Long non-coding RNAs

Dominic Rose
Bioinformatics Group, University of Freiburg

Bled, Feb. 2011

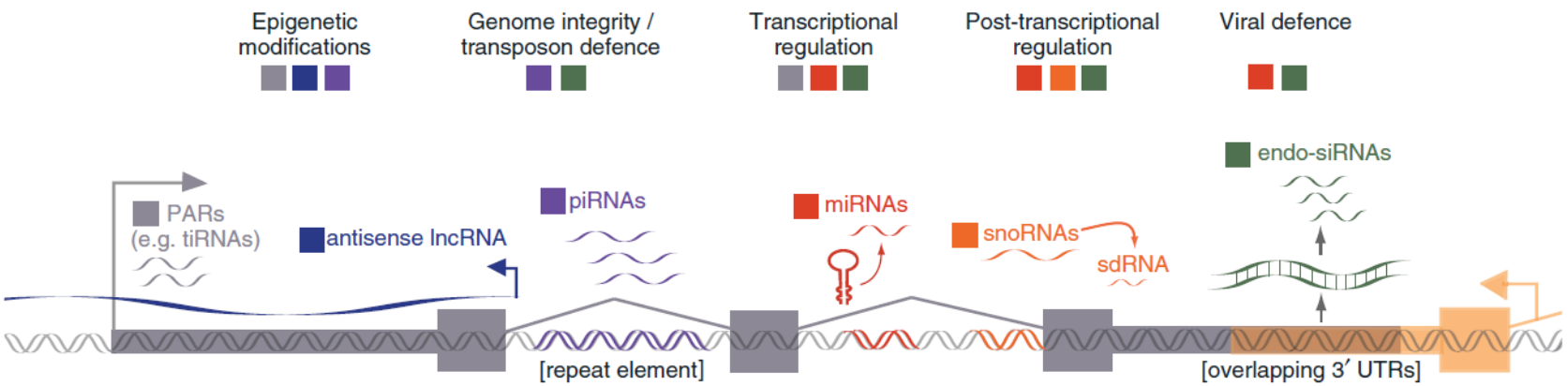
Outline

- *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 $\widehat{=}$ ncRNAs $>$ 200nt
- Capped, polyadenylated, often (alternatively) spliced (just like protein-coding genes), but lack discernible open reading frames
- Gene regulators: Evf-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

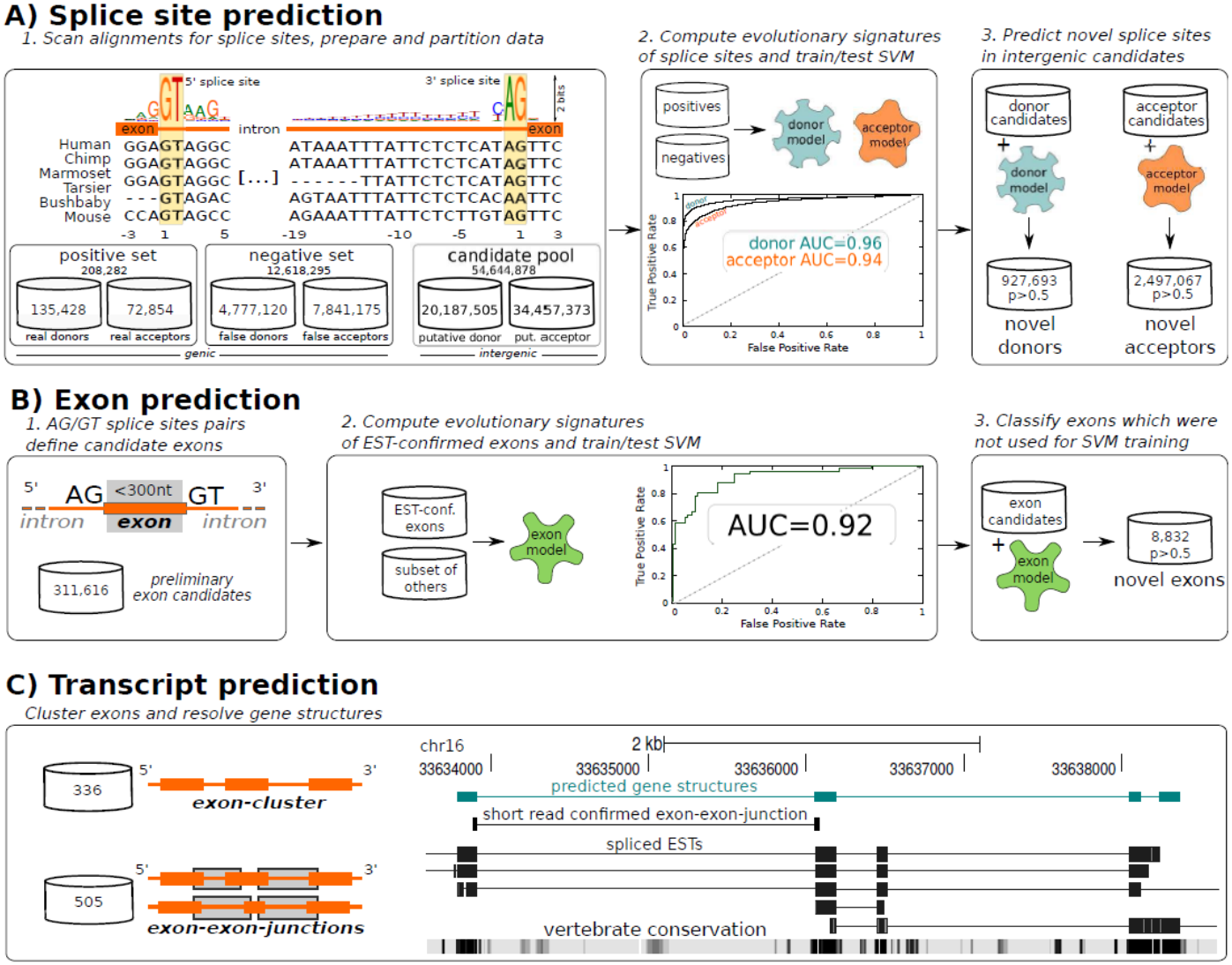


Eukaryotic genome organization

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
  - ...

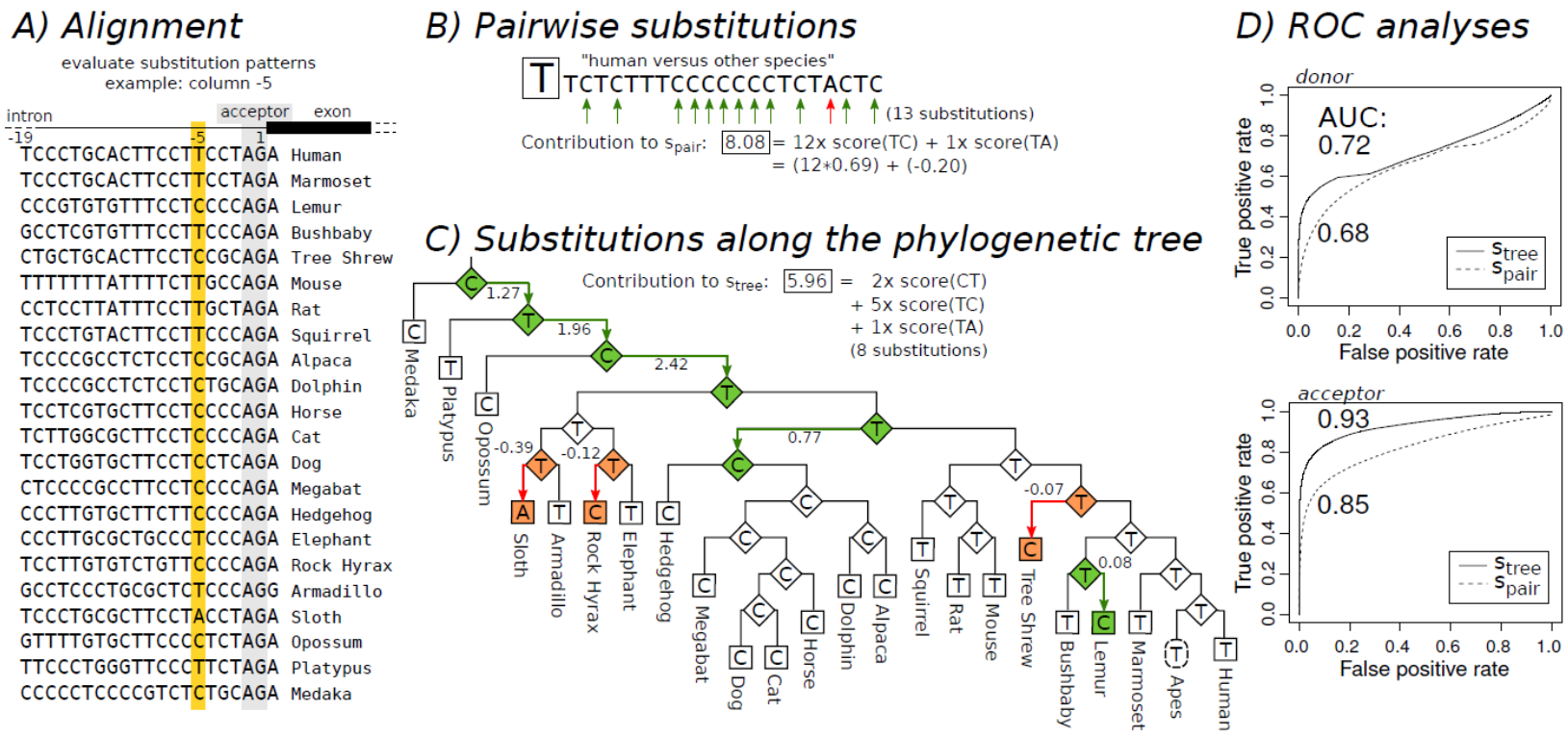
Long ncRNAs: a first (generic) gene-finding approach



