Systematic discovery of regulatory motifs in human promoters and 30 UTRs by comparison of several mammals

Xiaohui Xie[1], Jun Lu[1], E. J. Kulbokas[1], Todd R. Golub[1], Vamsi Mootha[1], Kerstin Lindblad-Toh[1], Eric S. Lander[1,2] & Manolis Kellis[1,3]

1Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, 2Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, 3Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

AIM: to create a catalogue of functional elements in the human genome

- Most functional elements are not protein coding--regulatory signals, RNA genes and structural elements.
- Goal here is to create a catalogue of 'common regulatory motifs' -- short, functional sequences (typically, 6–10 bases) that are used many times in a genome.

HOW ?

HOW ?

- Sequence comparison between different species.
- Works well with smaller genomes (proven by a study with 4 yeast species)

IN CASE OF HUMAN GENOME ?

HOW ?

- Sequence comparison between different species.
- Works well with smaller genomes (proven by a study with 4 yeast species)

IN CASE OF HUMAN GENOME ?

- Focused on limited subsets – Promoter regions, 3'UTR of protein-coding genes
- Aligned these regions across HUMAN, MOUSE, RAT and DOG genomes

3 Sets for Alignments :

- Studied a total of,17,700 well-annotated genes from the RefSeq database with a single mRNA as reference for each gene.

1. Promoter was defined as a 4kb sequence centered at TSS (Transcription Start Site) ~68 Mb
2. 3’ UTR was defined based on the annotation of the reference mRNA ~ 15 Mb
3. A CONTROL was also defined- last two introns from the genes(because these contain fewer regulatory elements). ~123 Mb

Alignment of the 3 sets

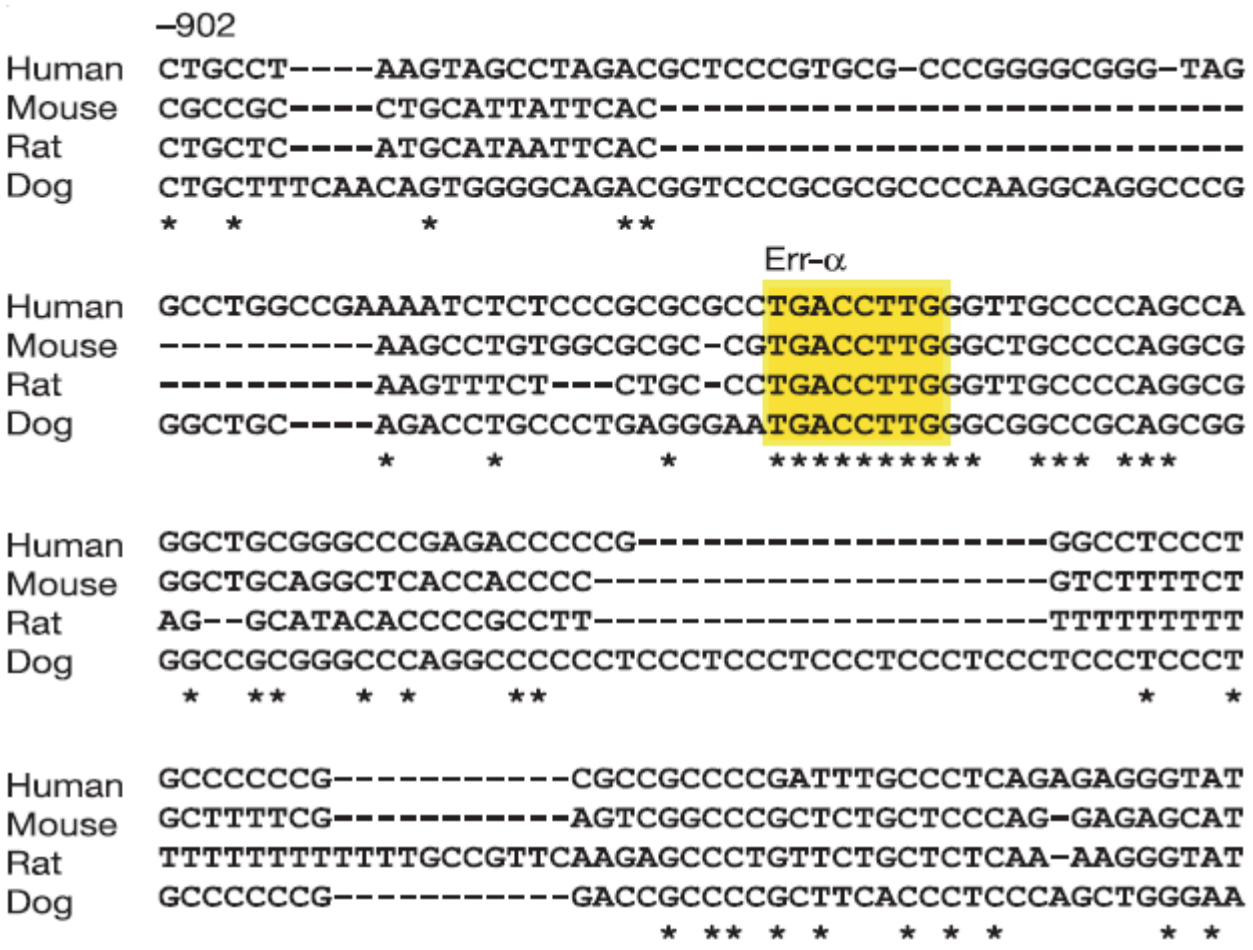
- Across HUMAN, MOUSE, RAT & DOG genomes
- Proportion of bases aligned across all 4 species:
  Promoter- 51%
  3’ UTR - 73%
- These are higher than that of control set- 34% or for the whole genome- 28%
- **This reflects the presence of important conserved elements in those regions**.

