100s of novel conserved short ORFs?

Irwin Jungreis

2016-10-04

Massachusetts
Institute of
Technology

Summary

- New PhyloCSF browser tracks
  - Novel coding predictions
  - Splice predictions

- 100s of novel conserved short ORFs?
  An analysis of Mackowiak-2015.

- A proposal for discussion:
  Ranking lncRNAs by coding potential

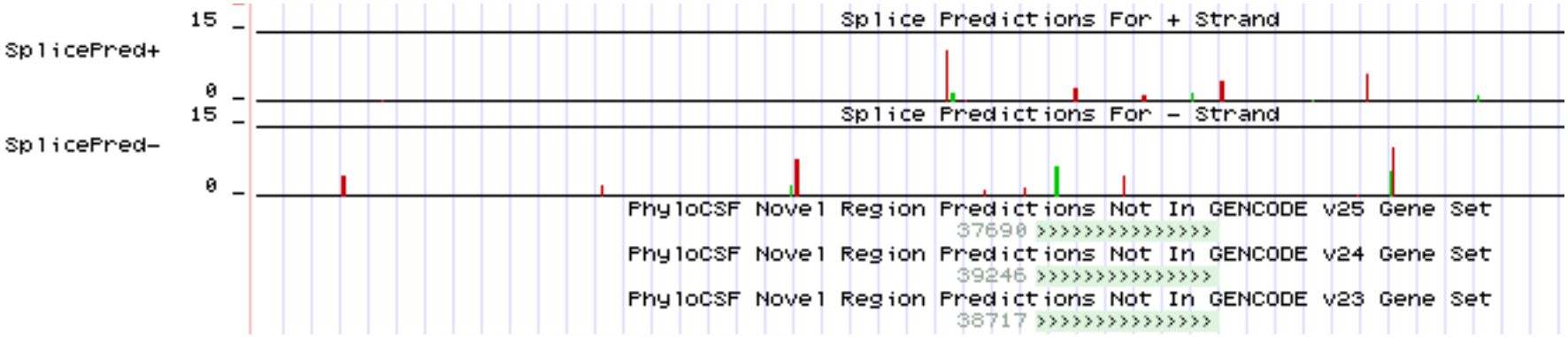
Summary

- **New PhyloCSF browser tracks**
  - **Novel coding predictions**
  - **Splice predictions**

- 100s of novel conserved short ORFs?
  An analysis of Mackowiak-2015.

