Extending the phylogenetic range of systematic surveys for ncRNA prediction to trypanosomatid species

Dominic Rose

Bioinformatics Group, Institute for Computer Science, University of Leipzig

Bled, Feb 2006

Outline

A short definition

Noncoding RNAs (ncRNAs) are

- Transcripts that are not translated into proteins
- Molecules that induce cellular activity without protein influence

Noncoding RNA variety

The Noncode DB provides:

- 5,339 public sequences
- 861 organisms
- 109 traditional classes
- 26 cellular process

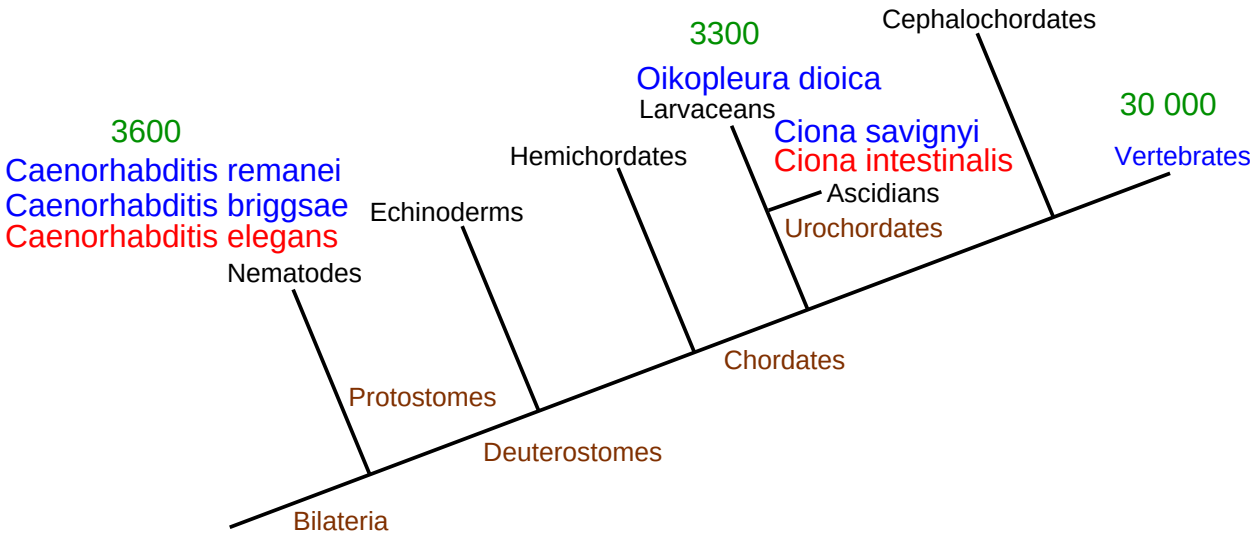
Noncoding RNA variety

Exemplary ncRNA classes and their activities

Class	Process	Function
XIST	Gene silencing	Required for X chromosome inactivation.
snRNA	RNA processing	Forming the core of the spliceosome, RNA splicing.
gRNA	RNA modification	RNA editing (insertion or deletion of uridylates).
miRNA	mRNA translation	Represses translation by pairing with 3' end of target mRNA