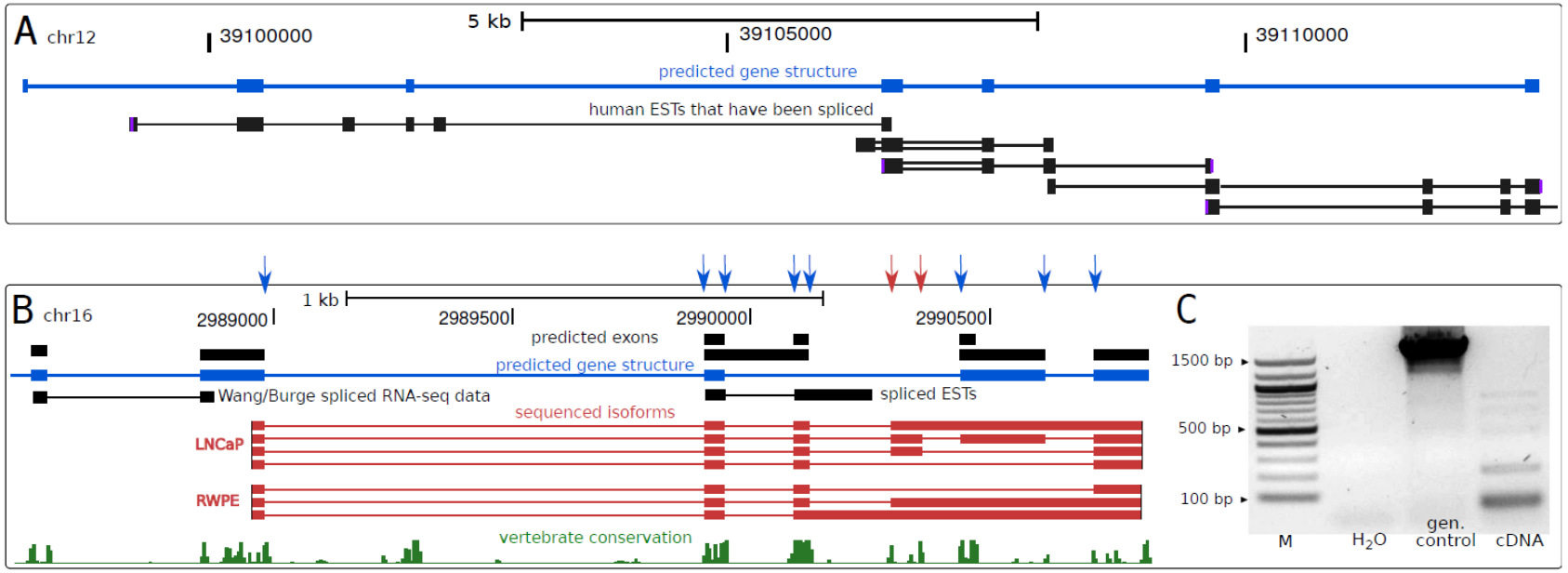
“Evolutionary signatures” of splice sites?

Novel method to capture compositional features of splice sites:

- evaluate substitution patterns
- derive log-odds substitution scores



Long ncRNAs: experimental evidence



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

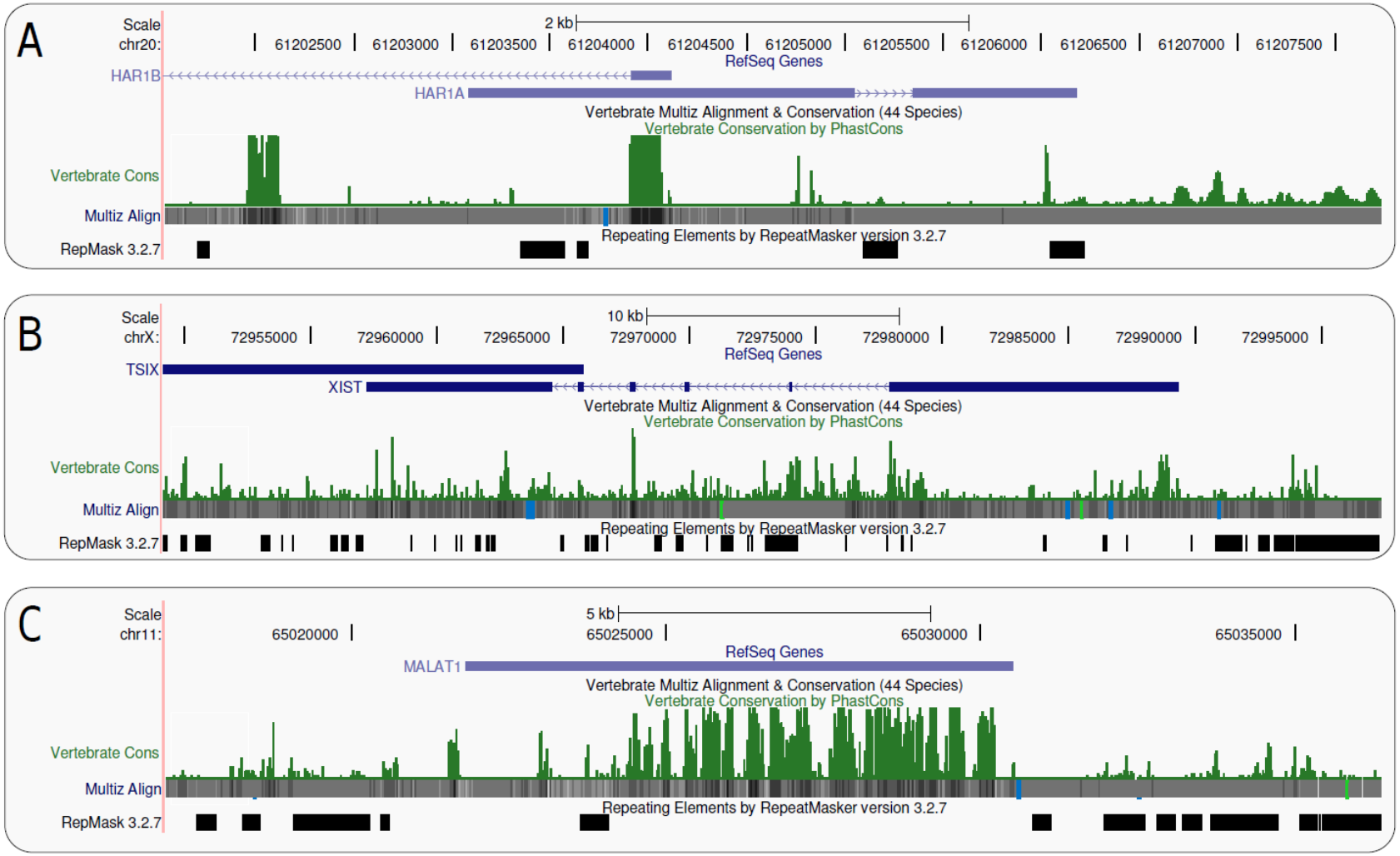
Long ncRNAs: further properties?

