The Brave New World of Non-Coding RNAs

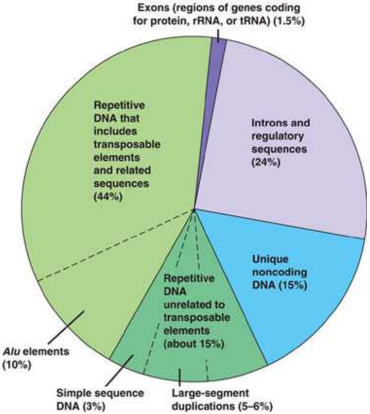
Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,
University of Leipzig
Max-Planck-Institute for Mathematics in the Sciences
RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
The Santa Fe Institute (external faculty)

Jena, Aug 2010

The Central Dogma

$$\text{DNA} \underbrace{\longrightarrow}_{\text{transcription}} \text{RNA} \underbrace{\longrightarrow}_{\text{translation}} \text{protein}$$



- only 3% of the non-repetitive part of genome codes for proteins
- Is all the rest **junk DNA**?
- Are all the repeats just genomic parasites?

Pervasive Transcription

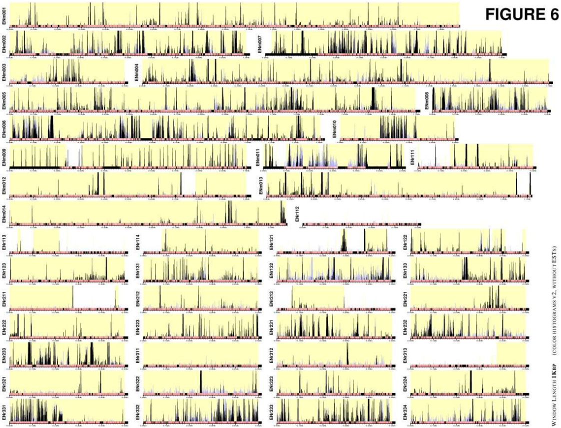
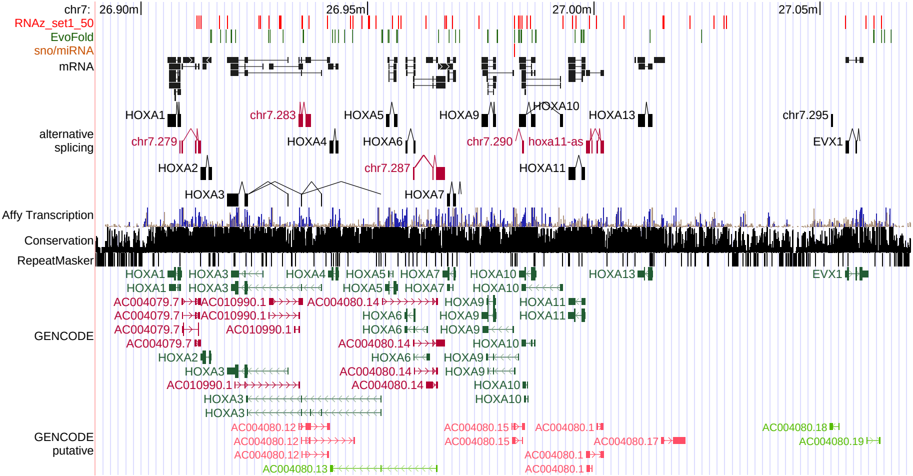


FIGURE 6

More than 90% of the non-repetitive genome shows evidence for transcription in at least one direction

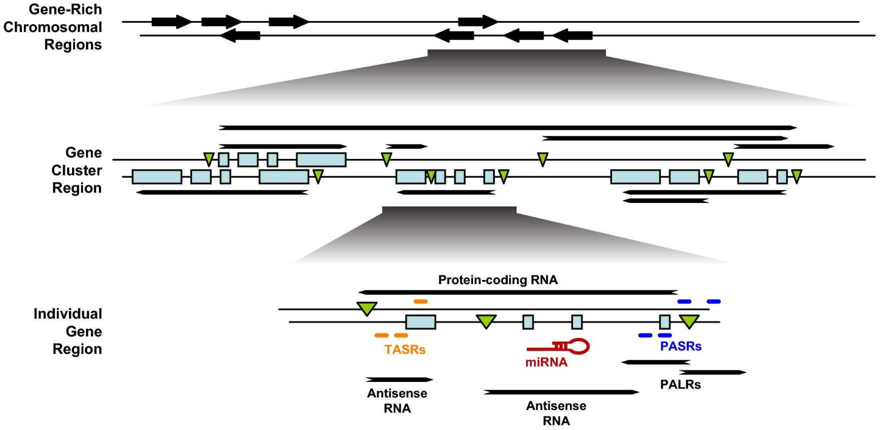
The ENCODE Consortium, *Nature*, 447: 779-816 (2007).

Transcriptome Complexity



Hox A cluster.

