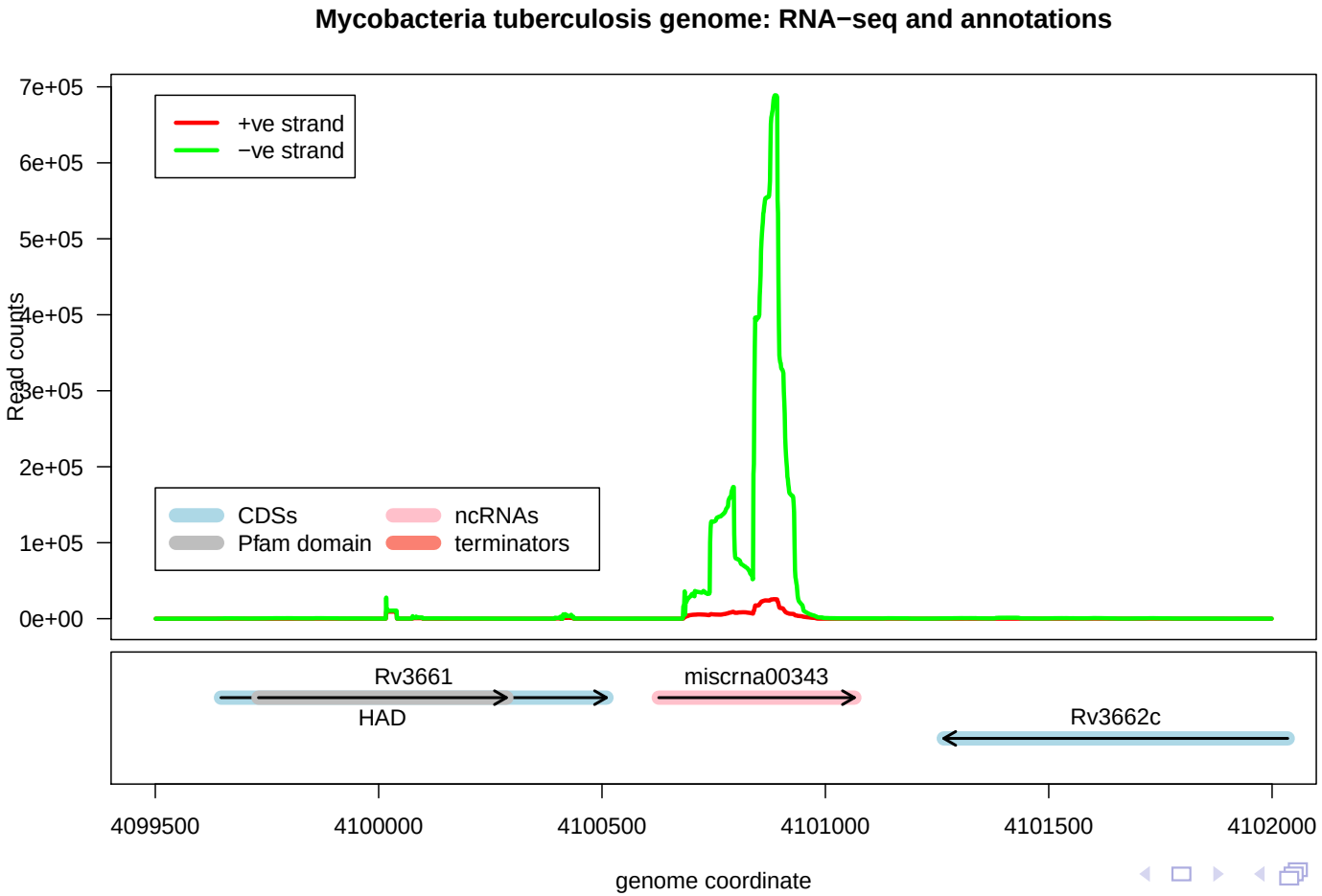
Hunting RNA motifs

Paul Gardner

July 23, 2012

Classification problems

- What is this?
  - The second most abundant *M. tuberculosis* transcript
  - No hits to Rfam, no homologs have been identified, no known function

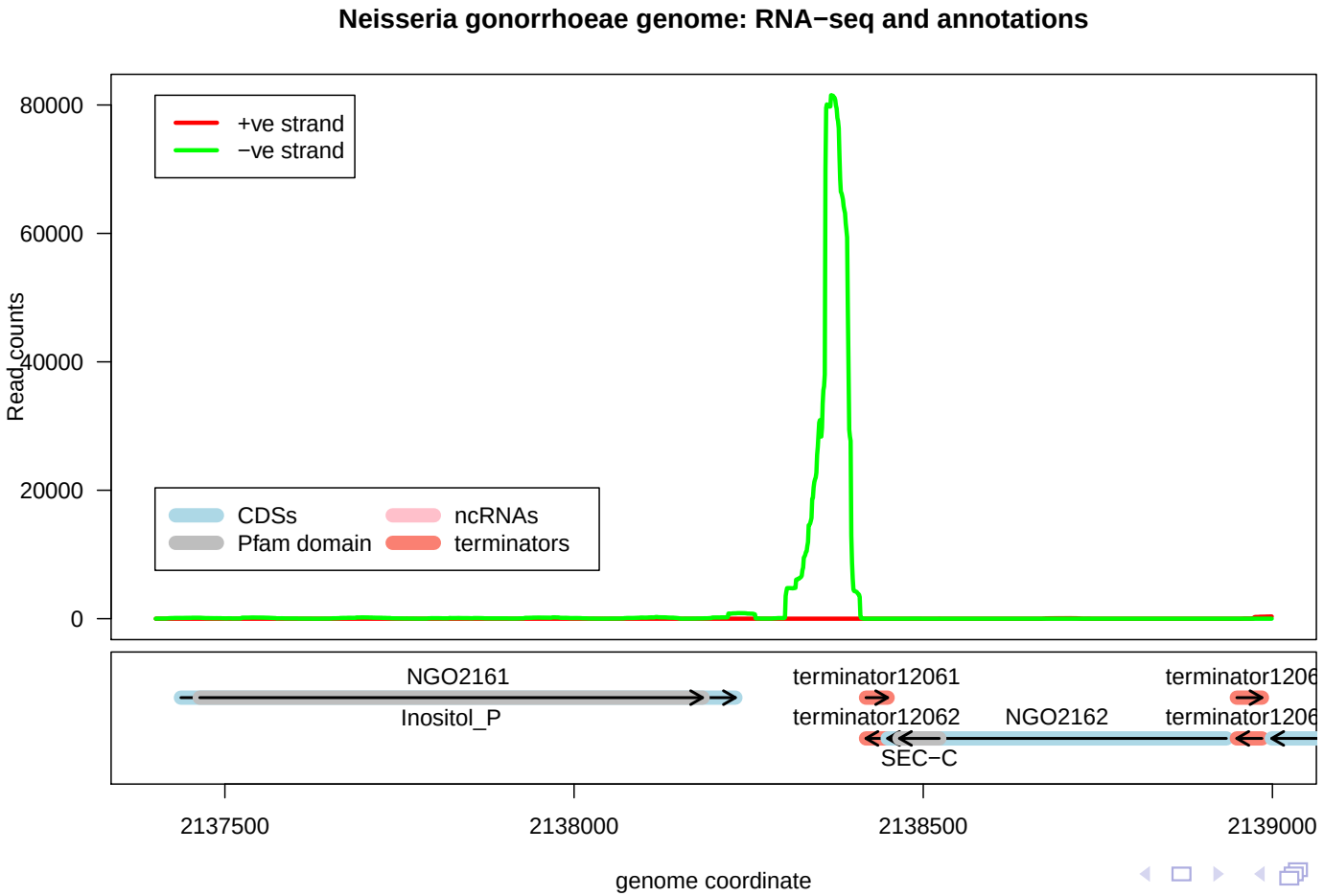


Rank	Reads	Gene
1	5,741K	SSU rRNA
2	689K	sRNA/RUF?
3	236K	RNaseP RNA
4	197K	LSU rRNA
5	119K	ATP synthase
6	119K	HSPX
7	115K	tmRNA
8	103K	Secreted protein

Arnvig *et al.* (2011) *PLoS Pathog.*

Classification problems

- What about this?
  - The seventh most abundant *N. gonorrhoeae* transcript
  - No hits to Rfam, no homologs have been identified, no known function

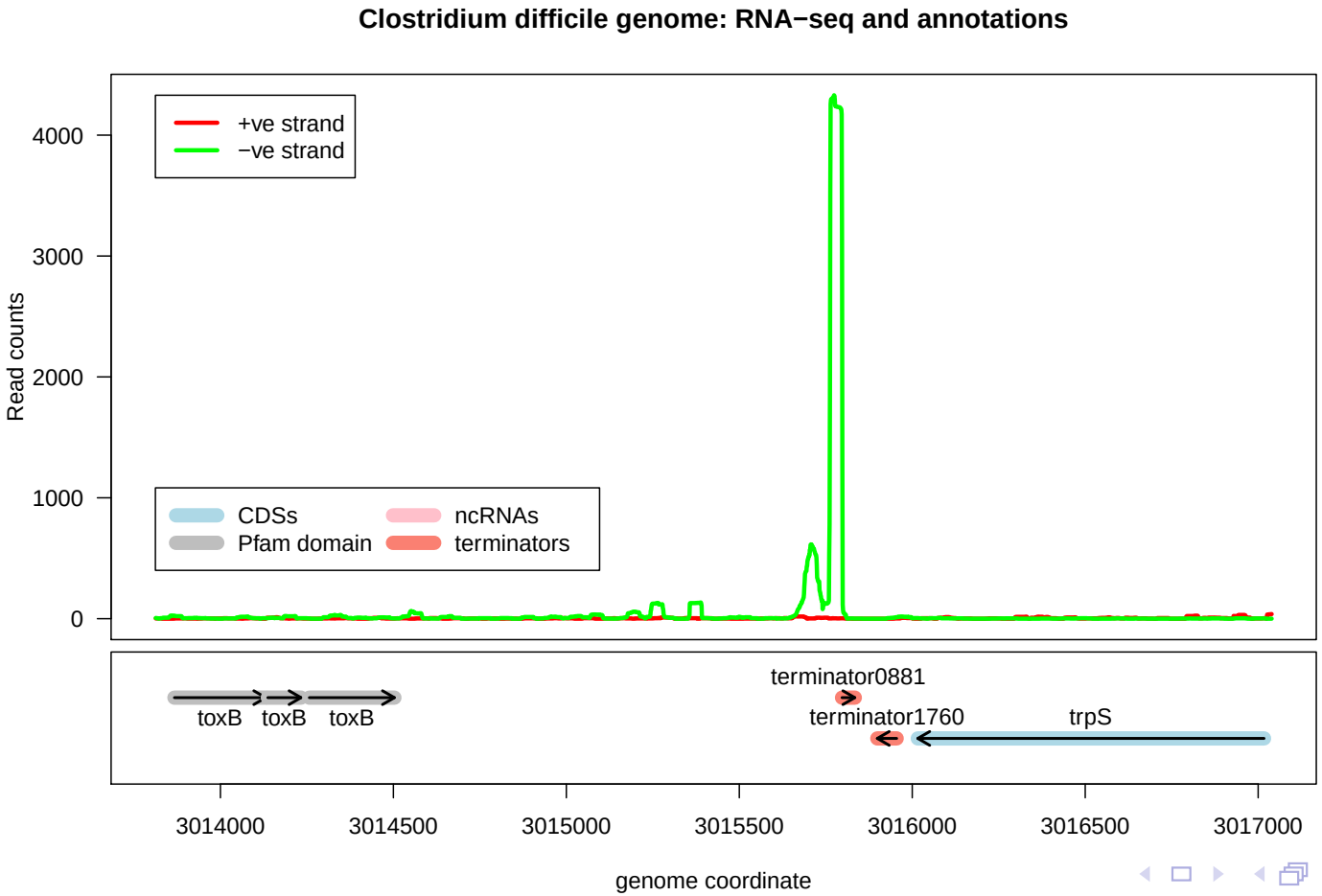


Rank	Reads	Gene
1	206K	RNaseP RNA
2	178K	tRNA-Ser
3	130K	ZapA 3‘ UTR?
4	120K	Phage protein 5‘ UTR?
5	97K	tmRNA
6	93K	BRO protein
7	81K	sRNA?
8	81K	Hsp70 protein
9	76K	tRNA-Asp