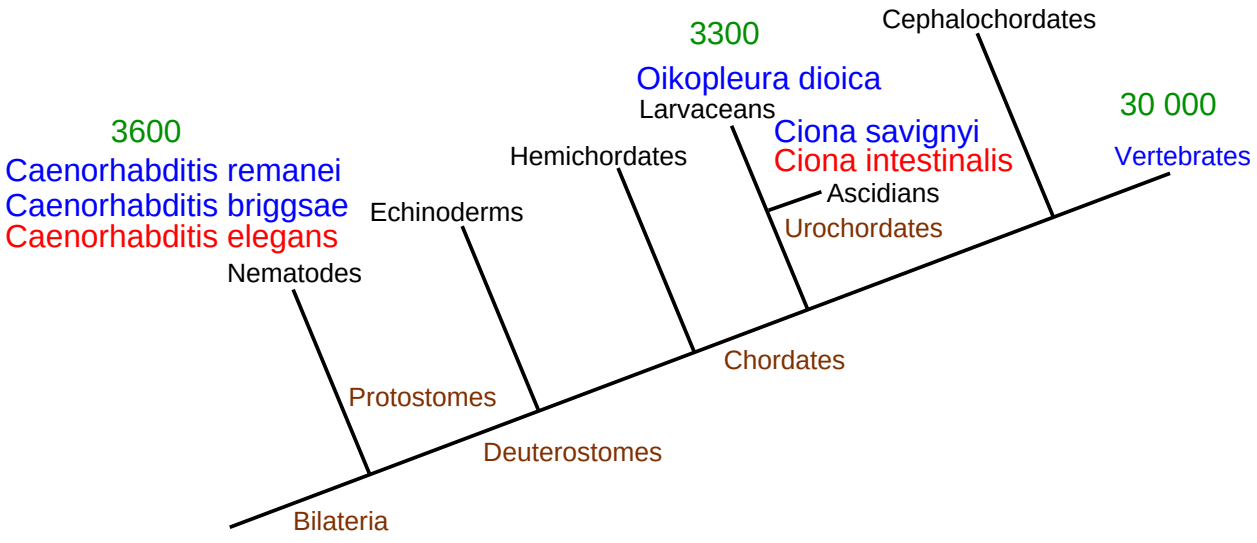
Phylogenetic range of ncRNA predictions

RNAz-based ncRNA predictions of the past:



Phylogenetic range of ncRNA predictions

RNAz-based ncRNA predictions of the past:



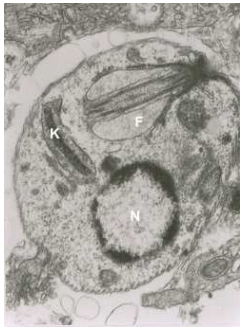
Recent screens:

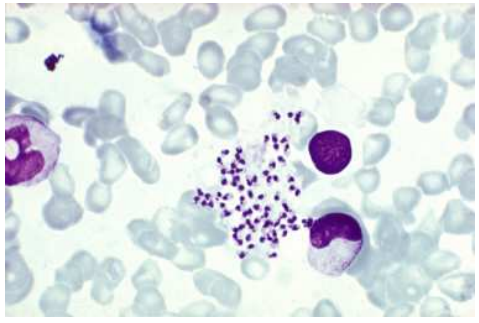
- Leishmania and Trypanosoma
- Teleost fishes
- Plants???

Objects of research

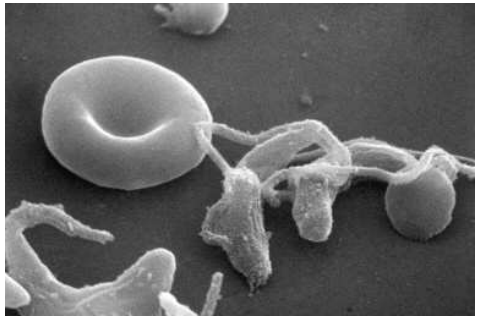
Taxonomic family: Trypanosomatidae

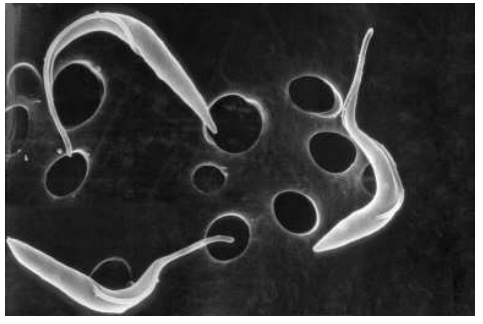
- Unicellular, flagellated protozoan parasites
- Leishmania: L. major (Lm), L. infantum (Li)






- Trypanosoma: T. brucei (Tb)



