

# Long non-coding RNAs

**Dominic Rose**

Bioinformatics Group, University of Freiburg

Bled, Feb. 2011



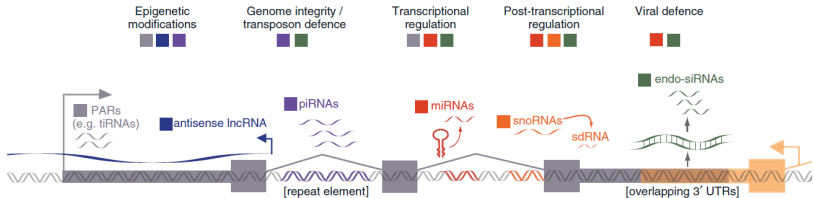
# Outline

- *De novo* prediction of long non-coding RNAs (lncRNAs)
- Genome-wide RNA gene-finding
- Intrinsic properties (sequence/structure) of lncRNAs?

# Long ncRNAs: introduction

- ENCODE: pervasive transcription in eukaryotes, large portion are lncRNAs.
- lncRNAs  $\hat{=}$  ncRNAs  $>$  200nt
- Capped, polyadenylated, often (alternatively) spliced (just like protein-coding genes), but lack discernible open reading frames
- Gene regulators: Evi-2, Xist, roX1, roX2, H19, ...
- Precursor for small RNAs: miRNAs, snoRNAs, ...
- Imprinting, epigenetics, disease-associated (expression correlates with viral insertion, carcinogenesis, ...)
- Functionally important ncRNA class
- **No general computational method for their detection**

# Non-coding RNA gene-finding



## Eukaryotic genome organization

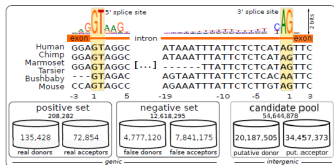
Taft *et al.*, J Pathol, 2010

- Challenging problem: heterogeneity, lack of features
- Short (structured) vs. long (unstructured?) ncRNAs
  - RNA secondary structure prediction
  - Splice site detection
  - Promoter recognition
  - ...

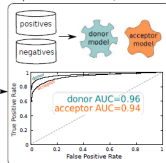
# Long ncRNAs: a first (generic) gene-finding approach

## A) Splice site prediction

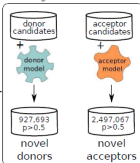
1. Scan alignments for splice sites, prepare and partition data



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

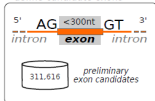


3. Predict novel splice sites in intergenic candidates

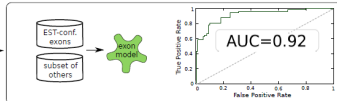


## B) Exon prediction

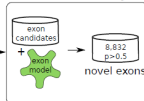
1. AG/GT splice sites pairs define candidate exons



