Computational RN(Az)omics of Drosophilids

Dominic Rose

Bioinformatics Group, Institute for Computer Science, University of Leipzig

Bled, Feb 2007

Outline

Introduction and Motivation

Results

Conclusion

A short introduction

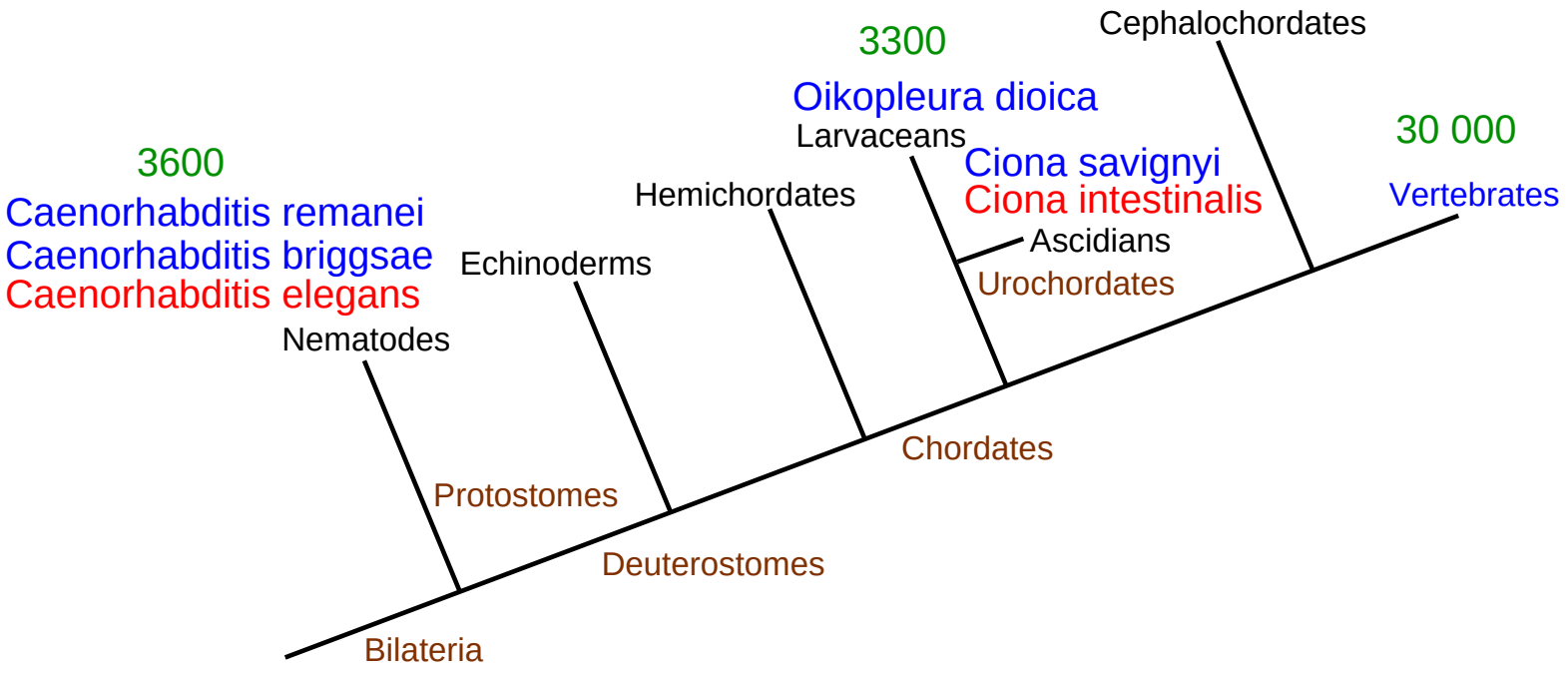
- Multiple regulatory layers of gene regulation are emphasized for eukaryotes
- Many involve non-protein-coding RNAs (ncRNAs)
- Vastly different structures, functions and evolutionary patterns
- Gene silencing, RNA processing/modification, . . .

Why Fly?