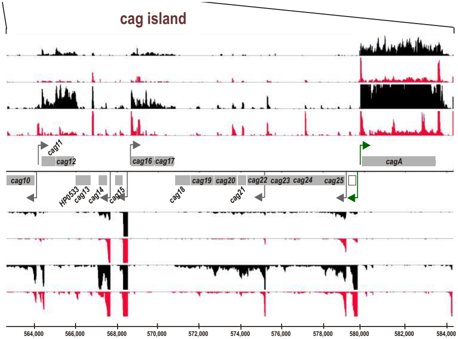
Transcriptome Complexity

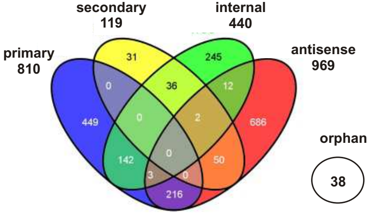


Science 316: 1484-1488 (2007)

H. phylori doesn’t read textbooks

mapping of transcription start sites in *Helicobacter pylori*
secondary start-sites and pervasive antisense transcription





Nature **464**: 250-255 (2010)

A New Paradigm of Molecular Biology!

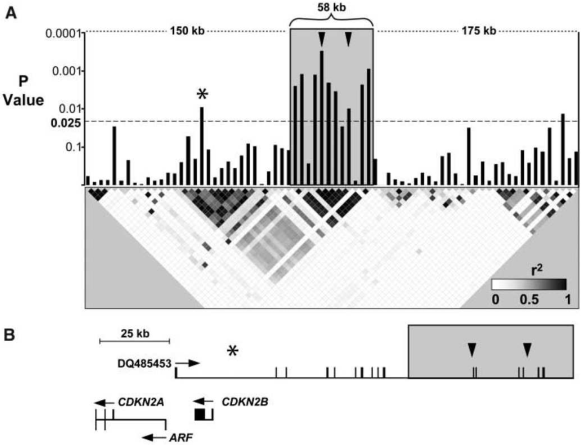
- There is no junk!
  Most of the human genome is transcribed, and there are good reasons to believe to most of the transcripts have function
- Most “genes” do not code for proteins
  We have to re-think — and maybe even abandon — the very notion of a gene
- Are these ncRNAs really functional????

Evidence for ncRNA function

- A small number of well-studied transcripts have functions identifyable by genetic methods (e.g. deletion/complementation)
- Statistical arguments:
  - differential regulation
  - Conservation at sequence level
  - Conservation of RNA structure
  - Conservation of splicing patterns
  - Association with (disease) phenotypes
  - Specific processing

CHD QTL Locus

The majority of QTLs for complex multi-genic diseases fall into non-coding regions



McPherson *et al.*, Science (2007)

Association of coronary heart disease (CHD) with a 58kb region on chr. 9p21

non-coding locus, produces the ANRIL transcript(s) ANRIL expression is associated with the atherosclerosis risk

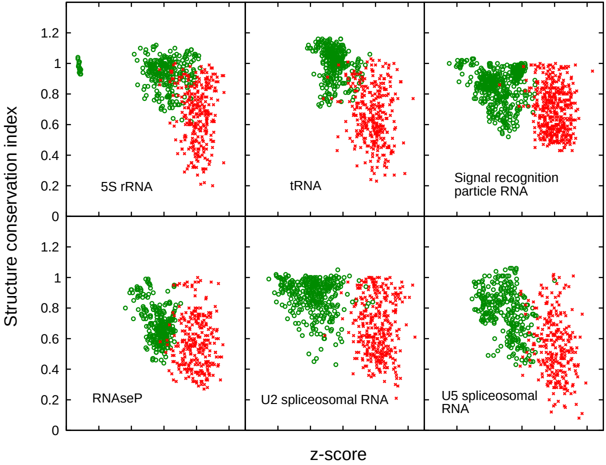
Holdt *et al.* *Arterioscler Thromb Vasc Biol.*, **30**, 620-627 (2010)

Computational RNA Gene Finding

- Many (but by no means all known functional RNAs are structured, i.e. certain base pairing patterns must be conserved
- This implies that substitutions are not random, but must be consistent with (GC→GU) or even compensate for base pairs (GC→AU)
- Empirical Observation: Known ncRNAs are (a little bit) more stable than genomic background with the same base composition.

IDEA: use this to build a gene finder

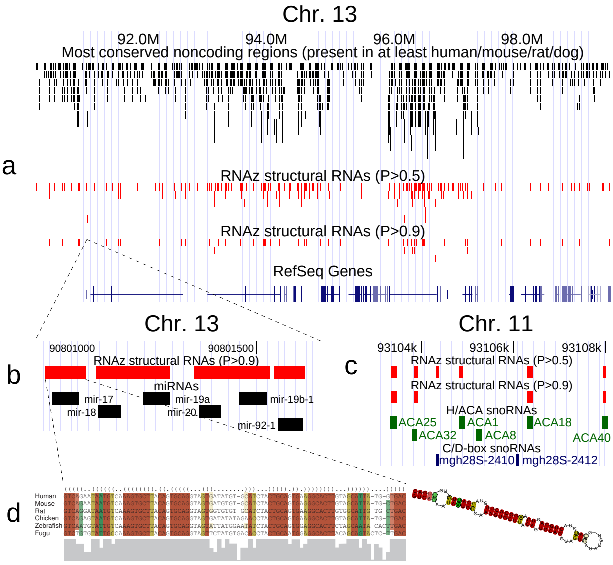
RNAz: a gene finder for structured RNA



Separation of native ncRNAs from random controls in two dimensions

Proc. Natl. Acad. Sci. USA **102**: 2454-2459 (2005)