- A proposal for discussion:
  Ranking lncRNAs by coding potential

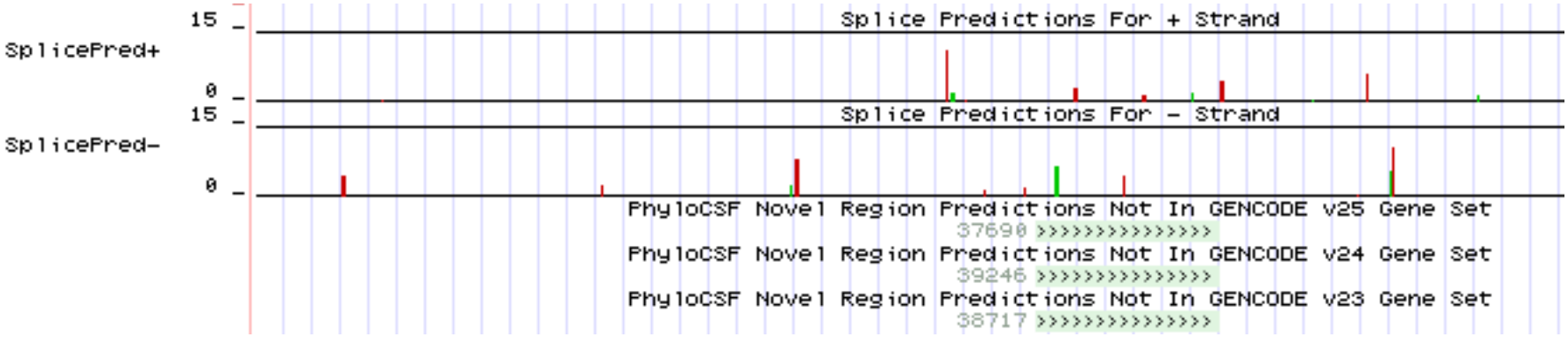
New PhyloCSF Tracks – Novel Regions



Novel Region Prediction tracks

- Relative to specific GENCODE version
- Available for:
  - hg38: GENCODE 23, 24, 25
  - hg19: GENOCODE 19
  - mm10: GENCODE M9
  - galGal4: Ensembl v4.82
  - AgamP4
- One track for both strands: green +, red –
- Intensity shows rank
- Click for position to be copied into CodAlignView (automate?)

New PhyloCSF Tracks – Splice Predictions



Splice Prediction tracks

- Use to predict ends of novel regions
- One track per strand
- Splice donors: green, Splice Acceptors: red
- Height indicates prediction score

New PhyloCSF Tracks – Example

- PhyloCSF predicts novel coding exon on chicken chr5.
- Strong splice predictions on + strand near start and end
- In this case, EST evidence confirms prediction

Summary

- New PhyloCSF browser tracks
  - Novel coding predictions
  - Splice predictions

- **100s of novel conserved short ORFs? An analysis of Mackowiak-2015.**

- A proposal for discussion:
  Ranking lncRNAs by coding potential

Mackowiak *et al. Genome Biology* (2015) 16:179
DOI 10.1186/s13059-015-0742-x

Genome Biology

RESEARCH Open Access

Extensive identification and analysis of conserved small ORFs in animals

Sebastian D. Mackowiak[1], Henrik Zauber[1], Chris Bielow[1,2], Denise Thiel[1], Kamila Kutz[1], Lorenzo Calviello[1], Guido Mastrobuoni[1], Nikolaus Rajewsky[1], Stefan Kempa[1], Matthias Selbach[1] and Benedikt Obermayer[1*]

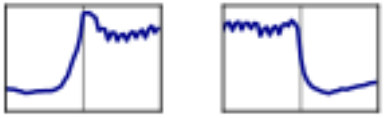
- Searched for novel *conserved* ORFs 8-100 AAs
  - human, mouse, zebrafish, fly, worm
- Primary tool: PhyloCSF
- Found about 2000, including 831 in human
  - lincRNA: 354
  - 3'-UTR: 229
  - 5'-UTR: 118
  - Other: 130
- Confirmed a few using mass spec and riboseq
- Considered stop codon readthrough and pseudogenes
- Analyzed properties of novel peptides

Are these real novel conserved sORFs?

- We focus on the 831 found in human
- Principles apply in other species as well

Mackowiak Pipeline

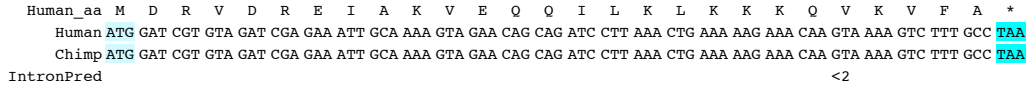
- Search for every ORF 8-100 AAs in every frame in:
  - GENCODE v19 coding and noncoding transcripts
  - 2 other lincRNA databases
- Exclude pseudogenes and overlaps with annotated coding regions
- Extract alignments using 46-vertebrates (hg19)
- Classify using SVM
  - Features:
    - PhyloCSF (most informative feature)
    - Frame conservation
    - Nucleotide level conservation profiles at start and end