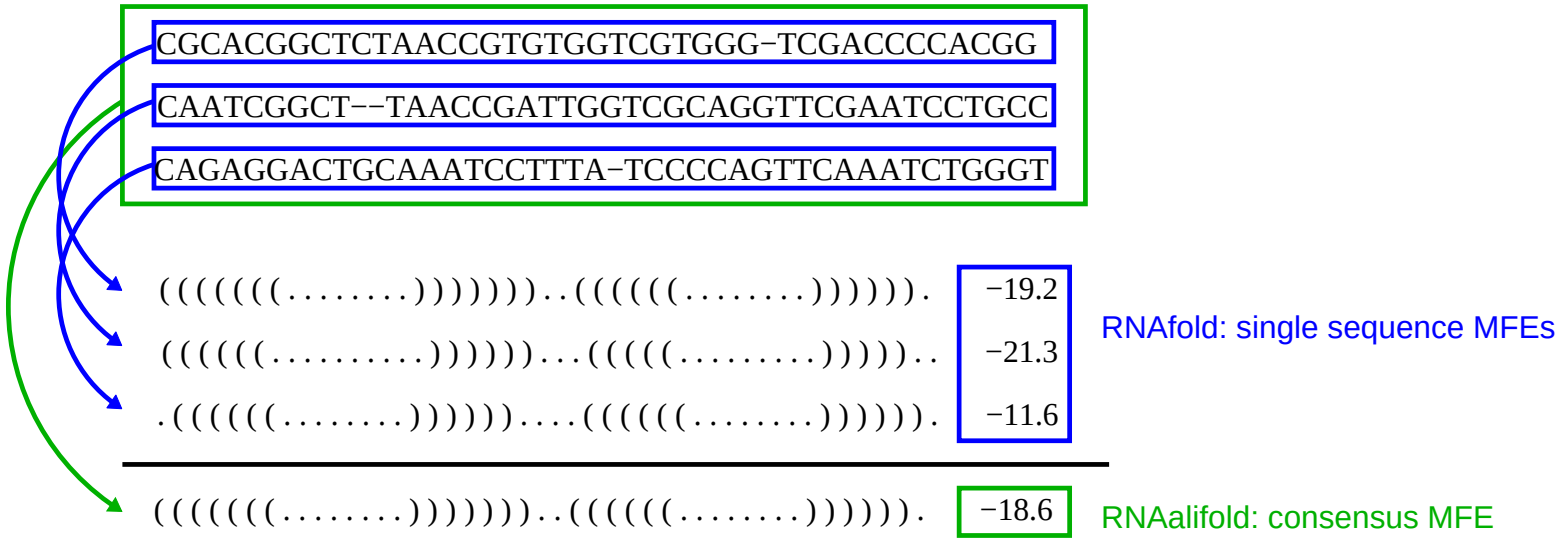
”Noncoding HCEs also show strong statistical evidence of an enrichment for RNA secondary structure.”

A. Siepel, G. Bejerano, J. S. Pedersen *et al.*:
Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.
Genome Res., 15(8):1034-1050, Aug 2005

RNAz surveys of the past



The RNAz approach



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

Screen design

- 12 drosophilid genomes (CAF1-”Freeze”)
- *D. melanogaster* serves as reference
- Use existing Pecan alignments to apply RNAz pipeline
- Alignment pre-processing:
  - Filter for valid aligned regions
  - Filter for 3- to 6-way alignments
- Annotate predictions