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

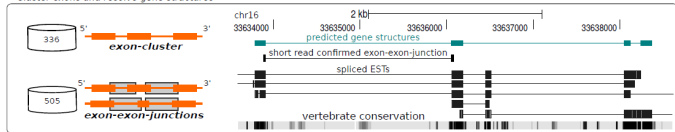


3. Classify exons which were not used for SVM training



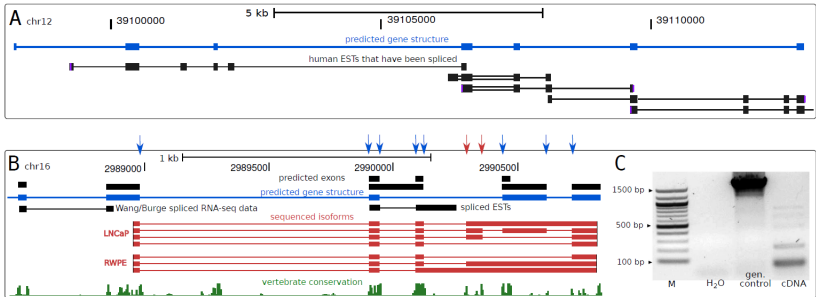
## C) Transcript prediction

Cluster exons and resolve gene structures





# Long ncRNAs: experimental evidence



**B: RT-PCR + sequencing confirms 10 SS, 8/9 predicted SS are true**



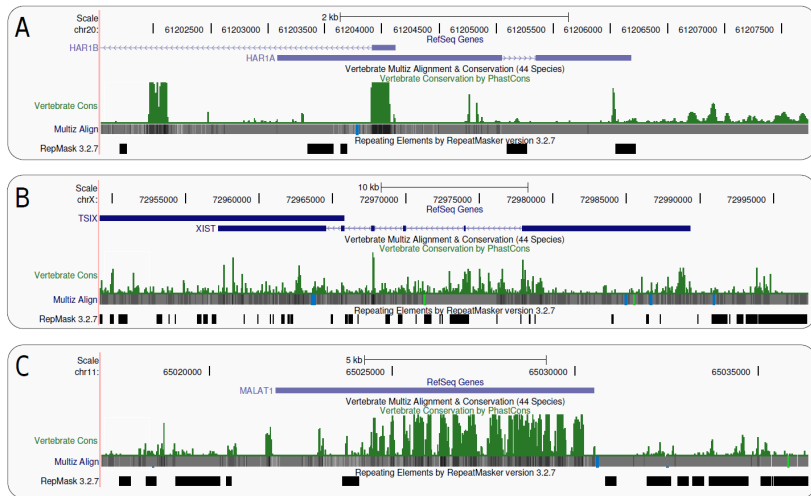
# Long ncRNAs: further properties?

- Recall: splice sites help to pin down lncRNAs
- Are there other features specific to lncRNAs?
- Do lncRNAs exhibit specific sequential or structural motifs?
- If so, are they conserved among species?
- We need sequence alignments . . .  
(or at least a set of orthologous lncRNAs)

# Long ncRNAs: orthologs

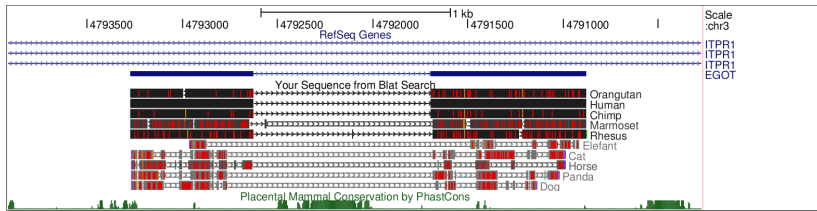
- Use existing genome-wide alignments?  
(UCSC 46-way, Ensembl Compara)
- Maybe, but for “large regions of low sequence similarity” (=lncRNA) these automatic pipelines have serious issues
  - Often very fragmented, short blocks are reported
  - Broken synteny ...

# Long ncRNAs: orthologs? problematic!



# Long ncRNAs: example EGO-B

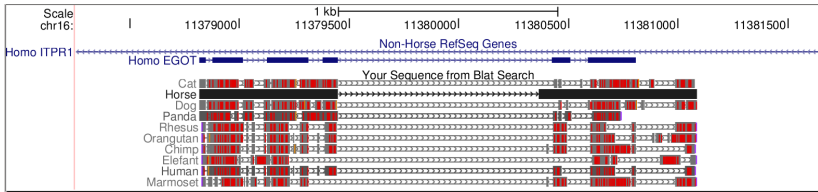
- EGO = eosinophil granule ontogeny
- Essential for the expression of major basic proteins in eosinophils (white blood cells, part of the immune system)
- Experimentally confirmed two-exon transcript structure
- Let's try to collect orthologs . . .



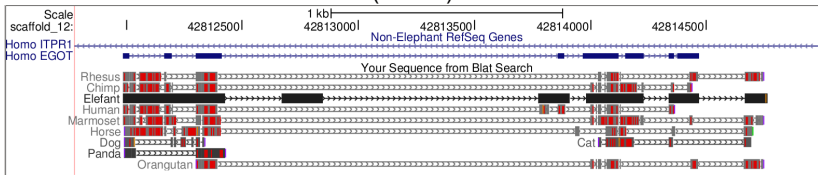
(human)

# Long ncRNAs: example EGO-B

- Black bars: hand curated annotation of EGO-B.
- Manual inspection reveals orthologs.
- Given annotation (here RefSeq) often wrong.