But, How can we say that the conservation represents regulatory motifs?

Conservation properties of Regulatory motifs

- TGACCTTG- binding site of Err-α protein
  - occurs 434 times in human promoter regions
  - 162 of these are conserved across all 4 species
  - Conservation Rate- 37%

- By contrast, a random 8-mer motif has a conservation rate of 6.8% in promoters


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

Motif Conservation Score(MCS)

- Represents the number of standard deviations (s.d.) by which the observed conservation rate of a motif exceeds the expected conservation rate for comparable random motifs
- Comparable random motifs are generated by sampling human genomic sequence in promoters (divided according to high or low CpG content) or 3’ UTRs in order to ensure equivalent sequence composition
- Err-α-binding motif, the MCS is 25.2 s.d. (the binomial probability of observing 162 conserved instances out of 434, given an expected rate of conservation of 6.8%)

Conservation of other known motifs in promoter regions

- The 446 motifs in TRANSFAC were clustered on the basis of sequence similarity, resulting in 123 motif clusters
  - 63% → MCS > 3
  - 47% → MCS > 5

- By contrast, a control set of random motifs has only 1.6% with MCS > 3 and none with MCS above 5

Conservation of other known motifs in 3'UTR regions

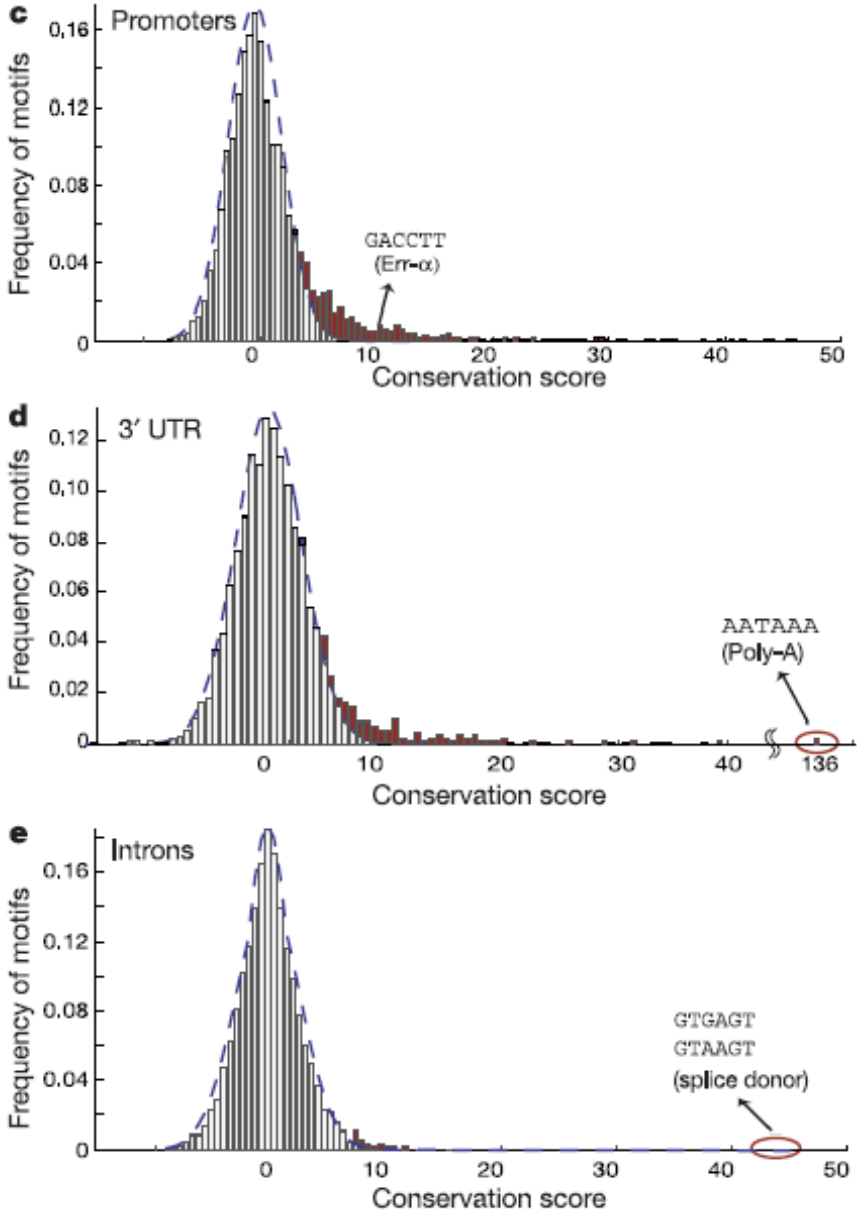
- No analogous database to TRANSFAC

- One example, AATAAA -polyadenylation signal
  - conservation rate = 46%, MCS =135

- For Control random motifs,
  - conservation rate = 10%

This reflects the conservation properties of regulatory motifs

Excess conservation among promoters and 3' UTRs unlike intronic regions



Discovery of new motifs

- Evaluated all possible motifs containing 6–18 bases (10^{12}) across genomic regions
- Motifs with MCS > 6 were clustered
- Motif with highest MCS in each cluster was selected- referred to as “highly conserved motifs”

New motifs in promoters

- 174 highly conserved motifs
- Off these, there were some already know or similar ones which had strong/weak(59/10) matches to the ones assembled from TRANSFAC database

- These accounted for 72% of 123 previously known motifs assembled from TRANSFAC

- The remaining 105 discovered motifs represent potentially new regulatory elements. For example, the newly discovered motif M4 (ACTAYRNNNCCCR) occurs 520 times →317 (61%) conserved, motif M8 (TMTCGCGANR) occurs 368 times, →236 (64%) conserved.

How to prove that these new motifs are biologically meaningful?

How to prove that these new motifs are biologically meaningful?

1). Correlate→ presence of motif with tissue specificity of gene expression(because genes controlled by a common regulator show enriched expression in specific set of tissues)

- For each motif, define 2 sets of genes,

S1→ genes with at least 1 conserved occurrence of motif

S2→ CONTROL SET OF *GENES* in which motif occurs in humans but is not conserved

- Using gene expression data from 75 human tissues, significant enrichment in one or more tissues was seen for 59 of the 69 (86%) known motifs and 53 of the 105 (50%) new motifs(when the motif is conserved)

- In contrast, the CONTROL SET OF *MOTIFS* show little or no enrichment across the same tissues