Outline

Introduction and Motivation

Results

Conclusion

Overall statistics of the RNAz screen

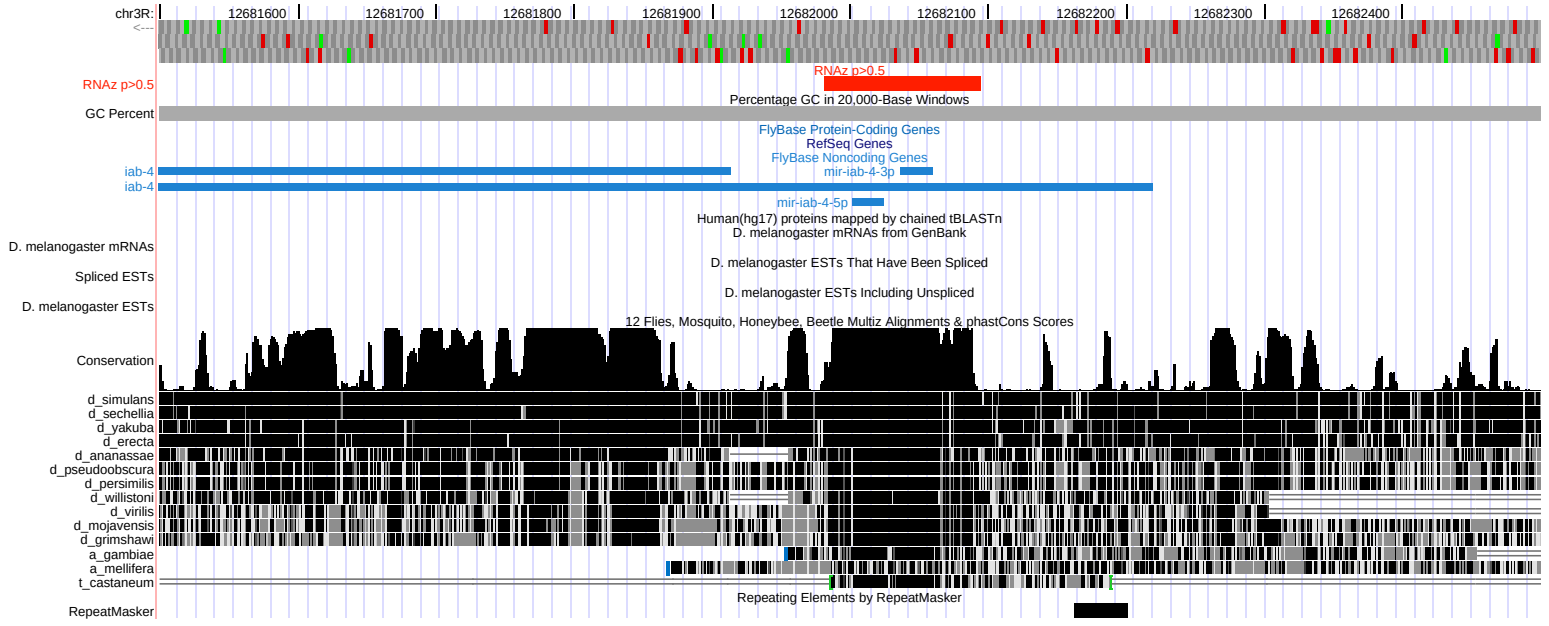
alignments	4077
aligned DNA [Mb]	117
screened by RNAz [Mb]	57.4
percentage	49
RNAz $p > 0.5$	42 482
[Kb]	5 079
RNAz $p > 0.9$	16 377
[Kb]	2 167
FDR $p > 0.5$ hits	56.5
sequence	52.8
FDR $p > 0.9$	45.3
sequence	40.2

Sensitivity on known ncRNAs

Class	RNAz	input	annotated	sensitivity (%)	
tRNA	171	250	297	69	
5S rRNA	0	0	99	—	not in input
SRP RNA	0	0	2	—	not in input
RNAse P	1	1	1	100	
snRNA	18	22	22	81	U6 not det.
snoRNA	96	202	250	48	
miRNA	75	78	85	96	

miRNAs

Bithorax 3R, 12681500-12682500 (ubx, abd-A, abd-B)
here: mir-iab–4-3p, mir-iab-4-5p



Evolutionary patterns of conservation

Blast comparison with our prior RNAz surveys (Mammals, Nematodes. Urochordates) reveals:

- 167 tRNAs
- 11 snRNAs
- 5 miRNAs
- New: U6atac

False discovery rates

Numbers of positive scored RNAz windows of the control screen:

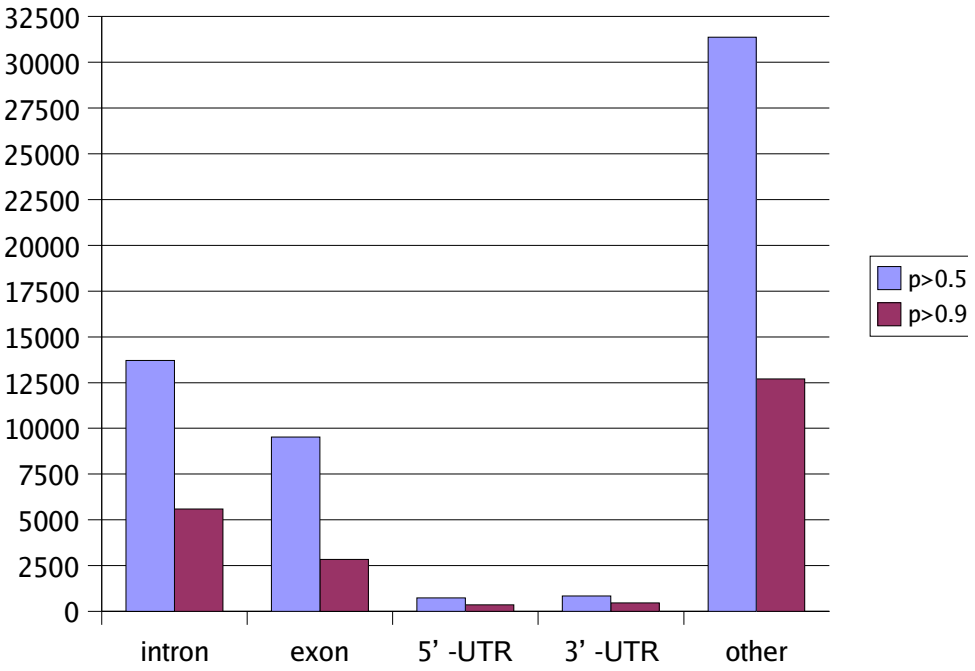
shuffling method	chromosomes						
	all	2L	2R	3L	3R	4	X
conservative	**29 938**	5 220	631	6 402	7 254	160	6 271
complete	**662**	123	99	132	155	1	152

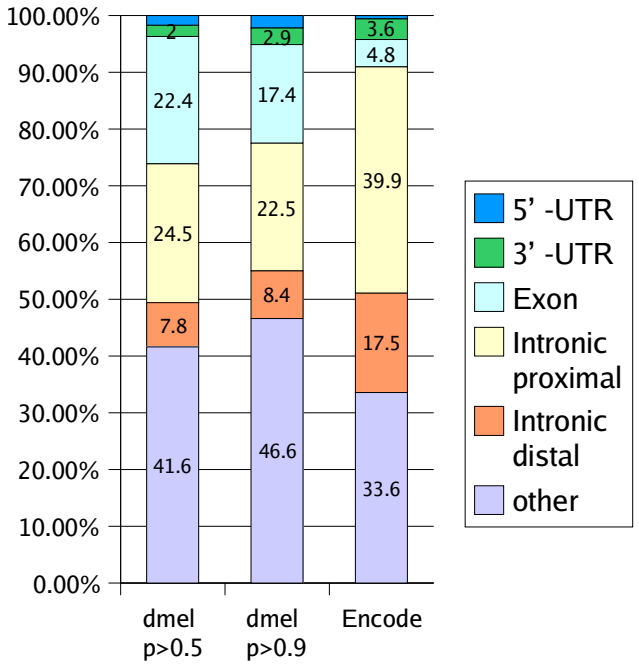
(68 562 windows overall)

Pessimistic estimates: **50 %** ($p > 0.5$), **40 %** ($p > 0.9$)
(preserving gap pattern and sequence conservation)
$\rightarrow$ intersection of true and control screen: **7.6 %**

Optimistic estimates : $\sim$**1-2 %**
(without preservation, ”complete” shuffling)