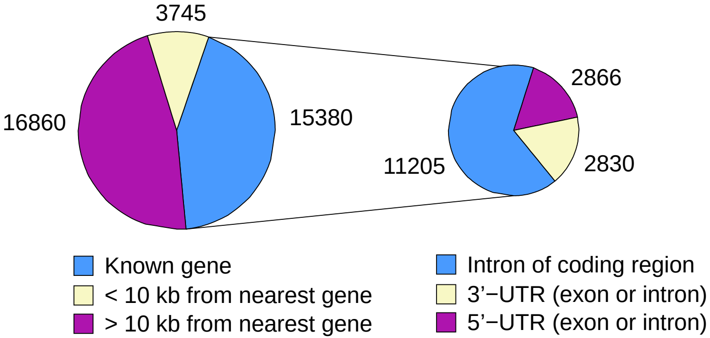
Structured RNAs in the Human Genome



Chr. 13
92.0M
94.0M
96.0M
98.0M
Most conserved noncoding regions (present in at least human/mouse/rat/dog)
a
RNAz structural RNAs (P>0.5)
RNAz structural RNAs (P>0.9)
RefSeq Genes
Chr. 13
b
90801000
90801500
RNAz structural RNAs (P>0.9)
miRNAs
mir-17
mir-18
mir-19a
mir-20
mir-19b-1
mir-92-1
Chr. 11
c
93104k
93106k
93108k
RNAz structural RNAs (P>0.5)
RNAz structural RNAs (P>0.9)
H/ACA snoRNAs
ACA25
ACA32
ACA1
ACA8
ACA18
ACA40
C/D-box snoRNAs
mgh28S-2410
mgh28S-2412
d
((((..((((((..((((((((.((.(((((...(((.........)))...))))).)).))))))))...))))))....)))))
Human GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATGT-GCATCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-GTGAC
Mouse GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATGTGT-GCATCTACTGCAGTGAGGGCACTTGTAGCATTA-TG-CTGAC
Rat GTCAGGATAATGTCAAAGTGCTTACAGTGCAGGTAGTGGTGTGT-GCATCTACTGCAGTGAAGGCACTTGTGGCATTG-TG-CTGAC
Chicken GTCAGAGTAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATATAGAACCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-TTGAC
Zebrafish GTCAATGTATTGTCAAAGTGCTTACAGTGCAGGTAGTATTATGGAATATCTACTGCAGTGGAGGCACTTCTAGCAATA-CACTTGAC
Fugu GTCTGTGTATTGCCAAAGTGCTTACAGTGCAGGTAGTTCTATGTGACACCTACTGCAATGGAGGCACTTACAGCAGTACTC-TTGAC

Structured RNAs in the Human Genome

Mammalian genomes contain $\sim 10^5$ structured RNA motifs
Statistics of the highest-confidence fraction ($\sim$ 36000):



Nature Biotech. **23** 1383-1390 (2005)

Finding mRNA-like ncRNAs

- long = contains at least one intron
- predict non-coding transcripts by predicting **conserved** **short** introns
- Why introns?

  • intron evolution is slow and essentially independent of the evolution of the mature sequence

  • splice sites are often conserved

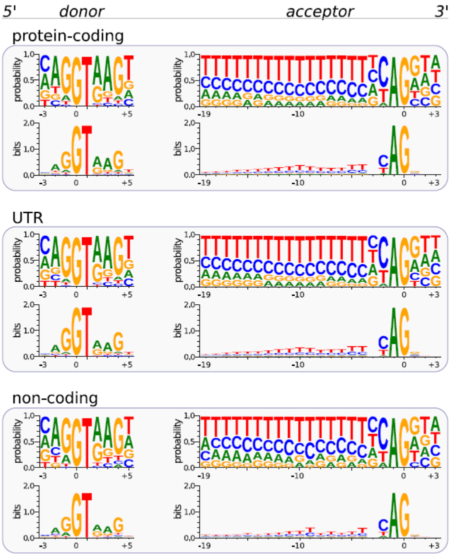
  • disruption of correct splicing usually destroys function

  ! non-coding transcrips do not have randomly placed large in/dels.
- Why short introns?

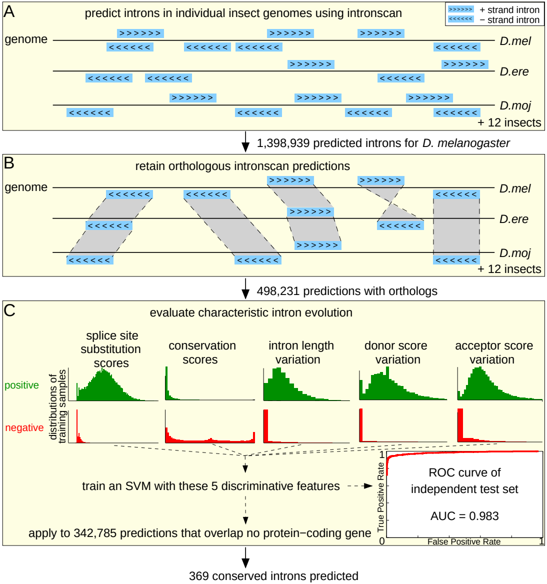
  • Most *Drosophila* introns are short.