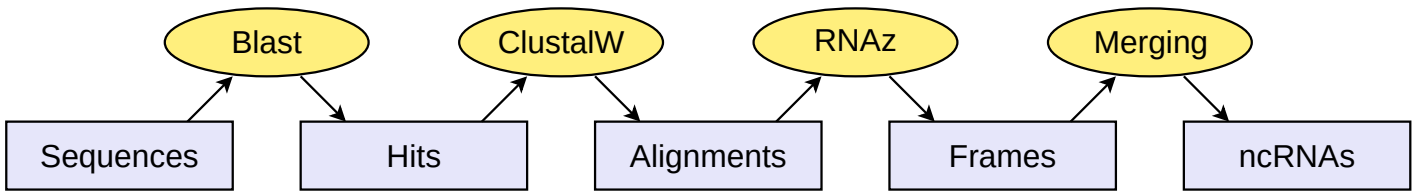


Outline

Give me the RNAs...

Retrieving structural ncRNAs out of blank sequence data:

- Start with genome-wide alignments of nc DNA
- Process them with your favorite ncRNA prediction tool
- Sort the output to annotate putative ncRNAs

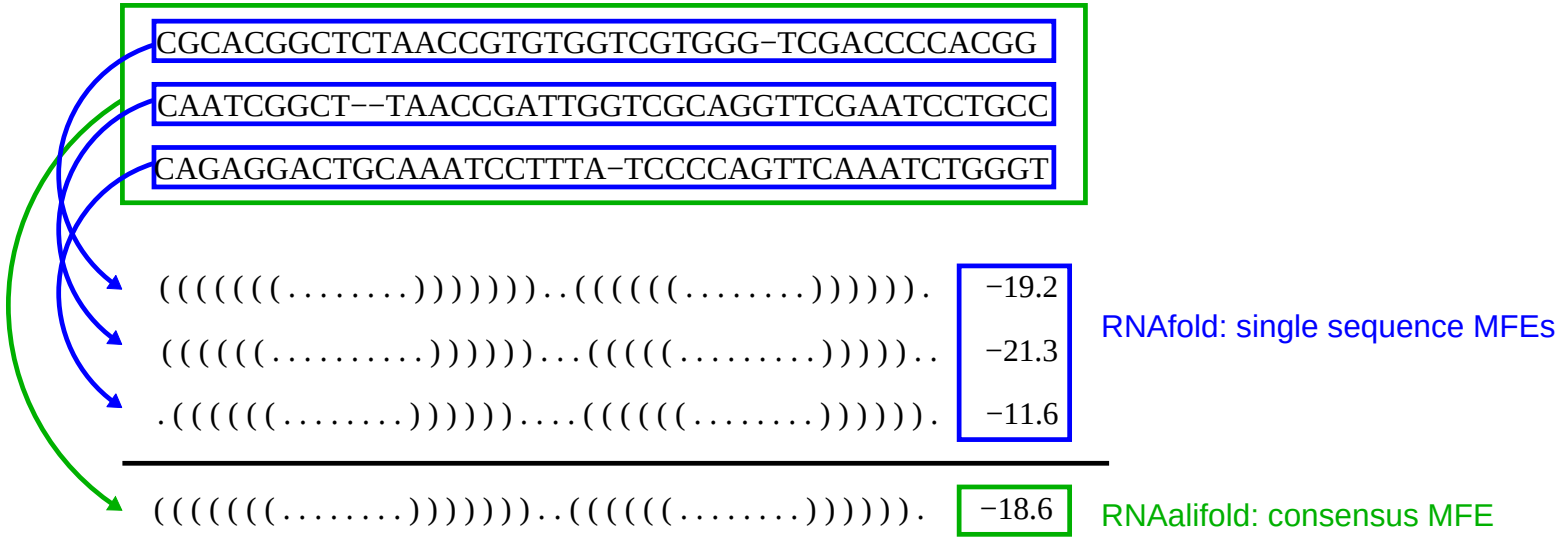


NcRNA characteristics

General ncRNA features:

- They lack sequence signals as found in protein coding genes
- Evolutionary conserved ncRNA secondary structures indicate functionality

The RNAz approach



- SVM classification:
  - SCI $= \frac{\text{consensus MFE}}{\text{mean single sequence MFE}}$ (structure conservation index)
  - z-score $= \frac{\sum \text{single sequence z-score}}{N}$ (thermodynamic stability)

Outline

NcRNA predictions for Leishmania species

screen	LiLm	LiLmTb
$p > 0.5$	45,329 (18%)	149 (0.05%)
$p > 0.9$	22,187 (9%)	66 (0.02%)
$p > 0.98$	11,496	35
$p > 0.99$	8,627	19

screen	LmLi	LmLiTb
$p > 0.5$	53,837 (22%)	291 (0.04%)
$p > 0.9$	26,030 (12%)	108 (0.01%)
$p > 0.98$	13,355	41
$p > 0.99$	10,109	27

Estimated sensitivities

type	N	n	N_a	N_g	s_{Na}	s_{Ng}	s_{na}	s_{ng}
	LiLmTb							
rRNA	5	12	11	21	0.46	0.24	1.09	0.57
tRNA	35	39	48	55	0.73	0.64	0.81	0.71
misc_RNA	0	0	0	9	-	0.00	-	0.00
snRNA	1	1	1	7	1.00	0.14	1.00	1.00
snoRNA	0	0	0	35	-	0.00	-	0.00
	LmLiTb							
rRNA	28	34	46	61	0.61	0.46	0.74	0.56
tRNA	45	50	65	67	0.69	0.67	0.77	0.77
snRNA	1	1	1	5	1.00	0.20	1.00	0.20

False positive rates and specifities

LiLmTb

p	> 0.5	> 0.9	> 0.98	> 0.99
FPR RNAz frames	54%	31%	24%	22%
Specificity	0.985	0.997	0.999	0.999

