Predict miRNA genes

The first 27 slides are modified from Anastasis Oulas

Outline

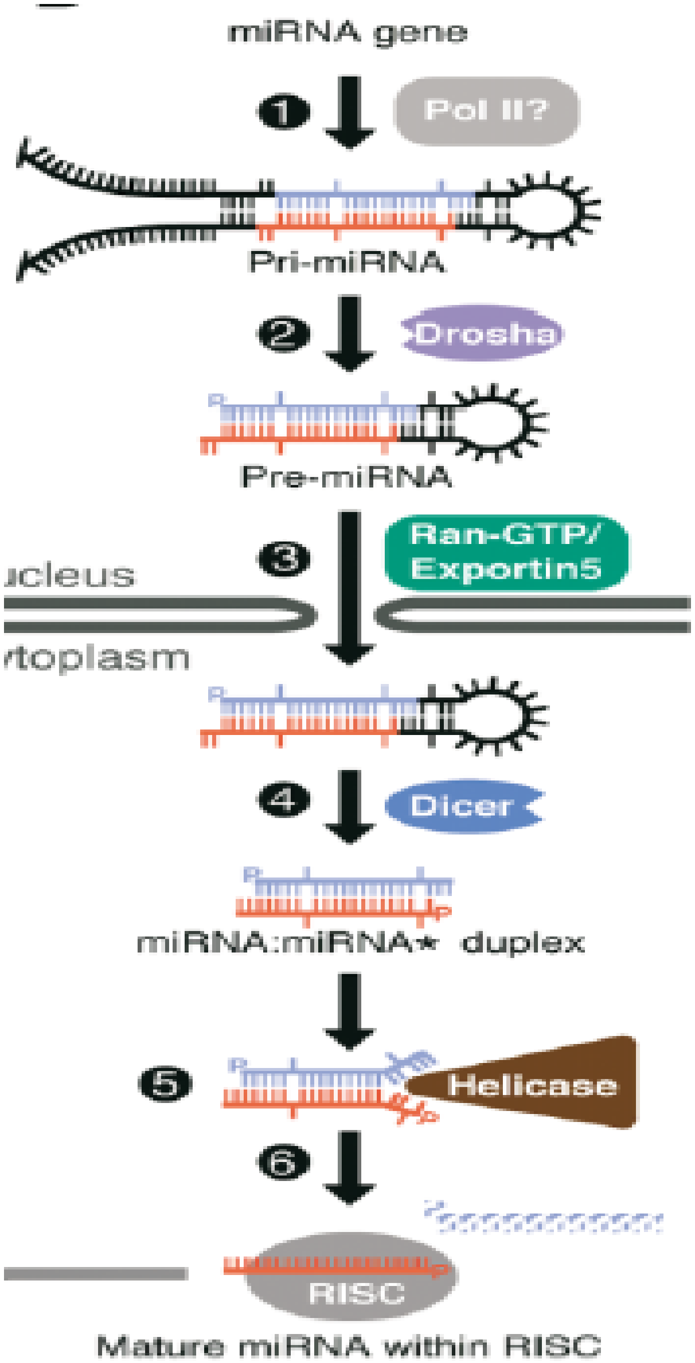
- Review
  - Brief history
  - miRNA Biogenesis
  - Why Computational Methods ?
- Computational Methods
  - Mature and precursor miRNA prediction

Brief history

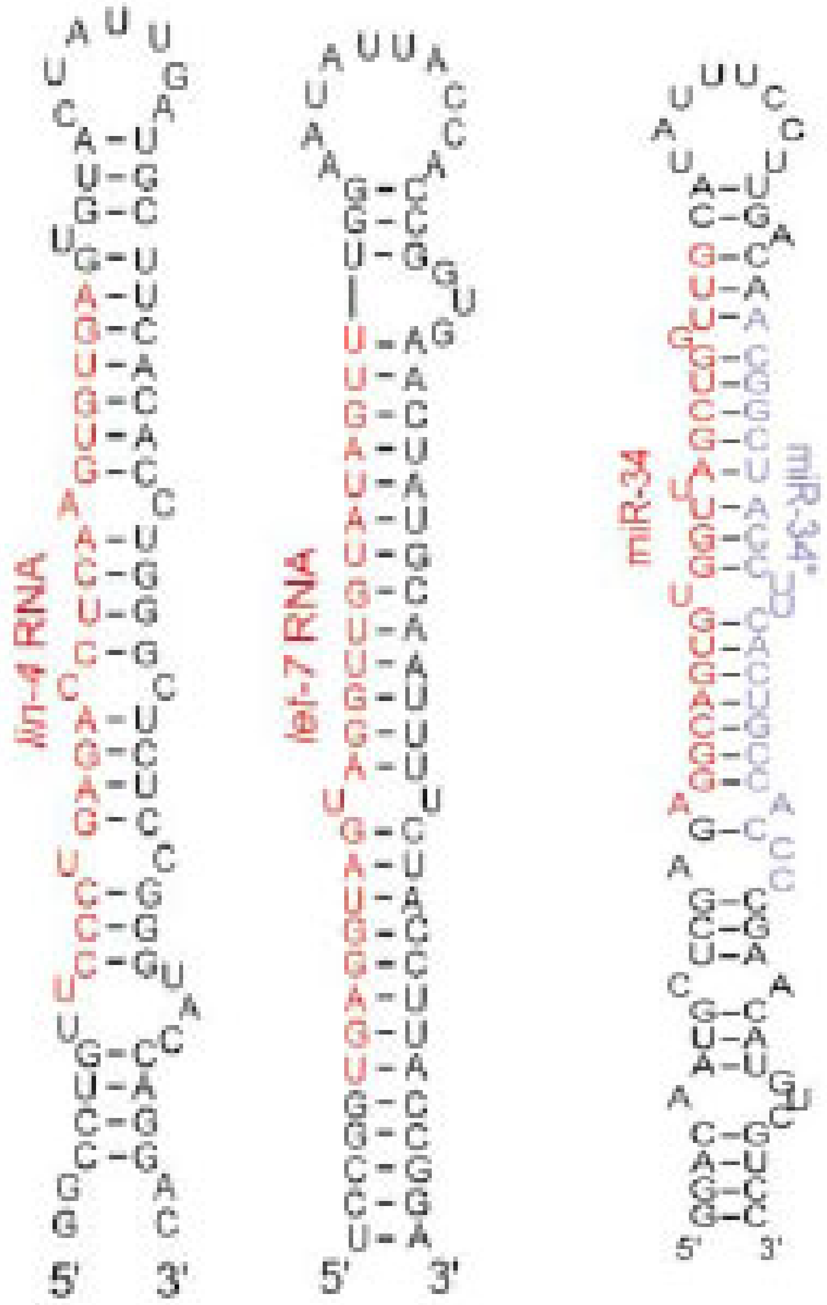
- MicroRNAs (miRNAs) are endogenous ~22 nt RNAs that play important roles in regulating gene expression in animals, plants, and fungi.
- The first miRNAs, lin-4, let-7, were identified in *C. elegans* (Lee R et al. 1993; Reihhart et al. 2000) when they were called small temporal RNAs (stRNA);
- The lin-4 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); actively (5’-p, ~2nt 3’overhang) transported to cytoplasm by Ran-GTP/Exportin5 (3); loop cut by dicer(RNaseIII)(4); *duplex is generally short-lived, by Helicase to single strand RNA, forming RNA-Induced Silencing Complex, RISC/maturation (5-6).



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



miRNA VS siRNA

- 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. Usually reside in introns.
  - siRNA
    - Less sequence conservation
    - Correspond to sequences of known or predicted mRNAs, or heterochromatin.

Computational methods to identify miRNA genes: Why?

- Significant progress has been made in miRNA research since the report of the lin-4 RNA (1993).
- 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.

Outline

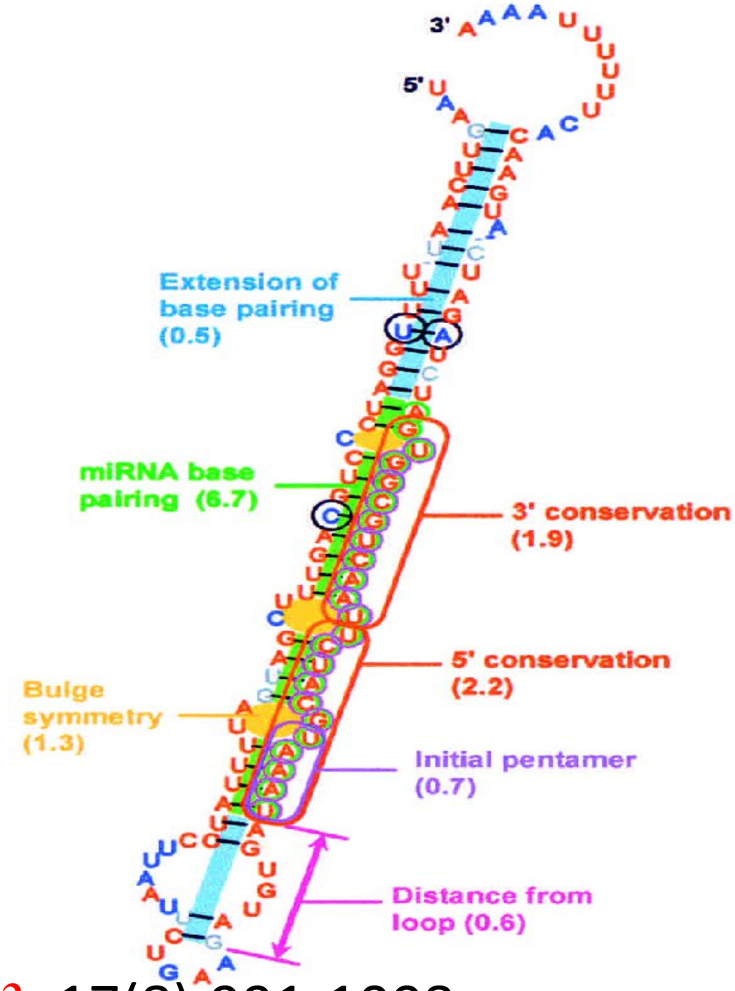
- Review
  - Brief history
  - miRNA Biogenesis
  - Why Computational Methods ?
- Computational Methods
  - Mature and precursor miRNA prediction