  • Can be accurately predicted (94% with both splice sites correct)
- Intron prediction (Lim & Burge 1999): machine learning using patterns of donor, acceptor, intron length, branch point, intron composition

mlncRNAs – splice sites



Intron-prediction pipeline



A predict introns in individual insect genomes using intronscan

1,398,939 predicted introns for *D. melanogaster*

B retain orthologous intronscan predictions

498,231 predictions with orthologs

C evaluate characteristic intron evolution

train an SVM with these 5 discriminative features

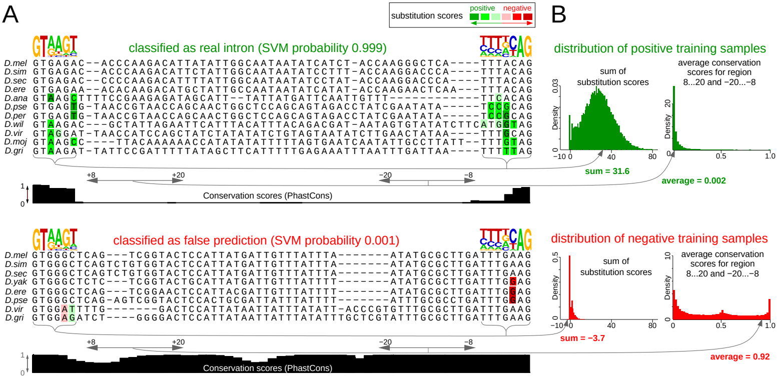
ROC curve of independent test set

AUC = 0.983

apply to 342,785 predictions that overlap no protein−coding gene

369 conserved introns predicted

Evaluation of Introns: Learn Species-Specific Patterns



classified as real intron (SVM probability 0.999)

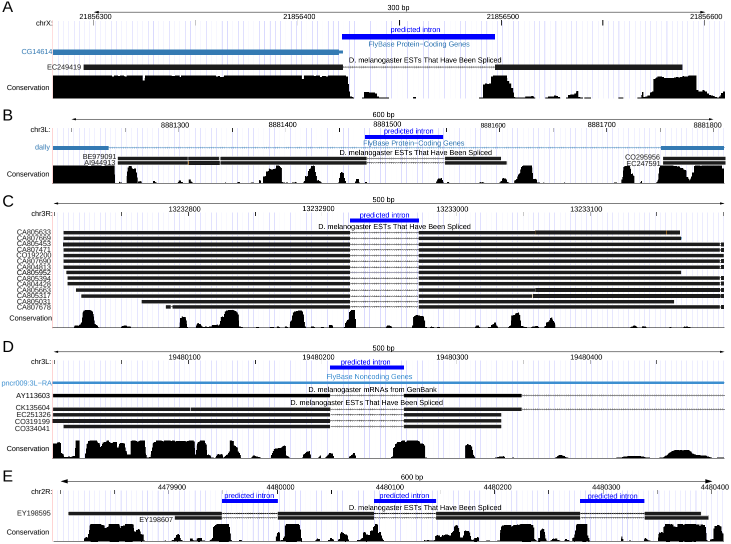
```
D.mel GTGAGAC--ACCCAAGACATTATATTGGCAATAATATCATCT-ACCAAGGGCTCA----TTTACAG
D.sim GTGAGAC--ACCCAAGACATTCTATTGGCAATAATATCCTTT-ACCAAGGACCCA----TTTACAG
D.sec GTGAGAC--CCCCAAGACATTTTATTGGCAATAATATCCTAT-ACCAAGGACCCA----TTTACAG
D.ere GTGAGAA--ACACAAGACATGCTATTGCCAATAATATCATAT-ACCAAGAACTCAA---TTTACAG
D.ana GTAAGCTTTTCCGAAGAGATAGCATT--TATTATGATTCAATTGTTT------------TTCACAG
D.pse GTGAGTG-TAACCGTAACCAGCAACTGGCTCCAGCAGTAGACCTATCGAATATA-----TCCGCAG
D.per GTGAGTG-TAACCGTAACCAGCAACTGGCTCCAGCAGTAGACCTATCGAATATA-----TCCGCAG
D.wil GTAAGAC---GCTATTAGAATTCATTTACATTTACAGACGAT-AATAGTGTATATCTTCATGGTAG
D.vir GTAGGAT-TAACCATCCAGCTATCTATATATCTGTAGTAATATCTTGAACTATAA----TTTGCAG
D.moj GTAAGCC----TTACAAAAAACCATATATATTTTTAGTGAATCAATATTGCCTTATT--TTTGTAG
D.gri GTAAGAT-TATTCCGATTTTTATAGCTTCATTTTTGAGAAATTTAATTTGATTAA----TTTTTAG
```

classified as false prediction (SVM probability 0.001)