- 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

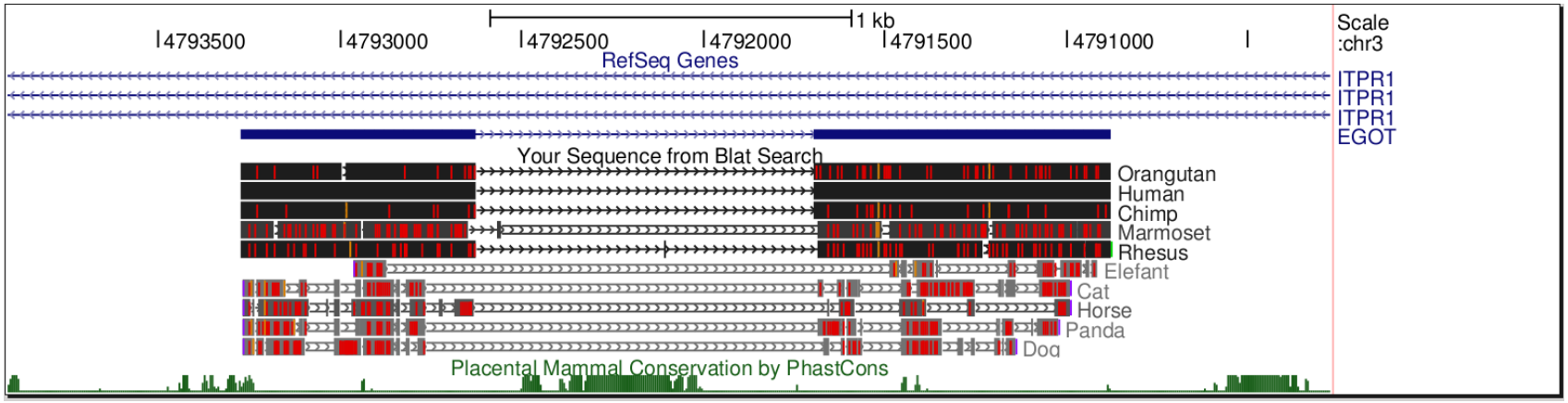
- Use existing genome-wide alignments?
  (UCSC 46-way, Ensembl Compara)
- Maybe, but for “large regions of low sequence similarity” (=lncRNA) these automatic pipelines have serious issues
  - Often very fragmented, short blocks are reported
  - Broken synteny . . .

Long ncRNAs: orthologs? problematic!



Long ncRNAs: example EGO-B

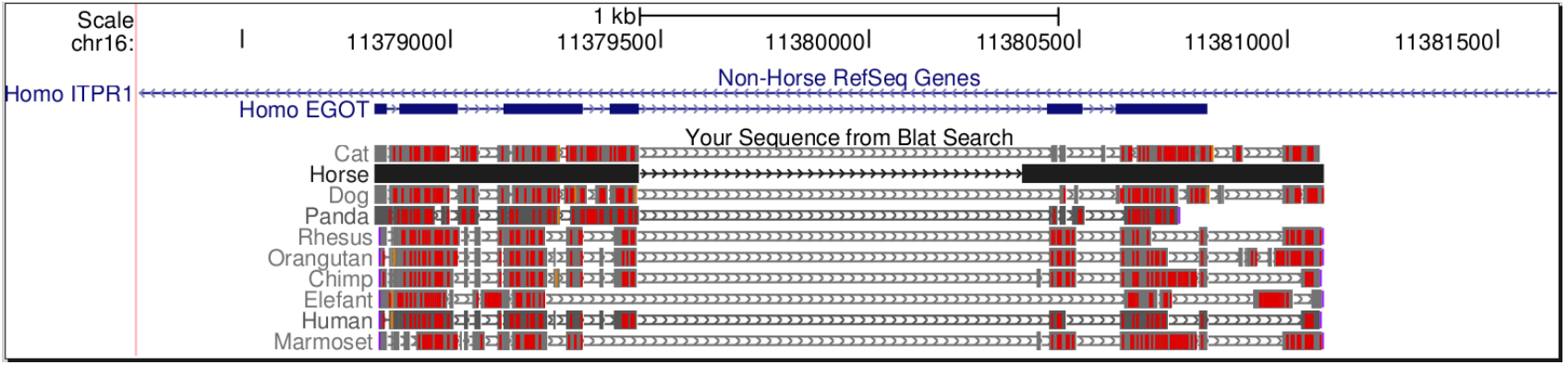
- EGO = eosinophil granule ontogeny
- Essential for the expression of major basic proteins in eosinophils (white blood cells, part of the immune system)
- Experimentally confirmed two-exon transcript structure
- Let's try to collect orthologs . . .



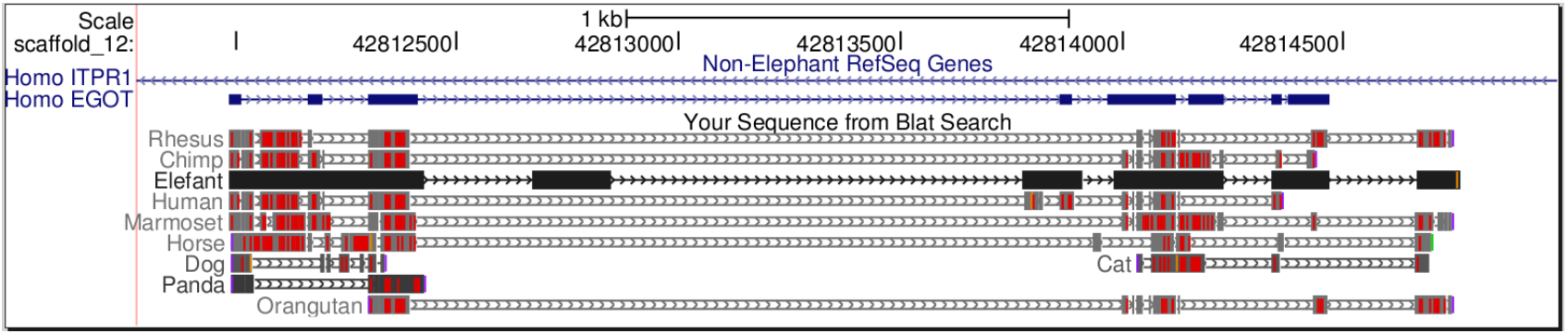
(human)

