The majority of long non-coding RNAs is conserved

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)
Center for non-coding RNA in Technology and Health, U. Copenhagen

The Santa Fe Institute (external faculty)

joint work with Anne Nitsche, Dominic Rose, Mario Fasold

Bled, Feb 13 2012

mRNA-like ncRNAs

over the last few years ncRNAs that otherwise look quite similar to mRNAs have become a major research topic
(using, as usual, a variety of acronyms) mlncRNA, lincRNAs,

- How well conserved are lncRNAs?
  Two answers:
  1. “relatively low degree of sequence constraint” (Marques & Ponting 2009)
  2. but ... some very well-conserved examples (Chodroff *et al.* 2010, ...)

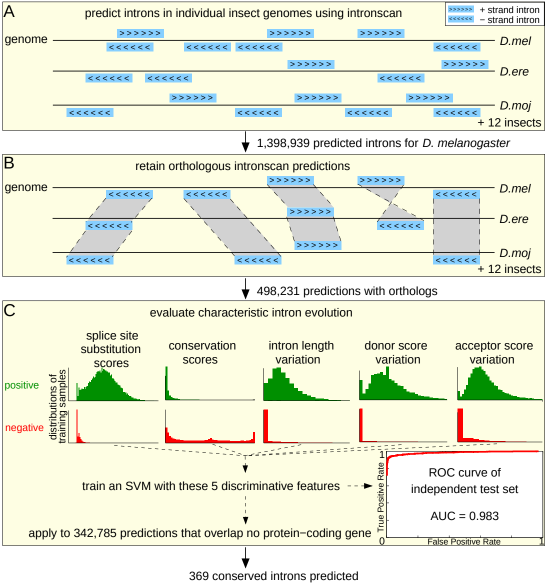
One problem: sequence conservation does not necessarily imply conservation of the ncRNA!

De novo Prediction of mRNA-like ncRNAs

- long ncRNA = 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

Intron-prediction pipeline



predict introns in individual insect genomes using intronscan

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

retain orthologous intronscan predictions

498,231 predictions with orthologs

evaluate characteristic intron evolution

train an SVM with these 5 discriminative features

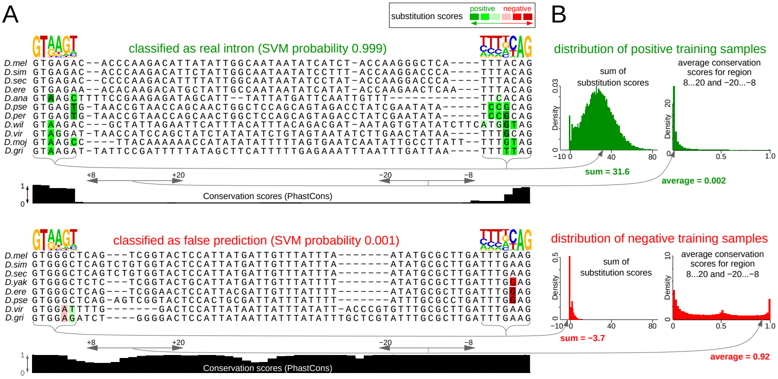
apply to 342,785 predictions that overlap no protein−coding gene