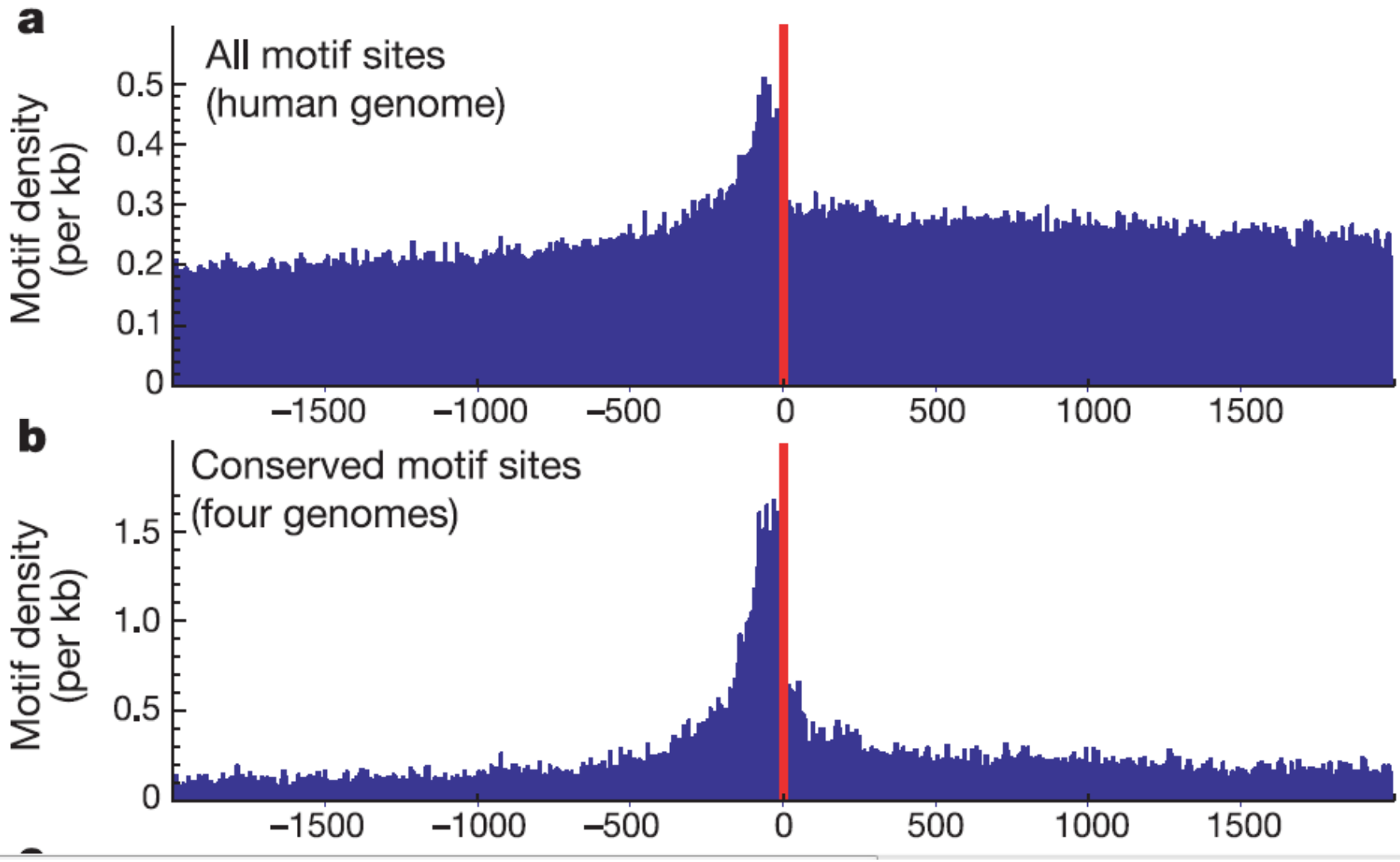
- Eg, new motifs M4 and M8 show enrichment in haematopoietic cells

2). Examined Positional Bias of motifs w.r.t TSS

- Discovered motifs preferentially occur in the human genome within 100 bases of the TSS(Fig a), and their conserved occurrences across all four species show an even more marked enrichment in this region(Fig b)

- 28% of the known motifs and 35% of the new motifs show significant positional preference

- For example, the two strongest new motifs (M4 and M8) tend to occur at distances centered around 289 and 262 bp upstream of the TSS, respectively


a
All motif sites
(human genome)
Motif density
(per kb)
0.5
0.4
0.3
0.2
0.1
0
–1500
–1000
–500
0
500
1000
1500
b
Conserved motif sites
(four genomes)
Motif density
(per kb)
1.5
1.0
0.5
0
–1500
–1000
–500
0
500
1000
1500

- Overall, 89% of the known motifs and 69%of the new motifs show tissue specificity, positional bias or both

- Taken together, these results strongly suggest that the new promoter motifs are likely to be biologically meaningful