Isabella & Clark (2011) *BMC Genomics.*

Classification problems

- And this?
  - The ninth most abundant *C. difficile* transcript
  - No hits to Rfam, no homologs have been identified, no known function

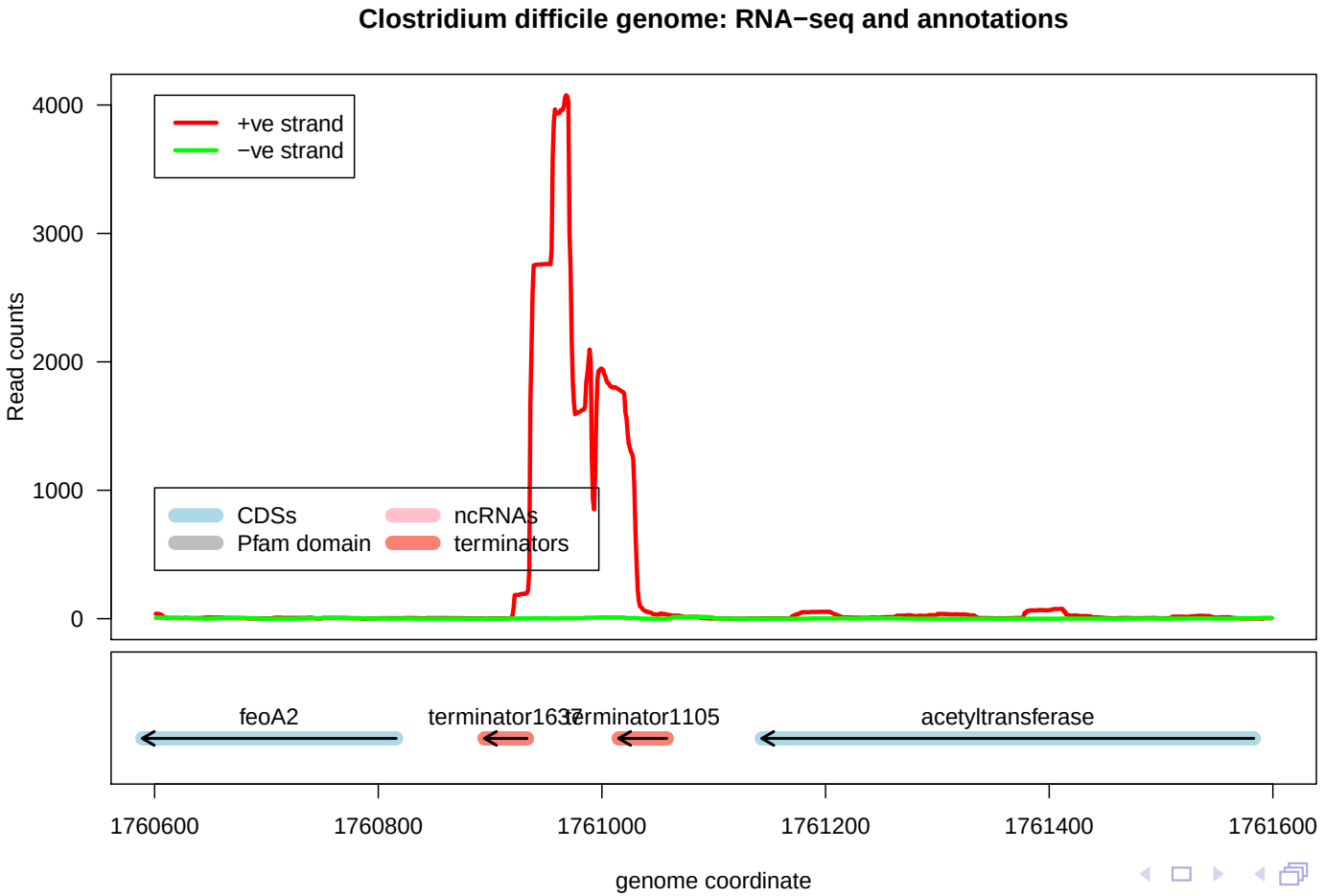


Rank	Reads	Gene
1	12K	RNaseP RNA
2	12K	6S RNA
3	12K	SRP RNA
4	11K	ldhA
5	10K	tmRNA
6	6K	spoVG
7	6K	Gly reductase
8	5K	slpA
9	4K	sRNA/RUF?
10	4K	hypothetical prot.
11	4K	sRNA/RUF?

Deakin, Lawley *et al.* (2012) *Unpublished.*

Classification problems

- And this?
  - The eleventh most abundant *C. difficile* transcript
  - No hits to Rfam, no homologs have been identified, no known function



Rank	Reads	Gene
1	12K	RNaseP RNA
2	12K	6S RNA
3	12K	SRP RNA
4	11K	ldhA
5	10K	tmRNA
6	6K	spoVG
7	6K	Gly reductase
8	5K	slpA
9	4K	sRNA/RUF?
10	4K	hypothetical prot.
11	4K	sRNA/RUF?

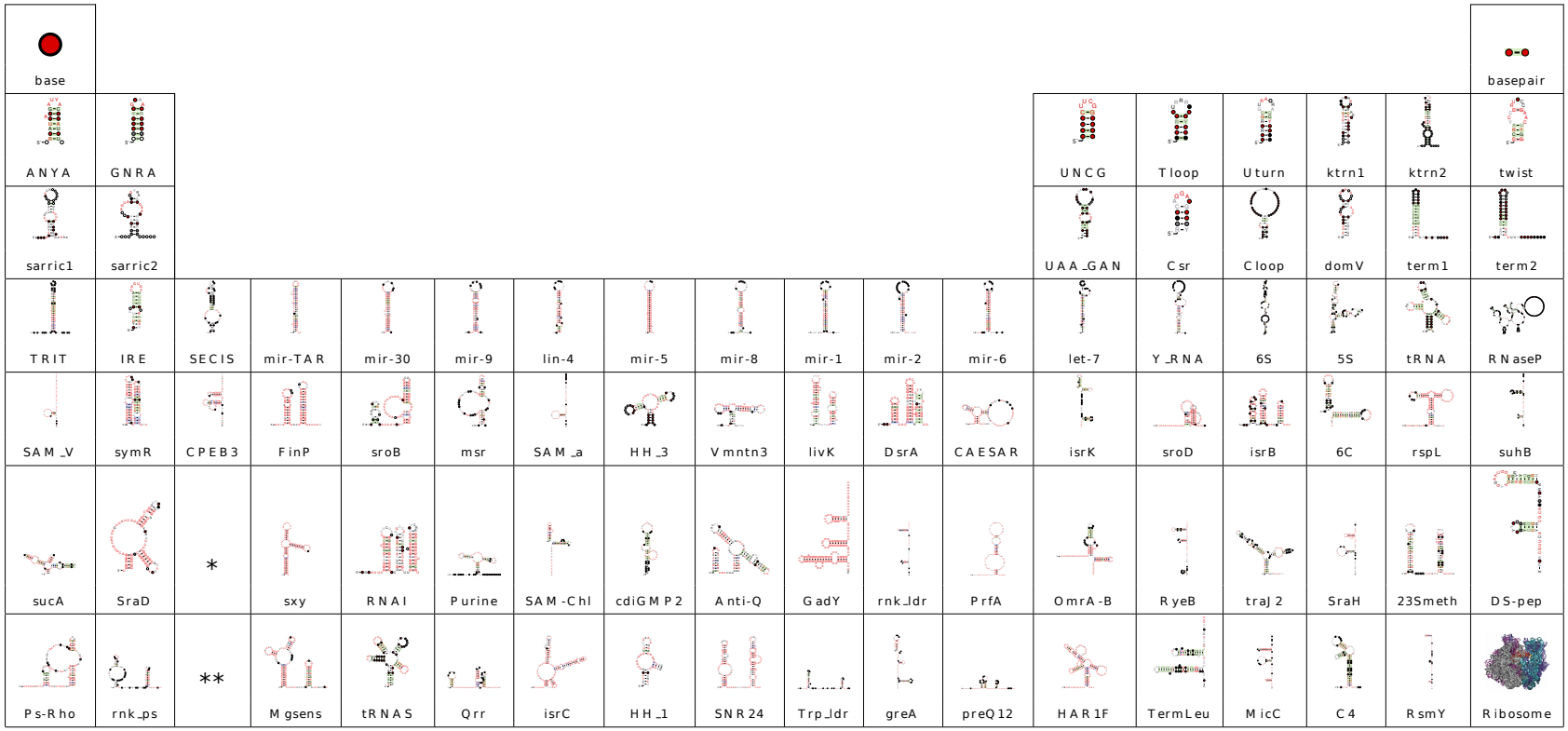
Deakin, Lawley *et al.* (2012) *Unpublished.*

What is our next big challenge?

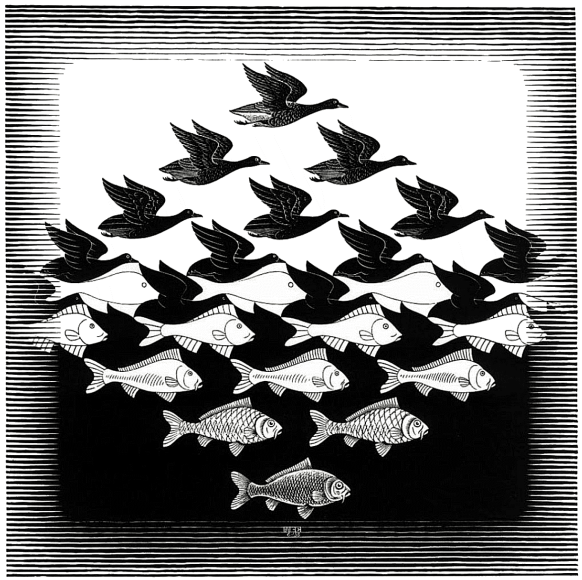
- Past:
  - How many non-coding RNA genes are there?
- Future:
  - Can we determine the functions, if any, for large sets of given RNAs?

Will a "periodic table of RNA" be useful?

- My aim is to build an analog to the Periodic Table for classifying RNA families and motifs, enabling researchers to predict function.



What is a motif?

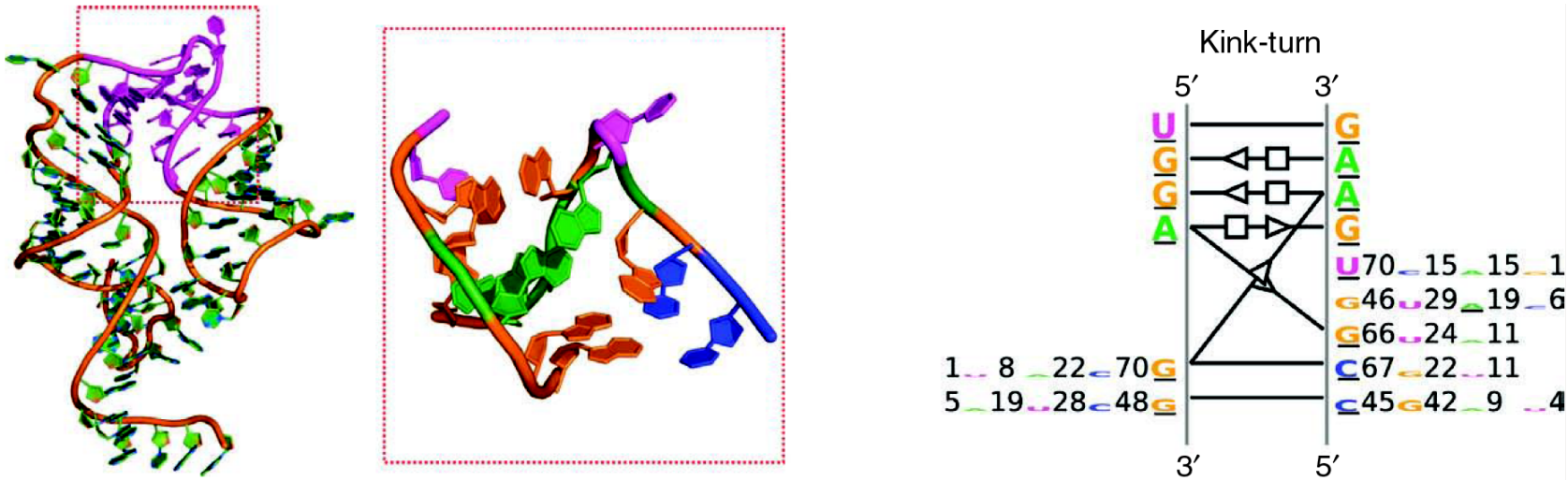


Escher, MC (1938) Sky and Water 1.