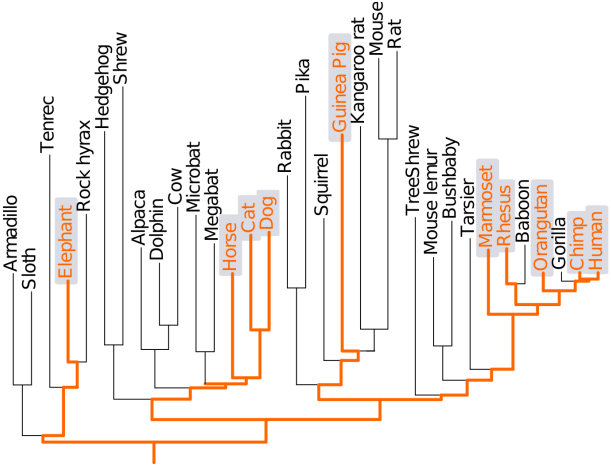


(horse)



(elefant)

# Long ncRNAs: example EGO-B





# Long ncRNAs: local structure motifs?

- Now use full spectrum of bioinformatic approaches
- e.g. search for **local** secondary structure motifs



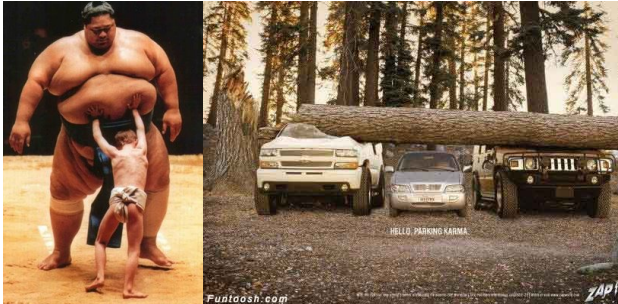
# Long ncRNAs: local structure motifs?

- Now use full spectrum of bioinformatic approaches
- e.g. search for **local** secondary structure motifs



# Long ncRNAs: local structure motifs?

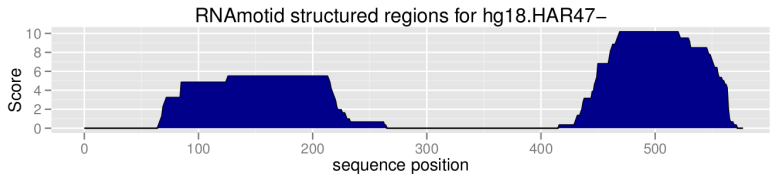
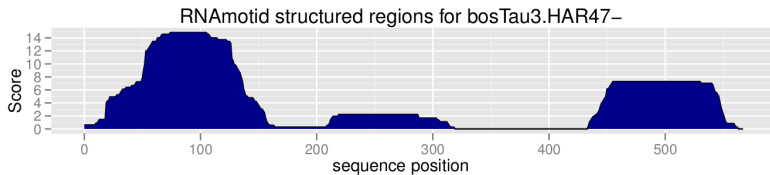
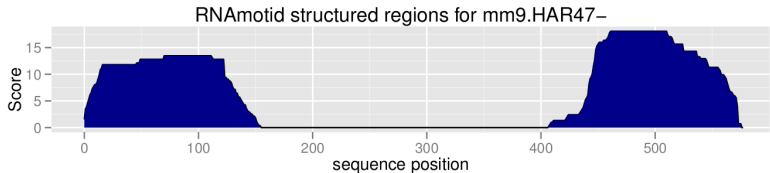
- Now use full spectrum of bioinformatic approaches
- e.g. search for **local** secondary structure motifs