Genomic distribution

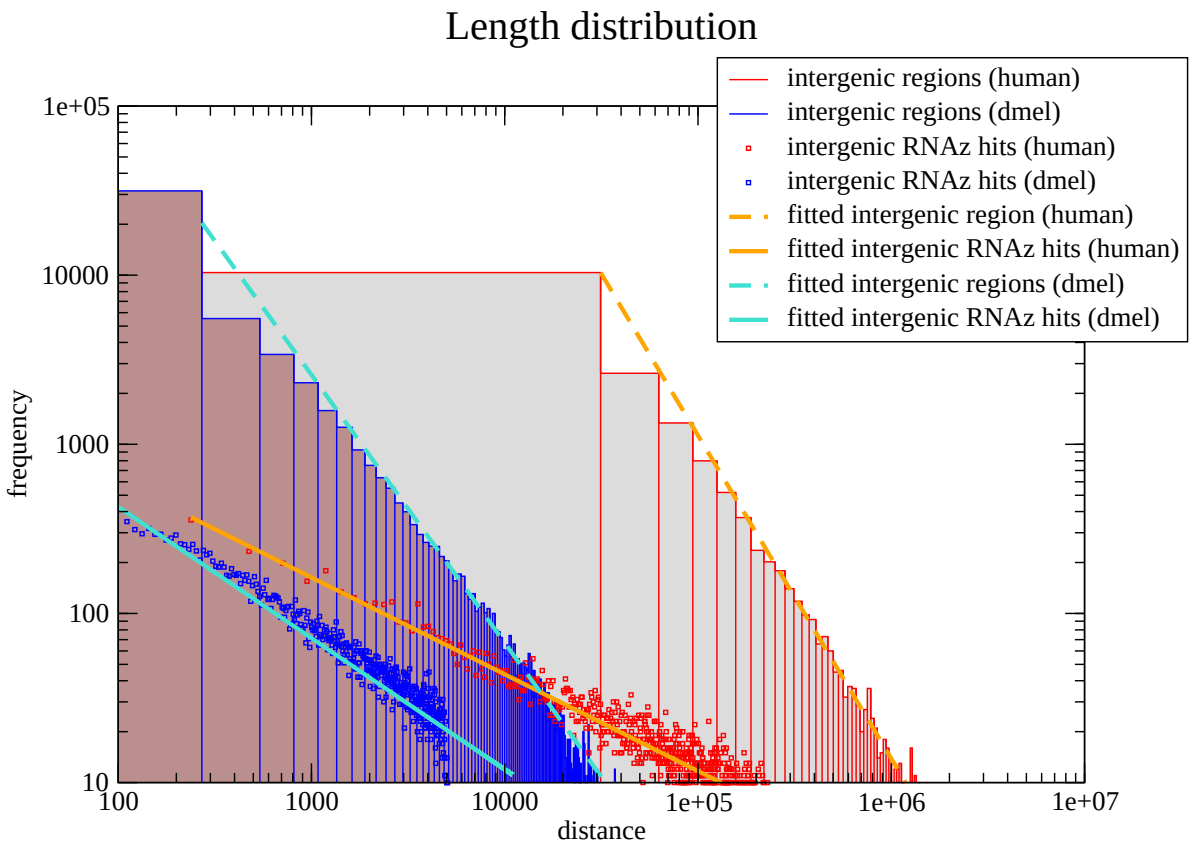




Distance boundaries:
"Distal": $>$ 5 kb away from next gene
"Proximal": $<=$ 5 kb

Genomic distribution

Comparison of the distribution of distances to nearest CDS of RNAz hits and intergenic regions in D. mel. and human.