Long ncRNAs: example EGO-B

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

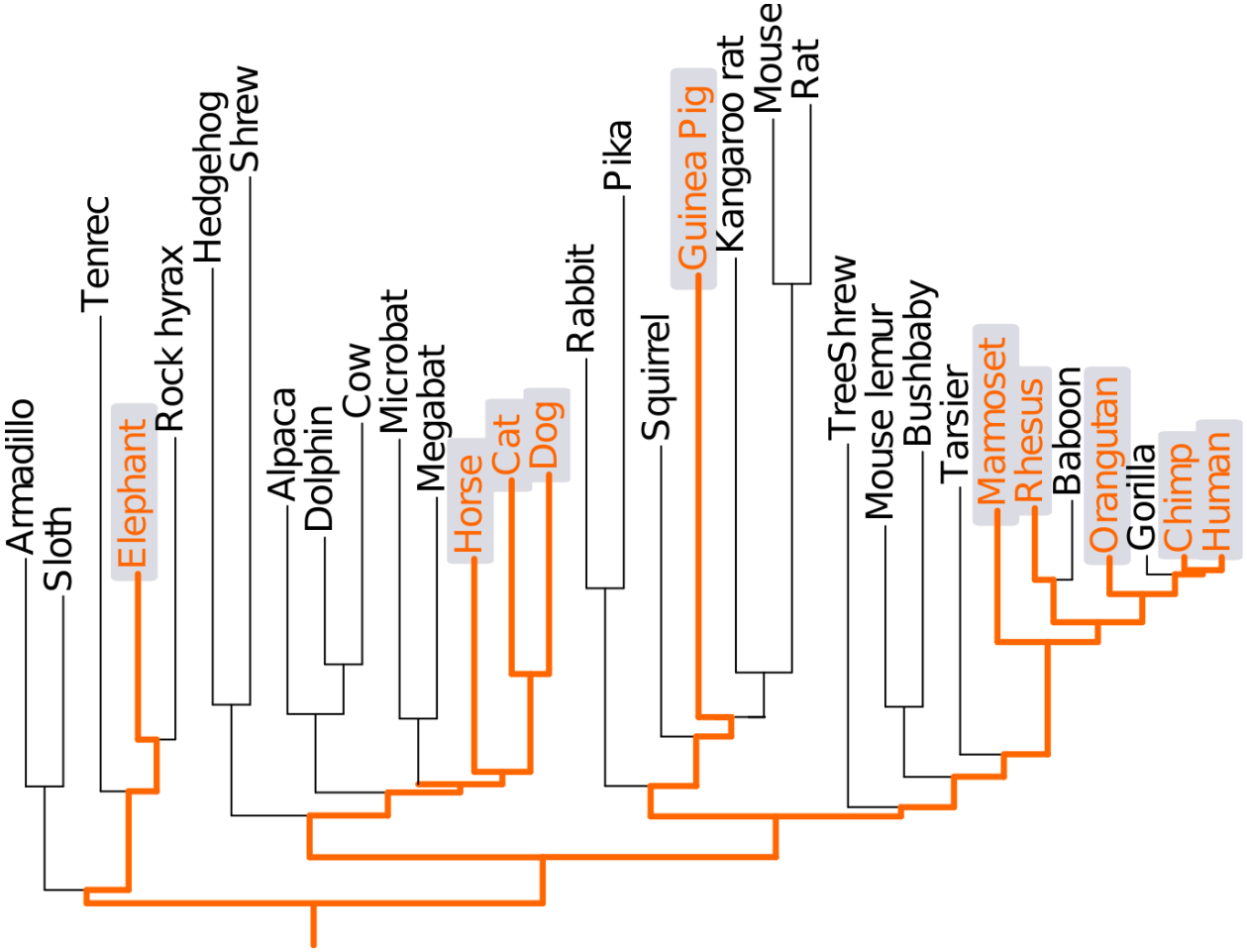


(horse)

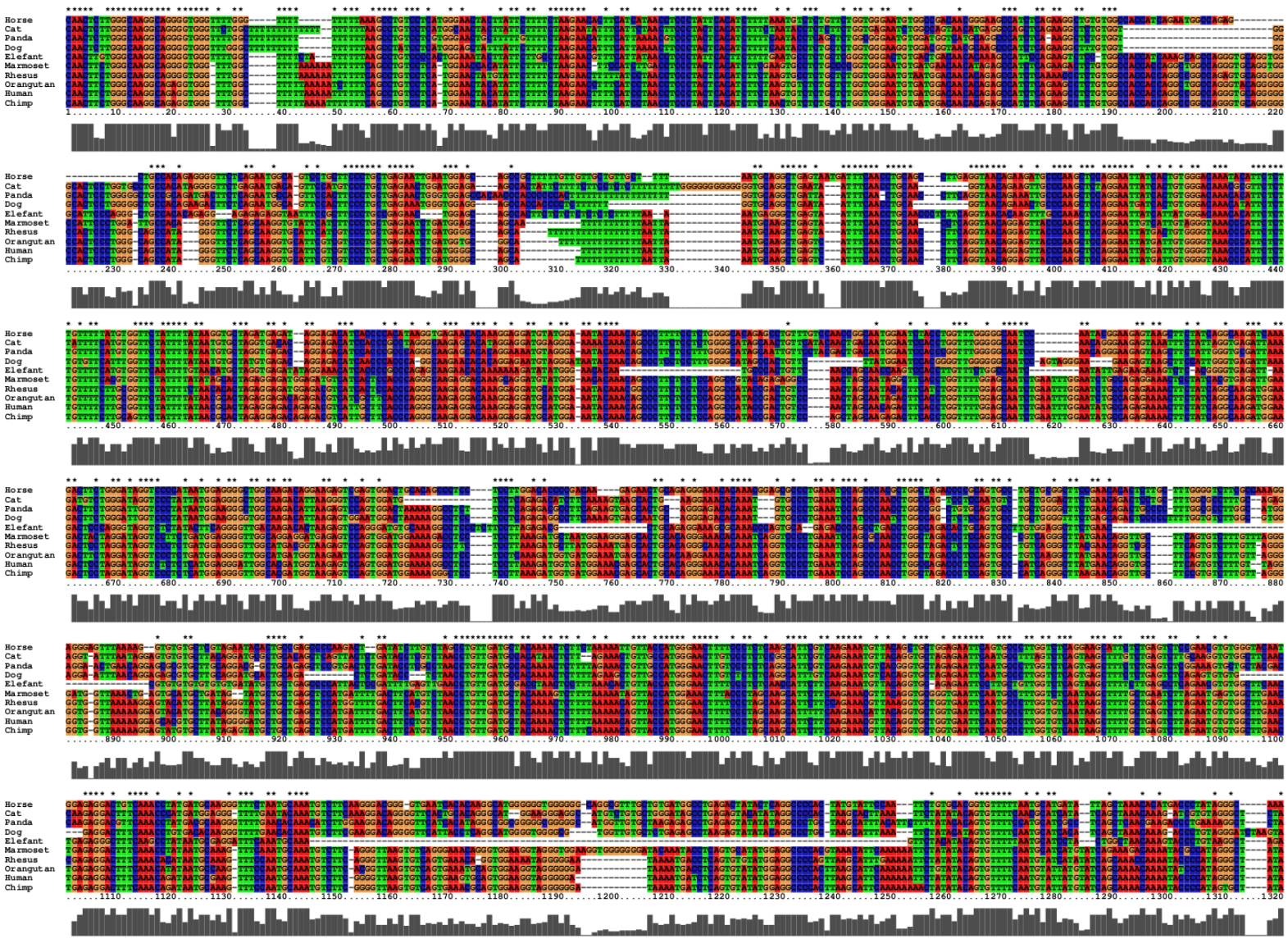


(elefant)

Long ncRNAs: example EGO-B



Long ncRNAs: example EGO-B



Long ncRNAs: local structure motifs?

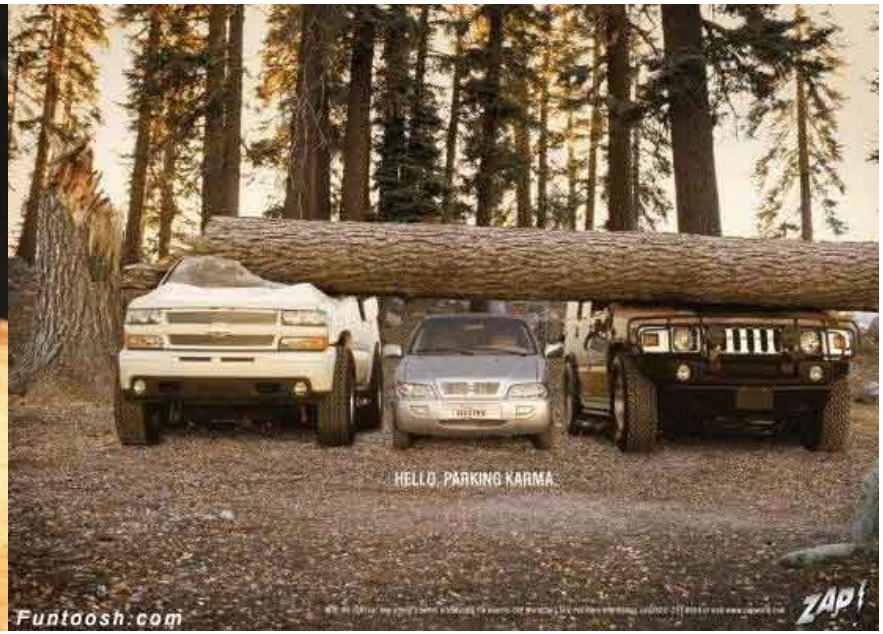
- Now use full spectrum of bioinformatic approaches
- e.g. search for local secondary structure motifs