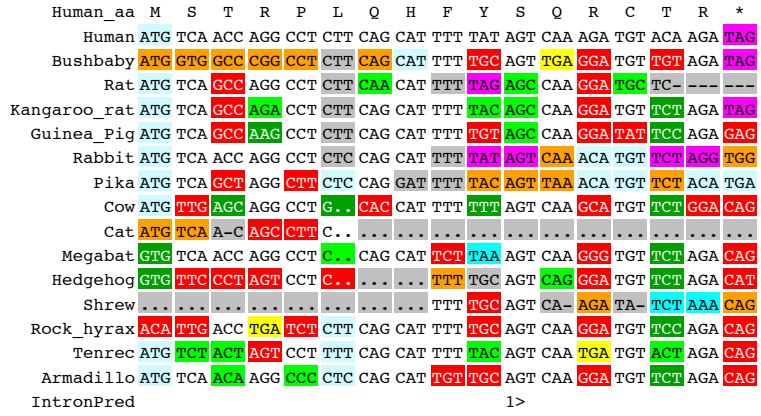
  - Training sets:
    - Positive: annotated Swiss-Prot sORFs with PhyloCSF > 0
    - Negative: sORFs in GENCODE noncoding transcripts other than lncRNAs
  - Cross-validated performance on training set:
    - 1-5% false negative rate
    - 0.1-0.5% false positive rate

Examined 20 chosen at random

- 1 looks good (has since been annotated)
- 19 are clearly not conserved coding ORFs
  - No alignment beyond apes (4)
  - Poorly conserved ORF (10)
    - non-conserved start and stop, internal stops, frame shifts
  - Antisense (1)
  - Pseudogene (1)
  - Real coding but ORF extends upstream of ATG (2)