- Wiktionary definitions for **motif**:
  - A recurring or dominant element; a theme.
- For the purposes of this work:
  - an RNA motif is a recurring RNA sequence and/or secondary structure found within larger structures that can be modelled by either a covariance model or a profile HMM.

What is a motif?

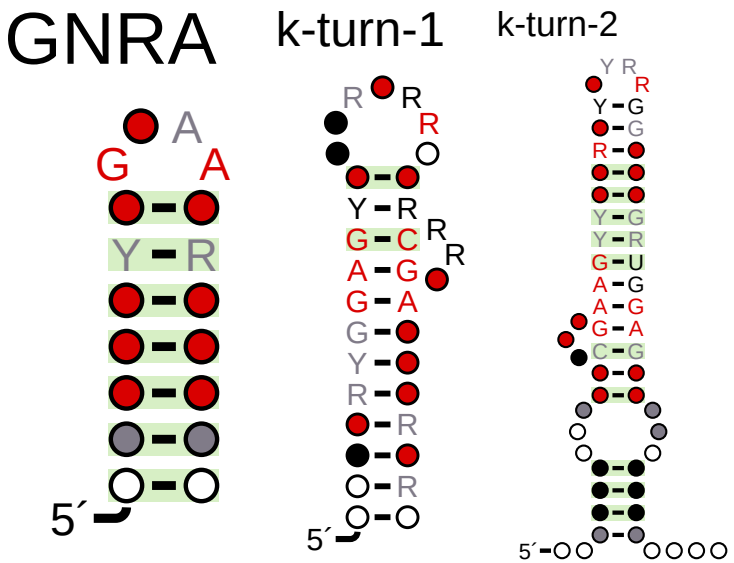
- The kink-turn RNA motif
  - Found in rRNA, riboswitches, RNase P, snoRNAs, ...
  - An asymmetric internal loop
  - Causes a sharp kink between two flanking helical regions

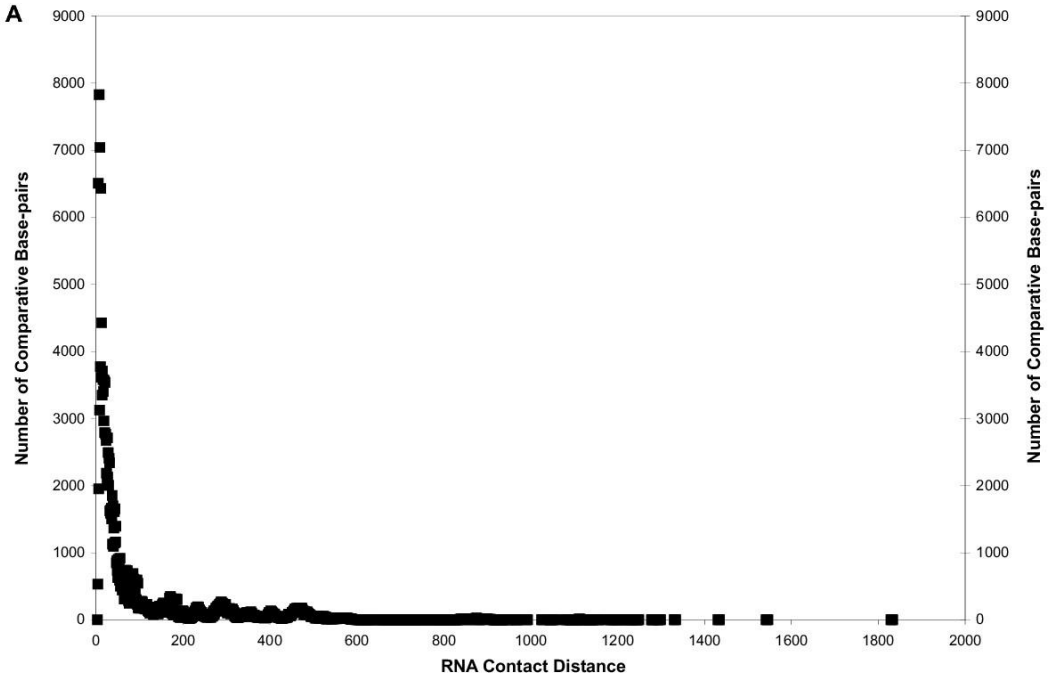


Cruz & Westhof (2011) Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nature Methods.*

Can we build a seed alignment and CM resource of these motifs?

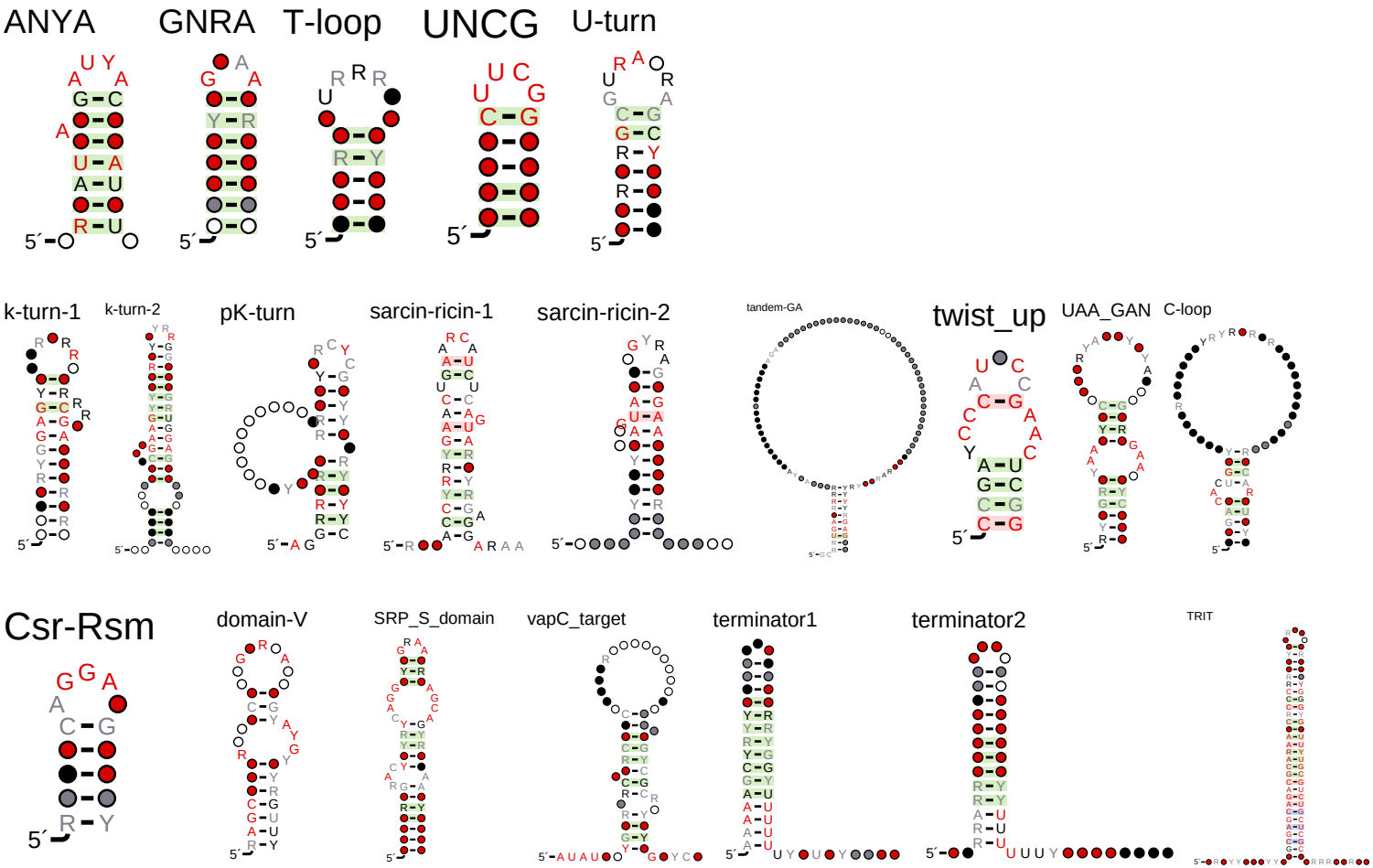
- For the hairpin motifs?
  - YES
- For the internal loop motifs?
  - MAYBE
- For the sequence motifs?
  - Not certain yet





Doshi *et al.* (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics.*

The RNA Motifs: RMfam

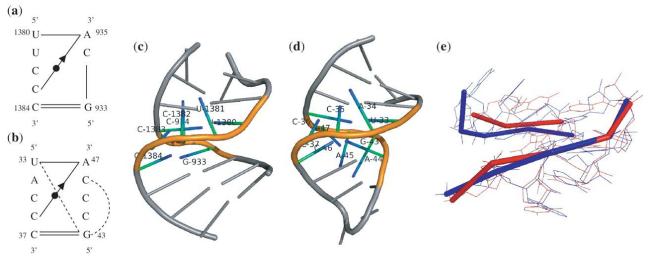


Breaking Rules

- Prefer motifs found in multiple families &/or multiple locations in the same family
  - Aligning **analogous** structures
- Motifs are allowed to overlap
- Models are non-specific
  - High false-positive rates
- Sequences are not edited, spliced or otherwise interfered with

- Sequence sources are diverse (PDB, EMBL & publications)

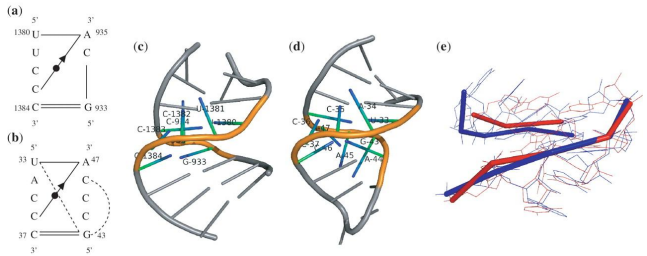
An example entry



```
# STOCKHOLM 1.0

#=GF ID    twist_up
#=GF AU    Gardner PP
#=GF SE    Gardner PP
#=GF SS    Published; PMID:21976732
#=GF GA    19.00
#=GF DR    URL;http://rnafrabase.cs.put.poznan.pl/?act=pdbdetails&id=1S72
#=GF RN    [1]
#=GF RM    21976732
#=GF RT    Clustering RNA structural motifs in ribosomal RNAs using
#=GF RT    secondary structural alignment.
#=GF RA    Zhong C, Zhang S
#=GF RL    Nucleic Acids Res. 2012;40:1307-17.

2zkr_Y/27-46            CCGAUCUCGUC-UGAUCUCGG
#=GR 2zkr_Y/27-46 SS    <<<<---<____>--->>>>
2gv3_A/3-20             CCAGACUC---CCGAAUCUGG
#=GR 2gv3_A/3-20  SS    (((((.(.---....)))))) 
3iz9_C/29-48            CGGAUCCCGUC-AGAACUCCG
#=GR 3iz9_C/29-48 SS    ((((...(...-.)...)))) 
3bbo_B/31-50            CCAA-UCCAUCCCGAACUUGG
#=GR 3bbo_B/31-50 SS    .(.....<.....>.....).
1nkw_9/33-53            CCCACCCCAUGCCGAACUGGG
#=GR 1nkw_9/33-53 SS    (((.............))) 
1c2x_C/31-51            CUGACCCCAUGCCGAACUCAG
#=GR 1c2x_C/31-51 SS    ((((...<.....>...)))) 
1giy_B/33-53            CCGUUCCCAUUCCGAACACGG
1S72_9/29-50            CCGUACCCAUCCCGAACACGG
#=GR 1S72_9/29-50 SS    ((((...(.....)...)))) 
#=GR 1S72_9/29-50 MT    ...xxxxx.....xxxxx...
#=GC SS_cons            ((((...(.....)...)))) 
//
```