Further Annotation

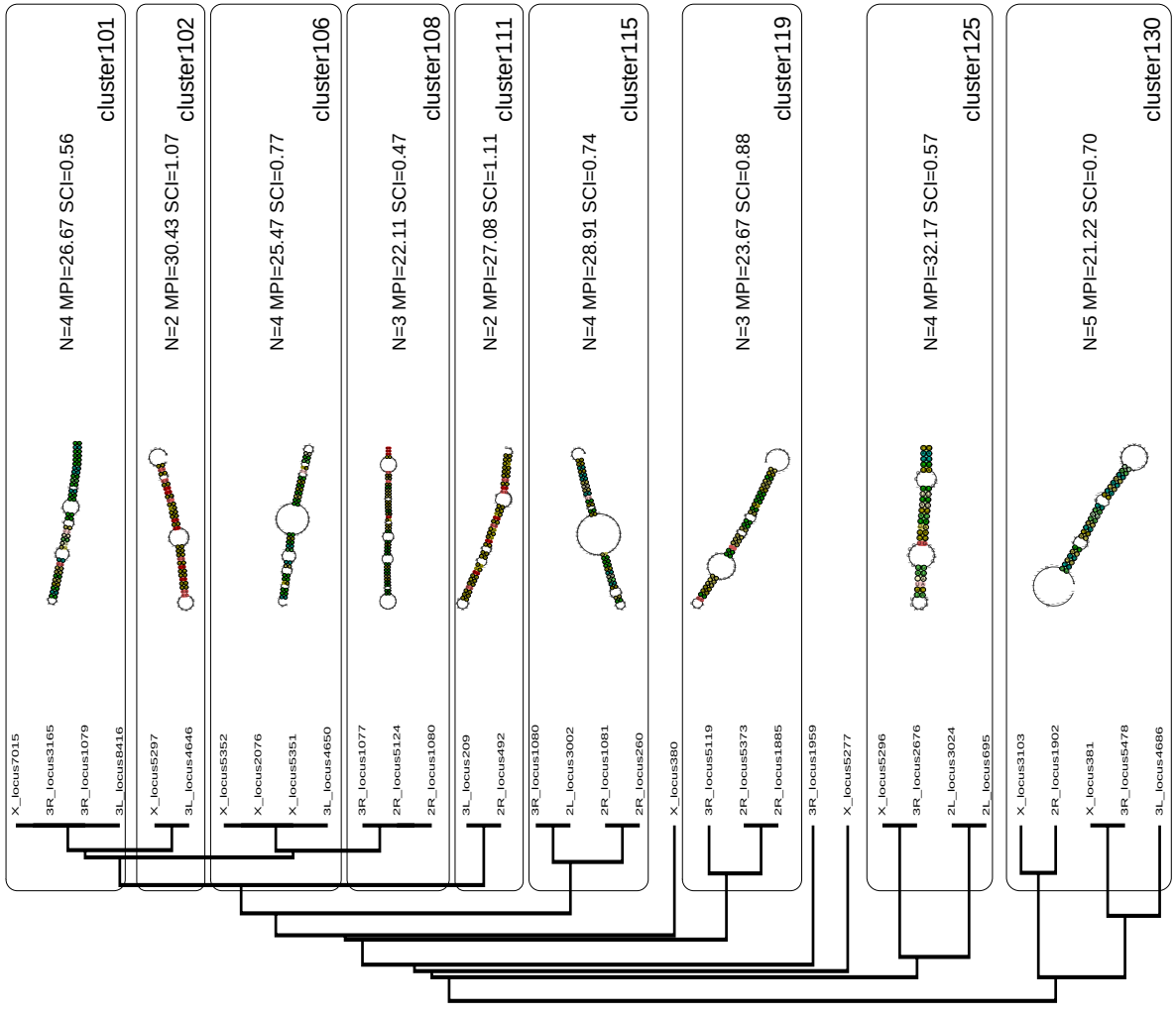
Isogai *et al.* (2006):

- Identification of TRF1/BRF binding sites using high-resolution tiling arrays
- Evidence that TRF1/BRF complex is responsible for initiation of known classes of Pol III transcription
- 3x enrichment of RNAz hits ($p > 0.9$) in these regions
- Mostly tRNAs, 7SL RNAs, and snoRNAs
- 197 unannotated RNAz hits $\rightarrow$ prime candidates for novel Pol-III transcripts