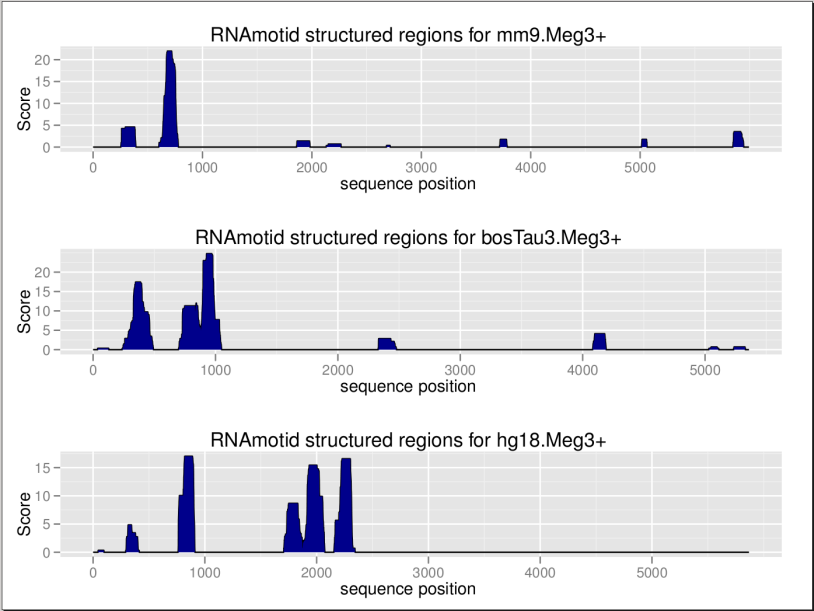
# RNAmotid: accuracy-based detection of local RNA elements

- RNAmotid = RNA motif identification
- Steffen Heyne, Essam Abdel Moaty Abdel Hady
- Identifies local RNA elements on a genome-wide scale
- Fast sparse algorithm to predict maximum expected accuracy structures
- Based on base-pairing/unpairing probabilities (RNAplfold)
- Relies on a novel accuracy function reflecting locality
- Allows genome-wide scans for structured regions that have high probabilities of containing significant local RNA motifs

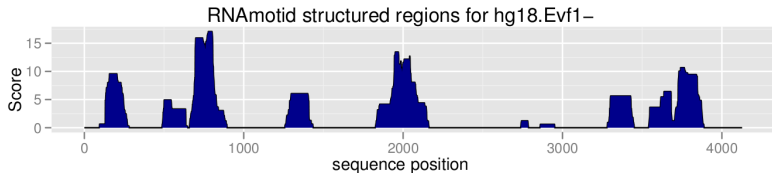
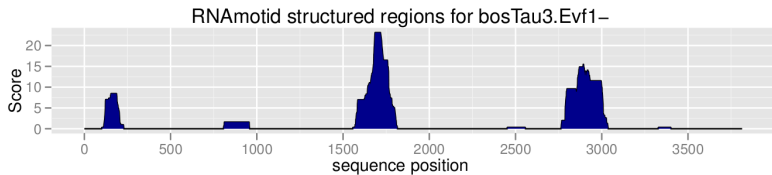
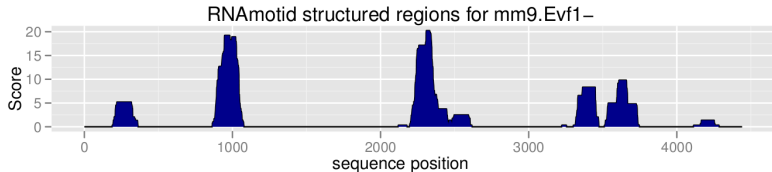
# RNAmotid: MEA-folding



# RNAmotid: MEA-folding

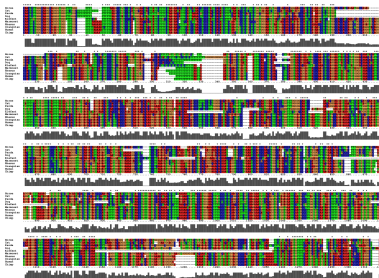


# RNAmotid: MEA-folding



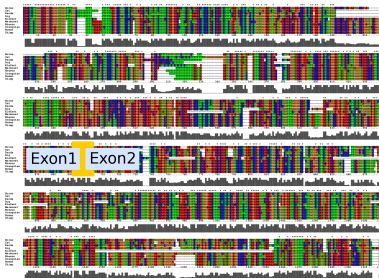
# Long ncRNAs: orthologs?