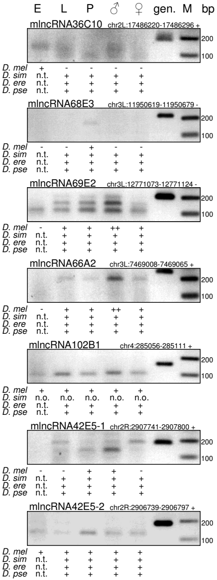
```
D.mel GTGGGCTCAG---TCGGTACTCCATTATGATTGTTTATTTA-------ATATGCGCTTGATTTGAAG
D.sim GTGGGCTCAGTCTGTGGTACTCCATTATGATTGTTTATTTA-------ATATGCGCTTGATTTGAAG
D.sec GTGGGCTCAGTCTGTGGTACTCCATTATGATTGTTTATTTA-------ATATGCGCTTGATTTGAAG
D.yak GTGGGCTCTC---TCGGTACTGCATTATGATTGTTTATTTT-------ATATGCGCTTGATTTGGAG
D.ere GTGGGCTCAG---TCGGAACTCCATTACGATTGTTTATTTT-------ATATGCGCTTGATTTGGAG
D.pse GTGGGCTCAG-AGTCGGTACTCCACTGCGATTATTTATTTT-------ATTTGCGCCTGATTTGGAG
D.vir GTGGATTTTG-------GACTCCATTATAATTATTTATATT-ACCCGTGTTTGCGCTTGATTTGAAG
D.gri GTGGAGATCT----GGGGACTCCATTATAATTATTTATATTTGCTCGTATTTGCGCTTGATTTGAAG
```

A
B
substitution scores
positive
negative
GTAAGT
TTTGCAG
+8
+20
−20
−8
1
0
Conservation scores (PhastCons)
distribution of positive training samples
sum of substitution scores
average conservation scores for region 8...20 and −20...−8
Density
0.03
0
−10
0
40
80
20
0.5
1.0
sum = 31.6
average = 0.002
distribution of negative training samples
0.5
10
sum = −3.7
average = 0.92

Validation with un-annotated ESTs



Novel conserved ncRNAs in *Drosophila*



11 out of 17 predictions verified by PCR and sequencing

Expression of transcripts

and existance of introns

also verified in 3 other fly species

Embryo

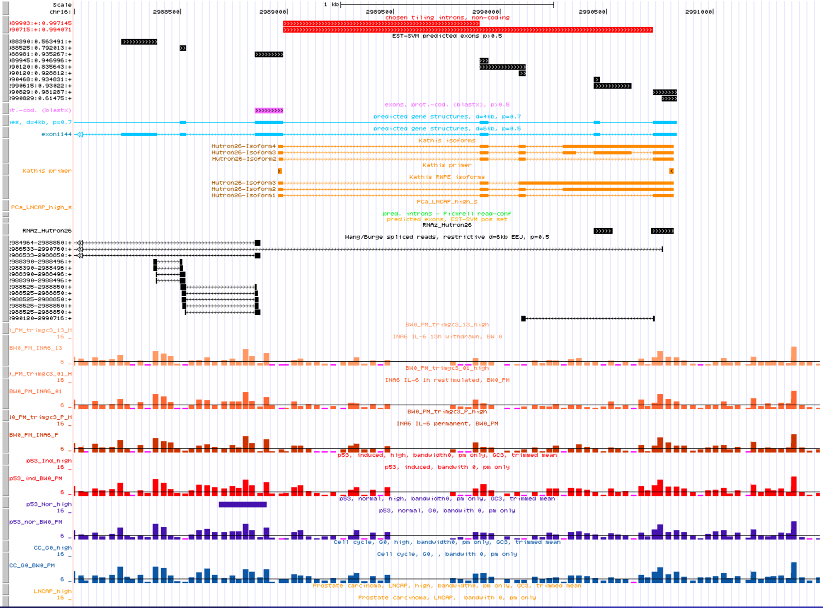
Larva

Pupa

male

female

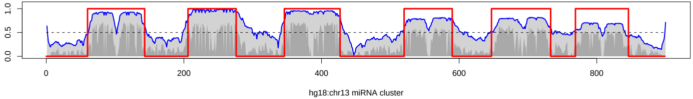
Extension to Mammals



Generation of small RNAs

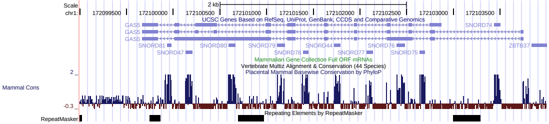
Many types of small RNAs are produced from longer precursors. In many cases, these precursors are mRNA-like pol-II transcripts.

1. primary microRNA precursors.
   miRNAs a processed out of either exons or introns



primary precursor of the *mir-17* cluster

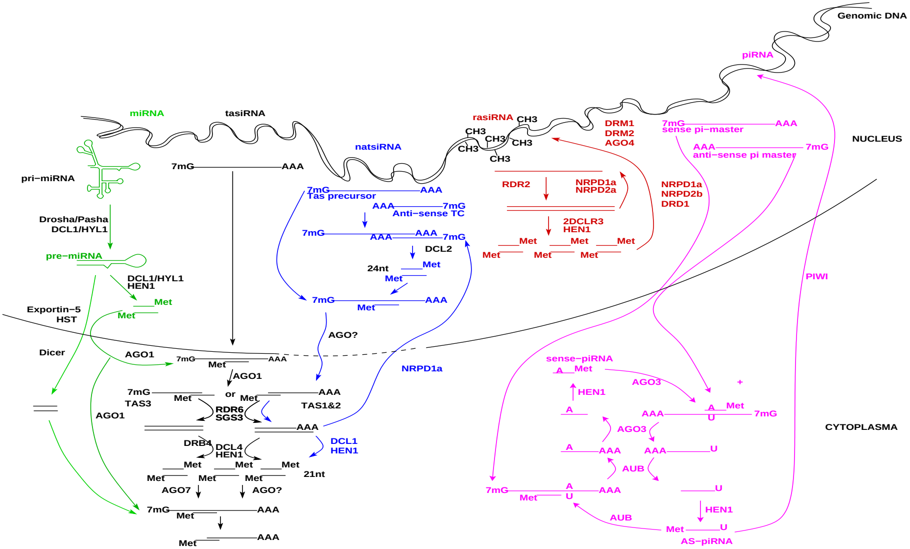
2. snoRNA host genes
   vertebrate snoRNAs are produced from introns