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

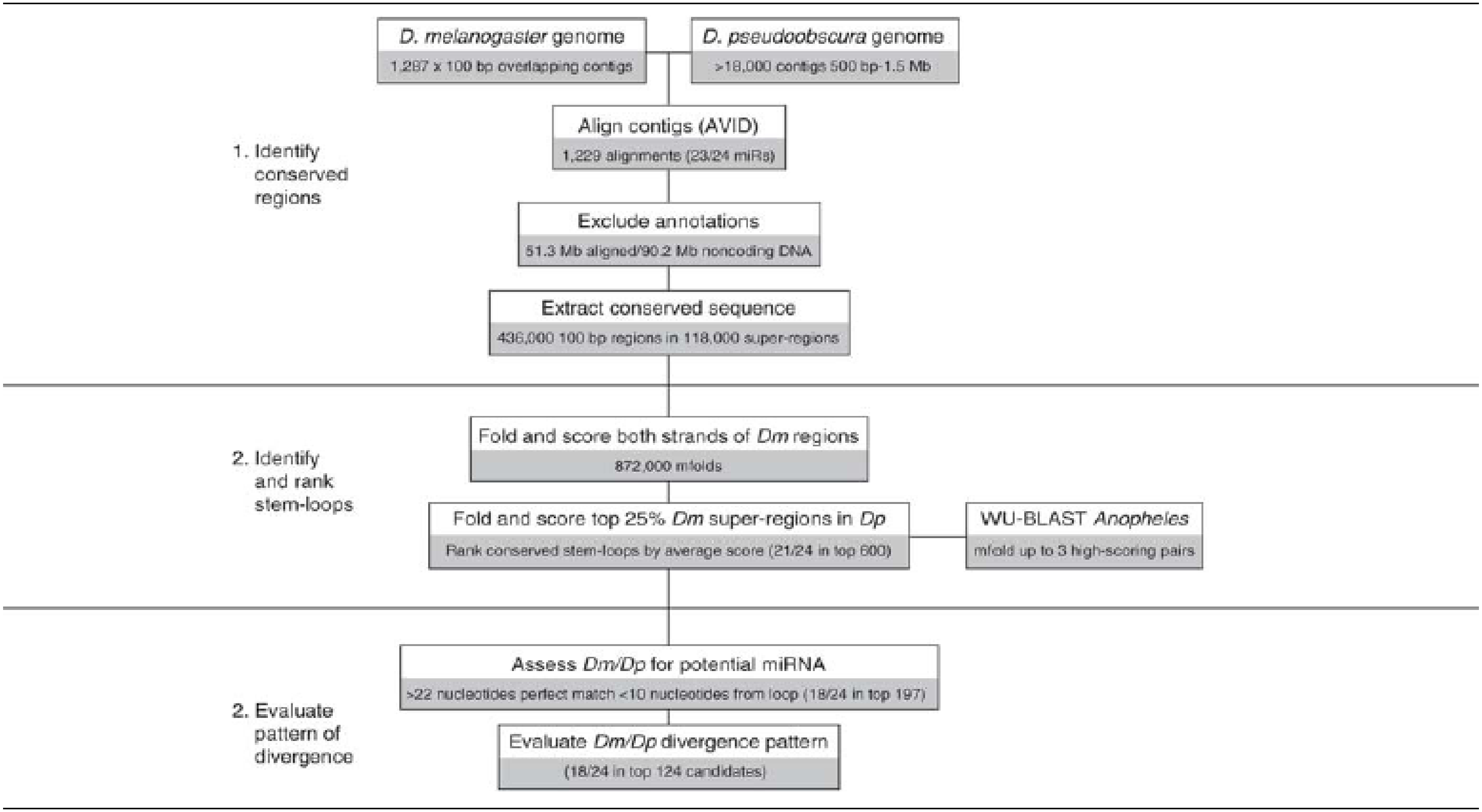


Lim et al, Genes and Development 2003, 17(8):991-1008

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”



1. Identify conserved regions

D. melanogaster genome
1,287 x 100 bp overlapping contigs

D. pseudoobscura genome
>18,000 contigs 500 bp-1.5 Mb

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)

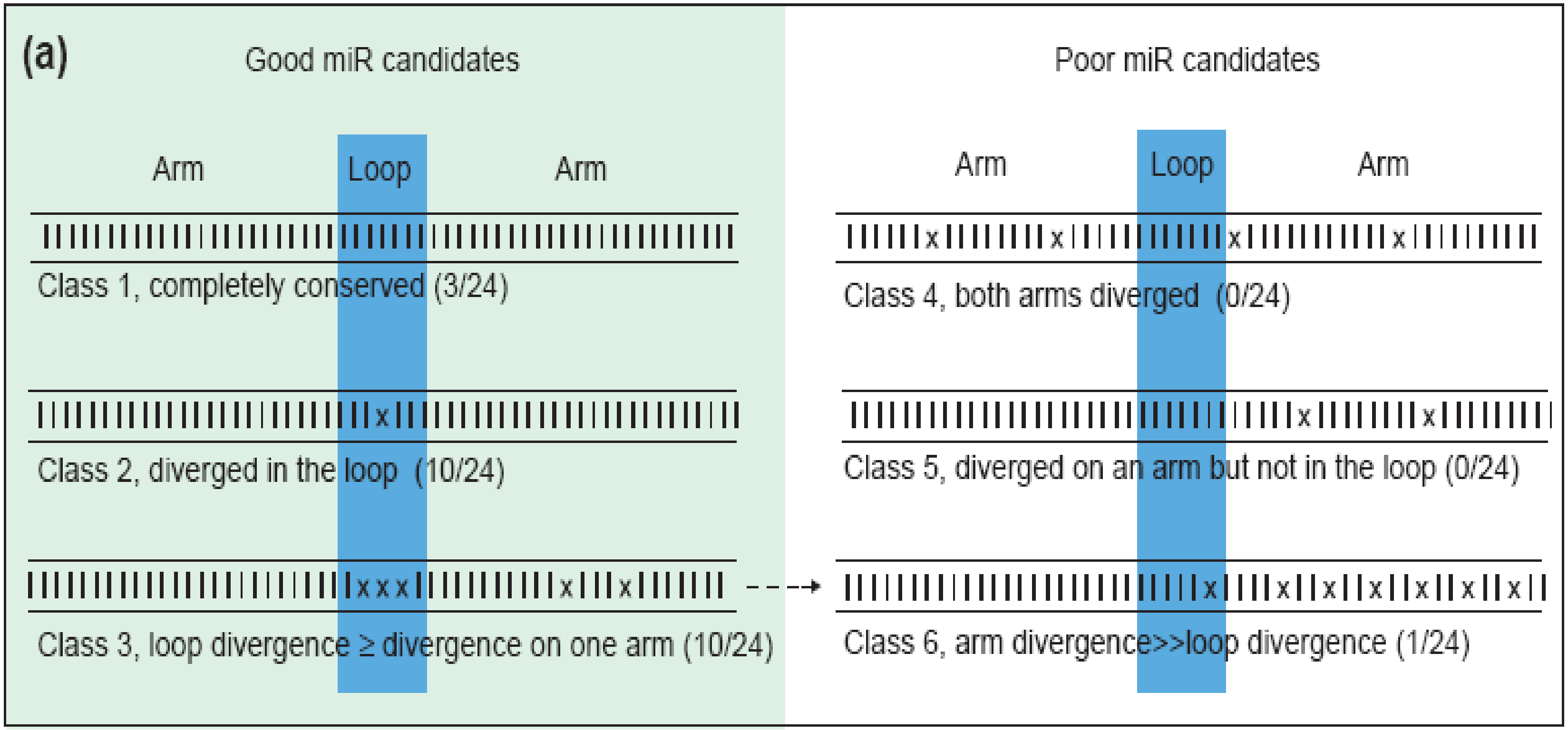
WU-BLAST *Anopheles*
mfold up to 3 high-scoring pairs

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

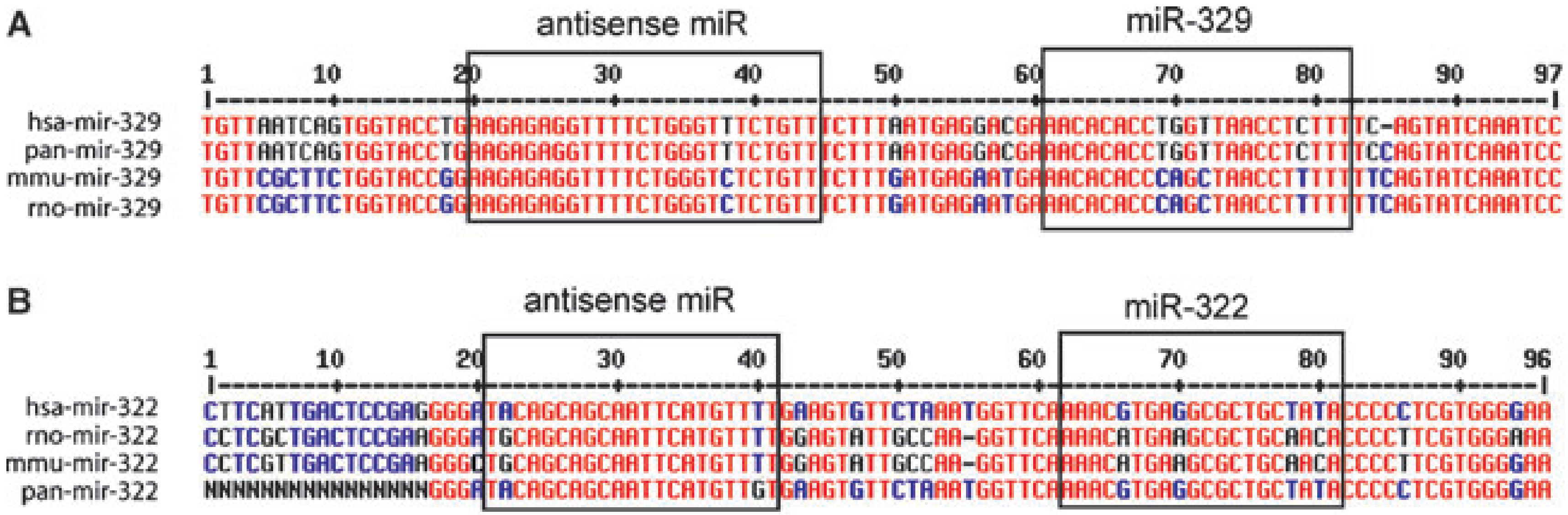
Results

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, (median score of 58 known miRNAs). Of these 35 were carried forward for experimental validation. 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.