TCONS_l2_00016218_chr20:25733316-25733399:+

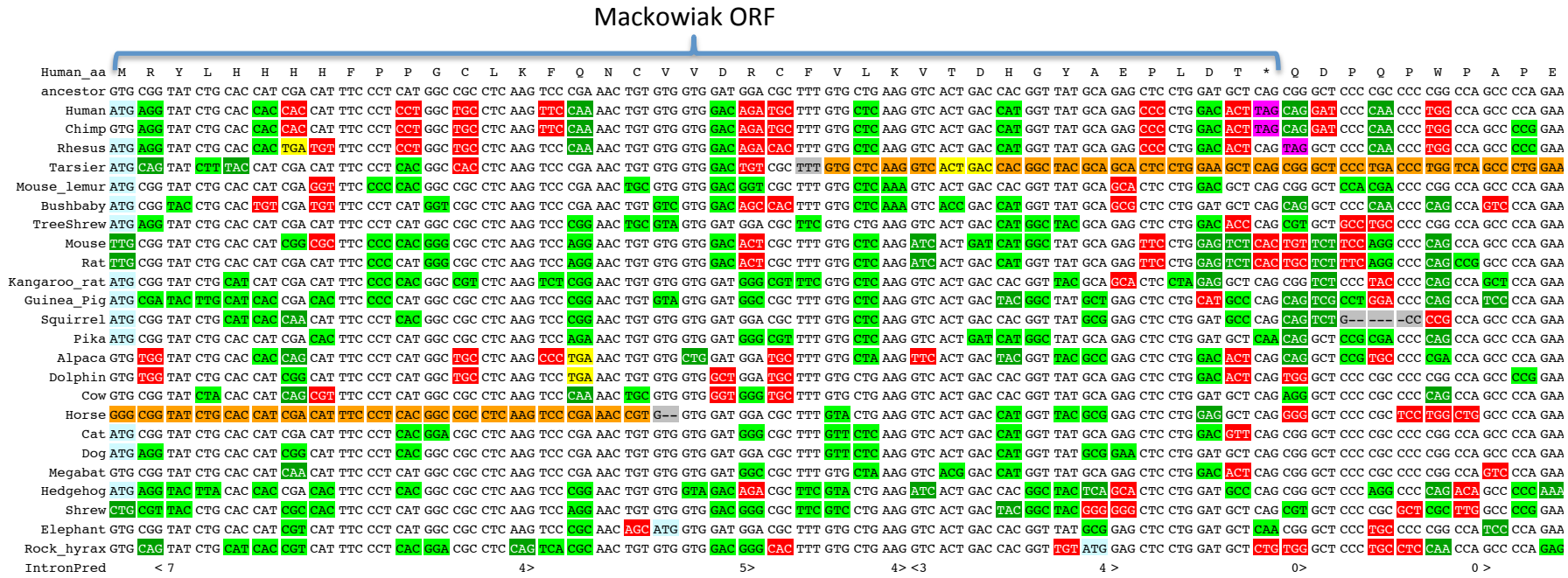


ENST00000321521_chr5:102544888-102544938:+

What went wrong?

- False Positive rate 0.1% - 0.5% not low enough
  - I estimate >750,000 ORF candidates
  - -> 750 - 3750 false positives
  - No reason to expect *any* of their 831 positives to be real
  - Probably set cutoff using MAE, but should have used weighted MAE
- Did not exclude ORFs with low alignment depth
  - PhyloCSF score is not meaningful without enough alignment
- No pseudogenes in SVM training set.
- No antisense regions in SVM training set.
- No regions starting at downstream ATGs in SVM training set.
  - Some of their sORFs are part of longer novel coding regions
- Did not check individual regions for stop codon readthrough
  - At least 3 of their sORFs are known readthrough regions
- Used PhyloCSF strategy “omega” instead of more accurate “mle”
  - Fewer than half of their sORFs have “mle” score > 0

Example: pseudogene

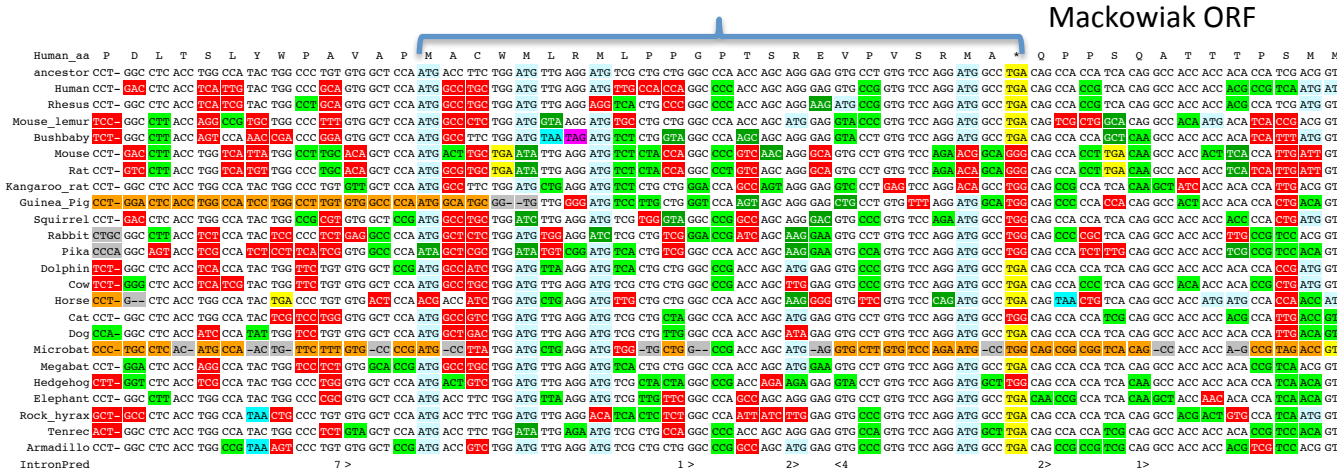


ENST00000525328_chr11:76423734-76423856:-

- Coding appears to continue beyond recent stop codon
- Looks like a unitary pseudogene
- Transcript ENST00000525328
  - lincRNA in v24
  - transcribed pseudogene in v25