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

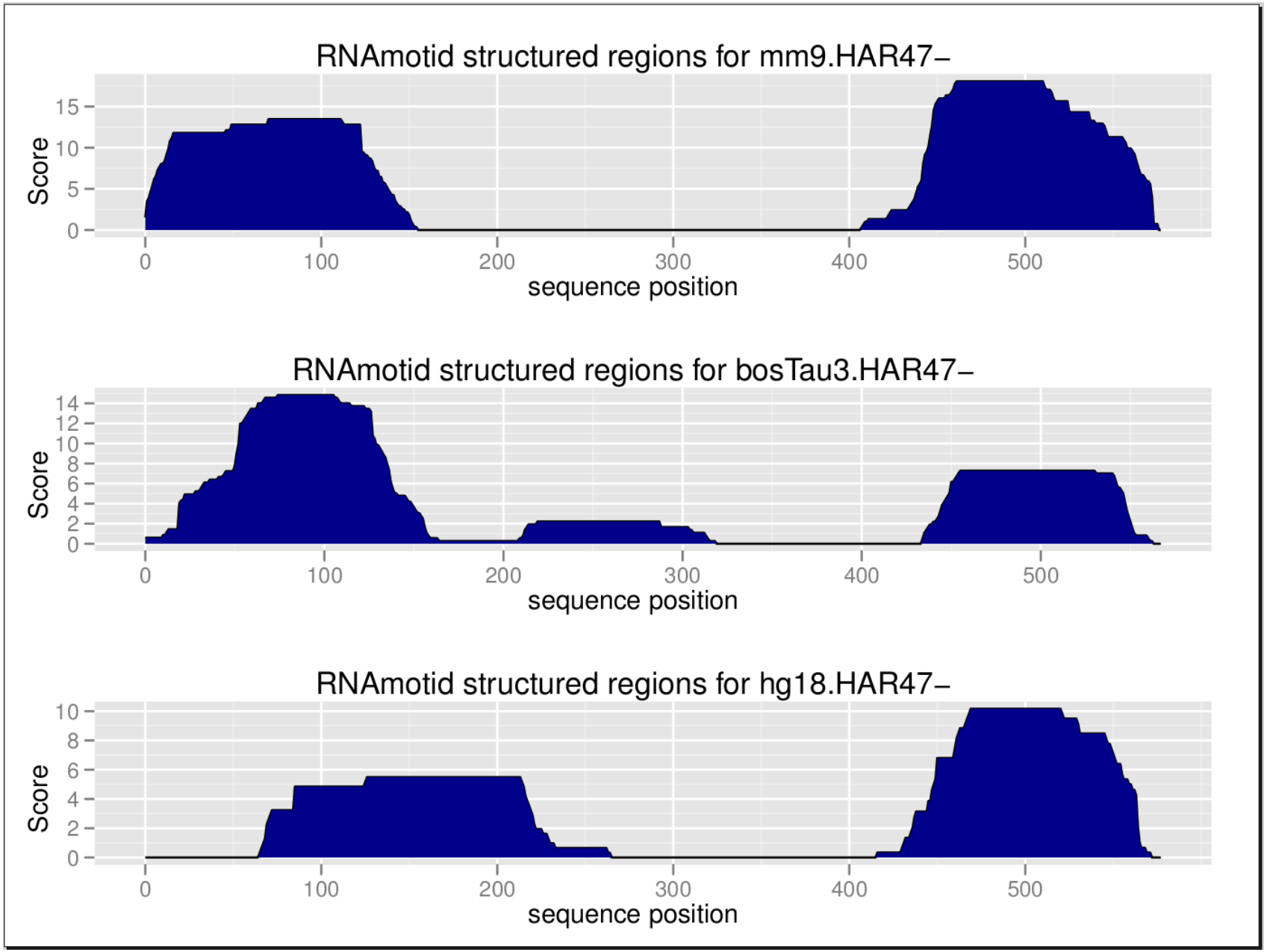


Paperpile

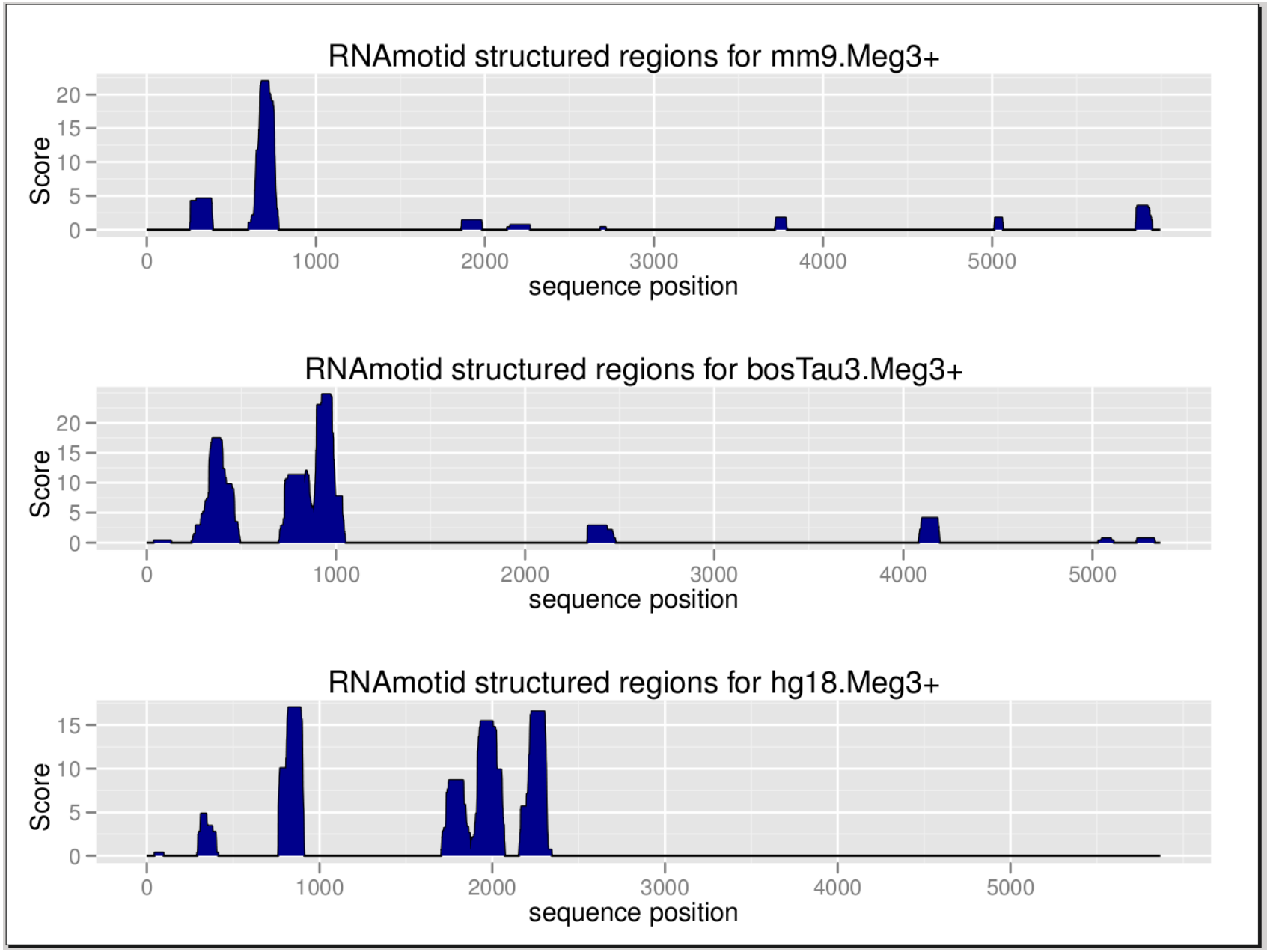
RNAmotid: accuracy-based detection of local RNA elements