Scale
2 kb
chr1:
172099500
172100000
172100500
172101000
172101500
172102000
172102500
172103000
172103500
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
GAS5
GAS5
GAS5
SNORD74
SNORD81
SNORD80
SNORD79
SNORD44
SNORD76
ZBTB37
SNORD47
SNORD78
SNORD77
SNORD75
Mammalian Gene Collection Full ORF mRNAs
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Basewise Conservation by PhyloP
2
Mammal Cons
-0.3
Repeating Elements by RepeatMasker
RepeatMasker

3. some piRNA precursors
4. Affymetrix high-density arrays showed that at least 1% of the human genome produces small RNAs

   Science **316**: 1484-1488 (2007) (joint work with Affymetrix)

microRNAs and their relatives

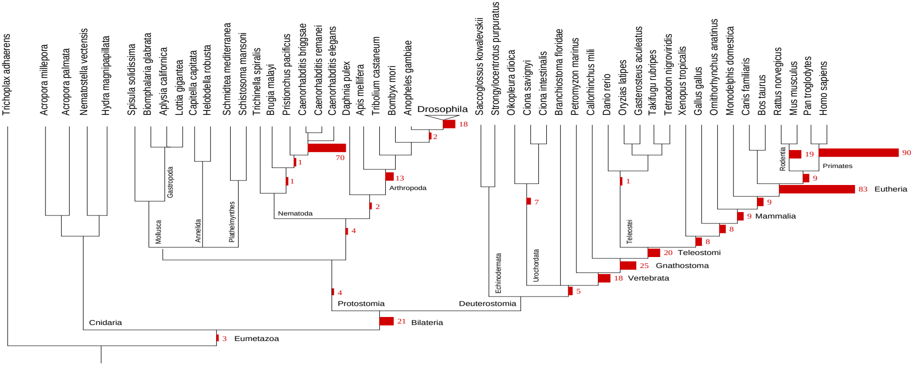


MicroRNAs: Innovation

- Most protein-coding genes are evolutionarily old. For instance, there are no or very few new transcription factor that were invented throughout vertebrate evolution
- Small RNAs, in particular microRNAs, however, are readily created *de novo*.
- Is there a link between ncRNA innovation and novelty at the organismal level?

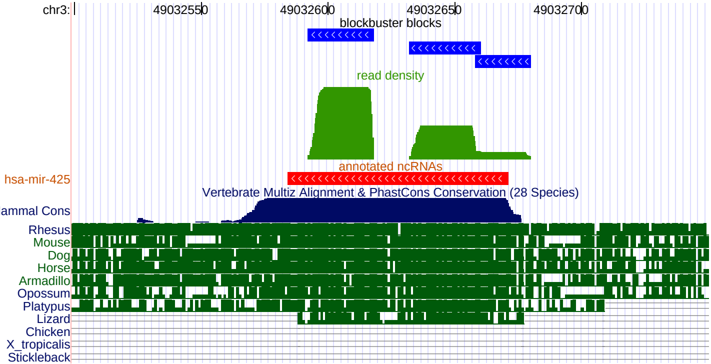
MicroRNAs: Innovation

Expansion of the Metazoan microRNA Repertoire



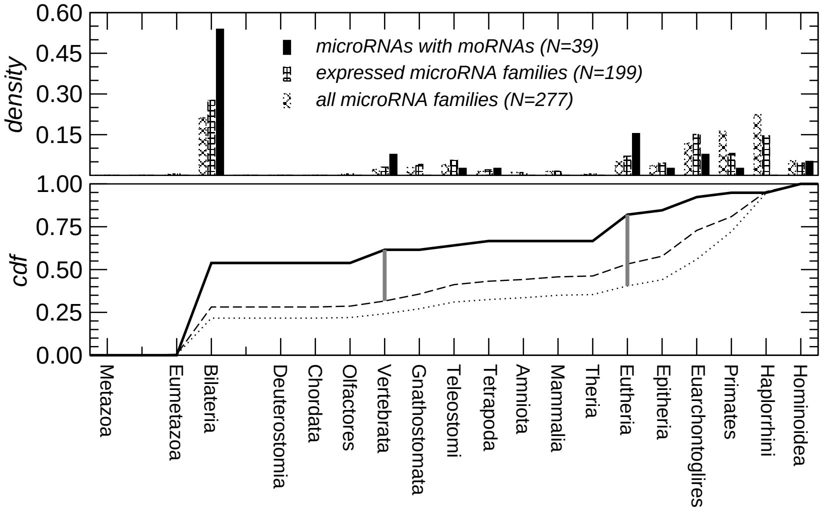
BMC Genomics **7**: 15 (2006), updated

MicroRNA Offset RNAs



Distribution of short reads at the *hsa-mir-425* locus. There are three clearly distinct blocks of reads: the two more abundant ones correspond to miR and miR*, the third one to the 5'moRNA.