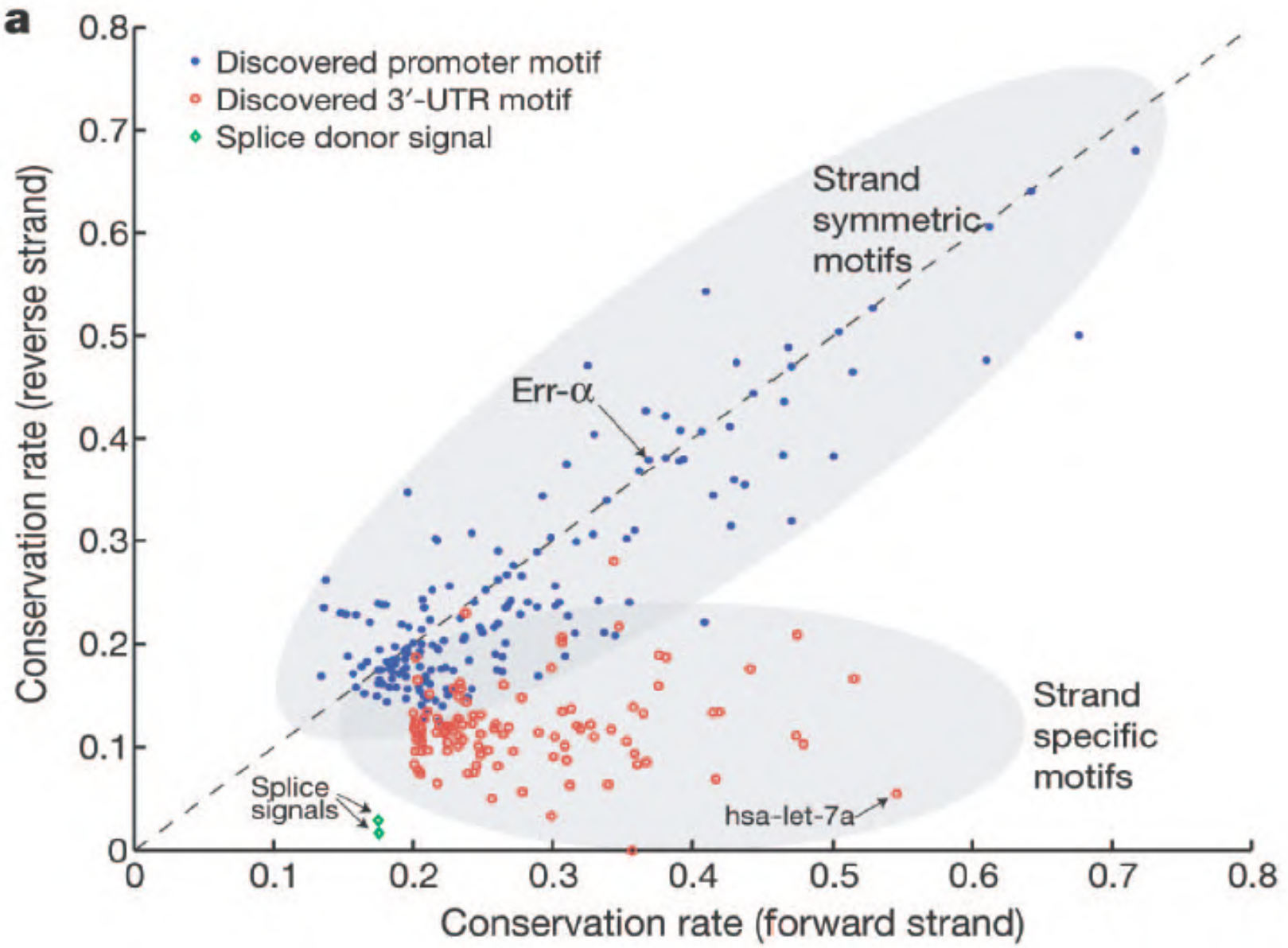
- In addition, several of the motifs tend to appear in multiple copies in the promoter

- For example, 17.5% of genes containing motif M4 have more than one copy of M4 within 200 bp of each other