ROC curve of independent test set

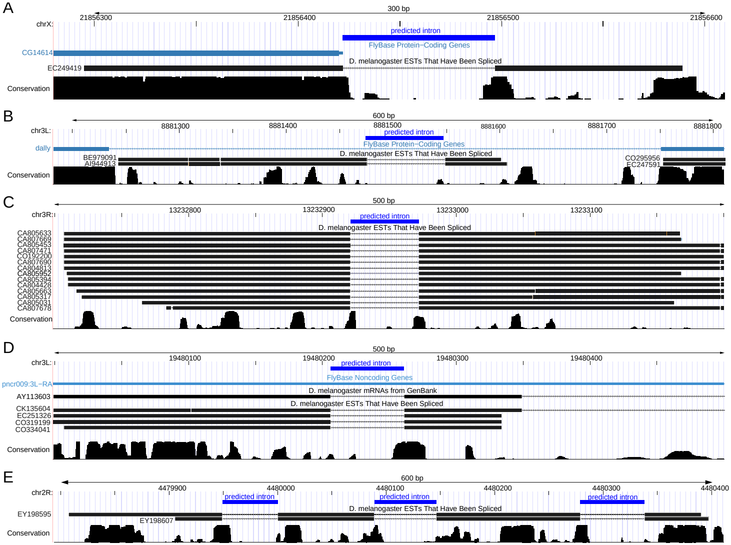
AUC = 0.983

369 conserved introns predicted

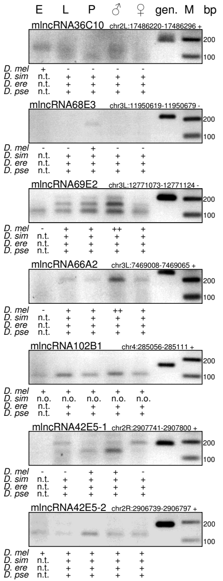
Evaluation of Introns: Species-Specific Patterns



Validation with un-annotated ESTs



Novel conserved ncRNAs in *Drosophila*



mlncRNA36C10 chr2L:17486220-17486296 +

	E	L	P	♂	♀
D. mel	+	-	-	-	-
D. sim	n.t.	+	+	+	+
D. ere	n.t.	+	+	+	+
D. pse	n.t.	+	+	+	+

mlncRNA68E3 chr3L:11950619-11950679 -

	E	L	P	♂	♀
D. mel	-	-	+	-	-
D. sim	n.t.	+	+	+	+
D. ere	n.t.	+	+	+	+
D. pse	n.t.	+	+	+	+

mlncRNA69E2 chr3L:12771073-12771124 -

	E	L	P	♂	♀
D. mel	-	+	+	++	+
D. sim	n.t.	+	+	+	+
D. ere	n.t.	+	+	+	+
D. pse	n.t.	+	+	+	+

mlncRNA66A2 chr3L:7469008-7469065 +

	E	L	P	♂	♀
D. mel	-	+	+	++	+
D. sim	n.t.	+	+	+	+
D. ere	n.t.	+	+	+	+
D. pse	n.t.	+	+	+	+

mlncRNA102B1 chr4:285056-285111 +

	E	L	P	♂	♀
D. mel	+	+	+	+	+
D. sim	n.o.	n.o.	n.o.	n.o.	n.o.
D. ere	n.t.	+	+	+	+
D. pse	n.t.	+	+	+	+

mlncRNA42E5-1 chr2R:2907741-2907800 +

	E	L	P	♂	♀
D. mel	-	-	+	+	-
D. sim	n.t.	+	+	+	+
D. ere	n.t.	+	+	+	+
D. pse	n.t.	+	+	+	+

mlncRNA42E5-2 chr2R:2906739-2906797 +

	E	L	P	♂	♀
D. mel	+	+	+	+	+
D. sim	n.t.	+	+	+	+
D. ere	n.t.	+	+	+	+
D. pse	n.t.	+	+	+	+