Offset RNAs are associated with old microRNAs

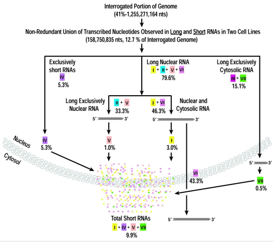


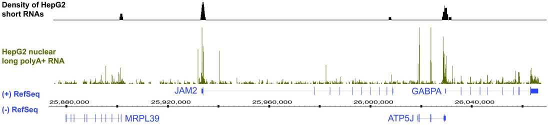
Bioinformatics **25**: 2298-2301 (2009)

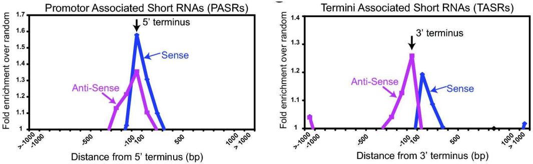
Promoter and Termini Associated RNAs

Examples of novel families of small RNAs

High-density tiling array screen of human small RNAs


Interrogated Portion of Genome
(41%-1,255,271,164 nts)
Non-Redundant Union of Transcribed Nucleotides Observed in Long and Short RNAs in Two Cell Lines
(158,750,835 nts, 12.7 % of Interrogated Genome)
Exclusively short RNAs
IV
5.3%
Long Nuclear RNA
I + II + V + VI
79.6%
Long Exclusively Cytosolic RNA
III + VII
15.1%
Long Exclusively Nuclear RNA
II + V
33.3%
I + VI
46.3%
Nuclear and Cytosolic RNA
5′
3′
IV
5.3%
V
1.0%
I
3.0%
Nucleus
Cytosol
VI
43.3%
VII
0.5%
Total Short RNAs
I + IV + V + VII
9.9%

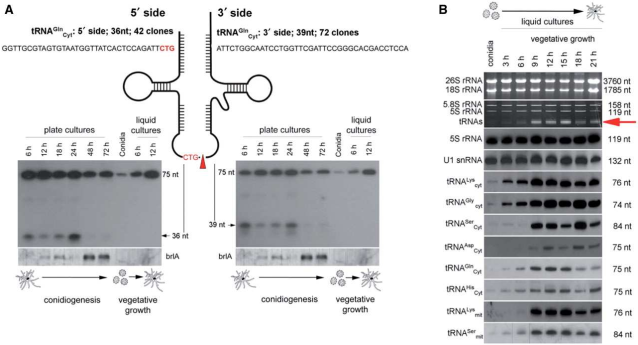




More than 1% of the human genome is transcribed into small RNAs.

Science **316**: 1484-1488 (2007) (joint work with Affymetrix)

Regulated specific cleavage of tRNAs

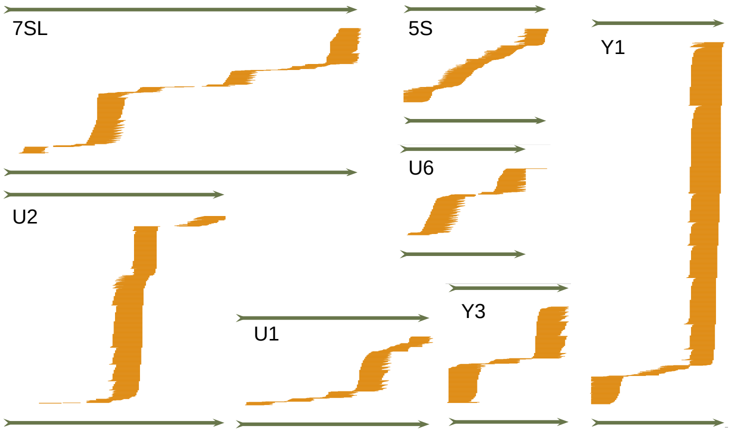


in *Aspergillus fumigatus*

Nucleic Acids Res. **36**, 2677-2689 (2008)

recently discovered by several groups also in mammals

Specific pattern of small RNAs



Block patterns are source specific

A first attempt:

random forrest classificator to distinguish the block patterns of microRNAs, the two snoRNA classes, and tRNAs.

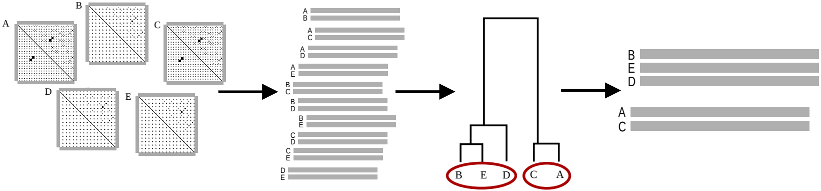
Confusion matrix: (10-fold crossvalidation)

	classified as				
class	miRNA	H/ACA	C/D	tRNA	other
miRNA	249	2	6	8	21
H/ACA	6	8	5	2	4
C/D	20	3	82	13	22
tRNA	7	0	12	310	41
other	25	4	16	56	312

New RNA Classes from Structural Clustering

Comparison of RNA secondary structures:

- Structure-enhanced alignments (e.g. `stral`)
- Tree-alignment or Tree-editing (e.g. `RNAforrester`, `MARNA`)
- `RNAshapes`
- Variants of the *Sankoff* algorithm