New motifs in 3’ UTRs

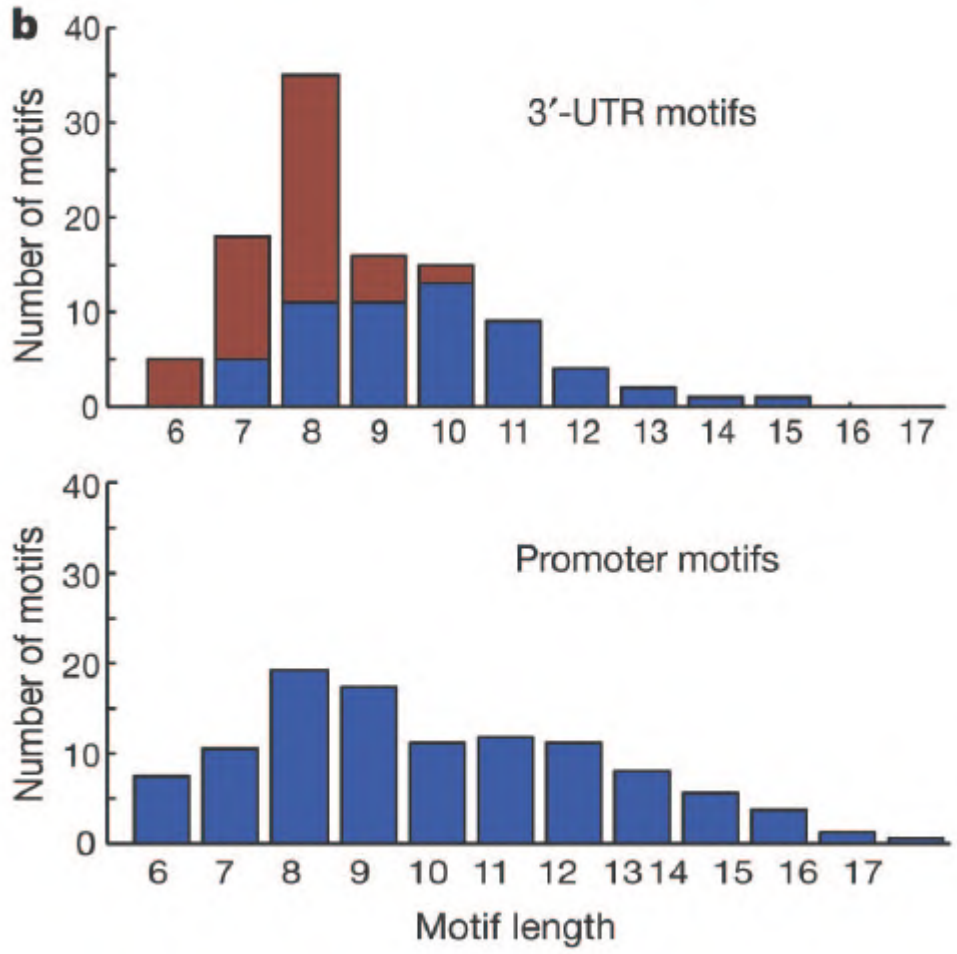
- 106 Highly conserved motifs (MCS > 6)
- Because 3’-UTR motifs have not been extensively studied, we could not compare the discovered motifs to a large collection of previously known motifs
- However, 2 properties give us insights:

1. 3’-UTR motifs show a strong directional bias with respect to DNA strand, preferentially