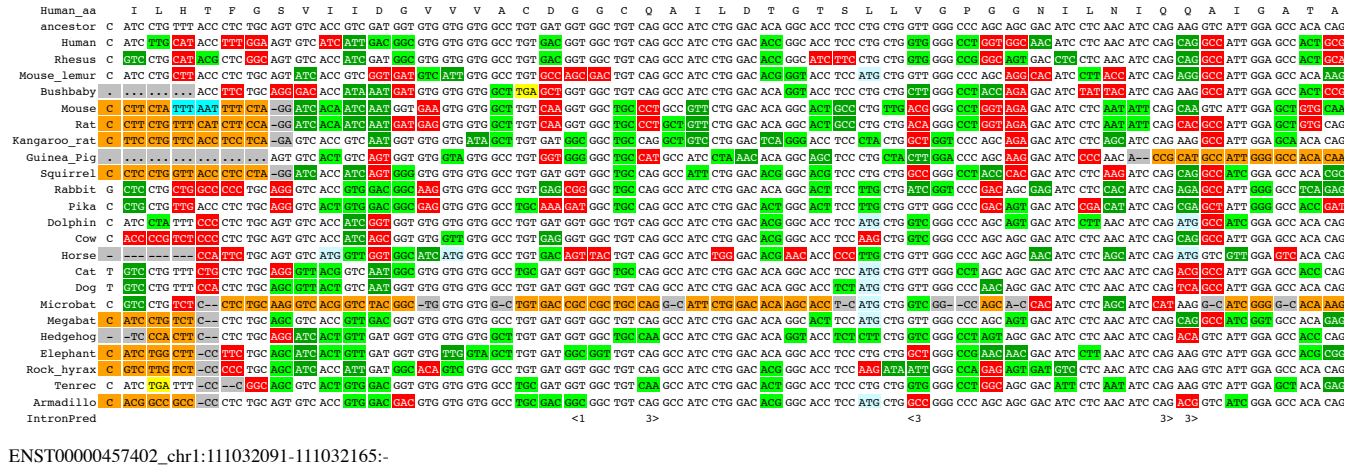
Example: antisense

Mackowiak ORF



ENST00000457402_chr1:111032091-111032165:-

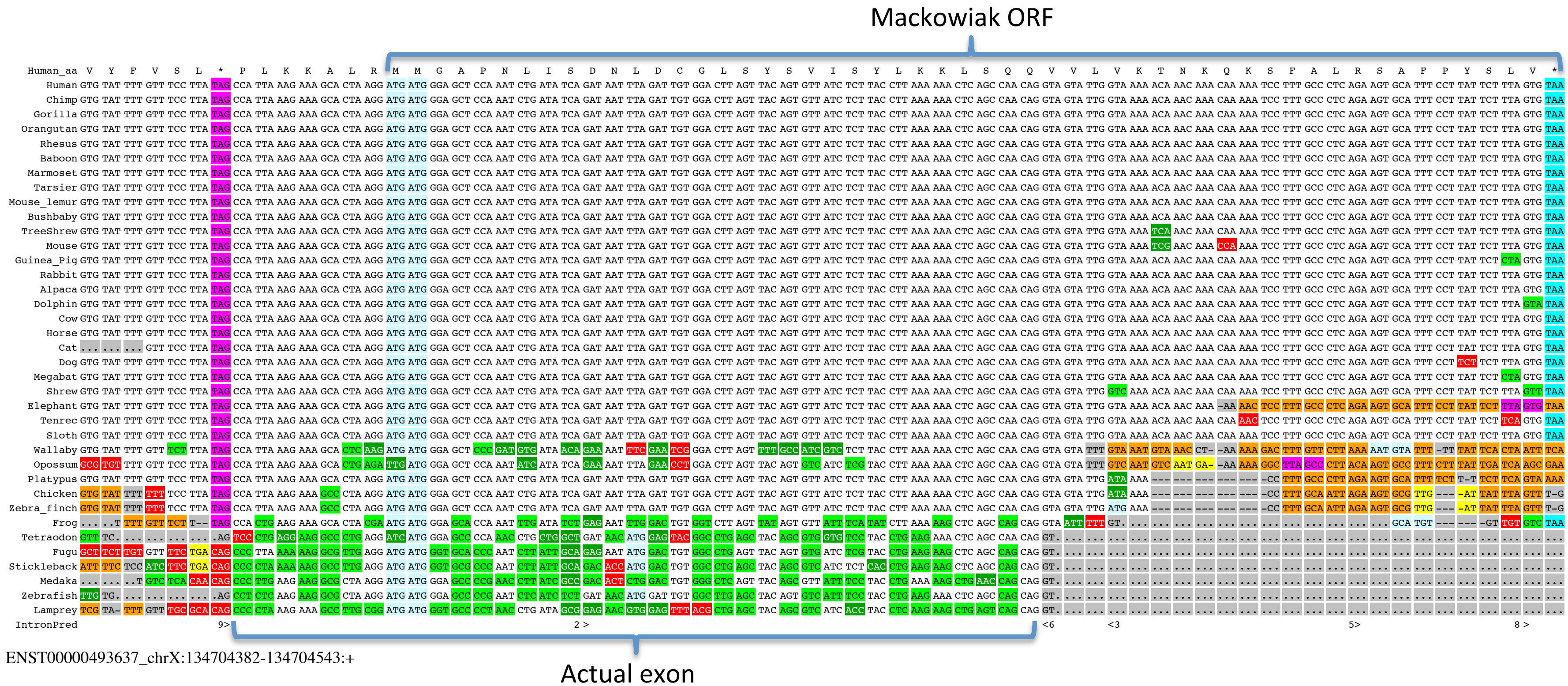
Antisense



ENST00000457402_chr1:111032091-111032165:-