11 out of 17 predictions verified by PCR and sequencing

Expression of transcripts

and existence of introns

also verified in 3 other fly species

Embryo

Larva

Pupa

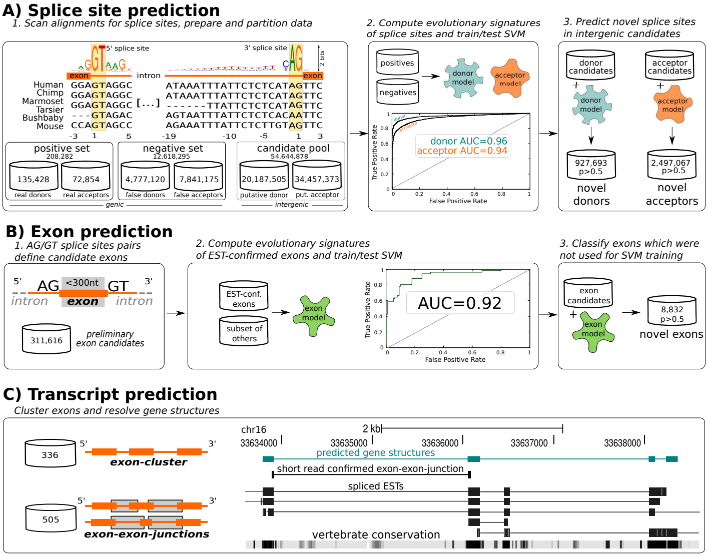
male

female

Functional mlncRNA in Vertebrates

- vertebrate introns $\neq$ insect introns
  (2% vs 54% short introns)
- predict single individual splice-sites
  (instead of introns)
- splice-site prediction ! $=$ intron prediction
  vertebrate exons are still short, so let's predict exons
  (novel pipeline, new SVM features, re-implementation)
  downside: single-intron genes are not visible

Conserved mlncRNA in Vertebrates



A) Splice site prediction

1. Scan alignments for splice sites, prepare and partition data

2. Compute evolutionary signatures of splice sites and train/test SVM

3. Predict novel splice sites in intergenic candidates

B) Exon prediction

1. AG/GT splice sites pairs define candidate exons

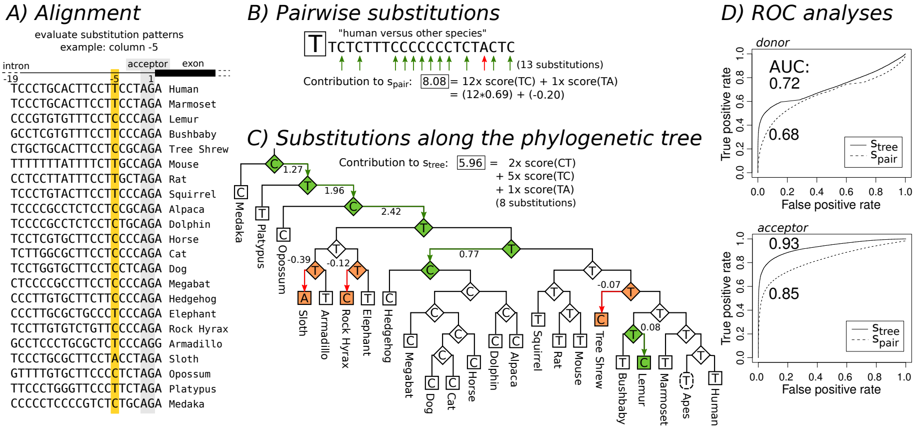
2. Compute evolutionary signatures of EST-confirmed exons and train/test SVM

3. Classify exons which were not used for SVM training

C) Transcript prediction

Cluster exons and resolve gene structures

Conserved mlncRNA in Vertebrates



A) Alignment

evaluate substitution patterns
example: column -5

```
intron                acceptor  exon
-19            -5         1
TCCCTGCACTTCCTTCCTAGA Human
TCCCTGCACTTCCTTCCTAGA Marmoset
CCCGTGTGTTTCCTCCCCAGA Lemur
GCCTCGTGTTTCCTTCCCAGA Bushbaby
CTGCTGCACTTCCTCCGCAGA Tree Shrew
TTTTTTTATTTTCTTGCCAGA Mouse
CCTCCTTATTTCCTTGCTAGA Rat
TCCCTGTACTTCCTTCCCAGA Squirrel
TCCCCGCCTCTCCTCCGCAGA Alpaca
TCCCCGCCTCTCCTCTGCAGA Dolphin
TCCTCGTGCTTCCTCCCCAGA Horse
TCTTGGCGCTTCCTCCCCAGA Cat
TCCTGGTGCTTCCTCCTCAGA Dog
CTCCCCGCCTTCCTCCCCAGA Megabat
CCCTTGTGCTTCTTCCCCAGA Hedgehog
CCCTTGCGCTGCCCTCCCAGA Elephant
TCCTTGTGTCTGTTCCCCAGA Rock Hyrax
GCCTCCCTGCGCTCTCCCAGG Armadillo
TCCCTGCGCTTCCTACCTAGA Sloth
GTTTTGTGCTTCCCCTCTAGA Opossum
TTCCCTGGGTTCCCTTCTAGA Platypus
CCCCCTCCCCGTCTCTGCAGA Medaka
```