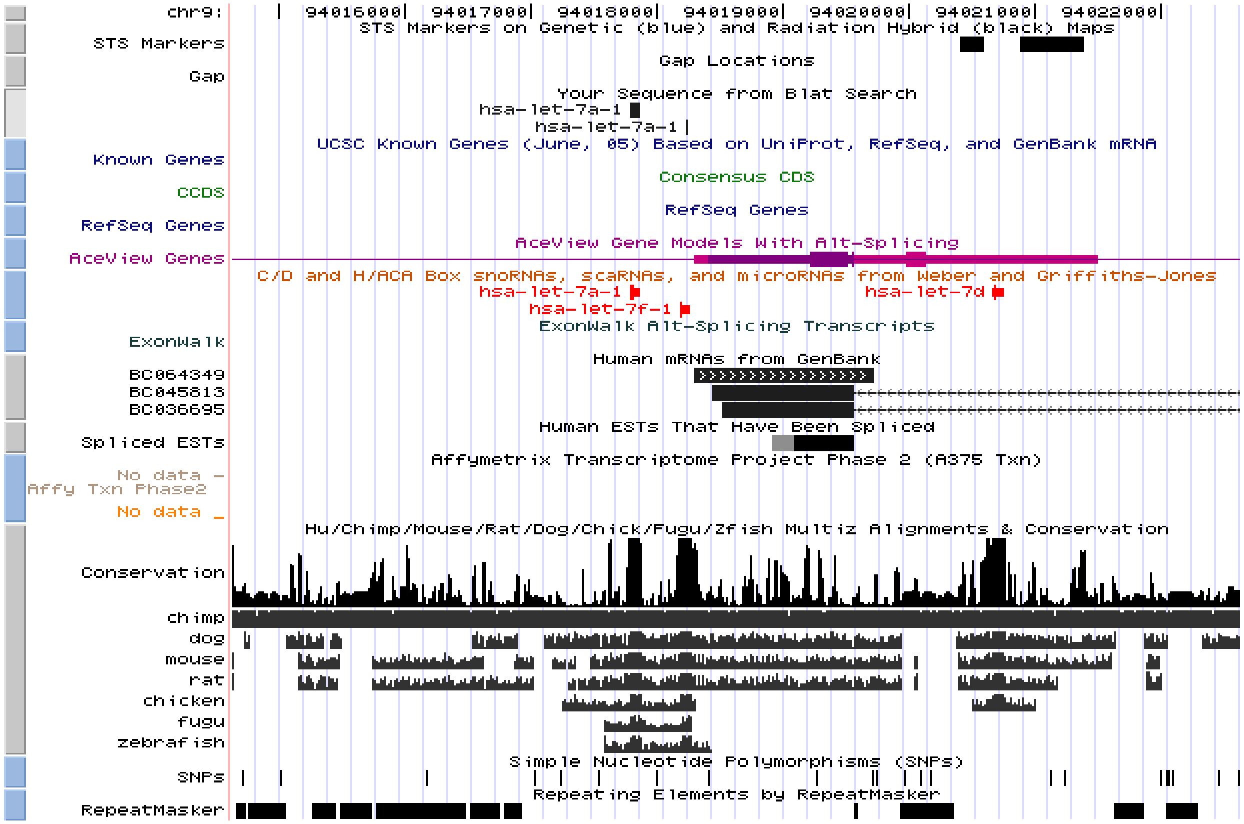
60 new potential miRNAs (15 for human and 45 for mouse)


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.

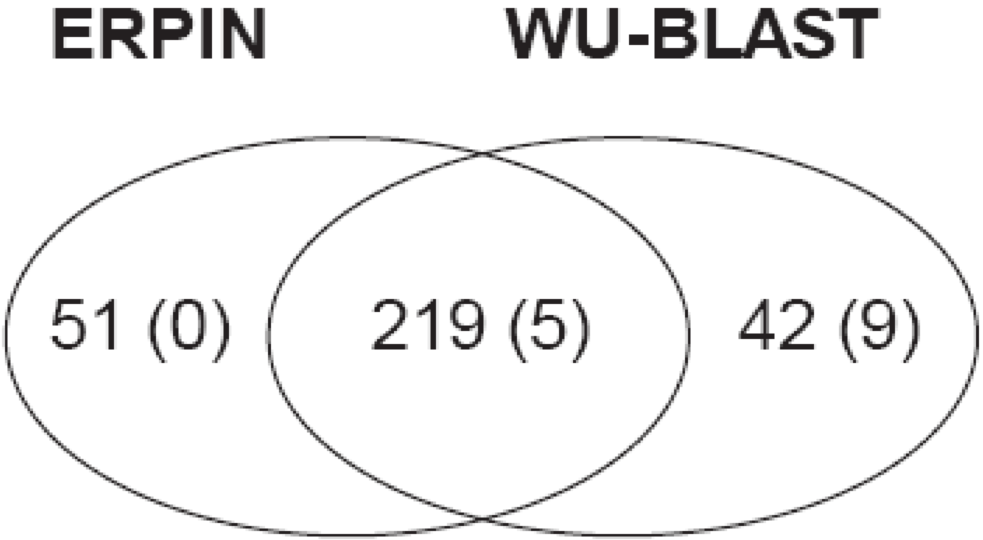
Limitations of methods so far

- 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).*

Profile-based detection of mRNAs

- 593 sequences form miRNA registry (513 animal and 50 plant)
- CLUSTALW 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: 270/593 top scoring ERPIN candidates previously un-identified



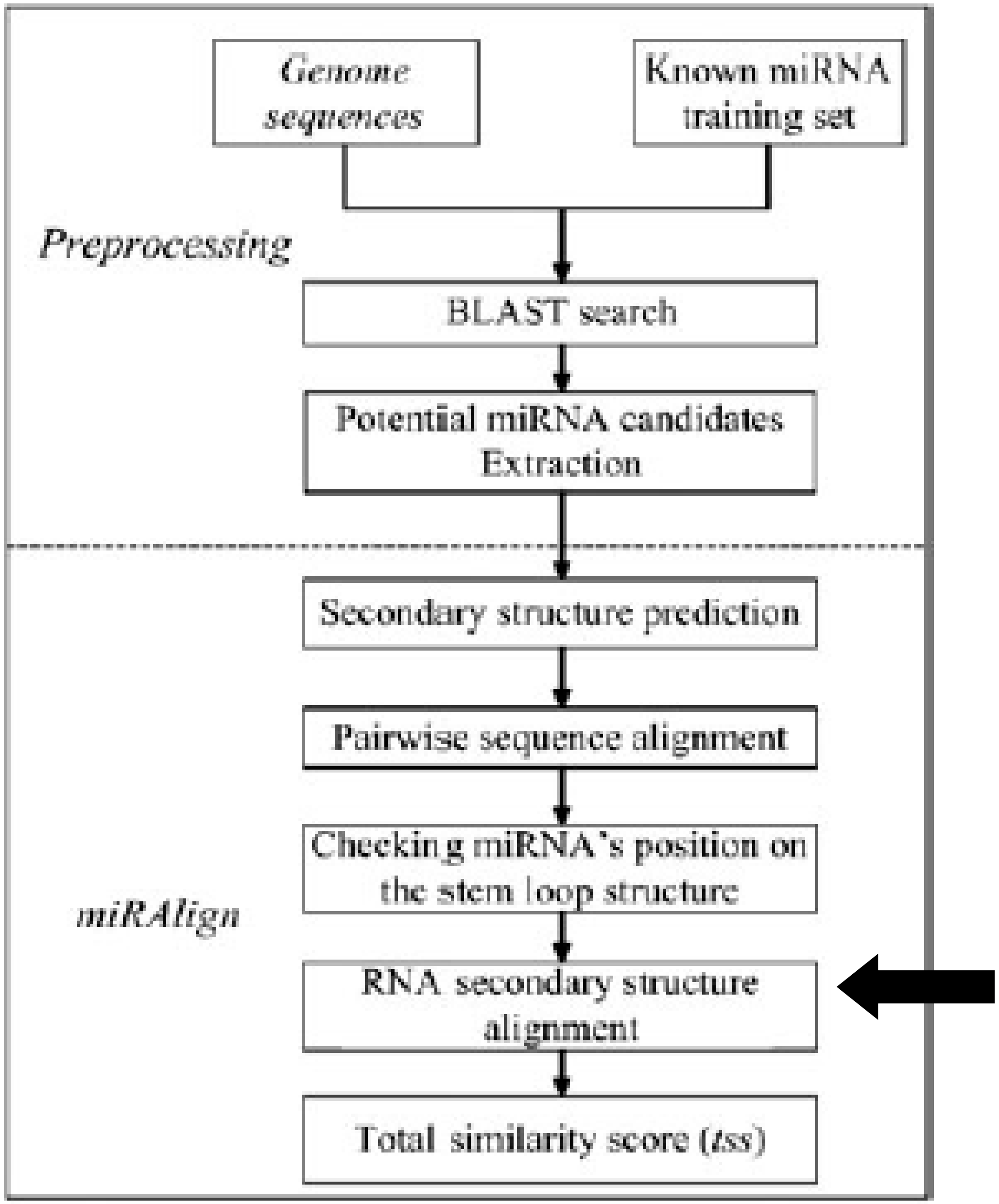
- *Adv:Takes into account 2ry structure conservation using Profiles.*
- *Disadv: Only applicable to miRNA families with sufficient known samples.*

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

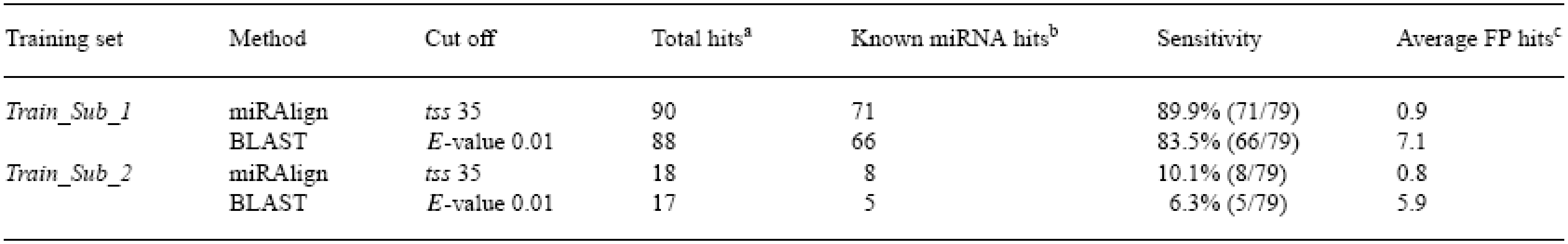
miRAlign

RNA secondary structure alignment

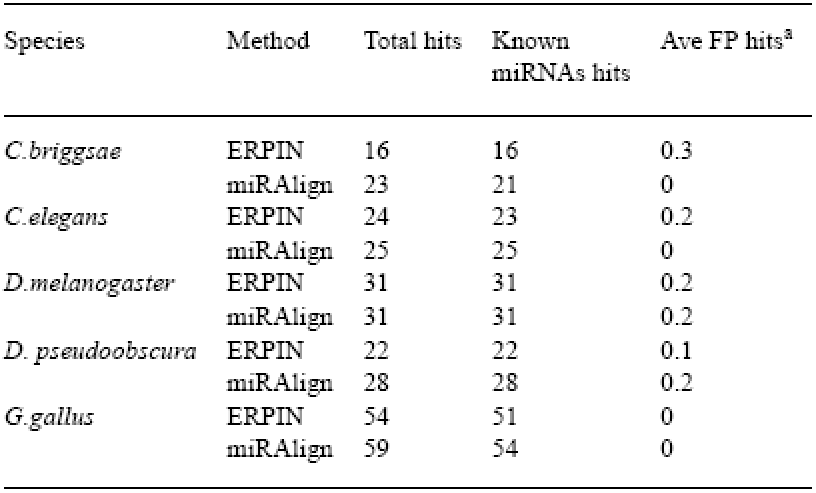
Total similarity score (*tss*)