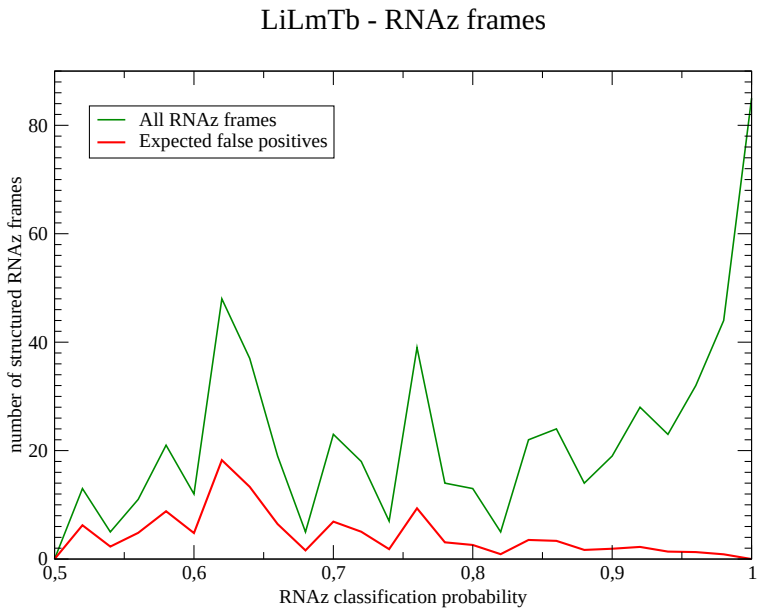
LmLiTb

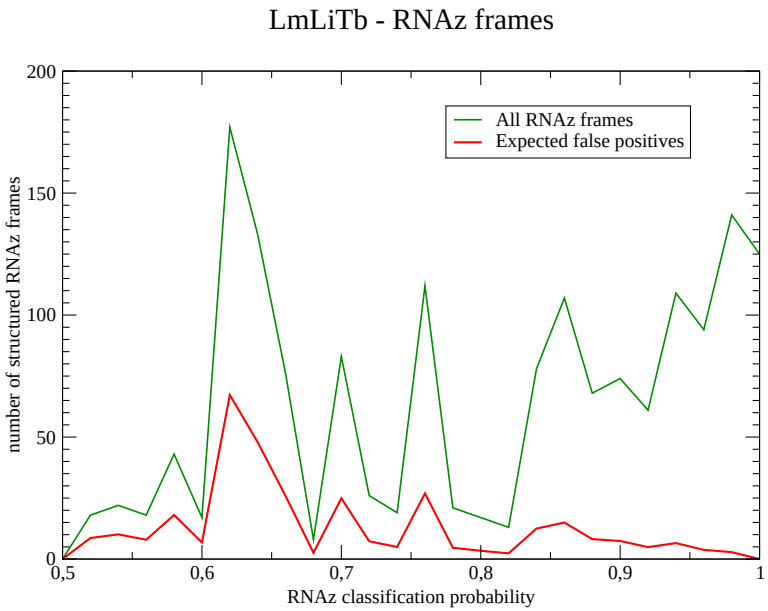
p	> 0.5	> 0.9	> 0.98	> 0.99
FPR RNAz frames	49%	29%	35%	18%
Specificity	0.975	0.996	0.999	0.999

Detection rates

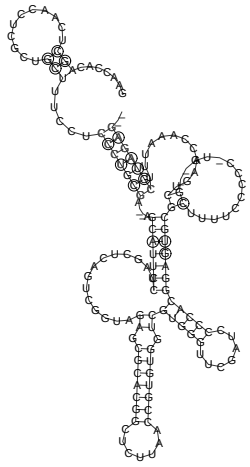
	per 1 mb alignment		per 1 mb nc region	
LiLmTb	normal	shuffled	normal	shuffled
$p > 0.5$	9.81	5.32	20.02	10.86
$p > 0.9$	3.58	1.1	7.31	2.24
$p > 0.98$	1.44	0.34	2.93	0.69
$p > 0.99$	0.76	0.17	1.55	0.34
LmLiTb	normal	shuffled	normal	shuffled
$p > 0.5$	23.18	11.26	56.26	27.32
$p > 0.9$	7.4	2.16	17.96	5.25
$p > 0.98$	1.75	0.61	4.24	1.49
$p > 0.99$	1.27	0.22	3.08	0.54

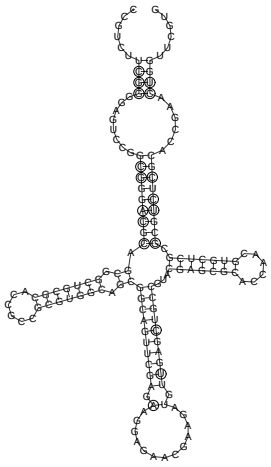
RNA classification probability

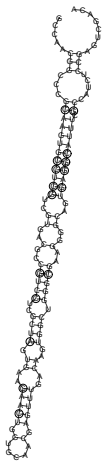


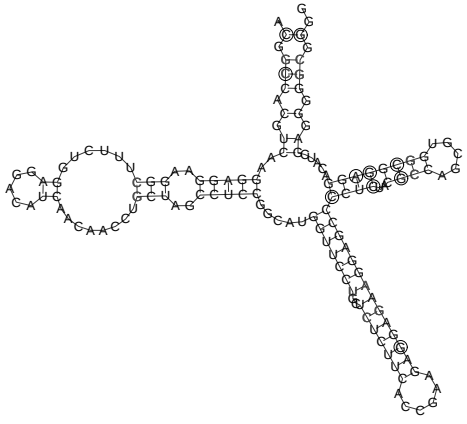


Exemplary consensus structures (LiLmTb)