B) Pairwise substitutions

"human versus other species"

T TCTCTTTCCCCCCCTCTACTC

(13 substitutions)

Contribution to s_{pair}: $\boxed{8.08}$ = 12x score(TC) + 1x score(TA) = (12∗0.69) + (-0.20)

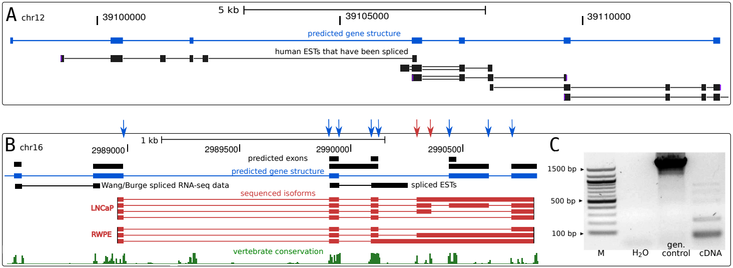
C) Substitutions along the phylogenetic tree

Contribution to s_{tree}: $\boxed{5.96}$ = 2x score(CT) + 5x score(TC) + 1x score(TA) (8 substitutions)

D) ROC analyses

Conserved mlncRNA in Vertebrates

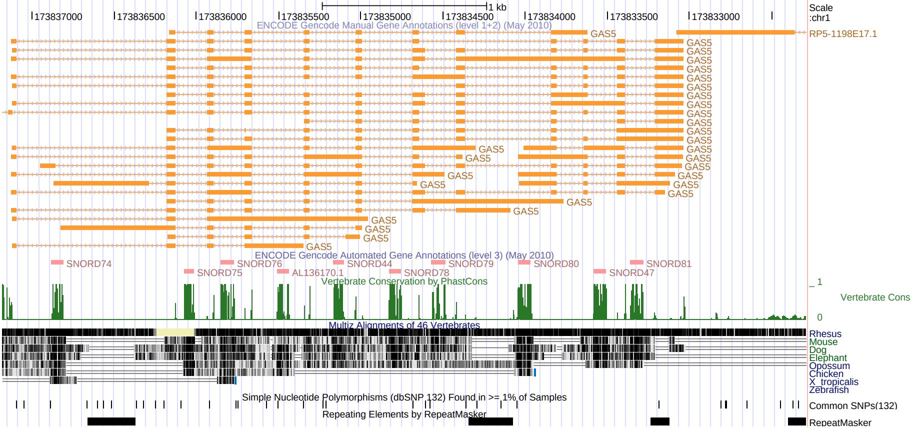
two validated examples



GAS5: A highly conserved host gene

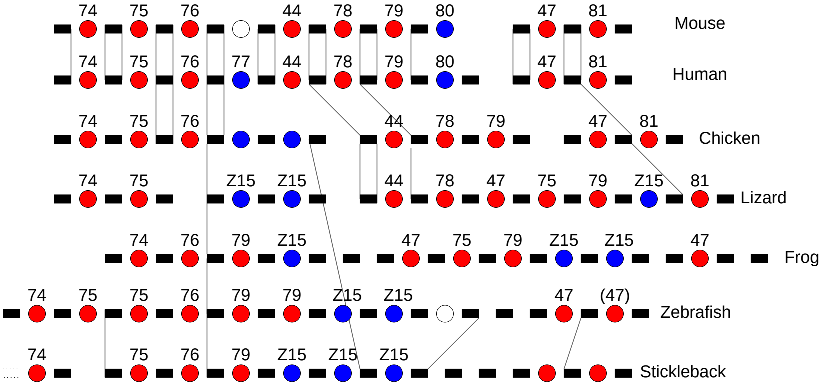
- most famous snoRNA host gene, 10 different snoRNAs are the payload in its introns
- The exonic part (“mRNA”) sequesters and inhibits the glucocorticoid receptor (Kino, Sci. Signaling 2010)
- conserved at least in gnathostomes
- very rapid evolution of exonic sequence, snoRNAs well conserved
- major changes in gene structure