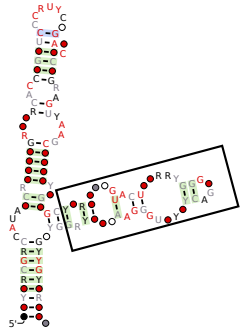
An example entry

```
# STOCKHOLM 1.0

#=GF ID   twist_up
#=GF AU   Gardner PP
#=GF SE   Gardner PP
#=GF SS   Published; PMID:21976732
#=GF GA   19.00
#=GF DR   URL;http://rnafrabase.cs.put.poznan.pl/?act=pdbdetails&id=1S72
#=GF RN   [1]
#=GF RM   21976732
#=GF RT   Clustering RNA structural motifs in ribosomal RNAs using
#=GF RT   secondary structural alignment.
#=GF RA   Zhong C, Zhang S
#=GF RL   Nucleic Acids Res. 2012;40:1307-17.

2zkr_Y/27-46           CCGAUCUCGUC-UGAUCUCGG
#=GR 2zkr_Y/27-46 SS   <<<<---<      >--->>>>
2gv3_A/3-20            CCAGACUC---CCGAAUCUGG
#=GR 2gv3_A/3-20  SS   (((((.(.---....)))))
3iz9_C/29-48           CGGAUCCCGUC-AGAACUCCG
#=GR 3iz9_C/29-48 SS   ((((...(...-.)...))))
3bbo_B/31-50           CCAA-UCCAUCCCGAACUUGG
#=GR 3bbo_B/31-50 SS   .(.....<.....>.....).
1nkw_9/33-53           CCCACCCCAUGCCGAACUGGG
#=GR 1nkw_9/33-53 SS   (((.............)))
1c2x_C/31-51           CUGACCCCAUGCCGAACUCAG
#=GR 1c2x_C/31-51 SS   ((((...<.....>...))))
1giy_B/33-53           CCGUUCCCAUUCCGAACACGG
1S72_9/29-50           CCGUACCCAUCCCGAACACGG
#=GR 1S72_9/29-50 SS   ((((...(.....)...))))
#=GR 1S72_9/29-50 MT   ...xxxxx.....xxxxx...
#=GC SS_cons           ((((...(.....)...))))
//
```



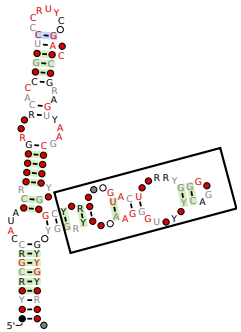
An example annotated Rfam alignment

```
# STOCKHOLM 1.0
#=GF ID   5S_rRNA
#=GF AC   RF00001
#...
#=GF WK   5S_ribosomal_RNA
#=GF SQ   712
#=GF MT.G GNRA
#=GF MT.s sarcin-ricin-2
#=GF MT.t twist_up

X62858.1/62-115             GCUGCG-U-UGUGGGUGUGUACUGCGGUUUU-UUG-CUGUGGGAA--GCCCACUUCA-CUG
X01588.1/64-115             UCACGU--UAGUGGG---GCCGUGGAUACCGUGAGGAUCC-GCAG-CCCCACU-AAG-CUG
#=GR X01588.1/64-115  MT.0  ........................GGGGGGGGGGGGGGGGG.....................
X05870.1/363-414            UCACGU--UGGUGGG---GCCGUGGAUACCGUGAGGAUCC-GCAG-CCCCACU-AAG-CUG
#=GR X05870.1/363-414 MT.0  ........................GGGGGGGGGGGGGGGGG.....................
L27170.1/63-116             CCAGCG--UCCGGCA--AGUACUGGAGUGCGCGAGCCUCUGGGAA-AUCCGGU-UCG-CCG
#=GR L27170.1/63-116  MT.0  ..ssss..sssssss..sssssssssssssssssssssssssss.sssssss.sss.ss.
L27163.1/61-115             CCAGCGUUCCGGGGA---GUACUGGAGUGCGCGACCCUCUGGGAA--ACCGGGUUCG-CCG
#=GR L27163.1/61-115 SS     ...(((((((((.........)))))))))(((((((...((......)))))).))).)..
#=GR L27163.1/61-115  MT.0  ..................................GGGGGGGGGGGG..GGGGGGG.......
#=GR L27163.1/61-115  MT.1  ..sssssssssssss...sssssssssssssssssssssssss..sssssssss.ss.
L27343.1/60-114             CCAGCGAACCAGCUA---GUACUAGAGUGGGAGACCCUCUGGGAG-CGCUGGU-UCG-CCG
#=GR L27343.1/60-114  MT.0  ........................GGGGGGGGGGGGGGGGG.....................
#=GR L27343.1/60-114  MT.1  ....sssssssssss...sssssssssssssssssssssssss.sssssss.sss.s..
M33891.1/66-117             CCAGCG--CCGGAGA---GUACUGCGC-GGGCAACCGCGUGGGAG--GCGAGG-CCGCUCG
#=GR M33891.1/66-117  MT.0  ........................GGGG.GGGGGGGGGGG.....................
X01484.1/62-114             GUCAGG--CCCAGUU--AGUACUGAGGUGGGCGACCACUUGGGAA-CACUGGG-U-G-CUG
#=GR X01484.1/62-114  MT.0  ........................GGGGGGGGGGGGGGGGG.....................
#=GR X01484.1/62-114  MT.1  ...sss..sssssss..sssssssssssssssssssssssss.sssssss.s.s.ss.
#=GC SS_cons                ::::::::::<-<<--------<<-----<<____>>----->>----->>->:::::::
#=GC RF                     cCaGcG..cccggga...GUAcUggGGUggGcgAcCcucUGGgAA.CaCccgg.uCG.CUG
//
```

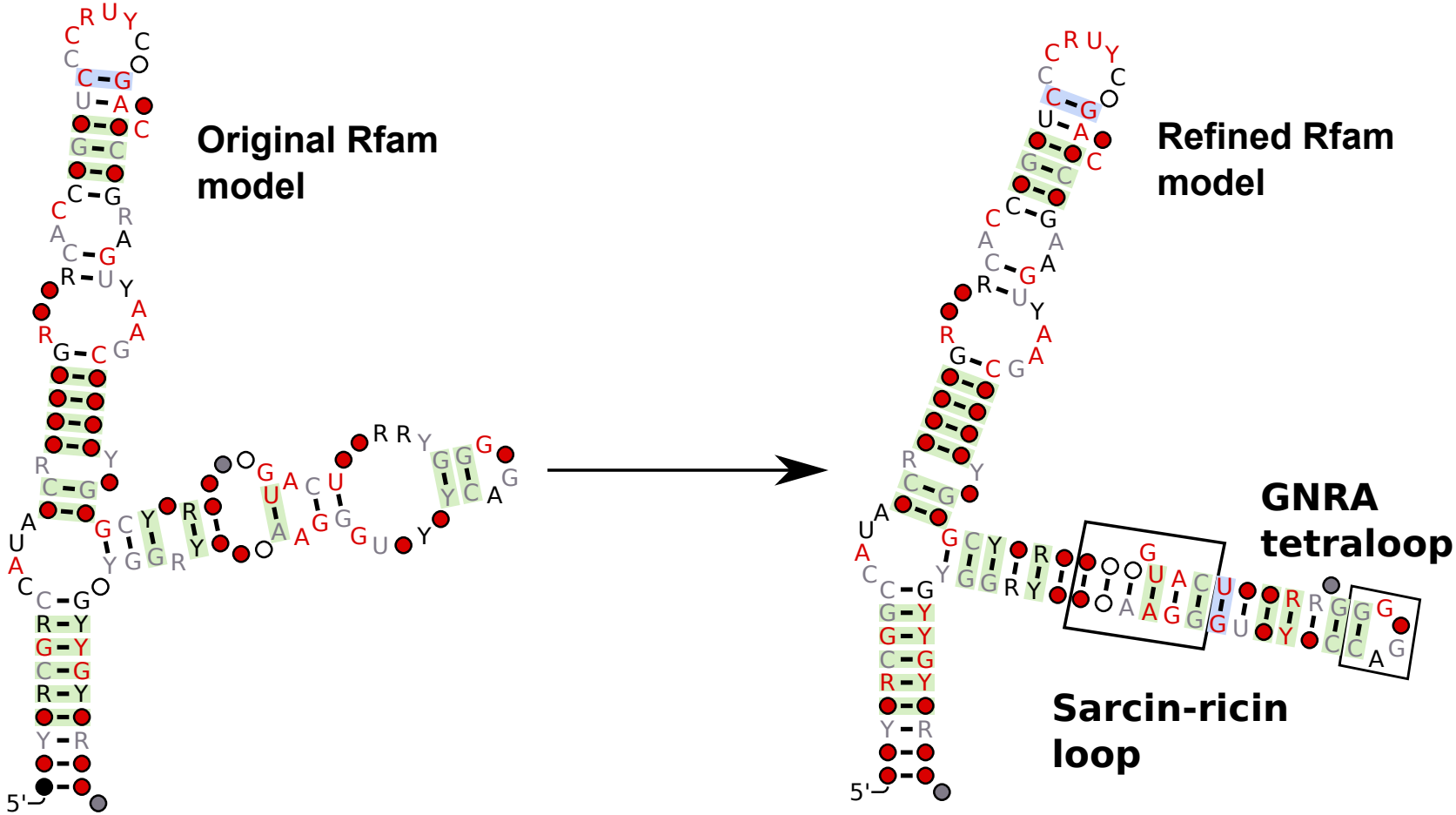


An example annotated Rfam alignment

```
# STOCKHOLM 1.0
#=GF ID   5S_rRNA
#=GF AC   RF00001
#...
#=GF WK   5S_ribosomal_RNA
#=GF SQ   712
#=GF MT.G GNRA
#=GF MT.s sarcin-ricin-2
#=GF MT.t twist_up

X62858.1/62-115             GCUGCG-U-UGUGGGUGUGUACUGCGGUUUU-UUG-CUGUGGGAA--GCCCACUUCA-CUG
X01588.1/64-115             UCACGU--UAGUGGG---GCCGUGGAUACCGUGAGGAUCC-GCAG-CCCCACU-AAG-CUG
#=GR X01588.1/64-115  MT.0  ........................GGGGGGGGGGGGGGGGG.....................
X05870.1/363-414            UCACGU--UGGUGGG---GCCGUGGAUACCGUGAGGAUCC-GCAG-CCCCACU-AAG-CUG
#=GR X05870.1/363-414 MT.0  ........................GGGGGGGGGGGGGGGGG.....................
L27170.1/63-116             CCAGCG--UCCGGCA--AGUACUGGAGUGCGCGAGCCUCUGGGAA-AUCCGGU-UCG-CCG
#=GR L27170.1/63-116  MT.0  ..ssss..sssssss..sssssssssssssssssssssssssss.sssssss.sss.ss.
L27163.1/61-115             CCAGCGUUCCGGGGA---GUACUGGAGUGCGCGACCCUCUGGGAA--ACCGGGUUCG-CCG
#=GR L27163.1/61-115 SS     ...(((((((((.........)))))))))(((((((...((......)))))).))).)..
#=GR L27163.1/61-115  MT.0  ..................................GGGGGGGGGGGG..GGGGGGG.......
#=GR L27163.1/61-115  MT.1  ..sssssssssssss...sssssssssssssssssssssssss..sssssssss.ss.
L27343.1/60-114             CCAGCGAACCAGCUA---GUACUAGAGUGGGAGACCCUCUGGGAG-CGCUGGU-UCG-CCG
#=GR L27343.1/60-114  MT.0  ........................GGGGGGGGGGGGGGGGG.....................
#=GR L27343.1/60-114  MT.1  ....sssssssssss...sssssssssssssssssssssssss.sssssss.sss.s..
M33891.1/66-117             CCAGCG--CCGGAGA---GUACUGCGC-GGGCAACCGCGUGGGAG--GCGAGG-CCGCUCG
#=GR M33891.1/66-117  MT.0  ........................GGGG.GGGGGGGGGGGG.....................
X01484.1/62-114             GUCAGG--CCCAGUU--AGUACUGAGGUGGGCGACCACUUGGGAA-CACUGGG-U-G-CUG
#=GR X01484.1/62-114  MT.0  ........................GGGGGGGGGGGGGGGGG.....................
#=GR X01484.1/62-114  MT.1  ...sss..sssssss..sssssssssssssssssssssssss.sssssss.s.s.ss.
#=GC SS_cons                ::::::::::<-<<--------<<-----<<___>>----->>----->>->:::::::
#=GC RF                     cCaGcG..cccggga...GUAcUggGGUggGcgAcCcucUGGgAA.CaCccgg.uCG.CUG
//
```

