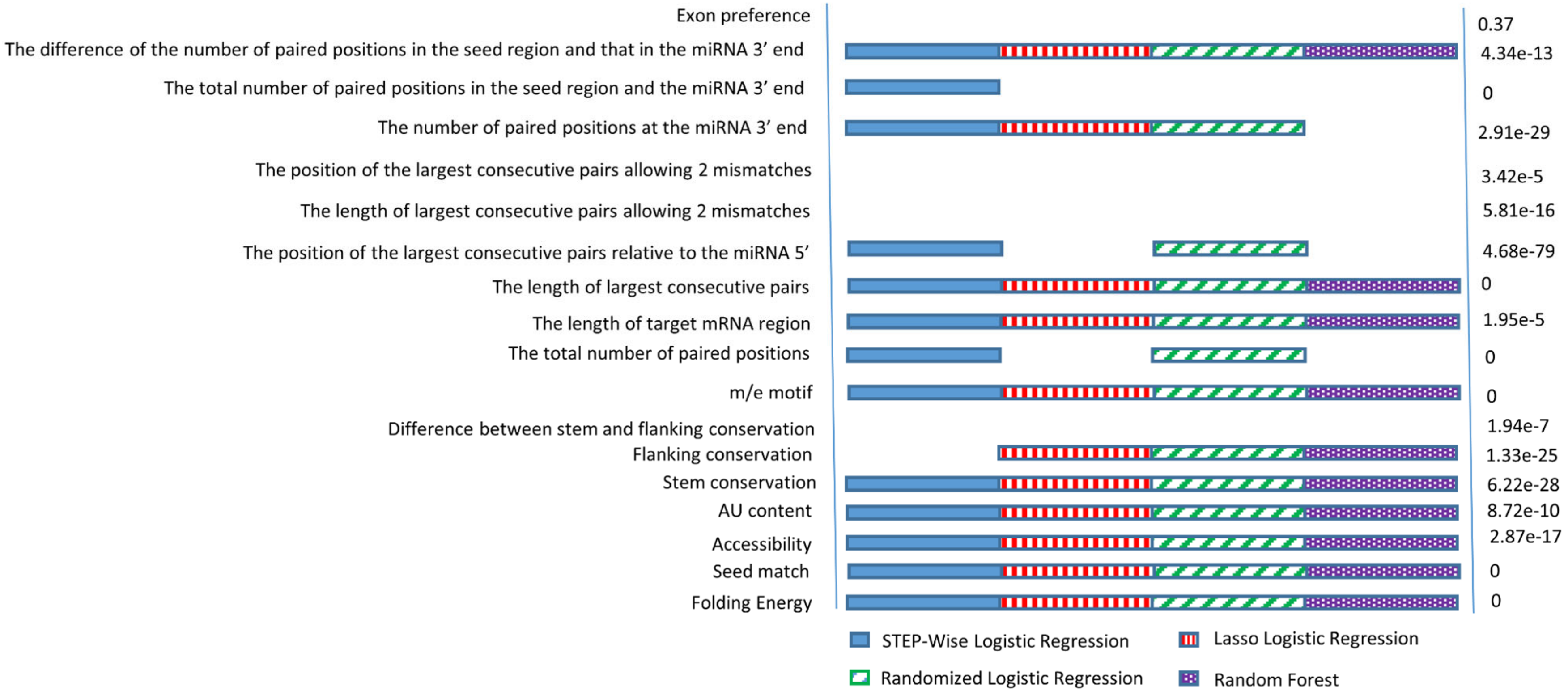


Table 2. Comparison of four methods on independent datasets

Data set	# of miRNAs input	Performance measurement	TarPmiR	miRanda	TargetScan V2010	miRmap	TargetScan V2015
I	60	# of predictions	240605	246311	219304	504447	215885
		% of correct predictions	11904/16041 =74.2%	7061/16041 =44.0%	6248/16041 =39.0%	7121/16041 =44.4%	7472/16041 =46.6%
		Recall	0.742	0.440	0.390	0.444	0.466
		Precision	0.0495	0.0287	0.0285	0.014	0.0346
	120	# of predictions	481135	476827	461280	906654	446074
		% of correct predictions	13846/16041 =86.3%	9683/16041 =60.4%	8969/16041 =55.9%	10342/16041 =64.5%	10614/16041 =66.2%
		Recall	0.863	0.604	0.559	0.645	0.662
		Precision	0.0288	0.0203	0.0194	0.0114	0.0238
II	60	# of predictions	469752	453880	437791	971238	399746
		% of correct predictions	34301/43251 =79.3%	20378/43251 =47.1%	17556/43251 =40.6%	20543/43251 =47.5%	19442/43251 =46.1%
		Recall	0.793	0.471	0.406	0.475	0.461
		Precision	0.0730	0.0449	0.0401	0.0211	0.0486
	120	# of predictions	961112	902611	922373	1952258	832842
		% of correct predictions	38821/43251 = 89.8%	23762/43251 =54.9%	24578/43251 = 56.8%	25667/43251 = 59.3%	27980/43251 =64.7%
		Recall	0.898	0.549	0.568	0.593	0.647
		Precision	0.0403	0.0263	0.0266	0.0131	0.0336
III	119	# of predictions	285491	439485	875442	341773	382173
		% of correct predictions	10766/11080 =97.2%	9069/11080 =81.8%	10084/11080 =91.0%	7840/11080 =70.8%	10334/11080 =93.3%
		Recall	0.972	0.818	0.910	0.708	0.933
		Precision	0.0377	0.0206	0.0115	0.0229	0.0270
IV	50	# of predicted interactions	184842	172256	141717	173378	149142
		% of correct predictions	31779/60818 =52.3%	25326/60818 =41.6%	19873/60818 =32.7%	19785/60818 =32.5%	23757/60818 =39.1%
		Recall	0.523	0.416	0.327	0.325	0.391
		Precision	0.172	0.147	0.140	0.114	0.159
	100	# of predicted interactions	412149	337863	286667	413213	298004
		% of correct predictions	52955/100608 =52.6%	41722/100608 =41.5%	32649/100608 =32.5%	33412/100608 =33.2%	37616/100608 =37.4%
		Recall	0.526	0.415	0.325	0.332	0.374
		Precision	0.128	0.123	0.114	0.081	0.126

Future directions?

Competition and cooperation

Non-seed-matching

……

4th in-class question

Please describe your understanding of the time process of how different research papers are produced on the same topic based on the two recent lectures.