Human GAS5 – a complex locus



Scale
:chr1
1 kb
173837000
173836500
173836000
173835500
173835000
173834500
173834000
173833500
173833000
ENCODE Gencode Manual Gene Annotations (level 1+2) (May 2010)
GAS5
RP5-1198E17.1
GAS5
ENCODE Gencode Automated Gene Annotations (level 3) (May 2010)
SNORD74
SNORD76
SNORD44
SNORD79
SNORD80
SNORD81
SNORD75
AL136170.1
SNORD78
SNORD47
Vertebrate Conservation by PhastCons
_ 1
Vertebrate Cons
0
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples
Common SNPs(132)
Repeating Elements by RepeatMasker
RepeatMasker

Evolution of GAS5



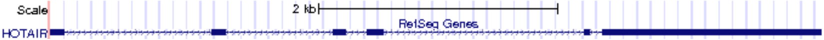
Two superimposed effects

- changes in the structure of the host gene itself
  gain & loss of splice sites
- snoRNAs can be behave like mobile elements

Evolution of mlncRNAs: HOTAIR

- transcribed from the HOXC cluster in antisense direction from the HoxC12-HoxC11 intergenic region
- directs PRC2 to the HOXD locus, silencing HoxD11-HoxD8. [Rinn et al 2007, Tsai et al 2010]
- however, the mouse homolog does not have this function [Schorderet & Duboule 2011]

Evolution of mlncRNAs: HOTAIR



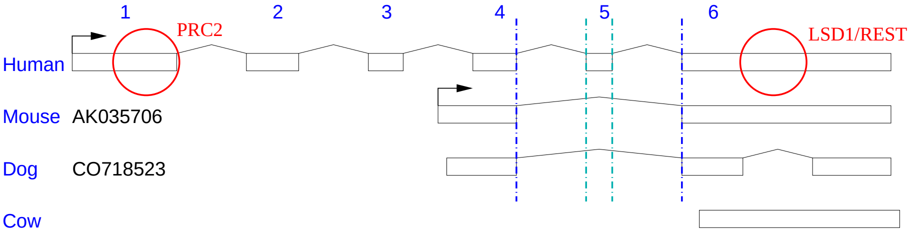
	Exon 1	Exon 2	Exon3	Exon 4	Exon 5	Exon 6
Mouse	74%	48%	75%	**86%**	**92%**	49%
Dog	73%	64%	**85%**	**91%**	**94%**	55%
Cow	76%	71%	**84%**	**91%**	79%	68%
Elephant	75%	75%	77%	**92%**	-	62%
Armadillo	76%	66%	-	-	**94%**	61%
Opossum	-	-	-	**83%**	-	-
Platypus	27%	-	47%	55%	-	-
Latimeria	-	-	-	57%	-	-

Jan Engelhart

Problem: conservation of sequence does not imply conservation of the transcript.

Sequence conservation could be caused e.g. by cis-acting DNA elements

Evolution of mlncRNAs: HOTAIR



Schorderet P, Duboule D. (2011): Mouse HOTAIR has a different structure, presumably lacks PRC2 binding domain

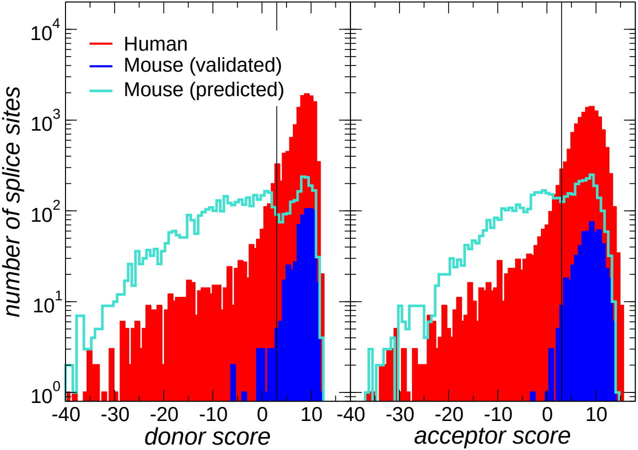
Comparative Map of Splice Sites

Simple idea:

1. use a genome-wide multiple sequence alignment
    1. UCSC 46-way multiz alignment
    2. ENSEMBL 12-way EPO alignment
2. map all splice sites that are experimentally known to the alignment RefSeq plus all ESTs

Prediction of functional splice sites

Limited coverage of transcript data limit sensitivity.
Use splice site scoring scheme (here `maxentscan` scores) to estimate whether a splice site is conserved.



Score distribution of all alignable position is bi-modal:

UCSC 46-way map of splicesites

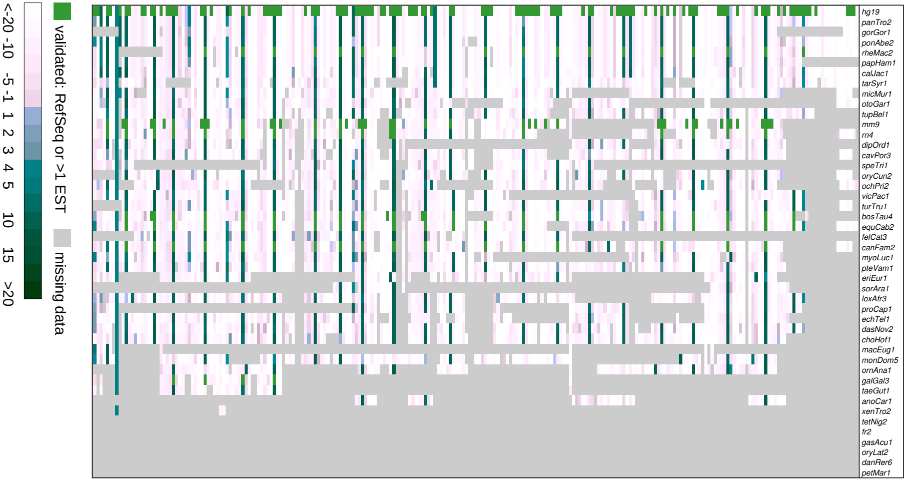
Splice site set: all ESTs and all RefSeq genes
1257773 SS of which 387318 are contained in RefSeq genes

Conservation of RefSeq splice sites between human and mouse

	all		coding		3' UTR		5' UTR		non-coding	
hg19	387318		353836		1103		15200		17179	
aligned	362258	93.5	340617	96.2	834	76	11910	78.3	8888	51.7
predicted	338510	87.3	324504	91.7	670	61	7933	52.1	5403	31.4
validated	336913	86.9	325458	91.9	599	54	6734	44.3	4122	23.9

. . . ncRNAs appear to be much less alignable than UTRs
hints at a problem with the genome-wide alignments

Splice Site Map for GAS5



we know GAS5 is conserved throughout vertebrates, but we have very little aligned sequence already in chicken & frog and nothing in teleotst.