conserved in only one strand (since they act at RNA level)

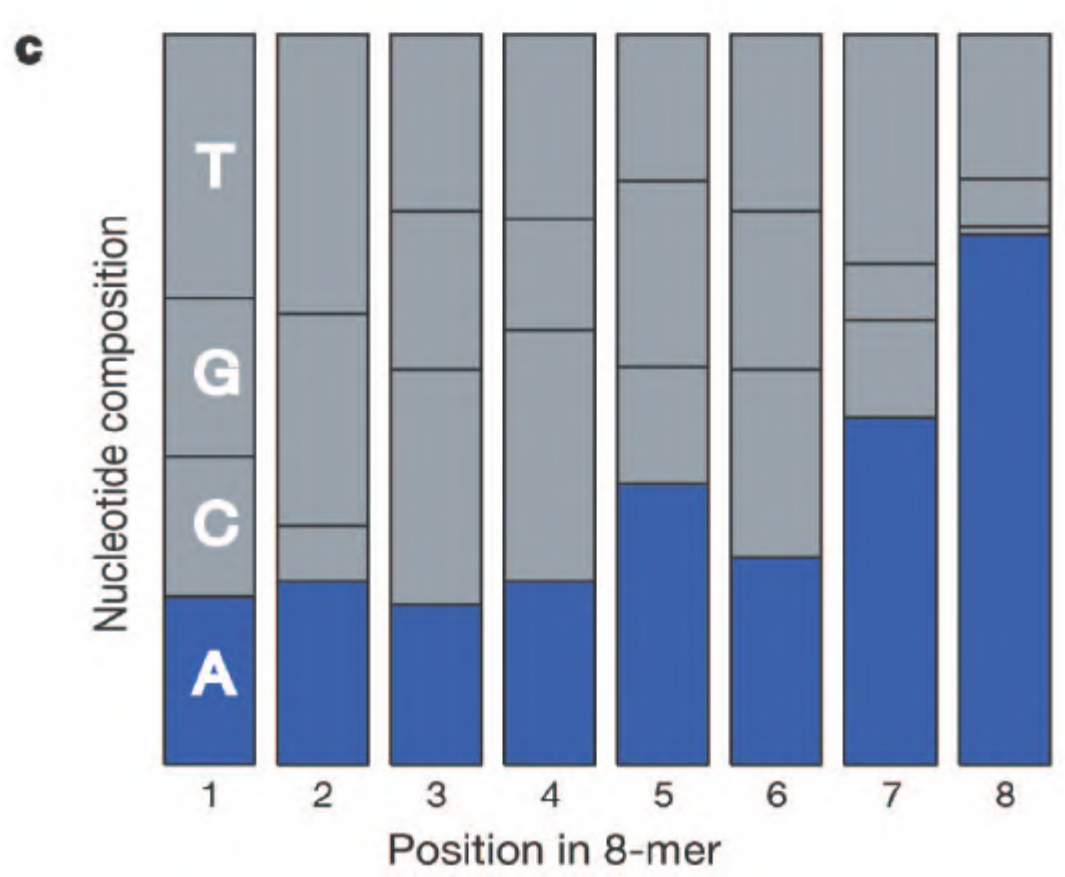

a
Discovered promoter motif
Discovered 3′-UTR motif
Splice donor signal
Strand symmetric motifs
Err-α
Strand specific motifs
Splice signals
hsa-let-7a
Conservation rate (reverse strand)
Conservation rate (forward strand)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

2).