RNAforeseter

Comparison to other programs



Training set	Method	Cut off	Total hits[a]	Known miRNA hits[b]	Sensitivity	Average FP hits[c]
Train_Sub_1	miRAlign	*tss* 35	90	71	89.9% (71/79)	0.9
	BLAST	*E*-value 0.01	88	66	83.5% (66/79)	7.1
Train_Sub_2	miRAlign	*tss* 35	18	8	10.1% (8/79)	0.8
	BLAST	*E*-value 0.01	17	5	6.3% (5/79)	5.9



Species	Method	Total hits	Known miRNAs hits	Ave FP hits[a]
C.briggsae	ERPIN	16	16	0.3
	miRAlign	23	21	0
C.elegans	ERPIN	24	23	0.2
	miRAlign	25	25	0
D.melanogaster	ERPIN	31	31	0.2
	miRAlign	31	31	0.2
D. pseudoobscura	ERPIN	22	22	0.1
	miRAlign	28	28	0.2
G.gallus	ERPIN	54	51	0
	miRAlign	59	54	0

Adv: Takes into account 2ry structure conservation by aligning 2ry structures. Applicable to all miRNA families

Disadv: Highly dependent on homology and BLAST, breaks down when more distantly related sequences are scanned

Wang et al, Bioinformatics 2005

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).*

Identification of miRNA genes on the genome scale

Some slides are from Mayukh Bhaowal

Two papers

- *Nature* **423**, 241-254 (15 May 2003)

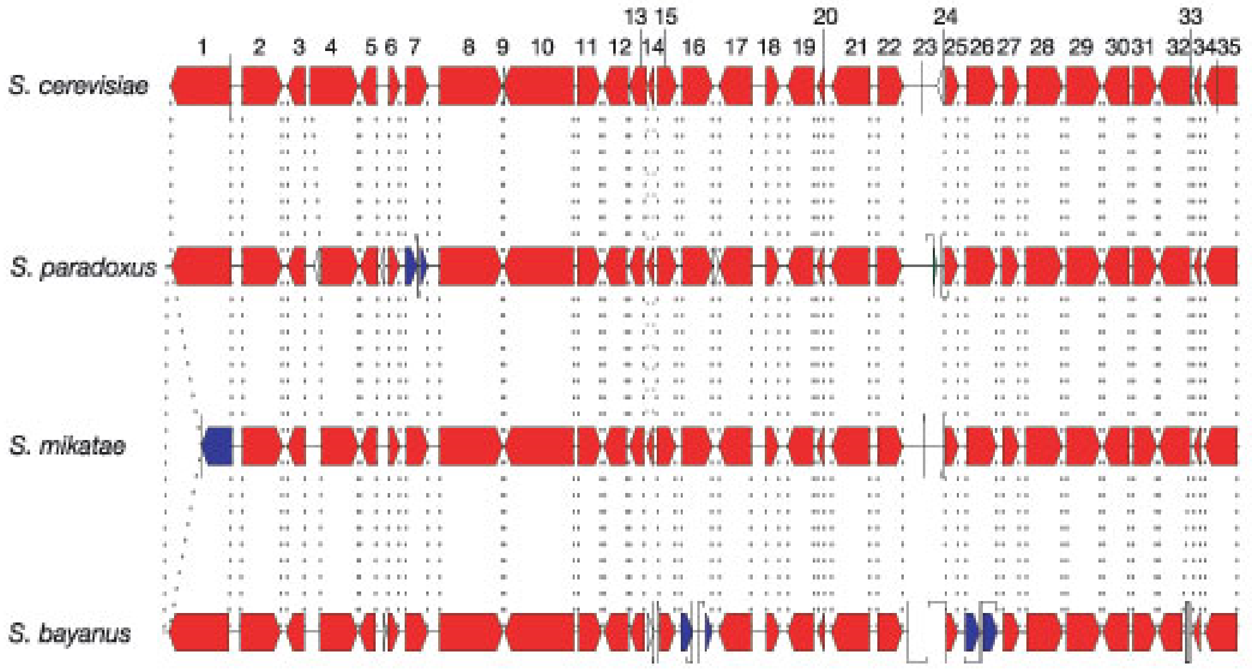
  Sequencing and comparison of yeast species to identify genes and regulatory elements

  Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, Eric S. Lander
- *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

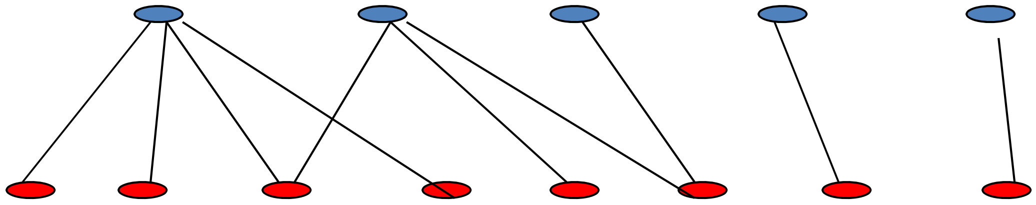
A block of conserved synteny



1	YGL140C
2	YGL139W
3	YGL138C
4	YGL137W (*SEC27*)
5	YGL136C
6	YGL135W (*RPL1B*)
7	YGL134W (*PCL10*)
8	YGL133W
9	YGL132W
10	YGL131C
11	YGL130W (*CEG1*)
12	YGL129C
13	YGL128C
14	YGL127C (*SOH1*)
15	YGL126W (*SCS3*)
16	YGL125W (*MET13*)
17	YGL124C
18	YGL123W (*RPS2*)
19	YGL122C (*NAB2*)
20	YGL121C
21	YGL120C (*PRP43*)
22	YGL119W (*ABC1*)
23	tW(CCA)G1
24	YGL118C
25	YGL117W
26	YGL116W (*CDC20*)
27	YGL115W (*SNF4*)
28	YGL114W
29	YGL113W
30	YGL112C (*TAF60*)
31	YGL111W
32	YGL110C
33	YGL109W
34	YGL108C
35	YGL107C

Bipartite graph to determine one-to-one orthologous ORFs

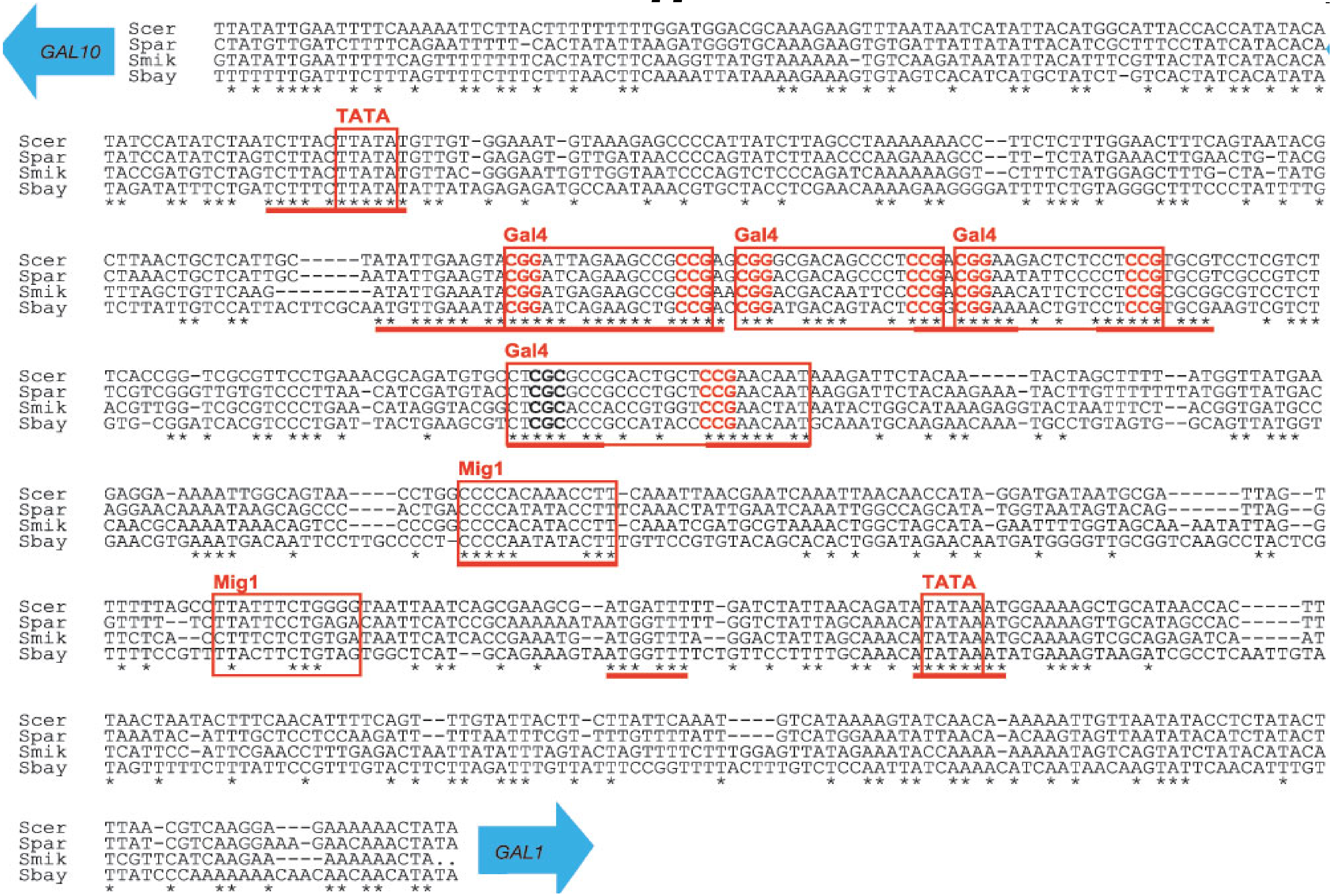
1. For each species, gather the complete set of all predicted ORFs starting with a methionine and having a length of at least 50 amino acids.
2. Connecte predicted ORFs to *S. cerevisiae* ORFs in a bipartite graph, based on blast hits.
3. Eliminate all edges that are less than 80% of the maximum-weight edge both in amino-acid identity and in length.
4. Built blocks of conserved gene order (synteny) on the basis of the resulting unambiguous matches.
5. Preferentially keep matches within synteny blocks and resolve additional matches.
6. Separate gene families into subfamilies by searching for best unambiguous subgraphs.
7. Report the connected components of the resulting graph as homology groups of unresolved genes.



Step 2: define mini motifs

Mini-motifs are sequences of the form XYZn(0–21)UVW, consisting of two triplets of specified bases interrupted by a fixed number (from 0 to 21) of unspecified bases. Examples are TAGGAT, ATAnnGGC, or the Gal4 motif itself. The total number of distinct mini-motifs is 45,760, if reverse complements are grouped together.