- 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

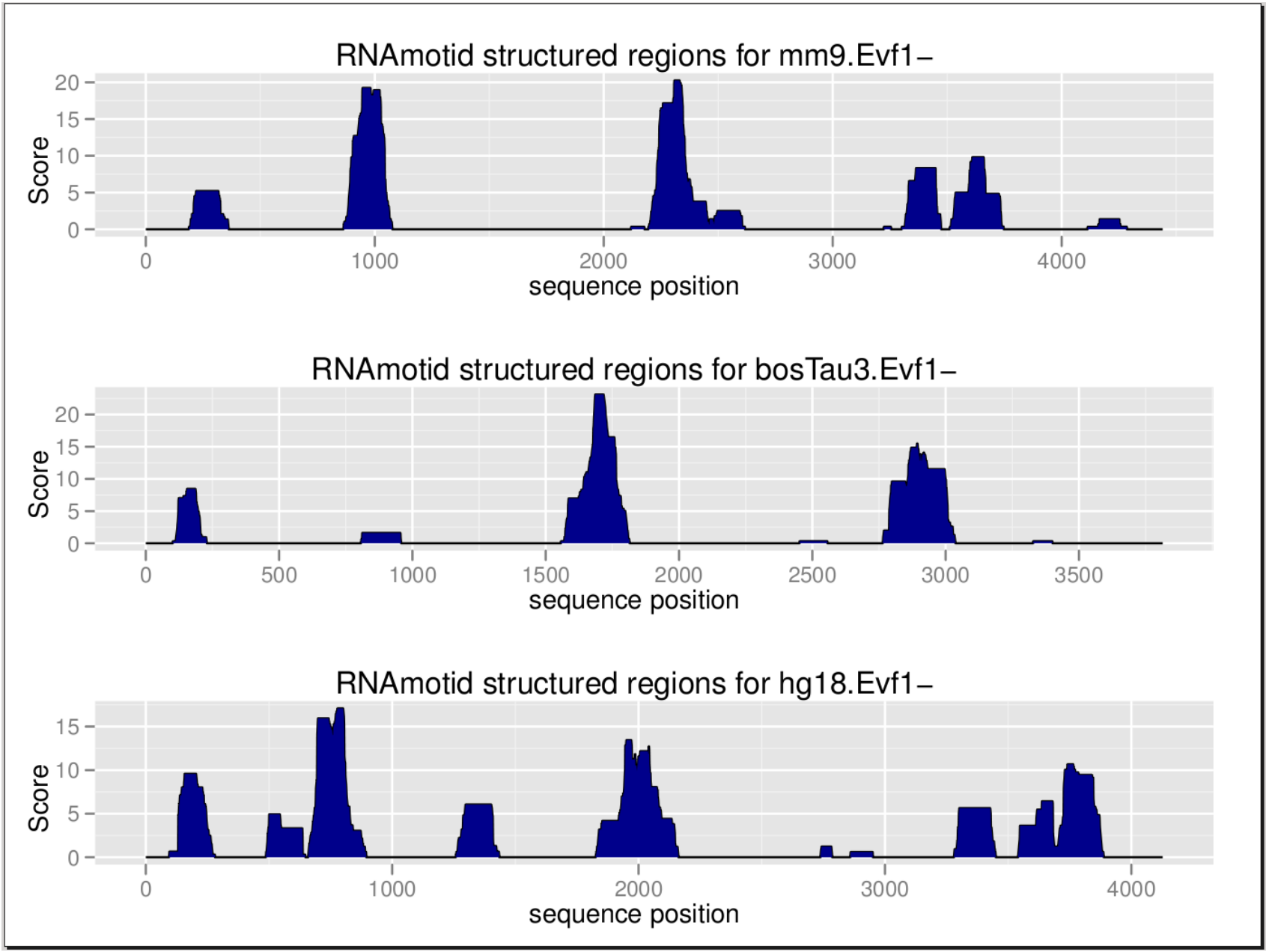
RNAmotid: MEA-folding



RNAmotid: MEA-folding

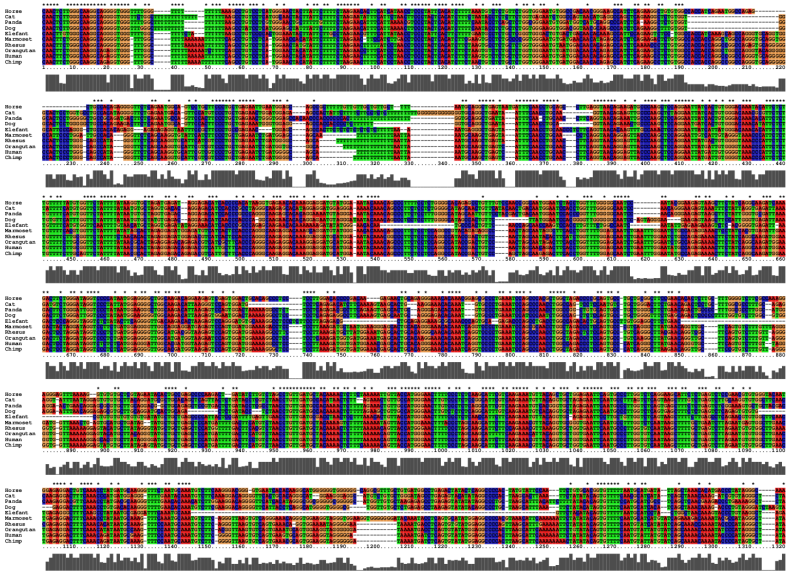


RNAmotid: MEA-folding



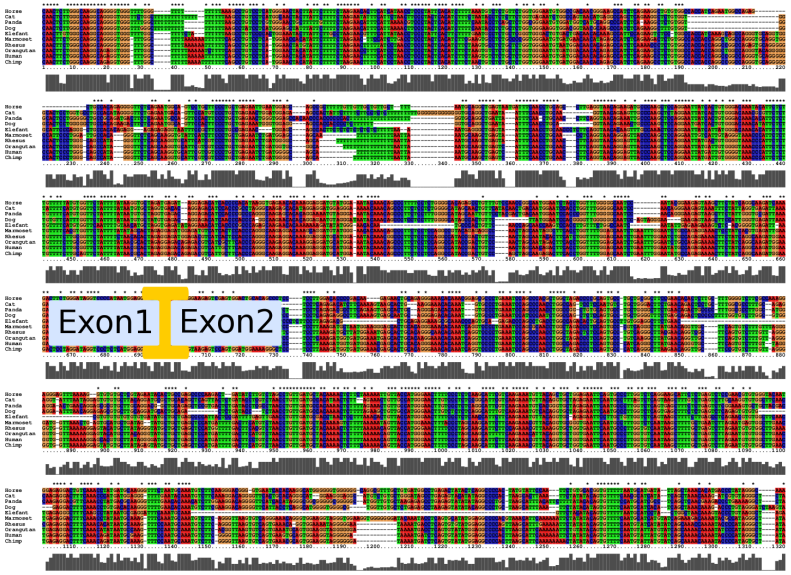
Long ncRNAs: orthologs?

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

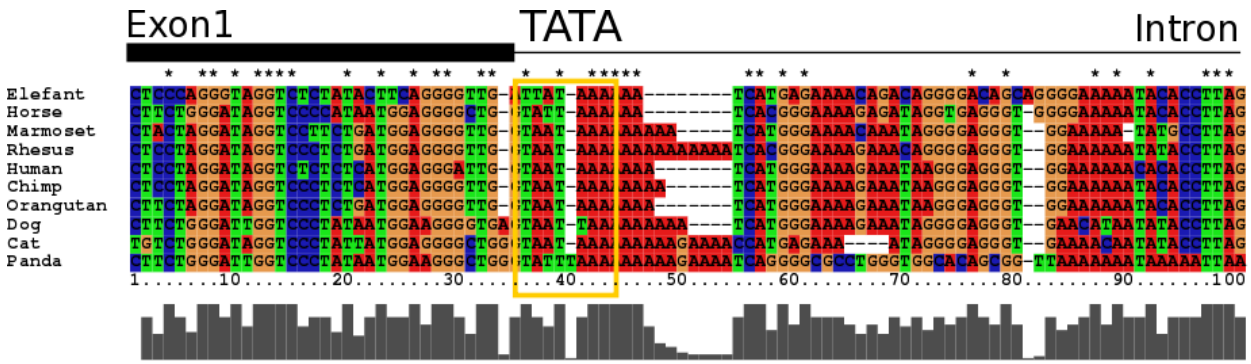


Long ncRNAs: orthologs?