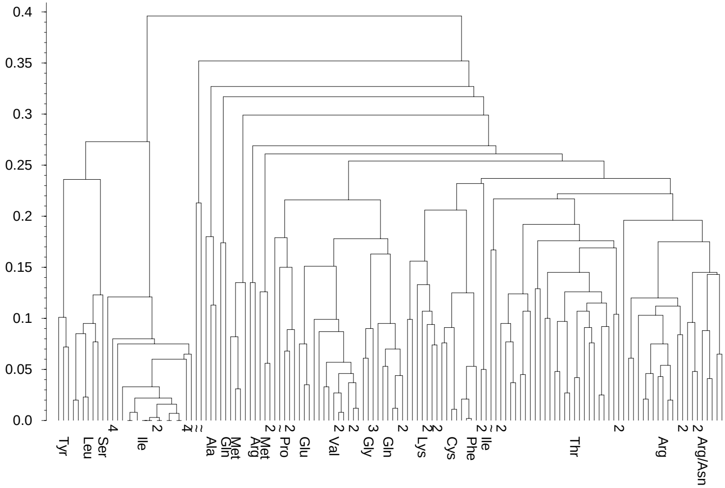


`locarna`: a Sankoff-based local structure alignment tool

Trick: use thermodynamically most plausible base-pairs only

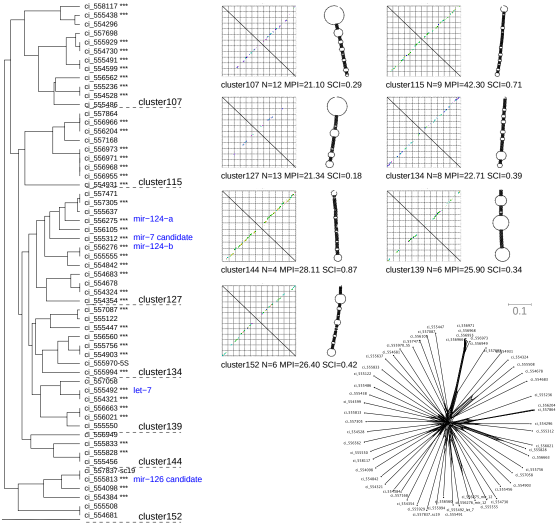
Joint work with Rolf Backofen's group

Clustering *Ciona intestinalis* RNAz Predictions



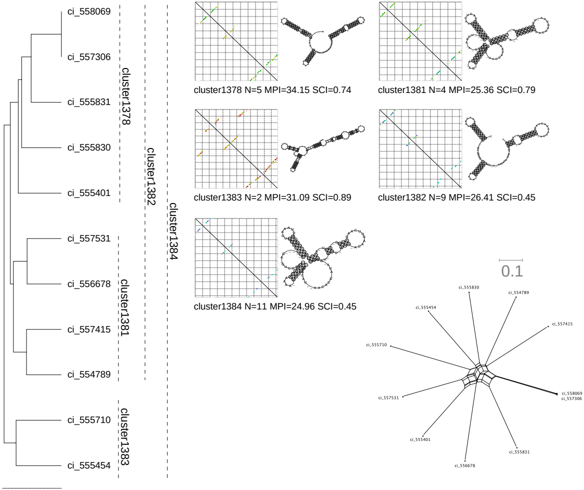
tRNAs subtree from a clustering 3332 ncRNA candidates

Putative Novel RNA Classes



Ciona intestinalis: microRNA subtree

Putative Novel RNA Classes



cluster1378 N=5 MPI=34.15 SCI=0.74

cluster1381 N=4 MPI=25.36 SCI=0.79

cluster1383 N=2 MPI=31.09 SCI=0.89

cluster1382 N=9 MPI=26.41 SCI=0.45

cluster1384 N=11 MPI=24.96 SCI=0.45

PLoS Comp. Biol. **3**: e65 (2007)

Many, many thanks . . .

- Leipzig: Sonja J. Prohaska, Dominic Rose, Jana Hertel, Manja Marz, Claudia & Roman Stocsits, Sven Findeiß, . . .
  FH RNomics group: Jörg Hackermüller, Antje Kretzschmar, Kristin Reiche, Kathy Schutt, Kerstin Ullmann, Tine Schulz
- Vienna: Stefan Washietl, Ivo L. Hofacker, Christoph Flamm, Andrea Tanzer, Stefan Bernhart, Hakim Tafer, Susanne Rauscher, Caroline Thurner, Christina Witwer, . . .
- Halle: Günter Reuter's Lab
- Freiburg: Rolf Backofen, Sebastian Will
- Tübingen: Kay Nieselt's Group
- Freiburg: Rolf Backofen's Lab
- Würzburg: Jörg Vogel, Cynthia Sharma
- Copenhagen: Jan Gorodkin, Stefan Seemann, Peter Menzel
- Affymetrix: Tom Gingeras, Phil Kapranov, *et al.*
- CAS Beijing: Wei Deng and all the others in Runsheng Chen's Lab
- PICB Shanghai: Axel Mosig and Phil Khaitovich and their students (PICB/SIBS)
- ASU Tempe: Julian L. Chen and his lab
- Stanford: Michael Hiller
- ENCODE: Ewan Birney and $10^{2.5}$ coauthors