Conservation in the *GAL1–GAL10* intergenic region



```
Scer  TTATATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACA
Spar  CTATGTTGATCTTTTCAGAATTTTT-CACTATATTAAGATGGGTGCAAAGAAGTGTGATTATTATATTACATCGCTTTCCTATCATACACA
Smik  GTATATTGAATTTTTCAGTTTTTTTTCACTATCTTCAAGGTTATGTAAAAAA-TGTCAAGATAATATTACATTTCGTTACTATCATACACA
Sbay  TTTTTTTGATTTCTTTAGTTTTCTTTCTTTAACTTCAAAATTATAAAAGAAAGTGTAGTCACATCATGCTATCT-GTCACTATCACATATA
       * * ****  * *  *   ** ** *  *   **           **  ** * *    *    **   **    *  *  * ** * * *

Scer  TATCCATATCTAATCTTACTTATATGTTGT-GGAAAT-GTAAAGAGCCCCATTATCTTAGCCTAAAAAAACC--TTCTCTTTGGAACTTTCAGTAATACG
Spar  TATCCATATCTAGTCTTACTTATATGTTGT-GAGAGT-GTTGATAACCCCAGTATCTTAACCCAAGAAAGCC--TT-TCTATGAAACTTGAACTG-TACG
Smik  TACCGATGTCTAGTCTTACTTATATGTTAC-GGGAATTGTTGGTAATCCCAGTCTCCCAGATCAAAAAAGGT--CTTTCTATGGAGCTTTG-CTA-TATG
Sbay  TAGATATTTCTGATCTTTCTTATATATTATAGAGAGATGCCAATAAACGTGCTACCTCGAACAAAAGAAGGGGATTTTCTGTAGGGCTTTCCCTATTTTG
      **   ** ***  **** ******* **    *  *   *     *  *    *  *       ** **       * *** *    ***    *  *  *

Scer  CTTAACTGCTCATTGC-----TATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGCGTCCTCGTCT
Spar  CTAAACTGCTCATTGC-----AATATTGAAGTACGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAATATTCCCCTCCGTGCGTCGCCGTCT
Smik  TTTAGCTGTTCAAG--------ATATTGAAATACGGATGAGAAGCCGCCGAACGGACGACAATTCCCCGACGGAACATTCTCCTCCGCGCGGCGTCCTCT
Sbay  TCTTATTGTCCATTACTTCGCAATGTTGAAATACGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGGCGGAAAACTGTCCTCCGTGCGAAGTCGTCT
            ** **          ** ***** ******* ****** ***** ***  ****   * *** ***** * *  ****** ***    * ***

Scer  TCACCGG-TCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
Spar  TCGTCGGGTTGTGTCCCTTAA-CATCGATGTACCTCGCGCCGCCCTGCTCCGAACAATAAGGATTCTACAAGAAA-TACTTGTTTTTTTATGGTTATGAC
Smik  ACGTTGG-TCGCGTCCCTGAA-CATAGGTACGGCTCGCACCACCGTGGTCCGAACTATAATACTGGCATAAAGAGGTACTAATTTCT--ACGGTGATGCC
Sbay  GTG-CGGATCACGTCCCTGAT-TACTGAAGCGTCTCGCCCCGCCATACCCCGAACAATGCAAATGCAAGAACAAA-TGCCTGTAGTG--GCAGTTATGGT
           ** *   ** *** *      *      ***** ** *  *   ****** **     *   * **     * *              ** ***

Scer  GAGGA-AAAATTGGCAGTAA----CCTGGCCCCACAAACCTT-CAAATTAACGAATCAAATTAACAACCATA-GGATGATAATGCGA------TTAG--T
Spar  AGGAACAAAATAAGCAGCCC----ACTGACCCCATATACCTTTCAAACTATTGAATCAAATTGGCCAGCATA-TGGTAATAGTACAG------TTAG--G
Smik  CAACGCAAAATAAACAGTCC----CCCGGCCCCACATACCTT-CAAATCGATGCGTAAAACTGGCTAGCATA-GAATTTTGGTAGCAA-AATATTAG--G
Sbay  GAACGTGAAATGACAATTCCTTGCCCCT-CCCCAATATACTTTGTTCCGTGTACAGCACACTGGATAGAACAATGATGGGGTTGCGGTCAAGCCTACTCG
             ****    *         *   ****     ***               * * *    *  * *    *     *           **

Scer  TTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG--ATGATTTTT-GATCTATTAACAGATATATAAATGGAAAAGCTGCATAACCAC-----TT
Spar  GTTTT--TCTTATTCCTGAGACAATTCATCCGCAAAAAATAATGGTTTTT-GGTCTATTAGCAAACATATAAATGCAAAAGTTGCATAGCCAC-----TT
Smik  TTCTCA--CCTTTCTCTGTGATAATTCATCACCGAAATG--ATGGTTTA--GGACTATTAGCAAACATATAAATGCAAAAGTCGCAGAGATCA-----AT
Sbay  TTTTCCGTTTTACTTCTGTAGTGGCTCATGCAGAAAGTAATGGTTTTCTGTTCCTTTTGCAAACATATAAATATGAAAGTAAGATCGCCTCAATTGTA
       * *      *    ***       * **   *  *     *** ***   *  *  **  ** * ********   ****    *

Scer  TAACTAATACTTTCAACATTTTCAGT--TTGTATTACTT-CTTATTCAAAT----GTCATAAAAGTATCAACA-AAAAATTGTTAATATACCTCTATACT
Spar  TAAATAC-ATTTGCTCCTCCAAGATT--TTTAATTTCGT-TTTGTTTTATT----GTCATGGAAATATTAACA-ACAAGTAGTTAATATACATCTATACT
Smik  TCATTCC-ATTCGAACCTTTGAGACTAATTATATTTAGTACTAGTTTTCTTTGGAGTTATAGAAATACCAAAA-AAAAATAGTCAGTATCTATACATACA
Sbay  TAGTTTTTCTTTATTCCGTTTGTACTTCTTAGATTTGTTATTTCCGGTTTTACTTTGTCTCCAATTATCAAAACATCAATAACAAGTATTCAACATTTGT
      *   *     *     *      * *  **  ***   *  *        *        *  ** *** ** * *  * *    * ***       *

Scer  TTAA-CGTCAAGGA---GAAAAAACTATA
Spar  TTAT-CGTCAAGGAAA-GAACAAACTATA
Smik  TCGTTCATCAAGAA----AAAAAACTA..
Sbay  TTATCCCAAAAAAACAACAACAACATATA
      *    *   **  *    ** **  **
```

Step 3: define conserved mini motifs

Conserved mini-motifs are then defined according to three conservation criteria (CC1–3). In each case, conservation rates are normalized to appropriate random controls.

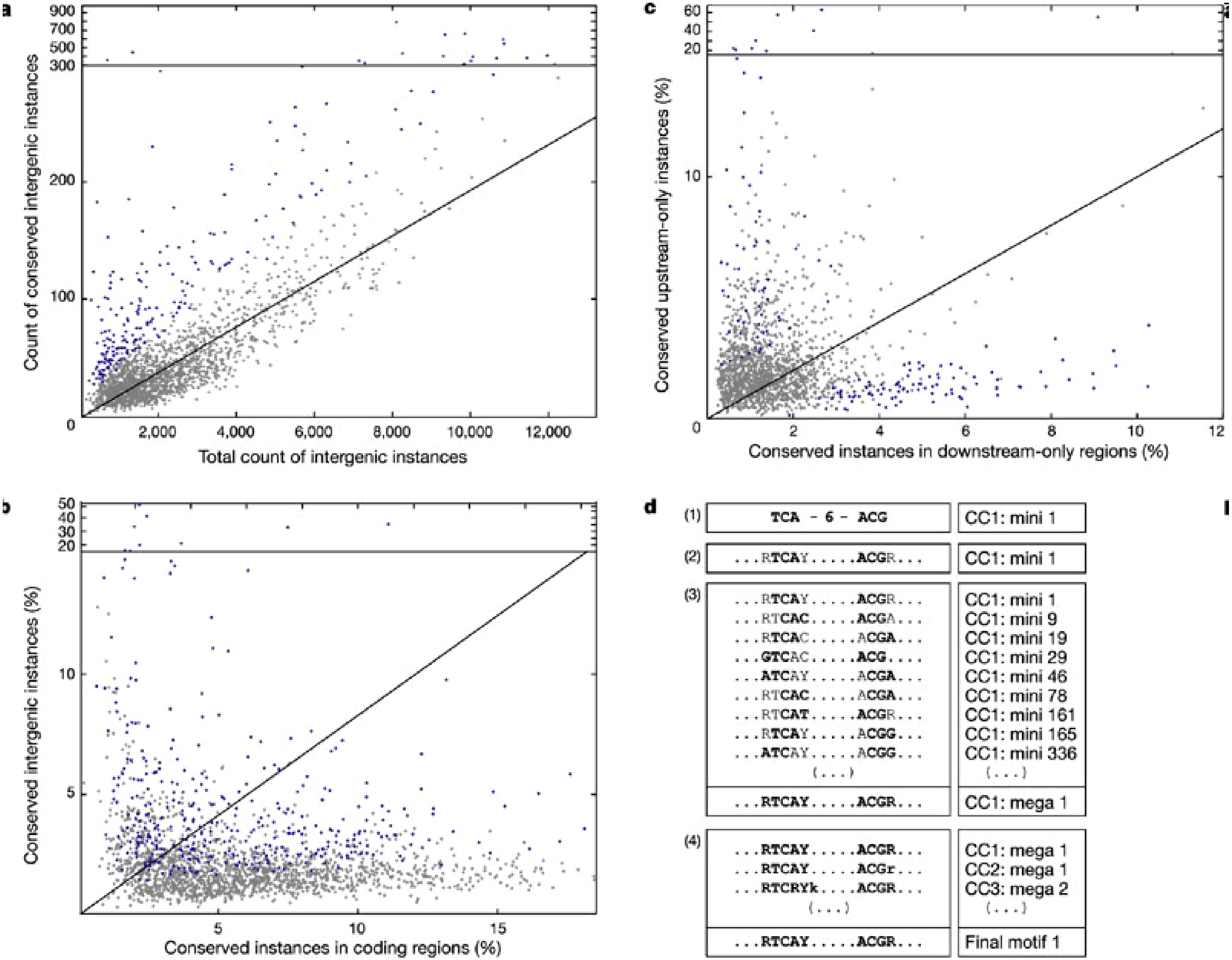
(1) intergenic conservation (CC1), the mini-motif shows a significantly high conservation rate in intergenic regions;

(2) intergenic–genic conservation (CC2), the mini-motif shows significantly higher conservation in intergenic regions than in genic regions;

(3) upstream–downstream conservation (CC3), the mini-motif shows significantly different conservation rates when it occurs upstream compared with downstream of a gene.