- 3’-UTR motifs show an unusual length distribution. They have a strong peak at an 8-base length, whereas no such bias was found for promoter motifs



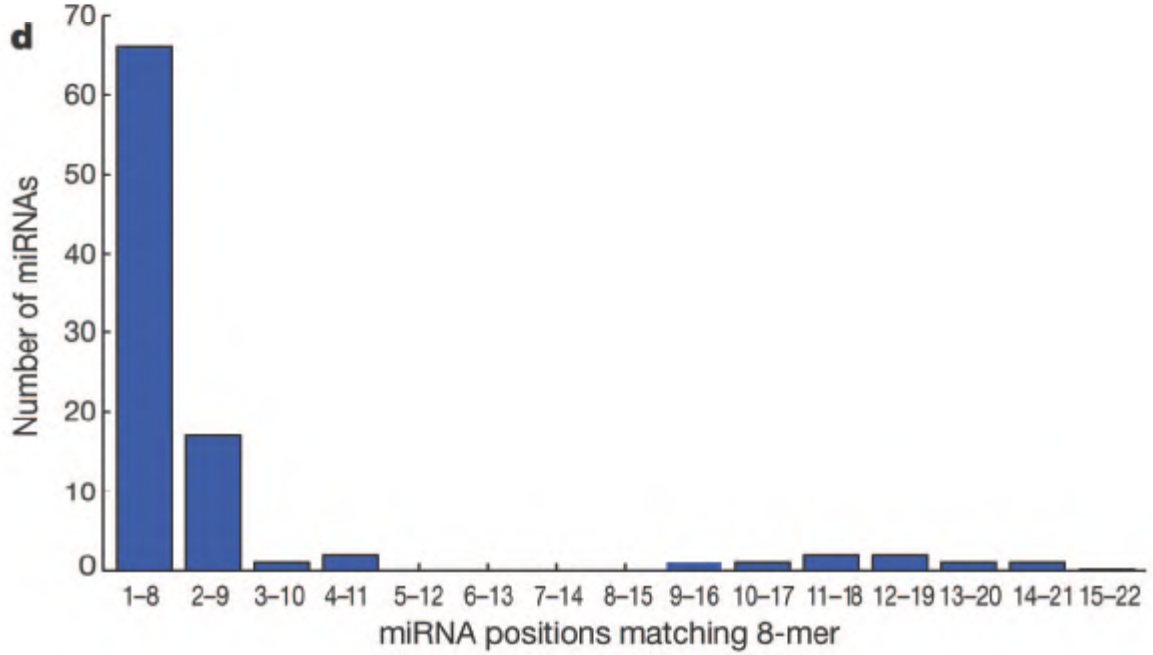
- Motifs of length 8 have a strong tendency to end with the nucleotide ‘A’. These properties are reminiscent of a feature of miRNA, many mature miRNAs start with a ‘U’ base followed by a 7-base ‘seed’ complementary to a site in the 3’ UTR of target mRNAs.
- So these motifs could be binding sites for such miRNAs



Relationship with miRNA

- The motif discovery procedure was repeated using only contiguous, non-degenerate 8-mers
- Identified the subset of 8-mers with conservation rate >18% (compared with the rate for a random 8-mer of 7.6%)
- Obtained “Highly Conserved 8-mer motifs” by clustering and choosing the highest MCS valued motif from each cluster (72 motifs)
- This formed 46% of the full set of 3’ UTR motifs
- Then searched for complementary matches of the 8-mer motifs to the 207 distinct human miRNAs listed in the current registry

- 43.5% of the known miRNAs can match through Watson–Crick pairing to the highly conserved 8-mer motifs
- When 1 mismatch was allowed- another 27 miRNAs paired to the 8-mer motifs
- Matches begin at nucleotide 1 or 2 of the miRNA in more than 95% of cases



- Among the known miRNAs, those that do not match highly conserved 8-mers show a rate of disruptive mutation that is fivefold higher than for known miRNAs that match the highly conserved 8-mers