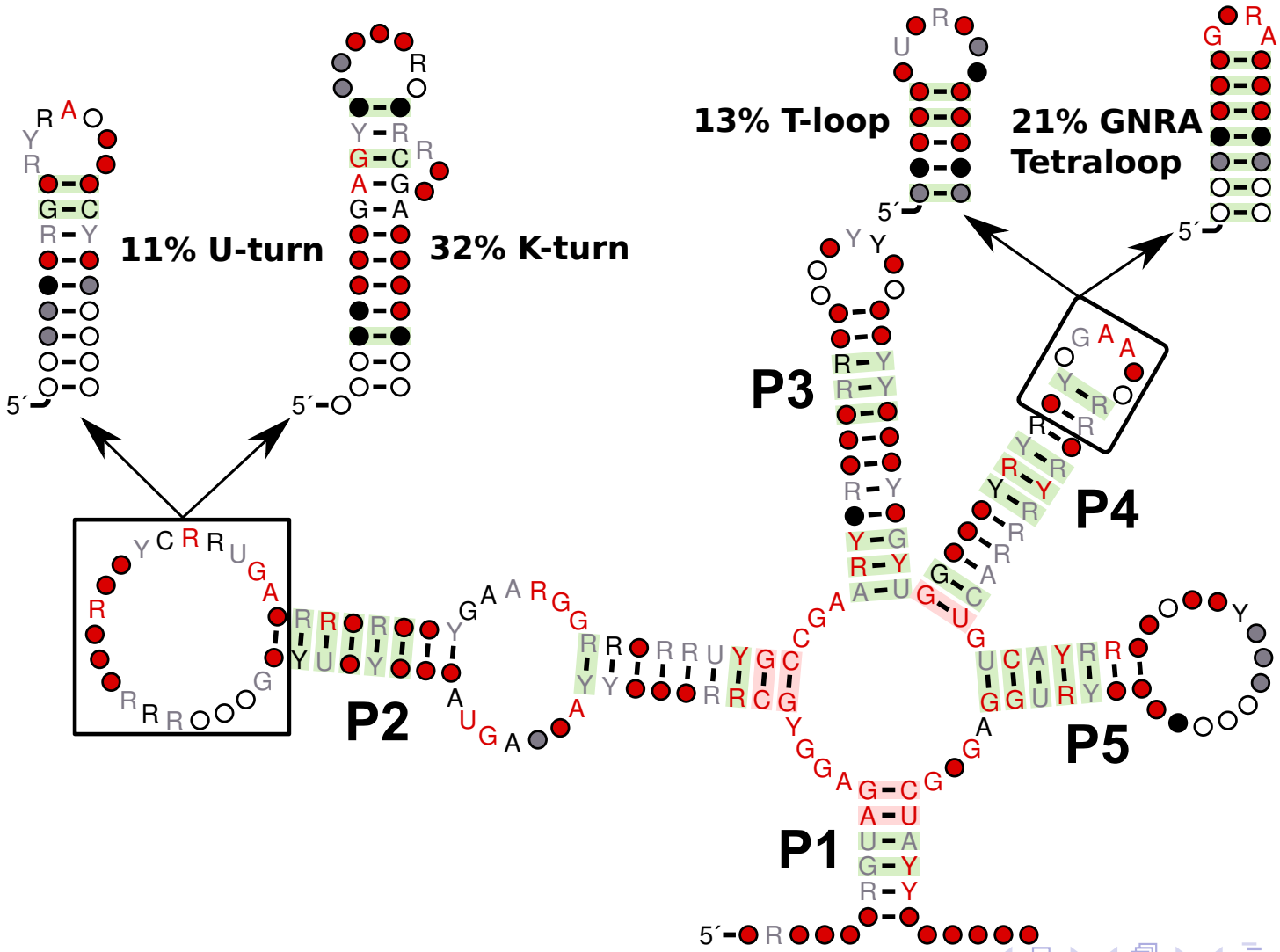


Uses for a motif DB (I): Improving a RNA alignments and structure prediction constraints



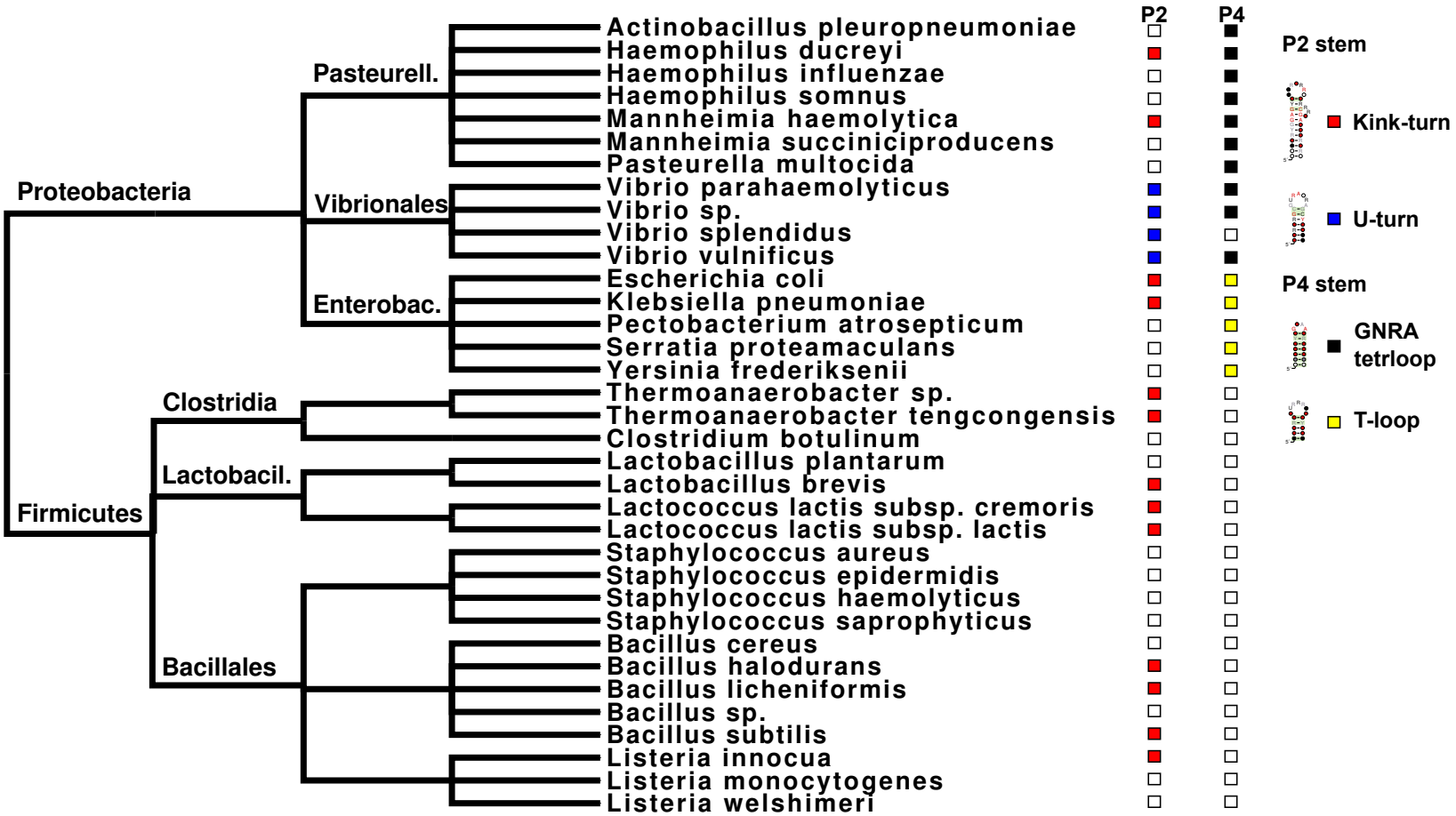
Uses for a motif DB (II): RNA structural motifs