Conservation criteria

For CC1, for every mini-motif, count *i*c, the number of perfectly conserved intergenic instances in all four species, and *i*, the total number of intergenic instances in *S. cerevisiae*. They found that the two counts seem linearly related for most of the patterns, which can be attributed to a basal level of conservation *r* given the total evolutionary distance that separates the four species compared. They estimated the ratio *r* as the log-average of non-outlier instances of *i*c/*i* within a control set of all motifs at a given gap size. They then calculated for every motif the binomial probability *p* of observing *i*c successes out of *i* trials, given parameter *r*. We assigned a *z*-score *S* to every motif corresponding to probability *p*. Similar methods were used for computing CC2 and CC3.


a
900
700
500
300
200
100
0
Count of conserved intergenic instances
2,000
4,000
6,000
8,000
10,000
12,000
Total count of intergenic instances
b
50
40
30
20
10
5
Conserved intergenic instances (%)
5
10
15
Conserved instances in coding regions (%)
c
60
40
20
10
0
Conserved upstream-only instances (%)
2
4
6
8
10
12
Conserved instances in downstream-only regions (%)
d
(1)
TCA - 6 - ACG
CC1: mini 1
(2)
...RTCAY.....ACGR...
CC1: mini 1
(3)
...RTCAY.....ACGR...
...RTCAC.....ACGA...
...RTCAC.....ACGA...
...GTCAC.....ACG....
...ATCAY.....ACGA...
...RTCAC.....ACGA...
...RTCAT.....ACGR...
...RTCAY.....ACGG...
...ATCAY.....ACGG...
(...)
CC1: mini 1
CC1: mini 9
CC1: mini 19
CC1: mini 29
CC1: mini 46
CC1: mini 78
CC1: mini 161
CC1: mini 165
CC1: mini 336
(...)
...RTCAY.....ACGR...
CC1: mega 1
(4)
...RTCAY.....ACGR...
...RTCAY.....ACGr...
...RTCRYk....ACGR...
(...)
CC1: mega 1
CC2: mega 1
CC3: mega 2
(...)
...RTCAY.....ACGR...
Final motif 1

Step 4: construct full motifs

The conserved mini-motifs are then used to construct full motifs. They are first extended by searching for nearby sequence positions showing significant correlation with a mini-motif. The extended motifs are then clustered, merging those with substantially overlapping sequences and those that tend to occur in the same intergenic regions. Finally, a full motif is created by deriving a consensus sequence (which may be degenerate).

MCS: motif conservation score

1. compute the table *F* containing the relative frequencies of twofold and threefold degenerate bases, given the *S. cerevisiae* nucleotide frequencies.
2. select 20 random intergenic loci in *S. cerevisiae* and used the order of nucleotides at each locus together with the order of degeneracy levels in the motif *m* to construct 20 random motifs that were used as controls.
3. count conserved and non-conserved instances of each control and computed *r*, the log-average of their conservation rates.
4. count the number of conserved and non-conserved intergenic instances of *m*, and evaluate the binomial probability *p* of observing the two counts, given *r*.
5. report the MCS of the motif as a *z*-score corresponding to *p*, the number of standard deviations away from the mean of a normal distribution that corresponds to tail area *p*.

Results of paper one

The conserved mini-motifs give rise to a list of 72 full motifs having MCS>=4. Most of the motifs show preferential enrichment upstream of genes, but six are enriched downstream of genes. These 72 discovered motifs, found with no previous biological knowledge, show strong overlap with 28 of the 33 known motifs having MCS>=4. They include 27 strong matches and one weaker match.

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

3’ UTR

- The **three prime untranslated region** (3' UTR) is a particular section of messenger RNA (mRNA).
- An mRNA codes for a protein through translation. The mRNA also contains regions that are not translated. In eukaryotes the 5' untranslated region, 3' untranslated region, cap and polyA tail.

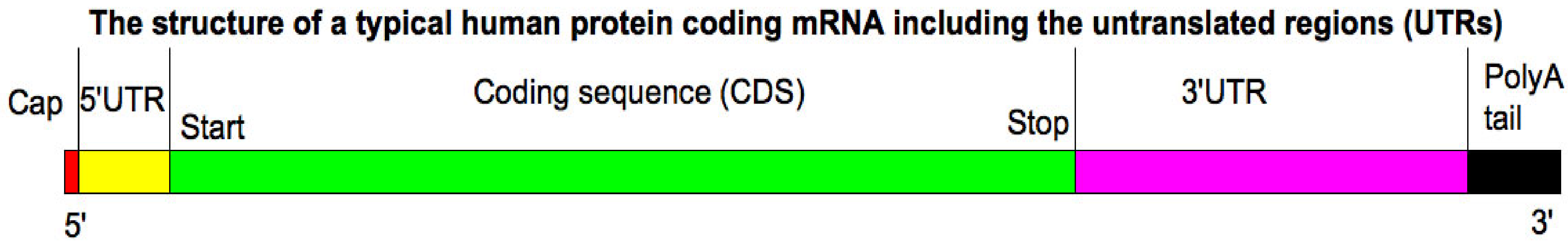
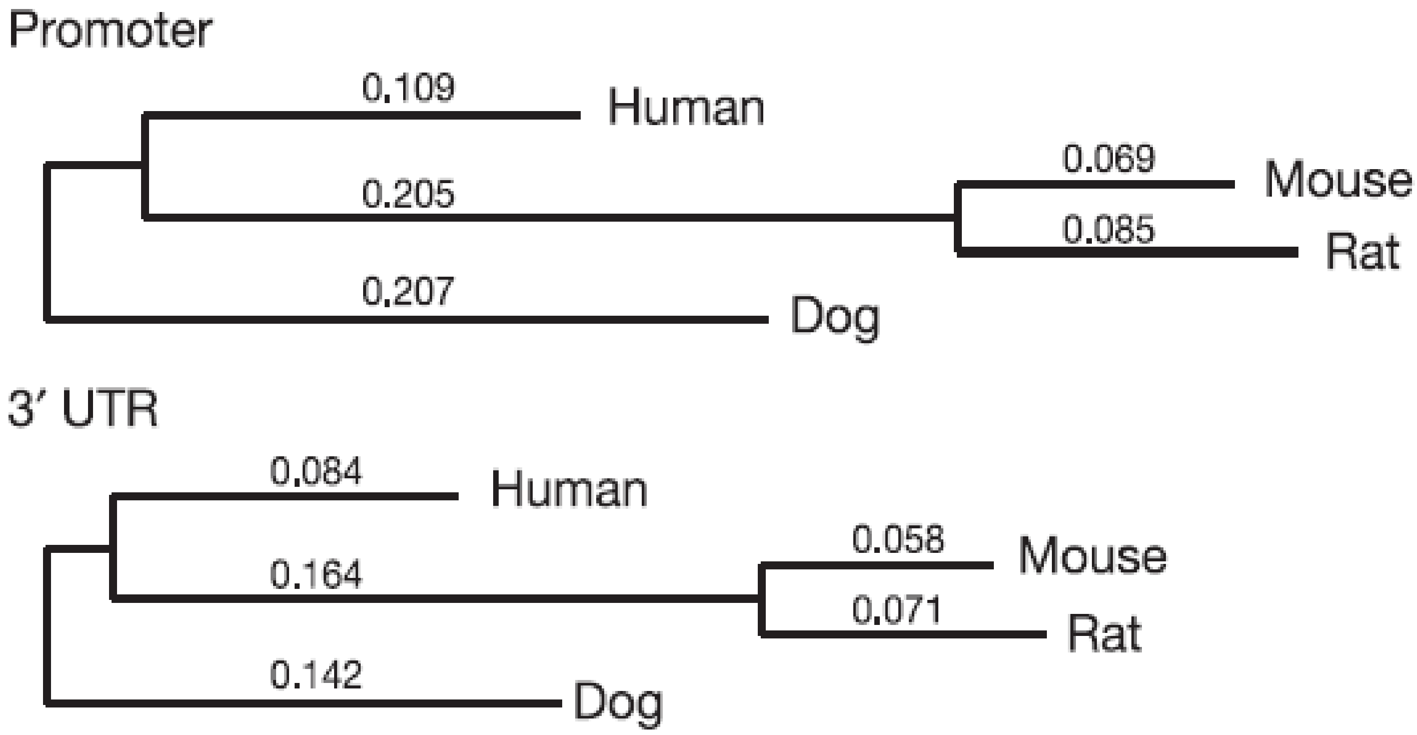

The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5'UTR
Start
Coding sequence (CDS)
Stop
3'UTR
PolyA tail
5'
3'

Image source : http://en.wikipedia.org/wiki/Image:MRNA_structure.png

What the paper proposes

- What? Discovering the regulatory motifs in human promoters and 3’ UTRs.
- How? By comparing sequence motifs of several mammals. That’s why it is called comparative motif finding.
- Which mammals? Human, mouse, rat, dog.

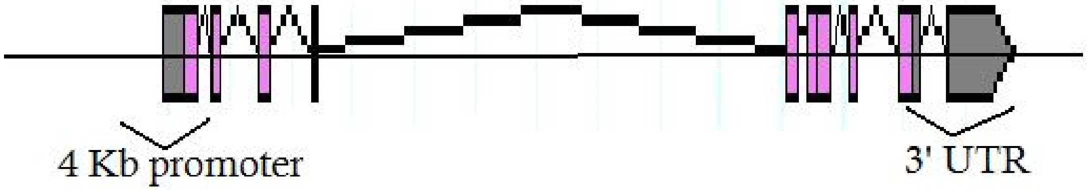
Conservation properties



Methods

- Chose 17,700 well annotated genes from RefSeq database.
- Promoters = 4kb centered at transcriptional start site (only noncoding)
- 3′-UTRs = based on annotation of reference mRNA
- Intronic sequences as a control (last two introns from each gene)

Type	Total Sequnce
Promoter	68 Mb
3' UTR	15 Mb
Intron Control	123 Mb



Motif Conservation Score

- A motif is said to be conserved when an exact match is found in all 4 species.
- Conservation = conserved occurrences/all occurrences
- MCS =

$$\frac{\text{Observed conservation} \quad - \text{random conservation}}{\text{Standard deviation}}$$

Known highly conserved motif

- Err α [TGACCTTG]
- Of the 434 times err α occurs in human promoter regions, 162 of them are conserved across all the 4 species.
- Conservation rate = 37%
- Random 8-mer motif shows only 6.8% conservation rate

b

```
          –902
Human     CTGCCT----AAGTAGCCTAGACGCTCCCGTGCG-CCCGGGGCGGG-TAG
Mouse     CGCCGC----CTGCATTATTCAC---------------------------
Rat       CTGCTC----ATGCATAATTCAC---------------------------
Dog       CTGCTTTCAACAGTGGGGCAGACGGTCCCGCGCGCCCCAAGGCAGGCCCG
          *  *        *        **

                                      Err-α
Human     GCCTGGCCGAAAATCTCTCCCGCGCGCCTGACCTTGGGTTGCCCCAGCCA
Mouse     ----------AAGCCTGTGGCGCGC-CGTGACCTTGGGCTGCCCCAGGCG
Rat       ----------AAGTTTCT---CTGC-CCTGACCTTGGGTTGCCCCAGGCG
Dog       GGCTGC----AGACCTGCCCTGAGGGAATGACCTTGGGCGGCCGCAGCGG
                    *    *       *    **********  *** ***

Human     GGCTGCGGGCCCGAGACCCCCG-------------------GGCCTCCCT
Mouse     GGCTGCAGGCTCACCACCCC---------------------GTCTTTTCT
Rat       AG--GCATACACCCCGCCTT---------------------TTTTTTTTT
Dog       GGCCGCGGGCCCAGGCCCCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCT
           *  **   * *    **                           *   *

Human     GCCCCCCG-----------CGCCGCCCCGATTTGCCCTCAGAGAGGGTAT
Mouse     GCTTTTCG-----------AGTCGGCCCGCTCTGCTCCCAG-GAGAGCAT
Rat       TTTTTTTTTTTTGCCGTTCAAGAGCCCTGTTCTGCTCTCAA-AAGGGTAT
Dog       GCCCCCCG-----------GACCGCCCCGCTTCACCCTCCCAGCTGGGAA
                                 * ** * *   * * *       * *

                                                      -693
Human     ---CGATCTTATTT-CTGGGTCTACGGCAAACTCCAAGGTCTACAAACGT
Mouse     TCACGGTCTTATTTAGTGAGCGTAAGGCAAATCTGAATACCCAGCAGGGC
Rat       TAACGGTCTTATTTATTGGGCGCAAAGCAAACTTTAATACCCAGCAGGGC
Dog       CCCCGGGCCTGAATACGGAGTCAGCCGCACACTTCACGGCCCAAACGCGG
             **  * *   *   * *      *** *    *    * *     *
```

Difference between the two papers

- Mini-motifs: paper 2 doesn’t use 3-fold degenerate sites.
- MCS>4 in paper 1 and MCS>6 in paper 2.