- RNAmotid scans single sequences
- Might be an option if lncRNA alignments are missing
- However, recall our nice EGO-B alignment . . .



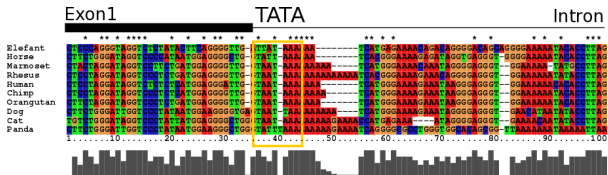
# Long ncRNAs: orthologs?

- RNAmotid scans single sequences
- Might be an option if lncRNA alignments are missing
- However, recall our nice EGO-B alignment . . .

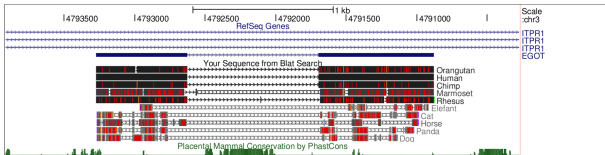
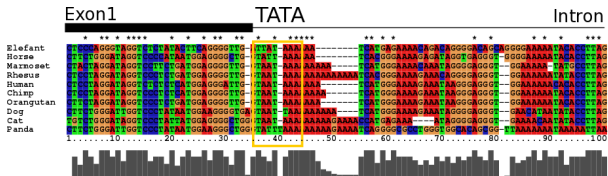




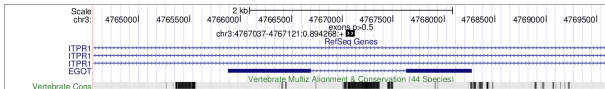
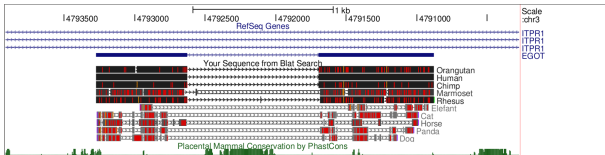
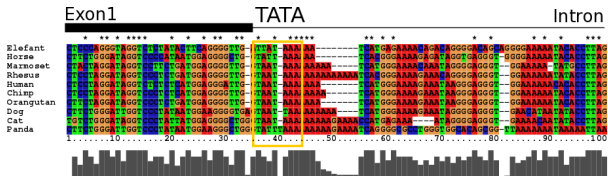
# Long ncRNAs: EGO-B



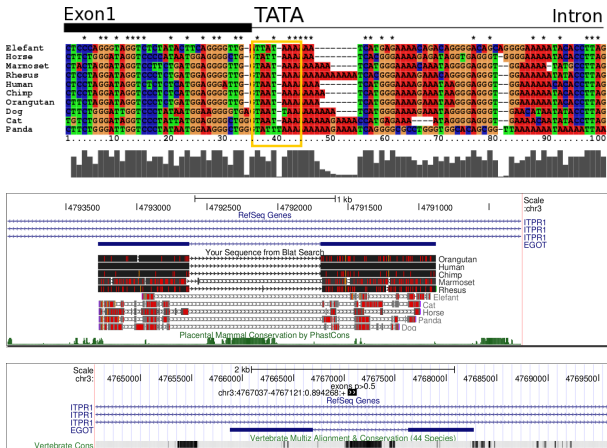
# Long ncRNAs: EGO-B



# Long ncRNAs: EGO-B



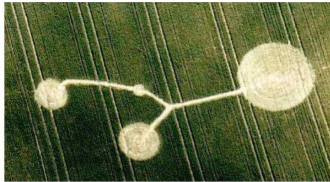
# Long ncRNAs: EGO-B



... which brings us full circle with regard to our initial topic, predicting novel transcripts via conserved splice sites.