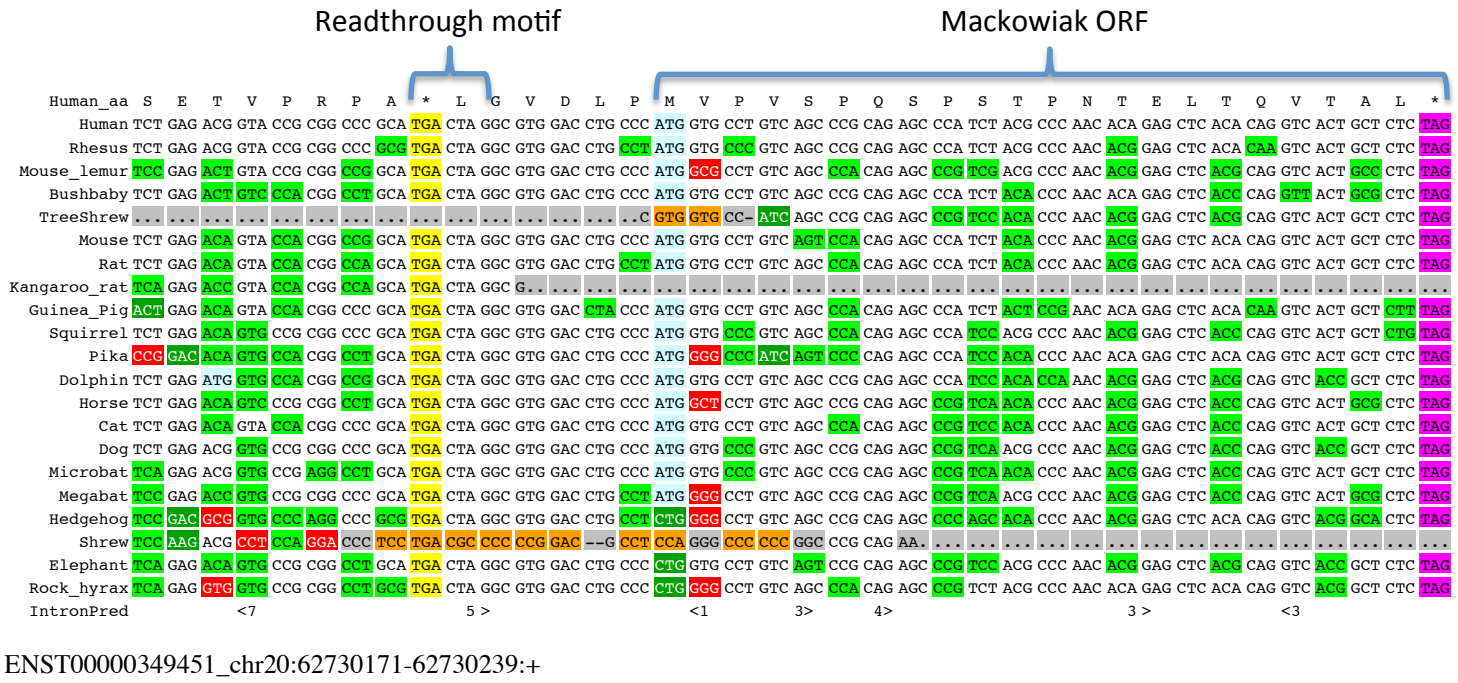
- Mackowiak proposed ORF:
  - many in-frame stops
  - poorly conserved start and stop.
- Antisense region is longer and has better score
- Antisense region is now annotated as unitary pseudogene CYMP