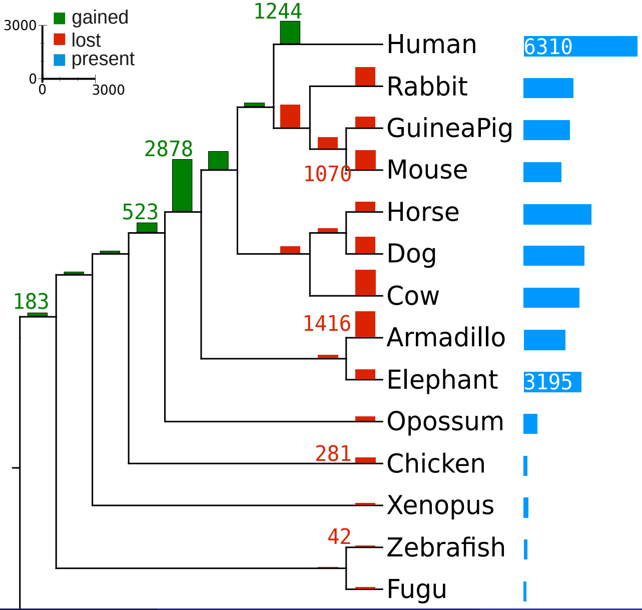
Conservation of lncRNAs

GENCODE set of 25644 splice sites in 6310 lncRNA transcripts

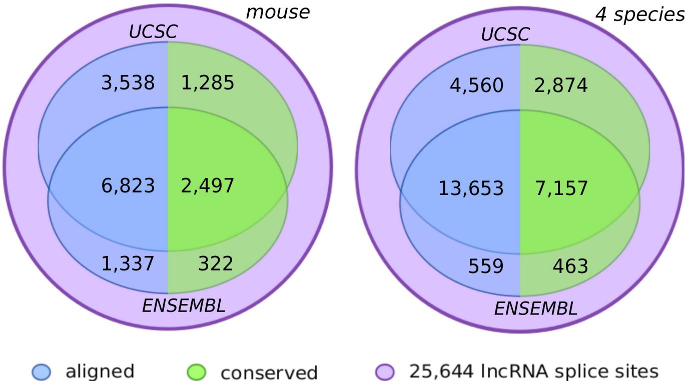
(removed all overlaps with coding regions and RNAcode hits)

Species	splice sites		transcripts	
	cons.	val.	cons.	val.
mouse	3,809	1,155	2,098	362
rat	3,508	752	1,968	224
cow	6,214	1,047	3,096	351
dog	6,828	694	3,382	238
union 5	10,057	1,876	4,292	585
union 15	13,720	2,098	5,072	635

Gains and Losses



Comparison of UCSC and ENSEMBL Alignments



human/mouse

4 eutheria

Conservation of special mlncRNA sets

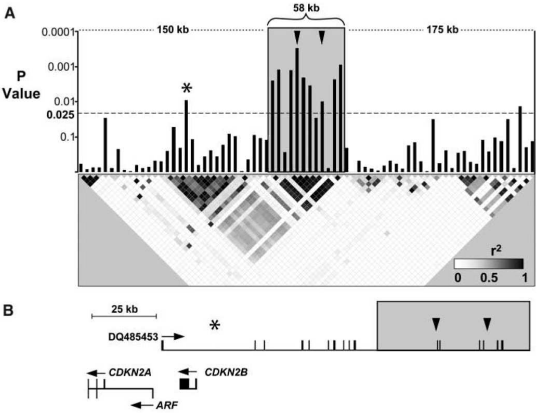
	aligned	predicted	known
	213 human transcripts hosting microRNAs		
mouse	151	91	15
dog	185	144	5
5 Eutheria	191	164	23
	94 human transcripts hosting snoRNAs		
mouse	72	57	46
dog	81	70	40
5 Eutheria	84	74	51
	2,076 mouse lncRNAs from [1]		
human	1,770	1,113	446
dog	1,628	944	185
4 Eutheria	1,776	1,237	472
	1,508 zebrafish lncRNAs [2,3]		
Teleostei	953	513	112
Tetrapoda	476	170	56

[1] Guttmann *et al.* Nature 477: 295-300 (2011); [2] Pauli *et al.* Genome Res. 10.1101/gr.133009.111 (2011); [3] Ulitsky *et al.* Cell 147: 1537-1550(2011)

ANRIL

Most QTLs for complex multi-genic diseases hit noncoding regions

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

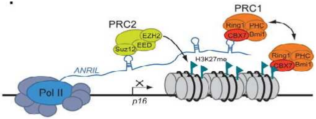


McPherson *et al.*, Science (2007)