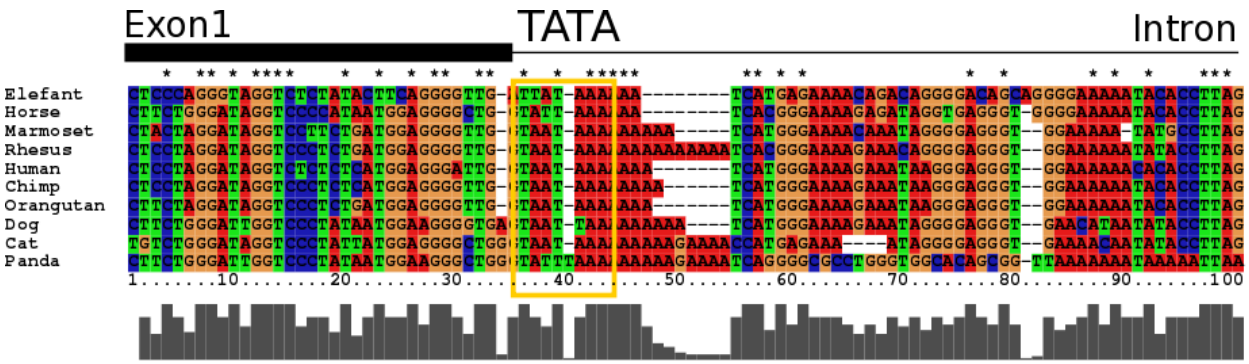
- RNAmotid scans single sequences
- Might be an option if lncRNA alignments are missing
- However, recall our nice EGO-B alignment . . .