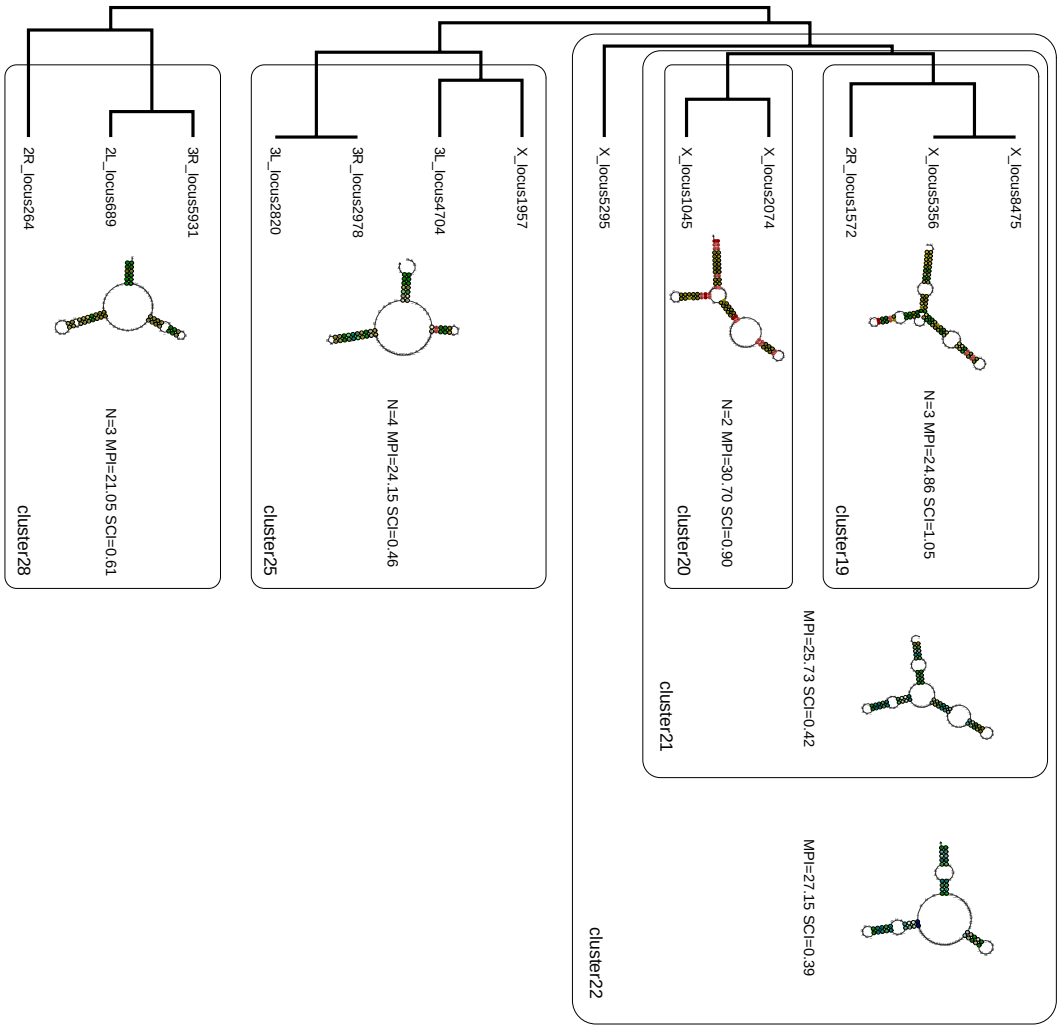
Moreover:

- 365 plausible miRNA predictions (RNAmicro)
- 1700 RNAz hits have direct evidence for expression through ESTs (not related to protein coding genes)