- The Lysine Riboswitch

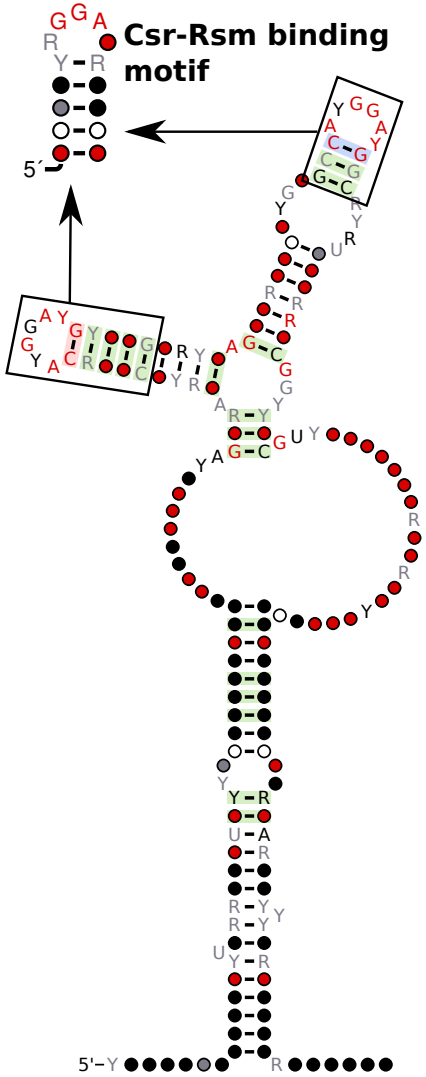


Uses for a motif DB (II): RNA structural motifs

- The Lysine Riboswitch

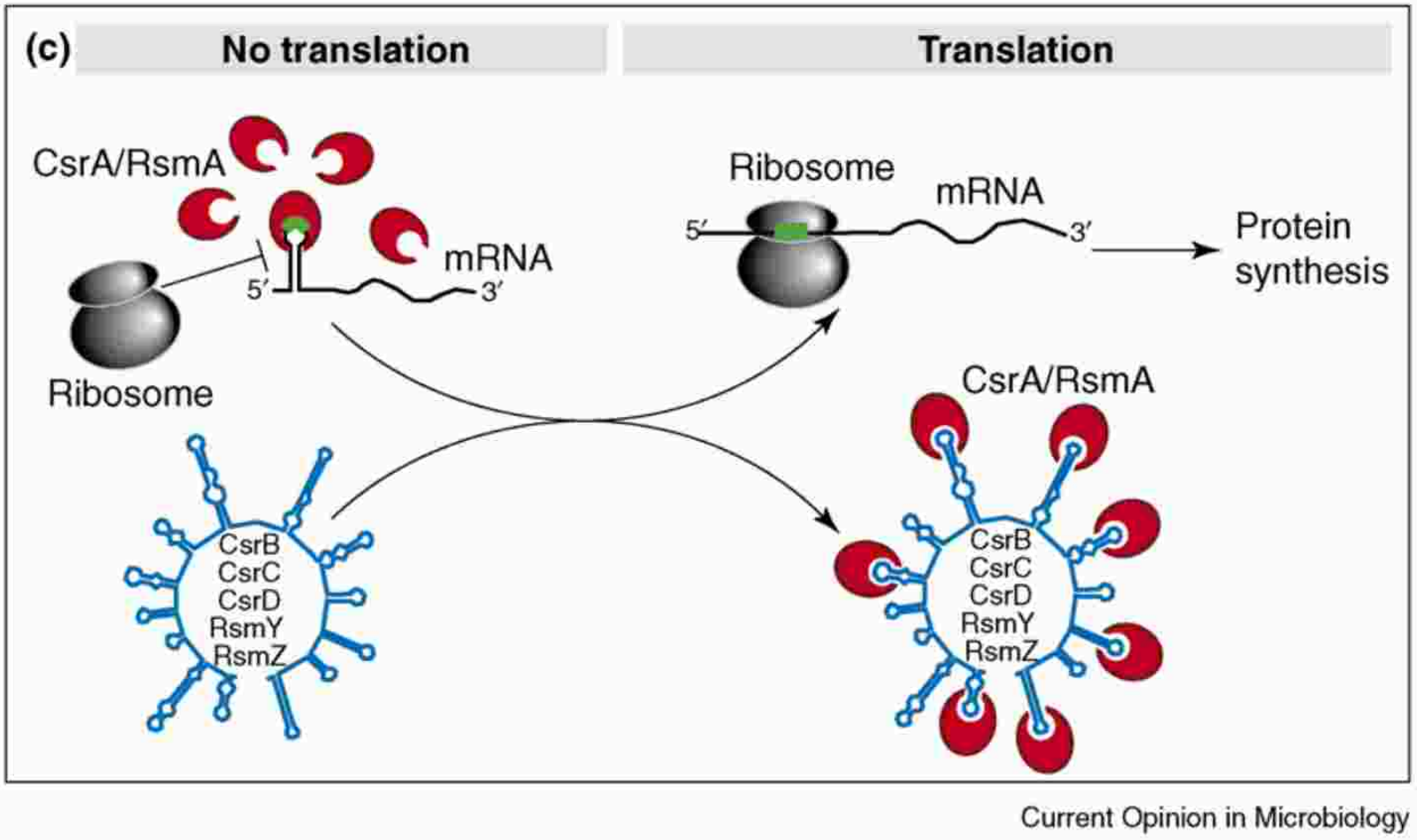


Uses for a motif DB (III): Functionally characterising novel RNAs



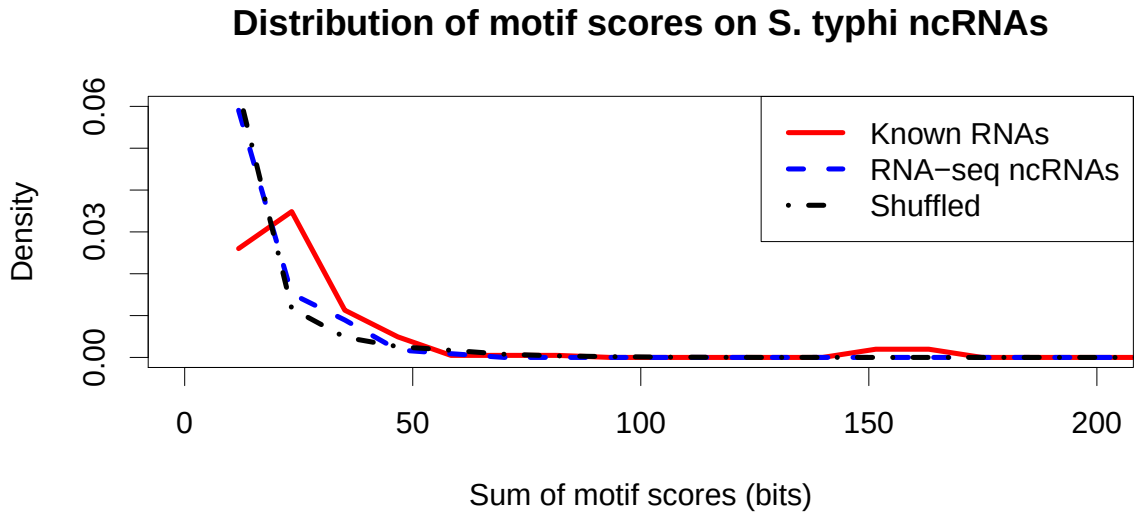
- TwoAYGGAY sRNA: 216 homologues found in Human gut metagenome sequences, γ-Proteobacteria and Clostridiales species
- Weinberg Z *et al.* (2010) "Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea and their metagenomes". *Genome Biol.*

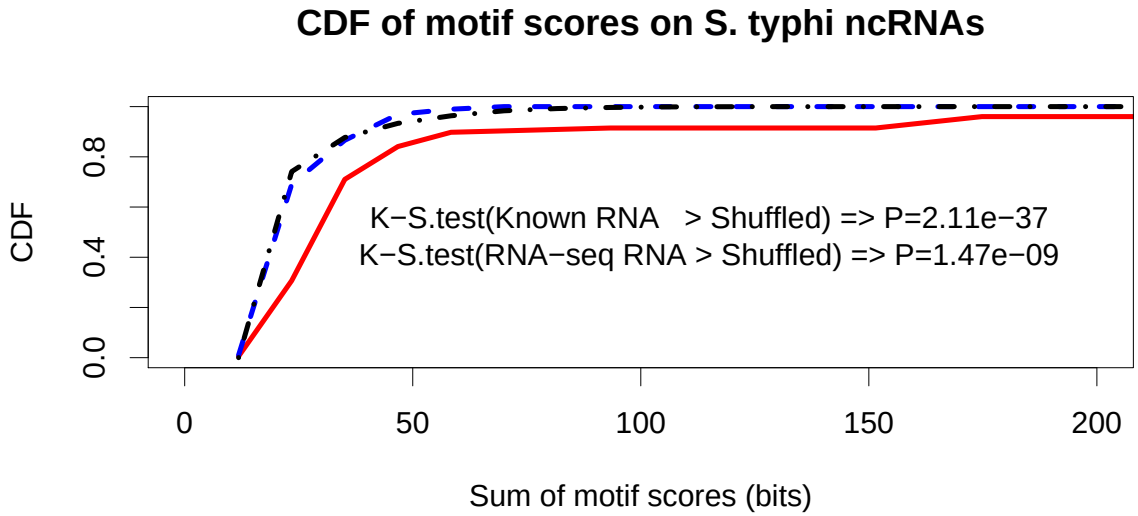
Csr-Rsm background



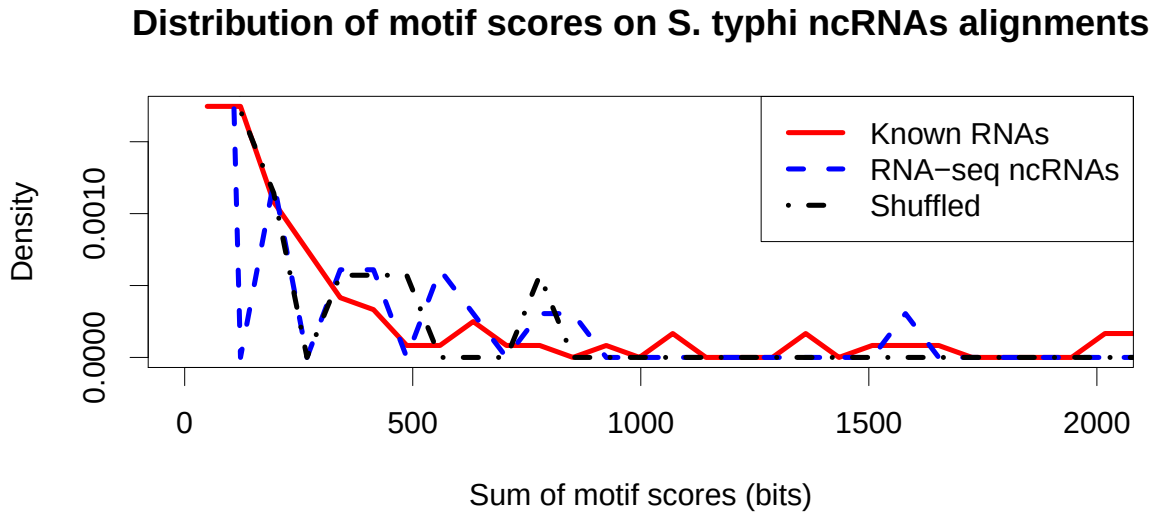
Toledo-Arana *et al.* (2007) Small noncoding RNAs controlling pathogenesis. *Current Opinion in Microbiology.*

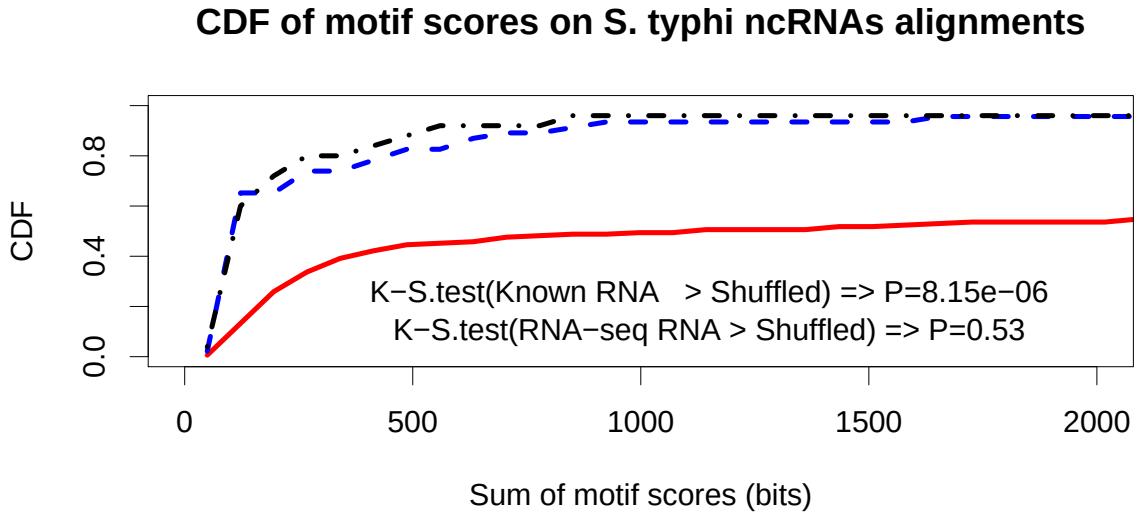
Uses for a motif DB (IV): Identifying and ranking ncRNAs of interest in transcriptome results



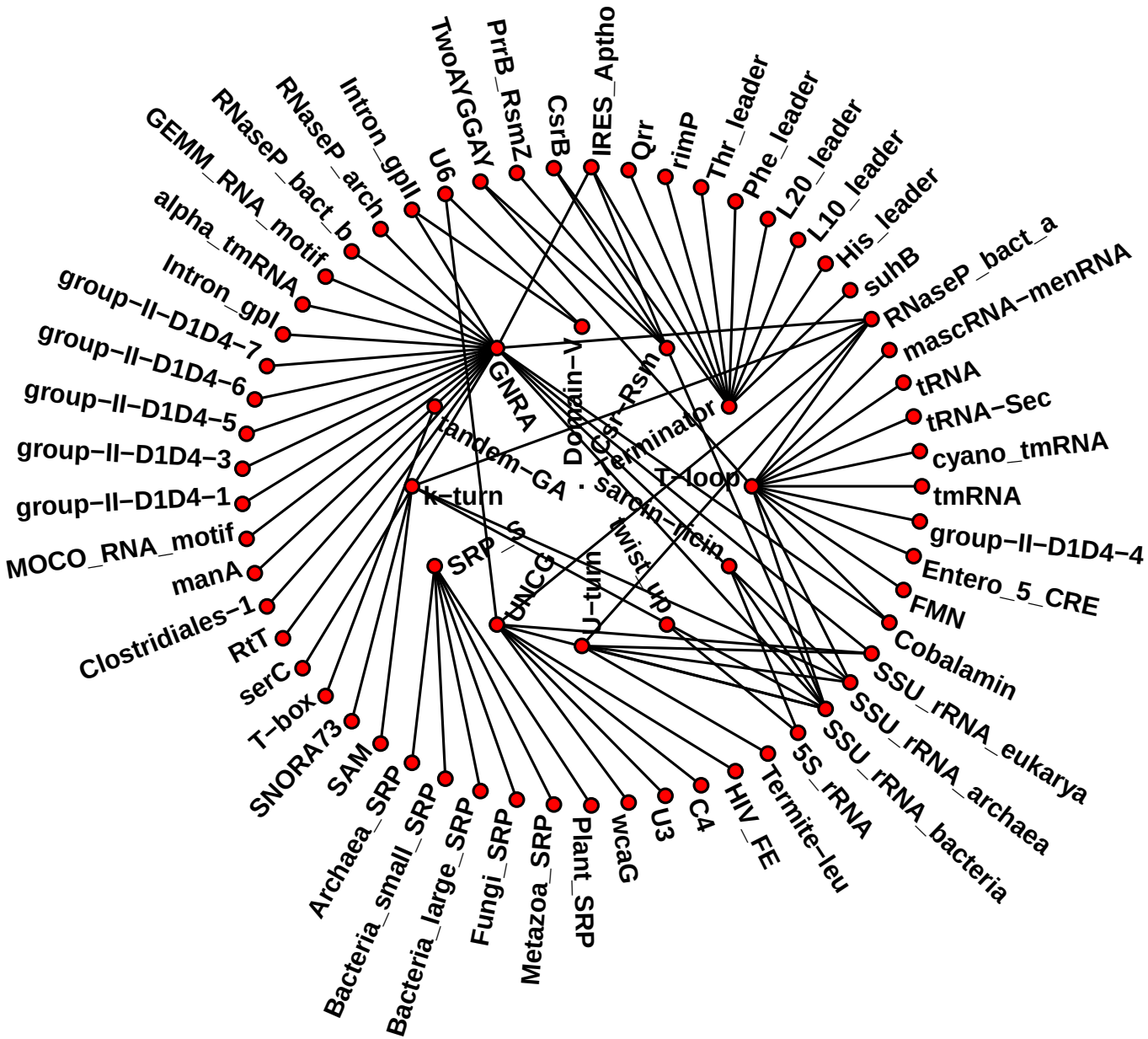


Uses for a motif DB (IV): Identifying and ranking ncRNAs of interest in transcriptome results