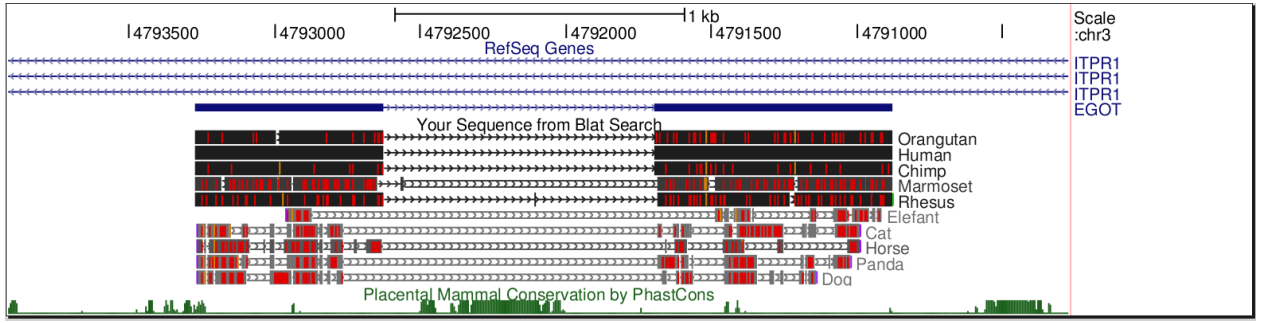


Long ncRNAs: EGO-B

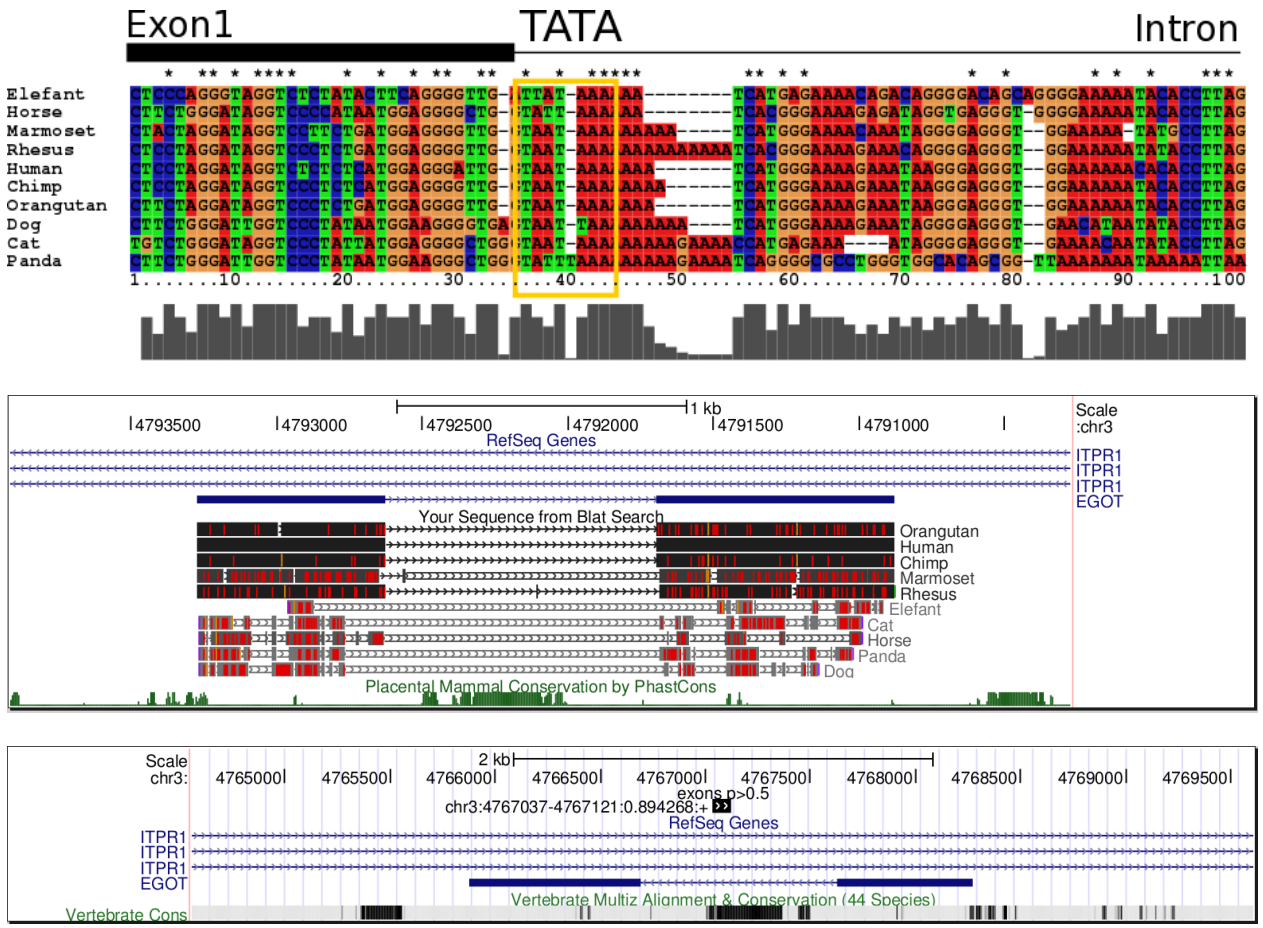


Long ncRNAs: EGO-B

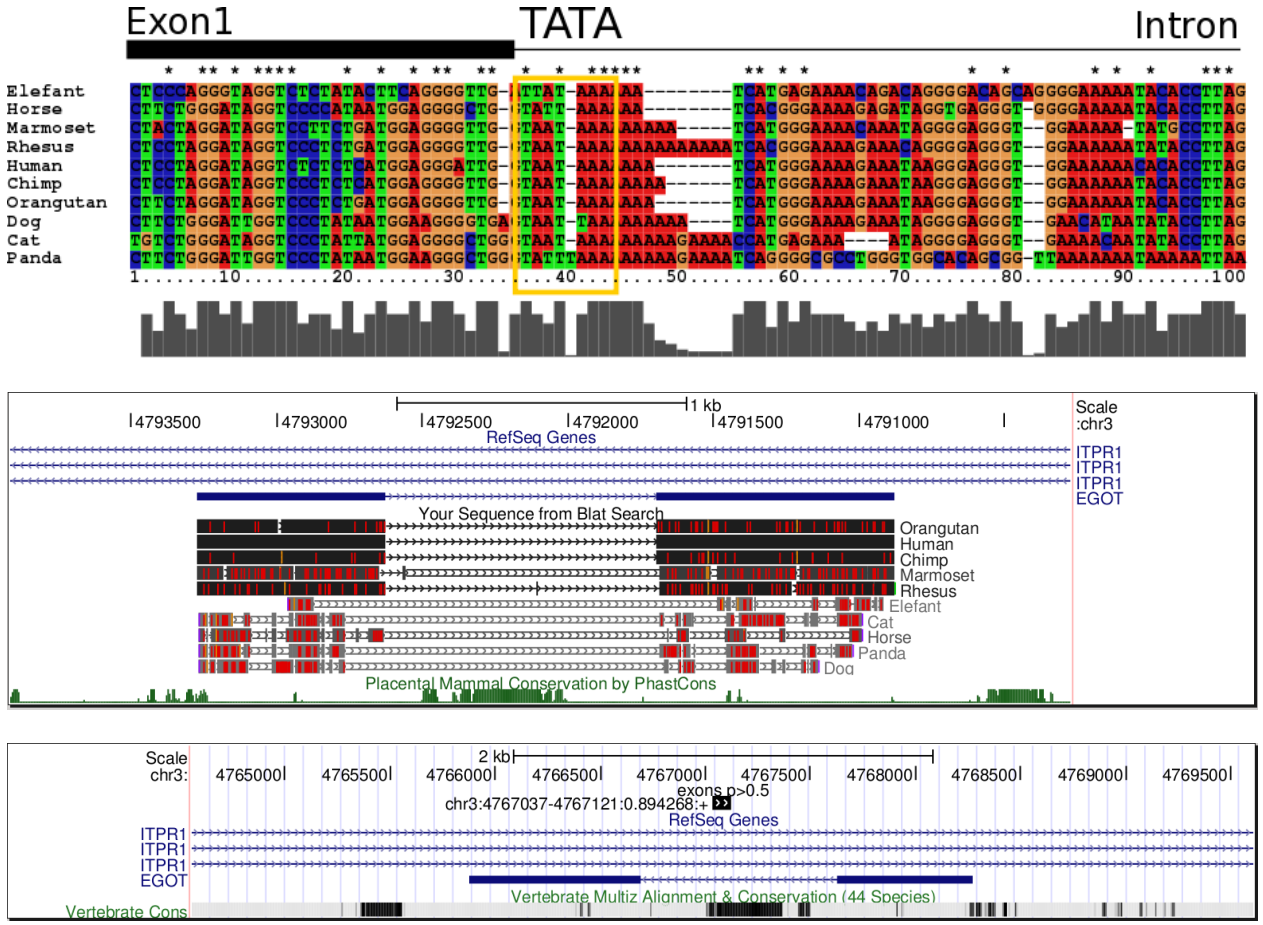




Long ncRNAs: EGO-B



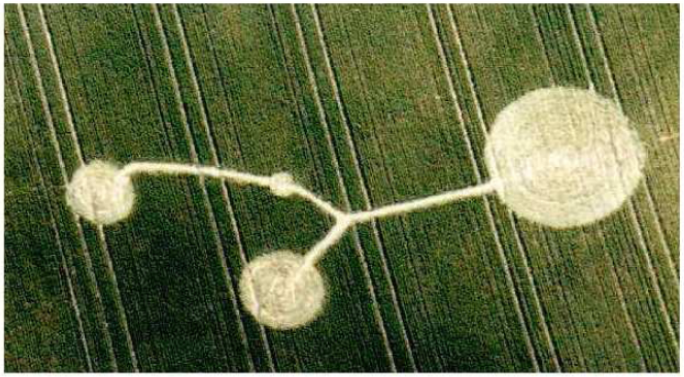
Long ncRNAs: EGO-B

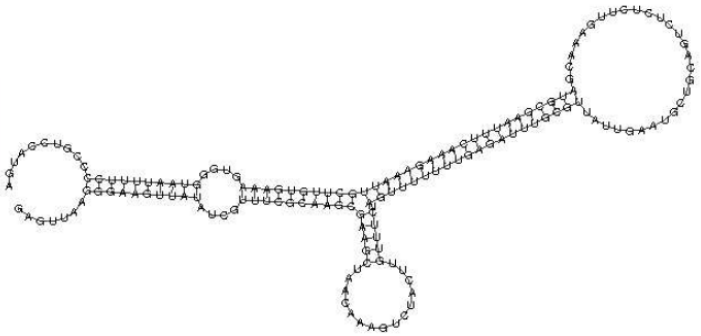


. . . which brings us full circle with regard to our initial topic, predicting novel transcripts via conserved splice sites.

Acknowledgements

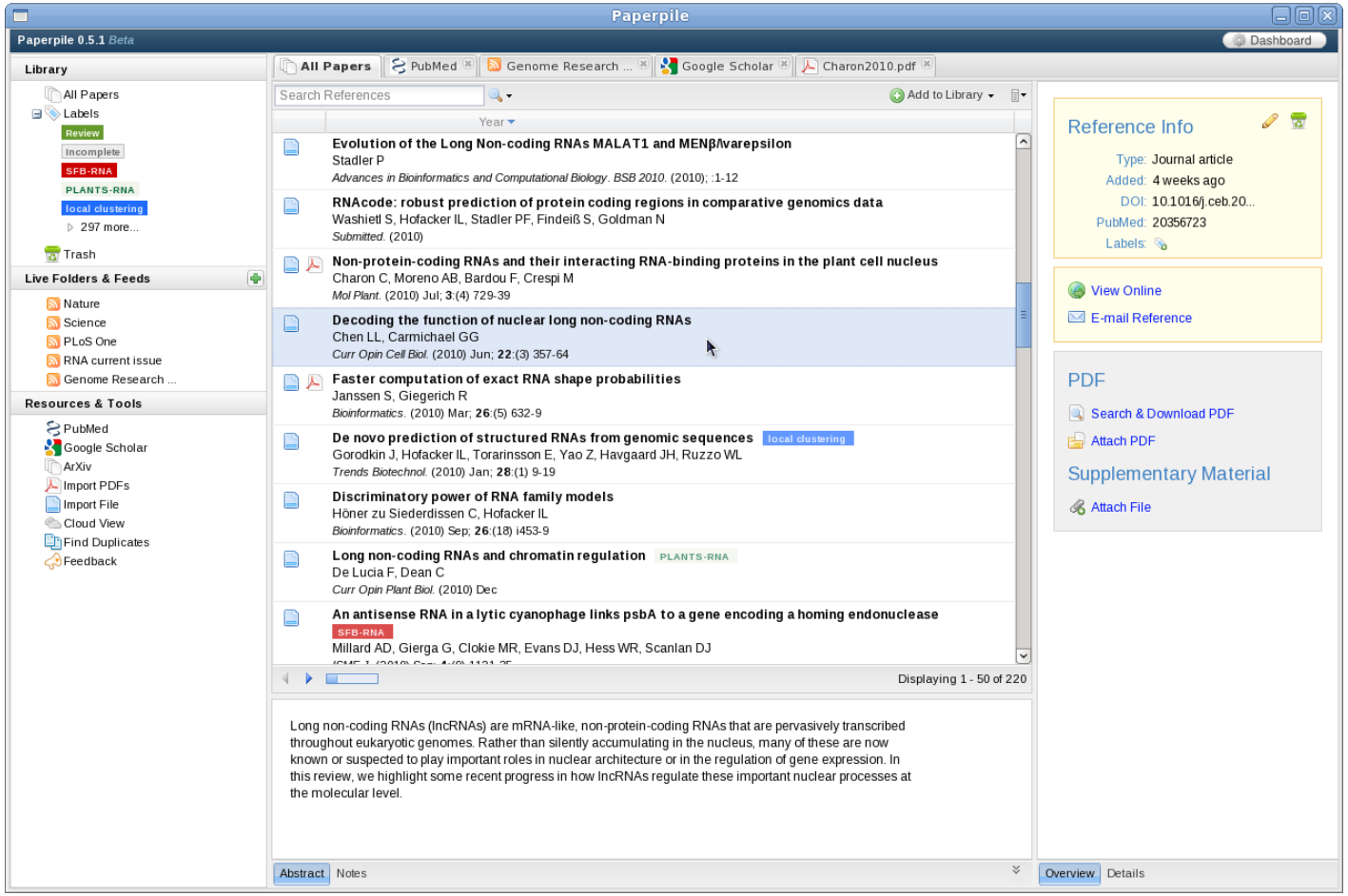
- Leipzig (Peter F. Stadler) $\leftrightarrow$ Freiburg (Rolf Backofen)
- Michael Hiller (Stanford University)





The truth is out there

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



http://paperpile.com/beta/

https://github.com/wash/paperpile