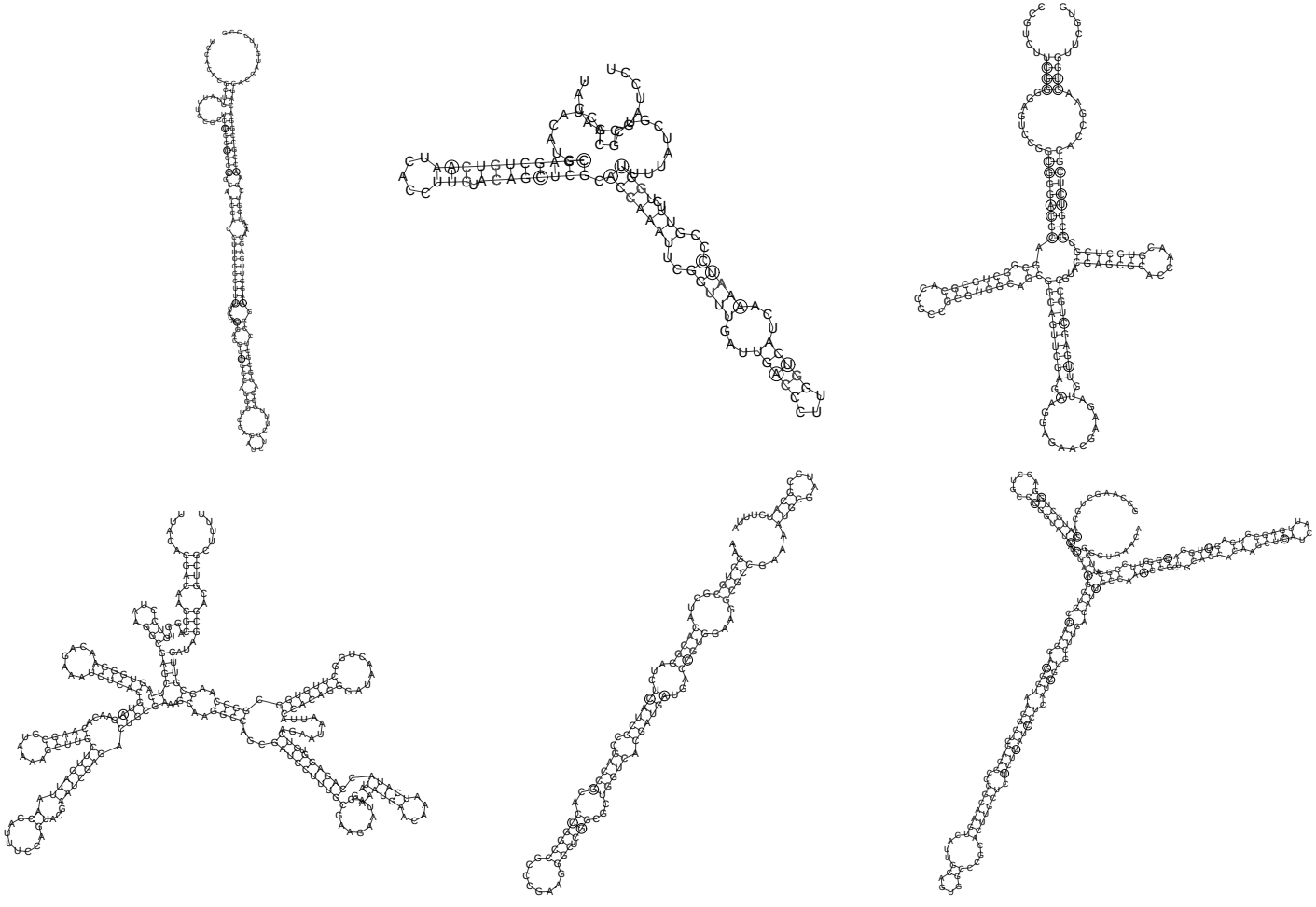






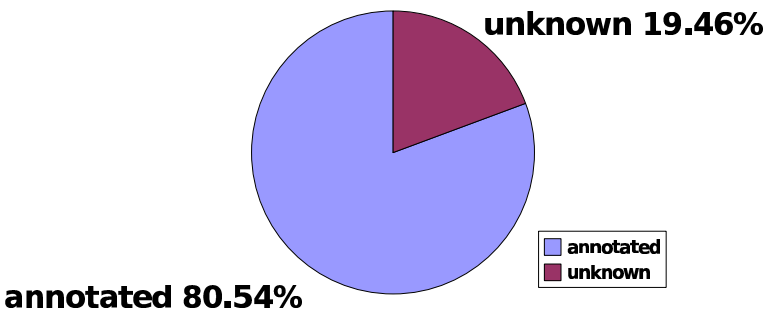


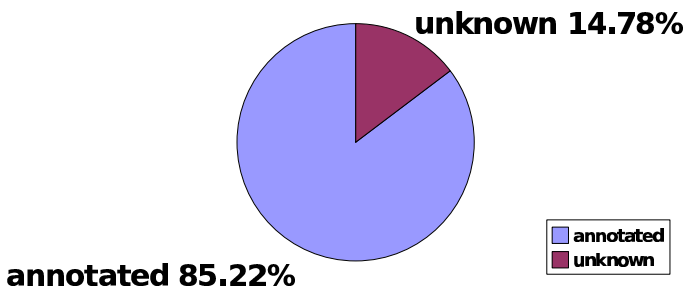
Exemplary consensus structures (LmLiTb)



Annotation

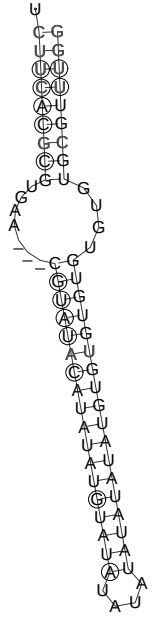
	LiLmTb	LmLiTb
tRNAscan-SE	56	49
RNAmicro	1	7
Noncode	2	2
Rfam	114	208
SMN (human)	1	7
SMN (L. seymouri)	2	7



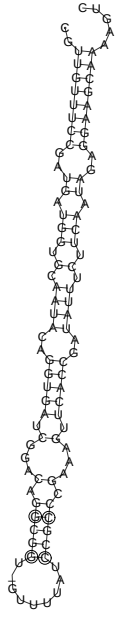


MiRNA examples obtained by RNAmicro

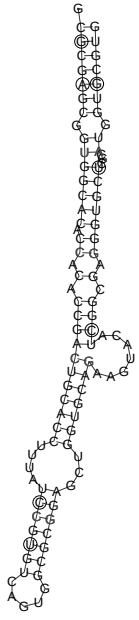
(Li)



(Lm)