Results: Promoter Region

- 174 highly conserved motifs (MCS > 6)
- 59 strong match to known motifs, 10 weaker match.
- 105 potential new regulatory motifs

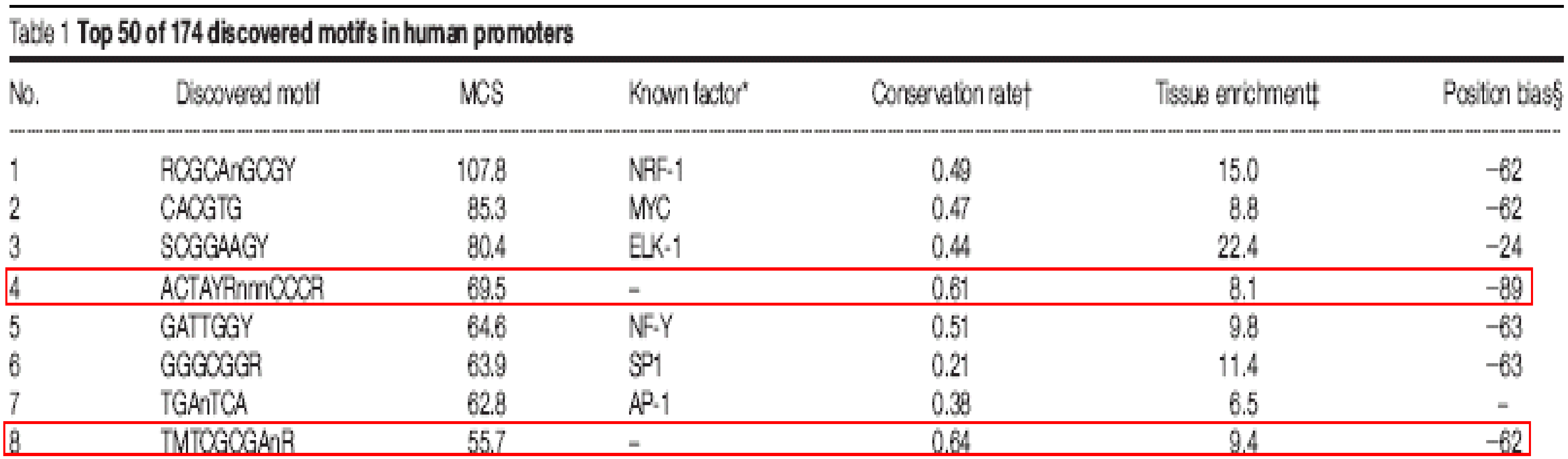


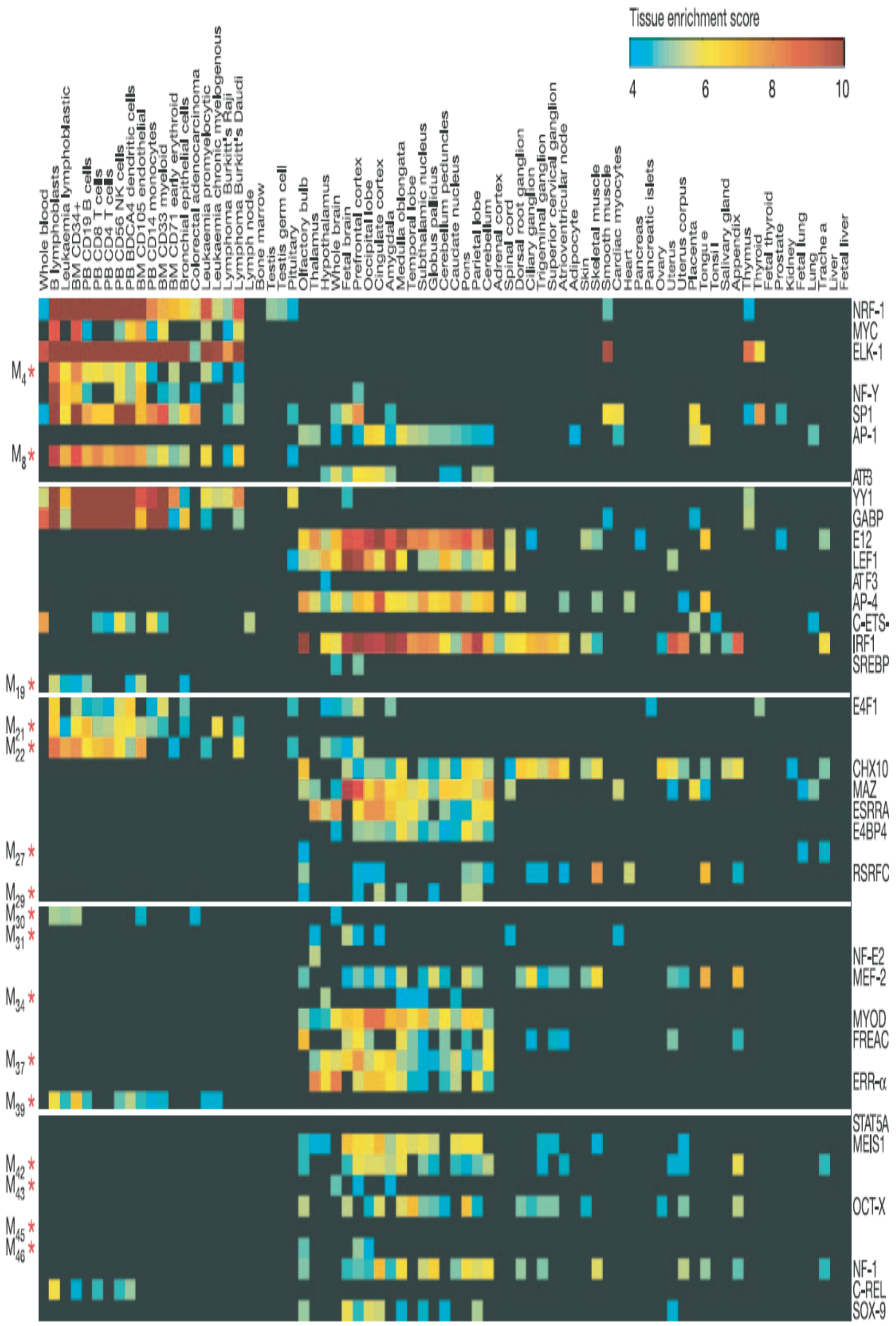
Table 1 **Top 50 of 174 discovered motifs in human promoters**

No.	Discovered motif	MCS	Known factor*	Conservation rate†	Tissue enrichment‡	Position bias§
1	RCGCAnGCGY	107.8	NRF-1	0.49	15.0	−62
2	CACGTG	85.3	MYC	0.47	8.8	−62
3	SCGGAAGY	80.4	ELK-1	0.44	22.4	−24
4	ACTAYRnnnCCCR	69.5	-	0.61	8.1	−89
5	GATTGGY	64.6	NF-Y	0.51	9.8	−63
6	GGGCGGR	63.9	SP1	0.21	11.4	−63
7	TGAnTCA	62.8	AP-1	0.38	6.5	-
8	TMTCGCGAnR	55.7	-	0.64	9.4	−62

Approaches to explore biological significance

- So why is the motif biologically significant?

1. tissue specificity
2. positional bias

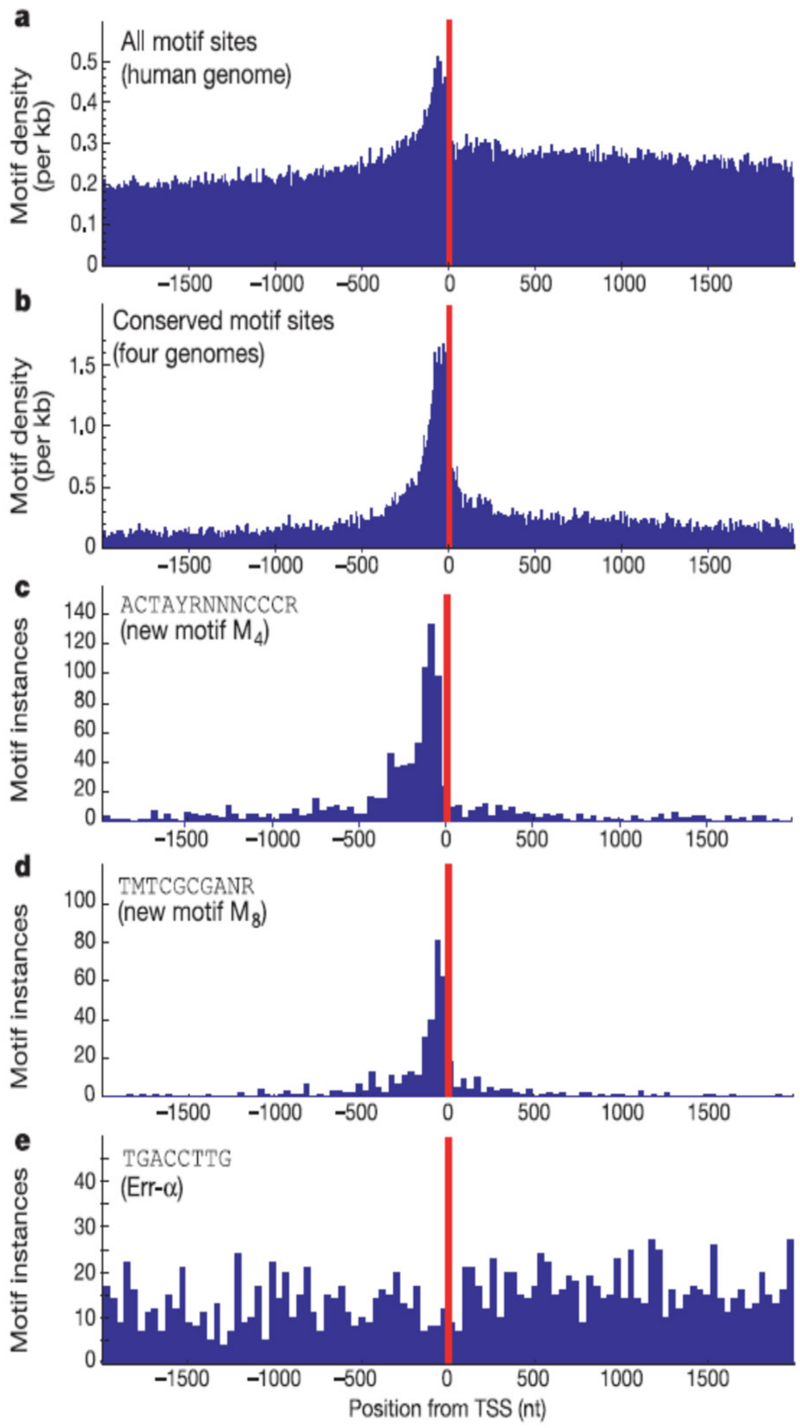


Tissue Specificity

- Tissue specificity of expression for genes containing discovered motifs
- Expression data for 75 tissues
- 59 of 69 known, and 53 of 105 unknown show tissue specificity