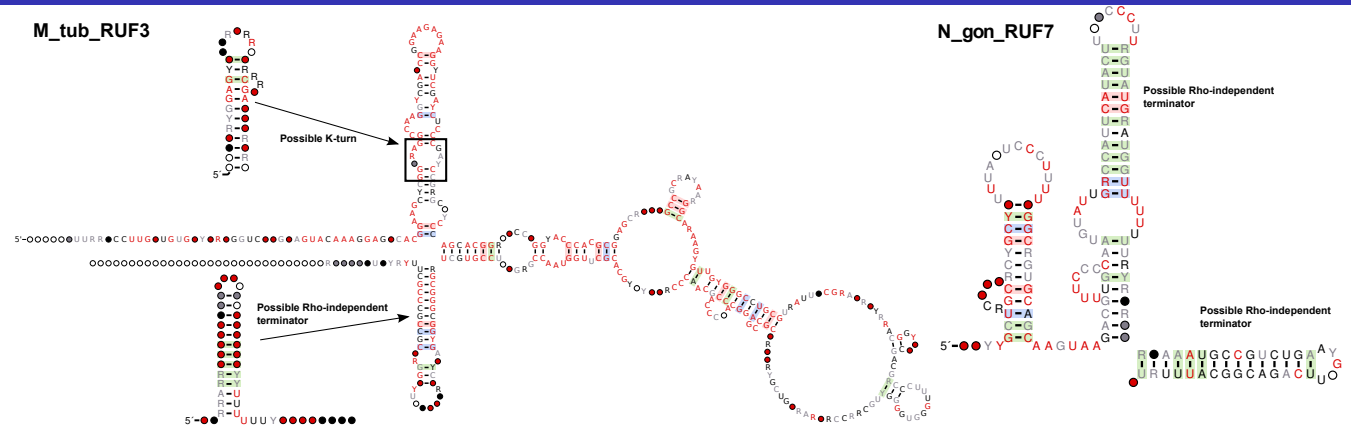




80 highest scoring matches between RMfam & Rfam



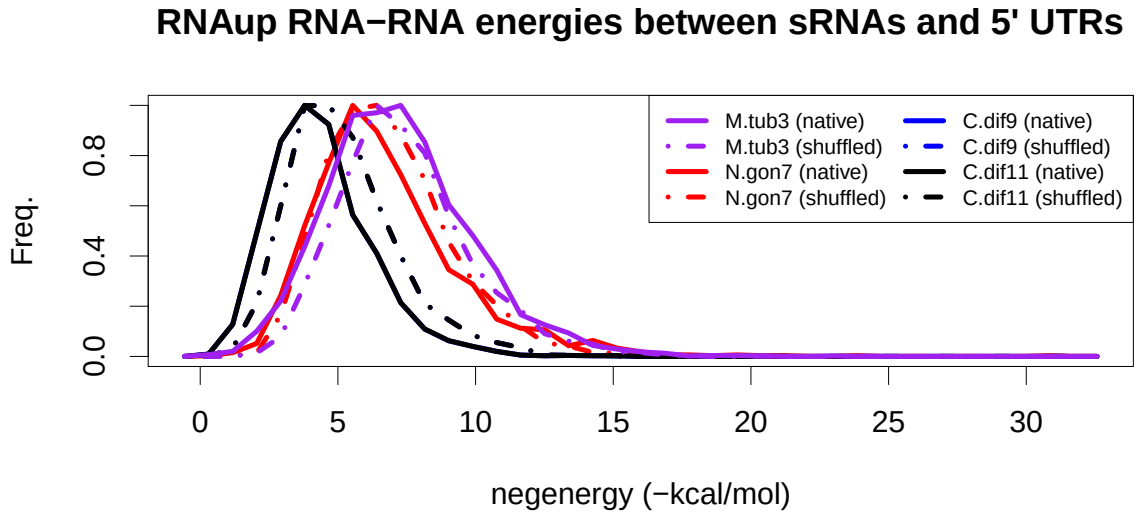
What about the highly expressed ncRNAs?

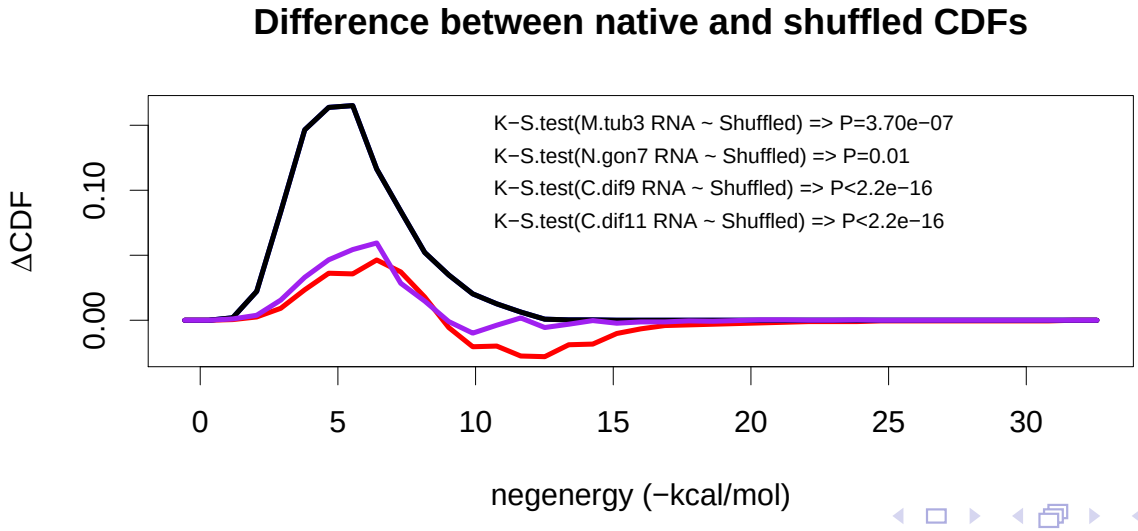


RNA	RNAcode	RNAz	Motifs	Low Seq. complexity	G+C sRNA (gen.)
M.tub. RUF3	No	RNA (prob=1.00)	Terminator,k-turn	No	59% (66%)
N.gon. RUF7	No	RNA (prob=0.94)	Terminator	No	43% (52%)
C.dif. RUF9	? (P=0.044)	RNA (prob=1.00)	Terminator	No	30% (29%)
C.dif. RUF11	? (P=0.079)	RNA (prob=1.00)	Terminator	Yes (23/170nts)	26% (29%)

What about the highly expressed ncRNAs?

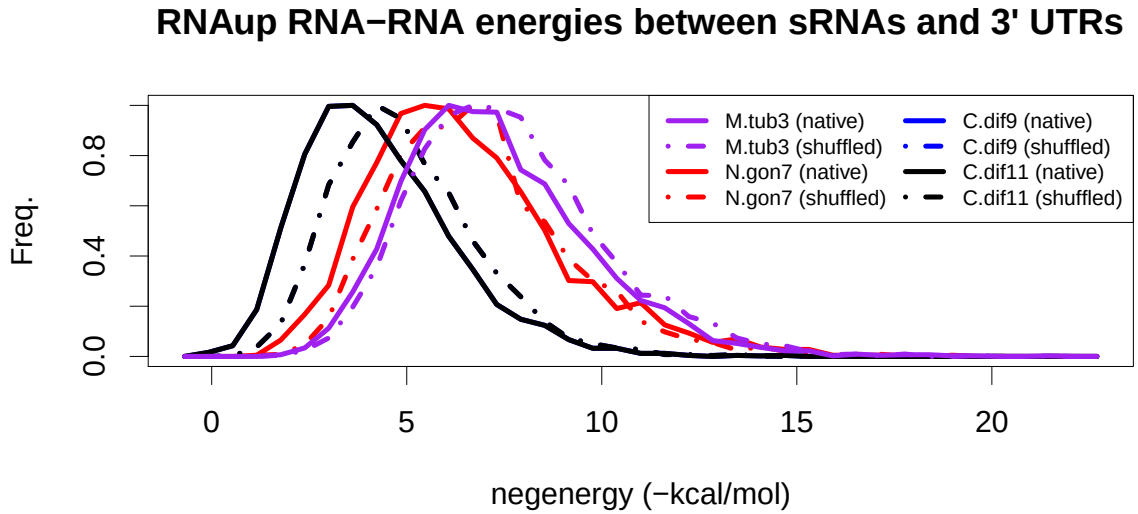
- Are they antisense to any 5‘ UTRs

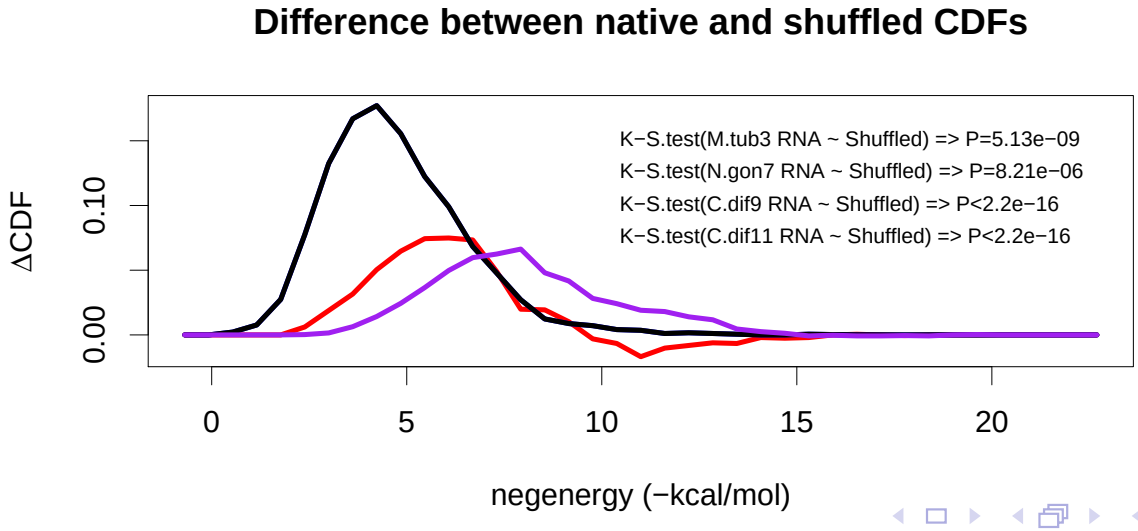




What about the highly expressed ncRNAs?