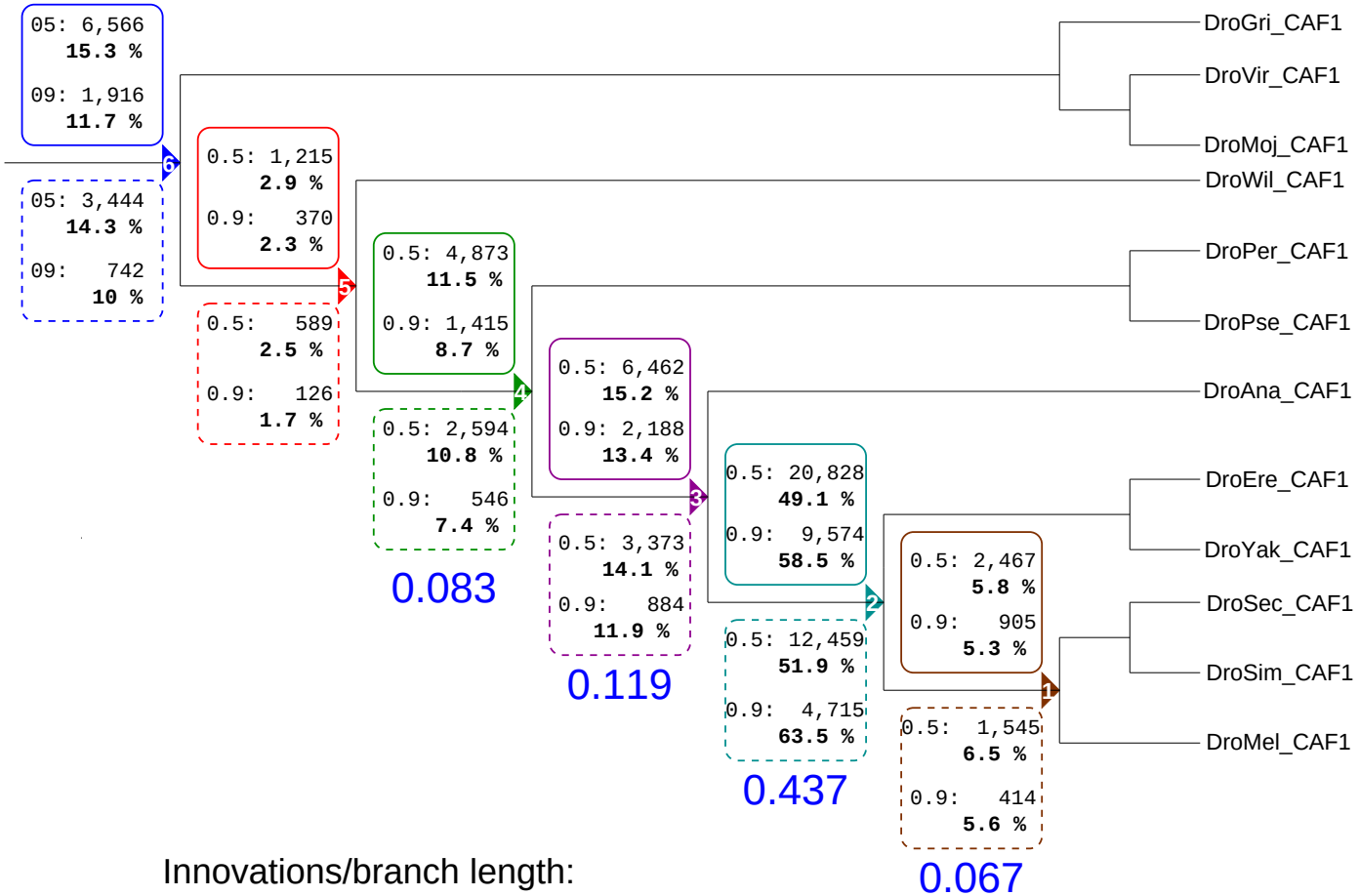
Structure-Based Clustering



Structure-Based Clustering



Phylogenetic Distribution



Innovations/branch length:

p>0.5	*1.39*	*1.27*	*1.12*	*0.87*
p>0.9	*1.05*	*1.12*	*1.33*	*0.79*

Outline

Introduction and Motivation

Results

Conclusion

Conclusion

- Submitted :-)
- High-quality ncRNA prediction for drosophilid species
- Insights into drosophilid genome organisation
- Reliable FDR?
- Very strong signals for true novel ncRNAs (Pol-III transcripts)

Thank you!!!

;-)

Jörg
Stefan
Kristin
Jana
Sven
Peter
Sonja