Example: upstream extension



- Coding signature extends upstream of Mackowiak ATG, and stops at predicted splice
- Splice sites supported by EST
- Novel exon of DDX26B
- Strong synonymous constraint in reptiles, birds, and mammals
- Synonymous constraint lost in marsupials (and in some softshell turtles but not others)
- We would not have found this one:
  - PhyloCSF regions were based on placental mammals
  - Synonymous constraint in placental mammals -> low PhyloCSF
  - Consider for future: PhyloCSF regions using full vertebrate alignments

Example: Known stop codon readthrough



ENST00000349451_chr20:62730171-62730239:+

- Mackowiak ORF in known readthrough region of OPRL1
- Readthrough motif TGA-CTAG is perfectly conserved
- ATG is not perfectly conserved

How many are true novel sORFs?

- Looks at alignments for 152 most promising
  - 25 real sORFs
    - 20 in lincRNAs
    - 5 in 5'-UTRs
    - 0 in 3'-UTRs
  - 18 maybes
- Unlikely to be very many in remaining 679
- Conclusion: a few dozen novel sORFs
  - Some additional novel coding regions, that are not sORFs
- Havana group looked at many lincRNAs, with similar conclusion
  - They had found most of the novel sORFs independently

Next steps

- This needs to be corrected:
  - Contradicts GENCODE count of protein-coding genes
  - We should make PhyloCSF issues explicit
- Several paragraphs in protein-coding genes paper.
- Pubmed comment
- Contact the editor
- Contact authors.
- Does anyone know these authors at Max Delbruck Center for Molecular Medicine in Berlin?
  - Sebastian Mackowiak
  - Benedikt Obermayer
  - Nikolaus Rajewesky

Summary

- New PhyloCSF browser tracks
  - Novel coding predictions
  - Splice predictions

- 100s of novel conserved short ORFs?
  An analysis of Mackowiak-2015.

- **A proposal for discussion:**
  Ranking lncRNAs by coding potential

Discussion with John Rinn

- Controversy about whether lincRNAs are coding continues
  - Some are claiming a large fraction are coding
- GENCODE is *perceived* as oblivious to the controversy
- He suggests:
  - Determine best PhyloCSF-based metric for scoring coding potential of lincRNAs (mouse and human)
  - GENCODE annotations will include rank for every GENCODE lincRNA
  - Publish a letter about this
  - Provide a tool for others to score their own lincRNA transcripts
  - Researchers interested in novel coding regions will look at top of list
  - Researchers interested in non-coding regions will look at bottom of list
- Thoughts?