ANRIL transcript(s) in many isoforms associated with the atherosclerosis risk

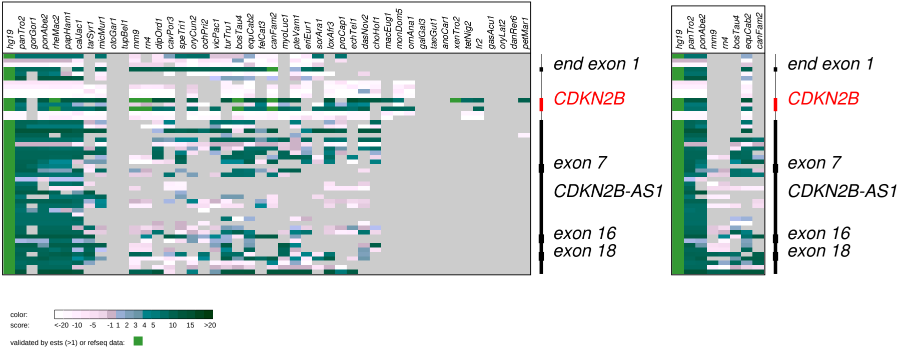
Holdt *et al.* (2010)

and it appears in many other GWAS studies

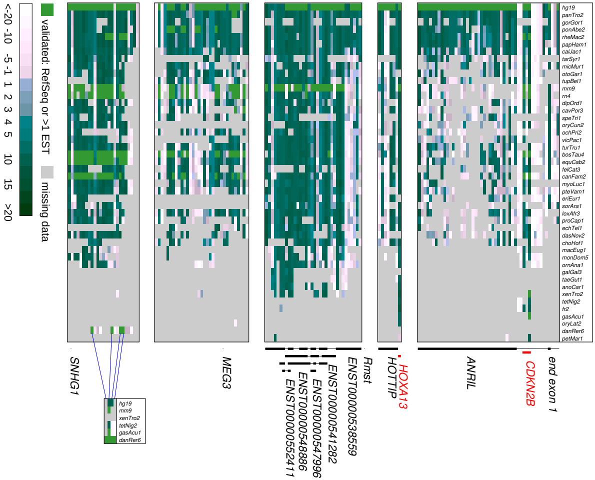


Yap *et al* (2010)

Conservation of ANRIL



Some more examples



Summary

- Conservation and substitution patterns of splice sites can be used to infer novel non-coding genes even in the absence of RNA secondary structure and
- As entitites, mlnRNAs are evolutionarily much older and better conserved than their sequence
- Most mlncRNAs are conserved only in parts with rapid changes in gene structure
- Large repertoire of unspliced lncRNAs whose evolutionary patterns we do not yet understand

Many, many thanks . . .

- Leipzig: Jana Hertel, Hakim Tafer, David Langenberger, Jan Engelhardt, Anne Nitsche, Sebastian Bartschat, Steffi Kehr, and many others
  Steve Hoffmann, Christian Otto
  Sonja J. Prohaska and her Comp. EvoDevo group
  FH RNomics group: Jörg Hackermüller, Kristin Reiche, Kathy Schutt, Kerstin Ullmann, ...
  FG ncRNAs: Friedemann Horn, Thomas Arendt, Kurt Engeland, Peter Ahnert, . . .
- Vienna: Ivo L. Hofacker, Christoph Flamm, Sven Findeiß, Stefan Bernhart, Andreas Gruber, and many others in Peter Schuster's Lab over the years
- Halle: Günter Reuter's Lab
- Dresden: Michael Hiller
- Marburg: Manja Marz and her group
- Freiburg: Rolf Backofen, Dominic Rose
- Copenhagen: Jan Gorodkin, Stefan Seemann, Peter Menzel, and the RTH
- Barcelona: Roderic Guigó, Andrea Tanzer
- Strassbourg: Catherine Florentz, Joern Pütz, Frank Jühling
- MIT: Stefan Washietl, Sebastian Will
- Affymetrix: Tom Gingeras, Phil Kapranov, *et al.*
- PICB Shanghai: Axel Mosig and Phil Khaitovich and their students (PICB/SIBS)
- ASU Tempe: Julian L. Chen and his lab
- ENCODE: Ewan Birney and $10^{2.5}$ coauthors
- Funding by the DFG, DAAD, the EU 6th and 7th framework programme, the VW Foundation