- This suggests that the miRNA genes that match the highly conserved 8-mers have many more targets and, as a result, are much more constrained in their evolution

Discovery of new miRNAs

- Searched the four aligned genomes for conserved sequences complementary to any of the 72 discovered 8-mer motifs
- Extracted the sequence flanking each conserved site, and used the published RNAfold program to test for a conserved RNA-folding pattern, characteristic of miRNA genes
- Characteristic: stem-loop structures with calculated folding free energy of at least 25 kcal mol21 in all four species

- Identified 242 conserved and stable stem-loop sequences, containing conserved instances of most (52 of 72) of the highly conserved 8-mer motifs
- These include 113 sequences encoding known miRNAs (52% of 222 known miRNA genes) and 129 sequences encoding predicted new miRNAs

VALIDATION OF PREDICTED miRNAs:

- A representative set of 12 predicted new miRNA genes was selected
- Steps: purification of small RNAs, ligation of adaptors, polymerase chain reaction (PCR) amplification, cloning and DNA sequencing to verify the precise sequence and junction

- Lacking prior information about tissue or temporal specificity of expression, we used pooled small RNAs prepared from ten human tissues(breast, pancreas, prostate, colon, stomach, uterus, lung, brain,liver and kidney)
- This may miss many miRNAs expressed primarily during development or at low levels in the adult, or not expressed in the tissues we tested
- Nonetheless, 6 of the 12 (50%) predicted miRNAs genes were found to be clearly expressed in the pooled adult tissues.
- Hence we can say many of the 129 candidates are likely to be bona-fide miRNA genes(however there may be more since our approach finds only those with a conserved 8-mer)

Table 2 **Top 50 conserved 8-mers in 3′ UTRs and corresponding miRNAs**

No.	Motif	Conservation rate	MiRNA*
1	GTGCAATA	0.55	miR-92, miR-32, miR-137, miR-367, miR-25, miR-217(†), new(12)
2	GTGCCTTA	0.54	miR-124a, miR-224(†), miR-208(†), miR-34b(†), miR-9*(†), miR-34c(†), miR-330(†), new(6)
3	CTACCTCA	0.53	miR-98, let-7i, let-7g, let-7f, let-7e, let-7c, let-7b, let-7a, let-7d, miR-196b, miR-196a, new(4)
4	ACCAAAGA	0.49	miR-9, new(11)
5	TGTTTACA	0.48	miR-30e-5p, miR-30d, miR-30c, miR-30b, miR-30a-5p, new(4)
6	GCACTTTA	0.48	miR-20, miR-106b, miR-18(†), miR-93, miR-372, miR-17-5p, miR-106a, miR-302d, miR-302c, miR-302b, miR-302a, miR-373, new(4)
7	TGGTGCTA	0.43	miR-29c, miR-29b, miR-29a, miR-107(†), miR-103(†), new(6)
8	CTATGCAA	0.42	miR-153, new(9)
9	TACTTGAA	0.42	miR-26b, miR-26a, new(4)
10	CGCAAAAA	0.42	New(2)
11	GTGCCAAA	0.41	miR-96, miR-182, miR-183, new(16)
12	GTACTGTA	0.40	miR-101, miR-199a‡, miR-144, new(2)
13	ATACGGGT	0.40	miR-99a, miR-100, miR-99b(†)
14	AAGCACAA	0.40	miR-218, new(8)
15	TTTGCACT	0.37	miR-19b, miR-19a, miR-301, miR-130b, miR-130a, miR-152, miR-148b, miR-148a,

miRNA Targets:

- 40% of human 3'UTRs contain a conserved occurrence of one of the miRNA associated 8-mer motifs, whereas only 25% contain conserved occurrences of a comparable control set.
- This suggests that at least 20% of 3' UTRs may be targets for conserved miRNA-based regulation at the 8-mer motifs
- With sequence from more mammalian genomes, it should be possible to distinguish the conserved target sites with high specificity and sensitivity
- There are likely to be additional miRNA target sites in the human genome that are not conserved across all mammals; possible to find most of these by genomic comparison with closer relatives such as primates

CONCLUSION

- Initial systematic catalogue of human regulatory motifs in promoters and 3’ UTRs
- Promoters: the approach automatically rediscovered many known motifs and discovered many new ones
- 3’ UTRs: the approach provided a first view of common regulatory motifs
- These motifs also led to the identification of numerous new miRNA genes

NEXT CHALLENGE

To develop systematic methods to discern the specific functions of these motifs in a genome-wide fashion

THANK YOU

Naveen Chandramohan