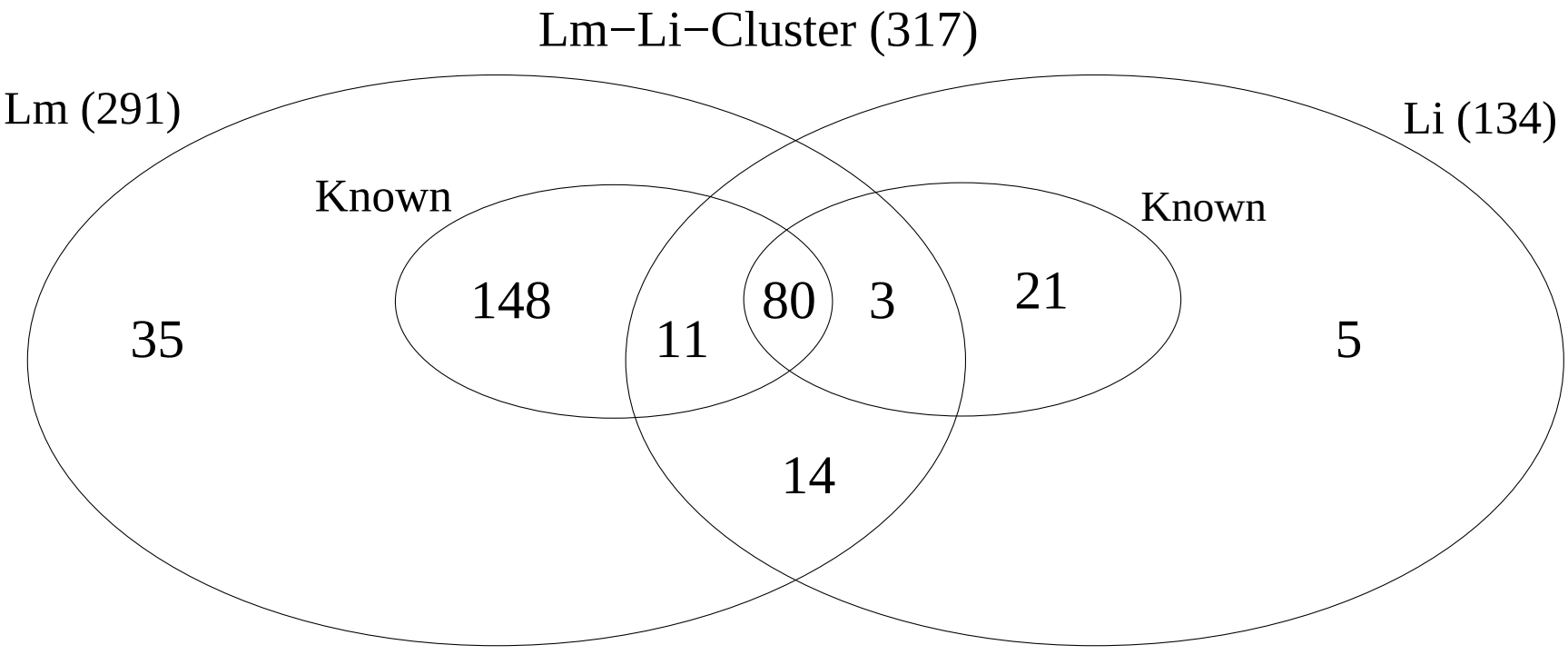


(Lm)



Combining the Leishmania screens

(which Lm of LmLiTb is identifiable in LiLmTb)



Browsing Leishmania ncRNA clusters

http://www.bioinf.uni-leipzig.de/~dominic/projects/leish

Leishmania ncRNA project

home | help | results | statistics | download | other ncRNA projects — coded by Dominic Rose

This is an overview of the ncRNAs we annotated on *Leishmania* species.

	LmLiTb screen				LiLmTb screen					
Cluster	ID	p-score	z-score	SCI	ID	p-score	z-score	SCI	Known as	Additional info
1.1	175680	0.73	-1.58	0.82	-	-	-	-	-	Rfam "Iron response element' [bit-score=1.43]"
2.1	175681	0.55	-1.59	0.59	-	-	-	-	-	Rfam "Iron response element' [bit-score=4.25]"
28.1	175713	0.94	-1.15	0.86	-	-	-	-	tRNA "tRNA Met anticodon CAT, Cove score 62.62"	tRNAscan-SE "Met [CAT]" Rfam "tRNA' [bit-score=67.80]"
28.2	-	-	-	-	151108	0.96	-1.26	0.76	tRNA "tRNA Met anticodon CAT, Cove score 62.62"	tRNAscan-SE "Met [CAT]" Rfam "tRNA [bit-score=67.80]" UTR "1000nt 5'UTR"
28.3	-	-	-	-	150913	0.94	-1.15	0.86	tRNA "tRNA Met anticodon CAT, Cove score 62.62"	tRNAscan-SE "Met [CAT]" Rfam "tRNA [bit-score=67.80]"

Outlook

Future work:

- RNAz will be integrated into DAS

Remember:

- RNAz is powerful but not the ultimative general ncRNA detection black box ;-) (uncertainity)
- Example: RNAz hits at CDS
  LiLmTb 149 ncRNAs, but 82 ~~non~~coding RNA signals
  LmLiTb 291 ncRNAs, but 66 ~~non~~coding RNA signals

Thank you!!!

;-)