Position Bias

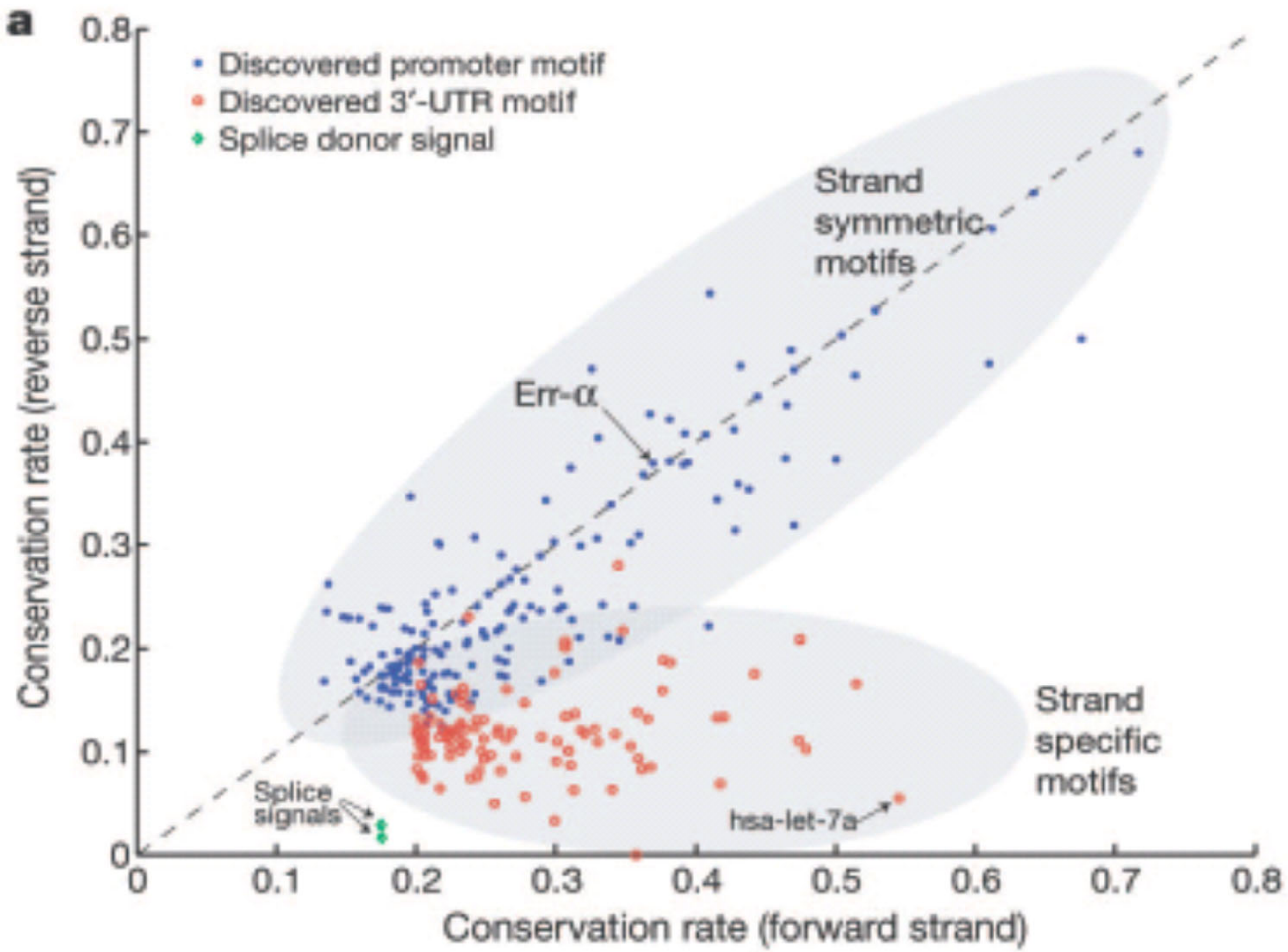
- Motifs show position bias
- Conserved motifs show strong position bias
- Preferential occurrence within 100bases of TSS



Results: motifs in 3’ UTRs

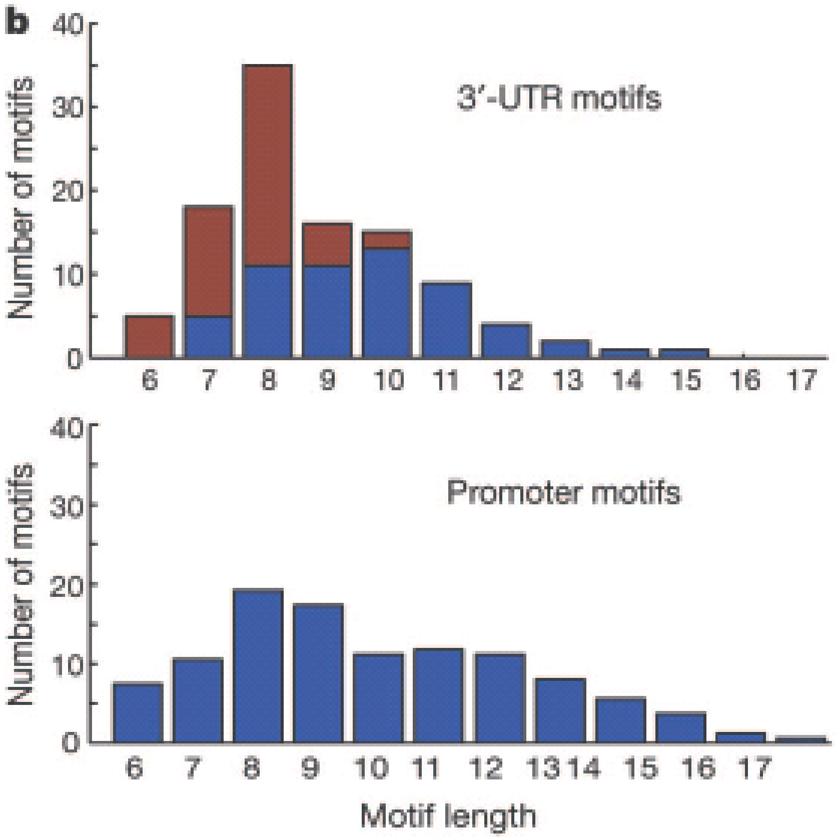
- In UTR 106 conserved motifs found (MCS>6)
- 3’-UTR motifs have not studied before
- Comparison of discovered motifs to a large collection of previously known motifs not possible
- Two unique properties
  - Strand specificity
  - Bias towards 8-mers

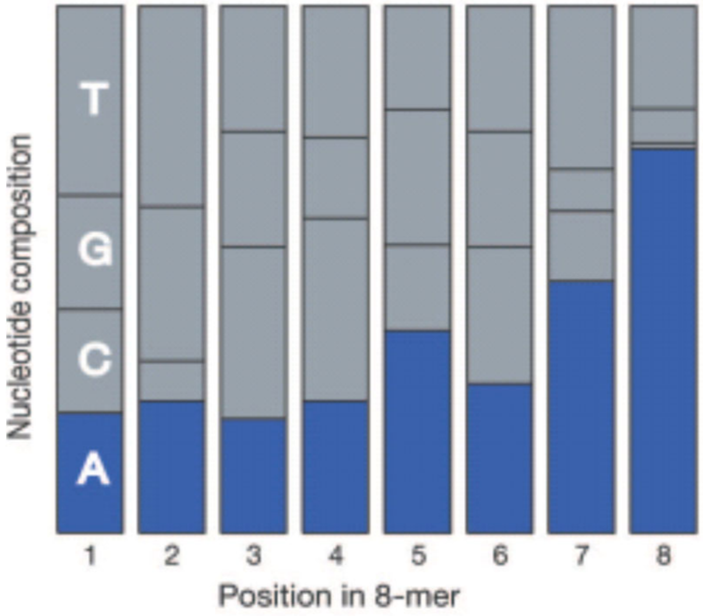
Property1: strand specificity



Xie, X. et al., Nature, 2005

Property2 : bias towards 8-mers

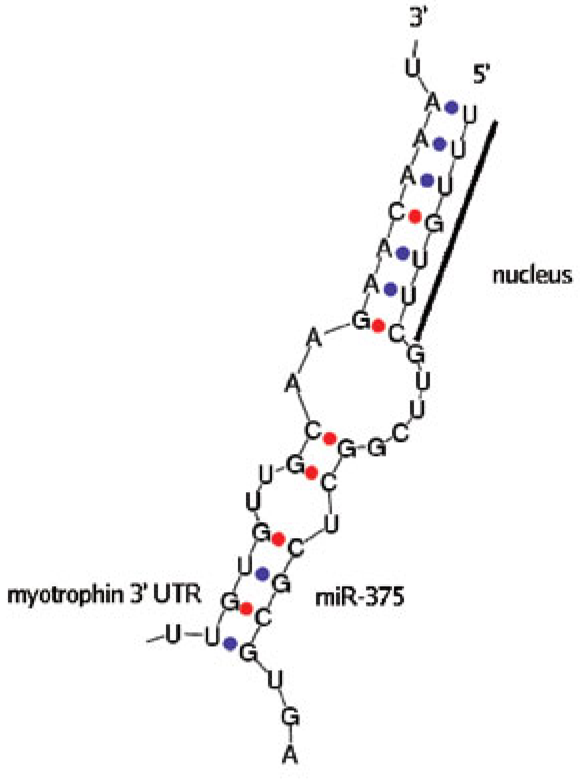




Xie, X. et al., Nature, 2005

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