- Are they antisense to any 3‘ UTRs



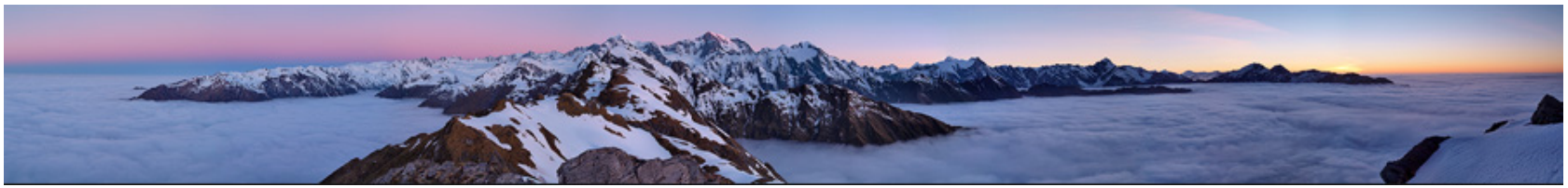


Discussion points

- A useful curation tool for building alignments and structures
- Can provide testable function hypotheses, sometimes
- Limited use outside of alignments based on the Salmonella Typhi results
- Even in alignments, false-positives & negatives are common
- More motifs (Terminators, Shine-Dalgarno, ...)
- Limit trivial motifs and prevent rediscovery of old motifs
- Snapshots of the data are available from https://github.com/ppgardne/RMfam
- Can anyone here do better on MTB3, NGON7, CDIF9 & CDIF11?

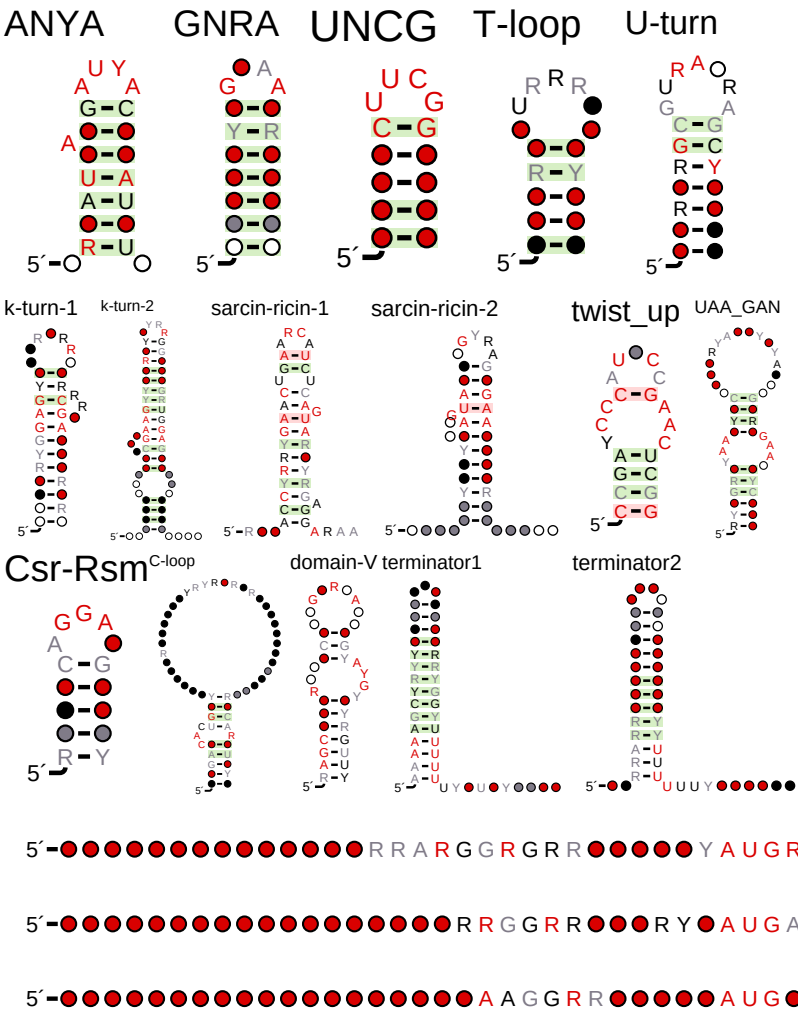
Thanks!

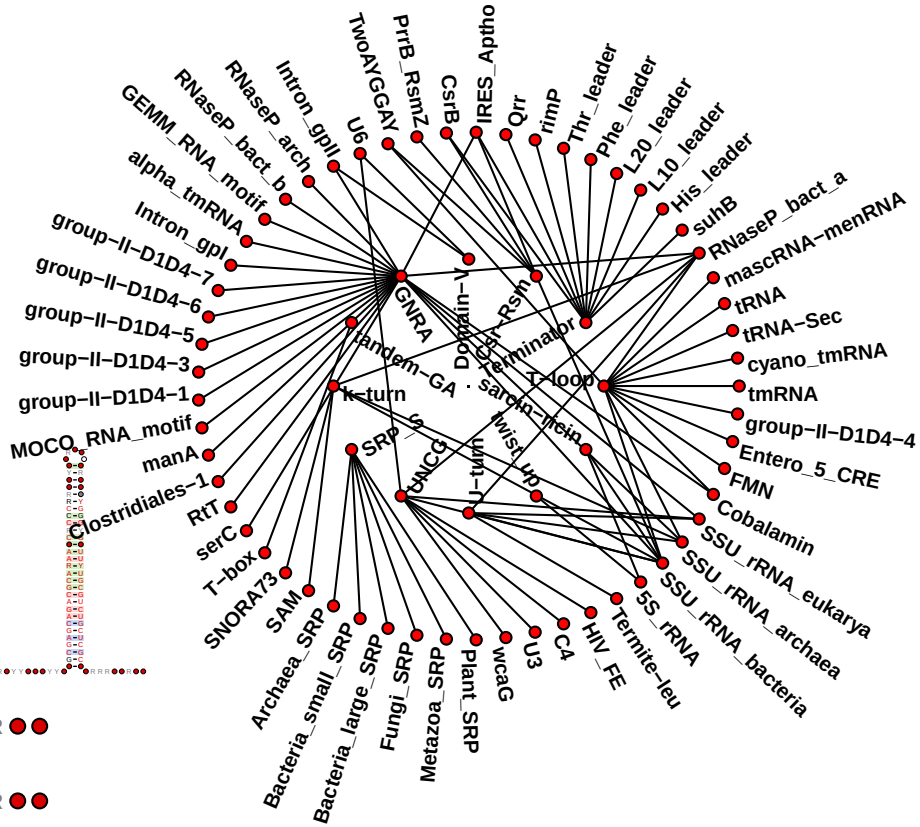
- Lars Barquist, Alex Bateman & Rob Knight



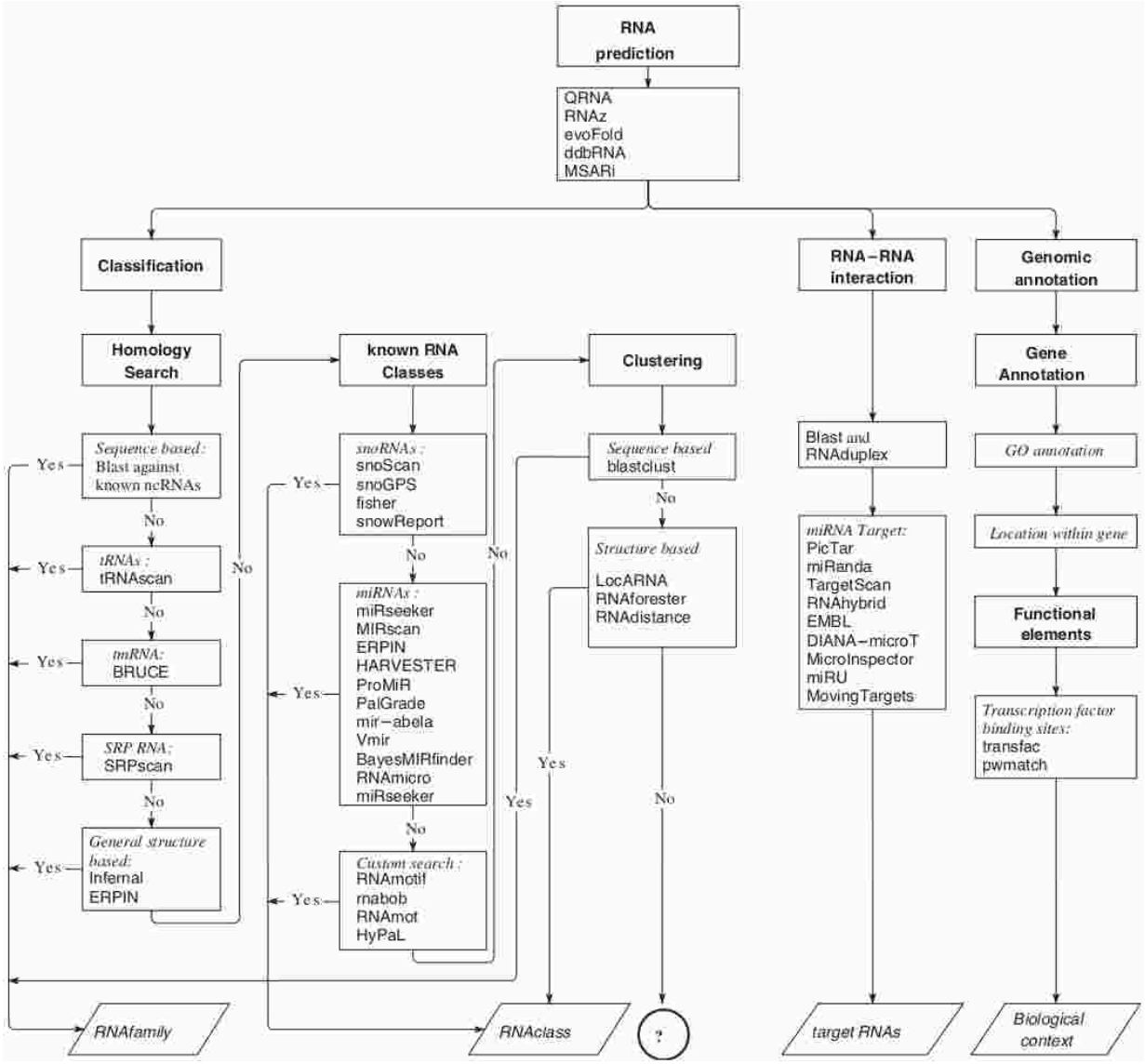
PPG is supported by a Rutherford Discovery Fellowship from Government funding, administered by the Royal Society of New Zealand.

RNA motifs


ANYA
GNRA
UNCG
T-loop
U-turn
k-turn-1
k-turn-2
sarcin-ricin-1
sarcin-ricin-2
twist_up
UAA_GAN
Csr-Rsm
C-loop
domain-V terminator1
terminator2
TRIT
5´-RRARGGRGRR-YAUGR-R
5´-RRGGRR-RY-AUGAR-R
5´-AAGGRR-AUG-R


GNRA
k-turn
SRP
UNCG
tandem-GA
sarcin-ricin
twist_up
T-loop
terminator
Csr-Rsm
U-turn
Domain-V
GEMM_RNA_motif
alpha_tmRNA
Intron_gpI
group-II-D1D4-7
group-II-D1D4-6
group-II-D1D4-5
group-II-D1D4-3
group-II-D1D4-1
MOCO_RNA_motif
manA
Clostridiales-1
RtT
serC
T-box
SNORA73
SAM
Archaea_SRP
Bacteria_small_SRP
Bacteria_large_SRP
Fungi_SRP
Metazoa_SRP
Plant_SRP
wcaG
U3
C4
HIV_FE
Termite-leu
5S_rRNA
SSU_rRNA_bacteria
SSU_rRNA_archaea
SSU_rRNA_eukarya
Cobalamin
FMN
Entero_5_CRE
group-II-D1D4-4
tmRNA
cyano_tmRNA
tRNA-Sec
tRNA
mascRNA-menRNA
RNaseP_bact_a
suhB
His_leader
L10_leader
L20_leader
Phe_leader
Thr_leader
rimP
Qrr
IRES_Aptho
CsrB
PrrB_RsmZ
TwoAYGGAY
U6
Intron_gpII
RNaseP_arch
RNaseP_bact_b

RNA classification



Backofen *et al.* (2007) RNAs everywhere: genome-wide annotation of structured RNAs. *Journal of Experimental Zoology.*