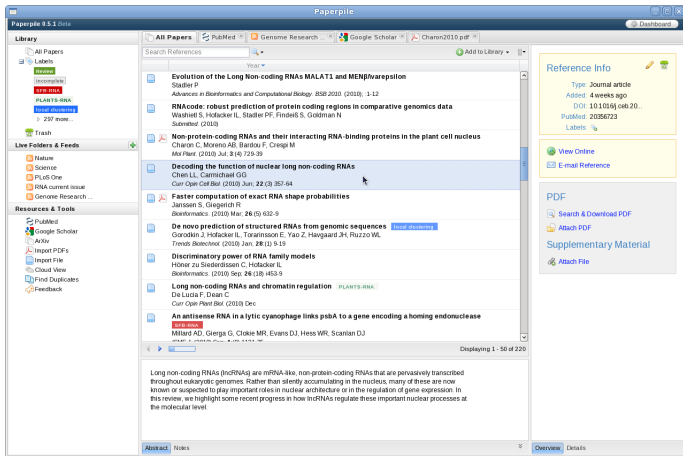
# Acknowledgements

- Leipzig (Peter F. Stadler) ↔ Freiburg (Rolf Backofen)
- Michael Hiller (Stanford University)



The truth is out there

# Your state-of-the-art reference manager: **Paperpile**



The screenshot displays the Paperpile 0.5.1 Beta web interface. The main content area shows a list of papers with the following titles and authors:

- Evolution of the Long Non-coding RNAs MALAT1 and MENP1/arepsilon**  
Stadler P  
*Advances in Bioinformatics and Computational Biology* 83B:2010 (2010): 1-12
- RNAcode: robust prediction of protein coding regions in comparative genomics data**  
Washietl S, Hofacker IL, Stadler PF, Finkels S, Goldman N  
Submitted (2010)
- Non-protein-coding RNAs and their interacting RNA-binding proteins in the plant cell nucleus**  
Charon C, Moreno AB, Baidou F, Crepiq M  
*Mol Plant* (2010) Jul; 3 (4): 729-39
- Decoding the function of nuclear long non-coding RNAs**  
Chen LL, Carmichael GG  
*Curr Opin Cell Biol* (2010) Jun; 22 (3): 357-64
- Faster computation of exact RNA shape probabilities**  
Janssen S, Giegerich R  
*Bioinformatics* (2010) Mar; 26 (5): 632-9
- De novo prediction of structured RNAs from genomic sequences**  
Gordkin J, Hofacker IL, Toraninsson E, Yao Z, Havgaard JH, Ruzzo WL  
*Trends Biotechnol* (2010) Jan; 28 (1): 5-19
- Discriminatory power of RNA family models**  
Honer zu Siedersissen C, Hofacker IL  
*Bioinformatics* (2010) Sep; 26 (18): 453-9
- Long non-coding RNAs and chromatin regulation**  
De Luca F, Dean C  
*Curr Opin Plant Biol* (2010) Dec
- An antisense RNA in a lytic cyanophage links psbA to a gene encoding a homing endonuclease**  
Millard AD, Gerga G, Ctokie MR, Evans DJ, Hess WR, Scanlan DJ  
*PLoS ONE* (2010) Dec 23; 5 (12): e12111

The interface also includes a left sidebar with navigation options like 'Library', 'Labels', 'PLANTS-RNA', and 'Resources & Tools'. A right sidebar provides 'Reference Info' for the selected paper, including 'View Online' and 'E-mail Reference' links. A 'PDF' section offers 'Search & Download PDF' and 'Attach PDF' options. A 'Supplementary Material' section includes an 'Attach File' link. At the bottom, there is a text block about long non-coding RNAs (lncRNAs) and a navigation bar with 'Abstract', 'Notes', 'Overview', and 'Details' buttons.

<http://paperpile.com/beta/>  
<https://github.com/wash/paperpile>