

100s of novel conserved short ORFs?

Irwin Jungreis

2016-10-04



**Massachusetts
Institute of
Technology**

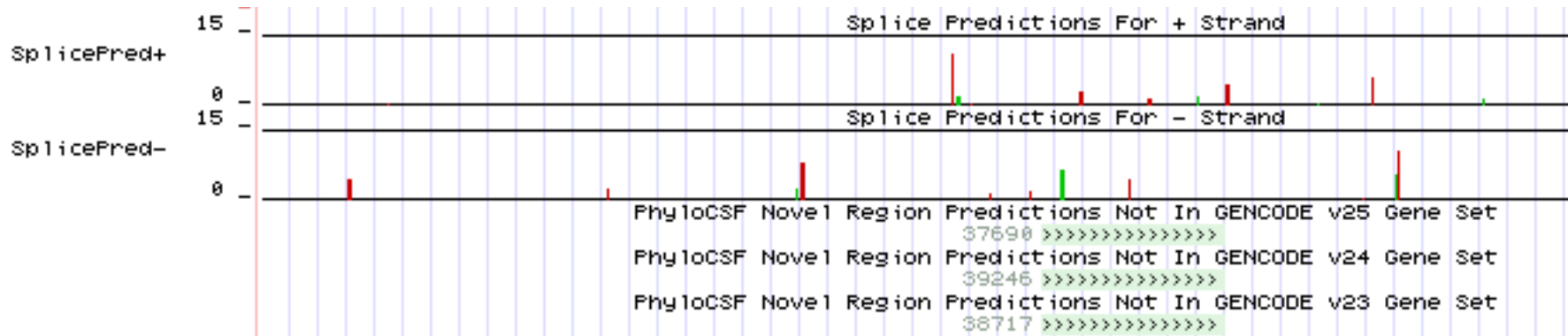
Summary

- 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

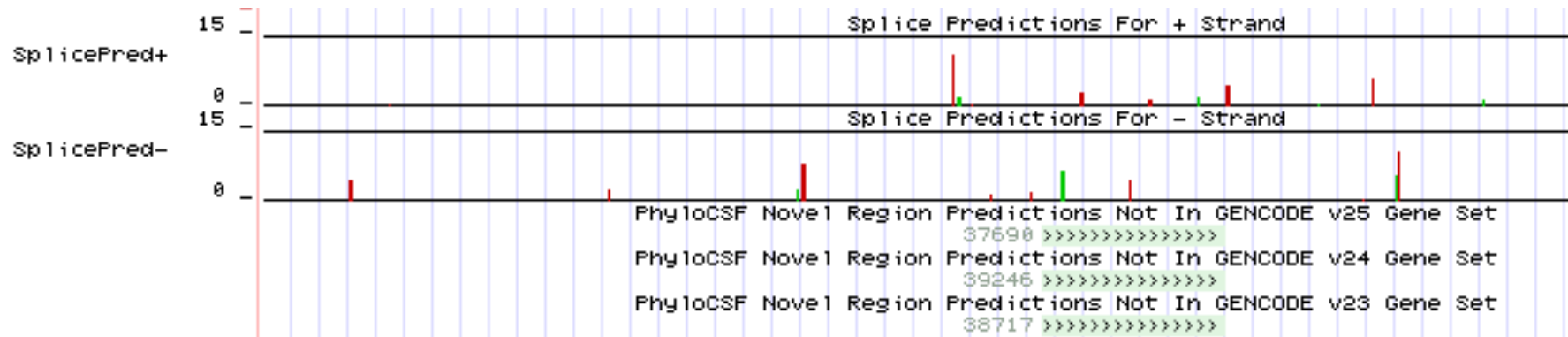
New PhyloCSF Tracks – Novel Regions



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?)

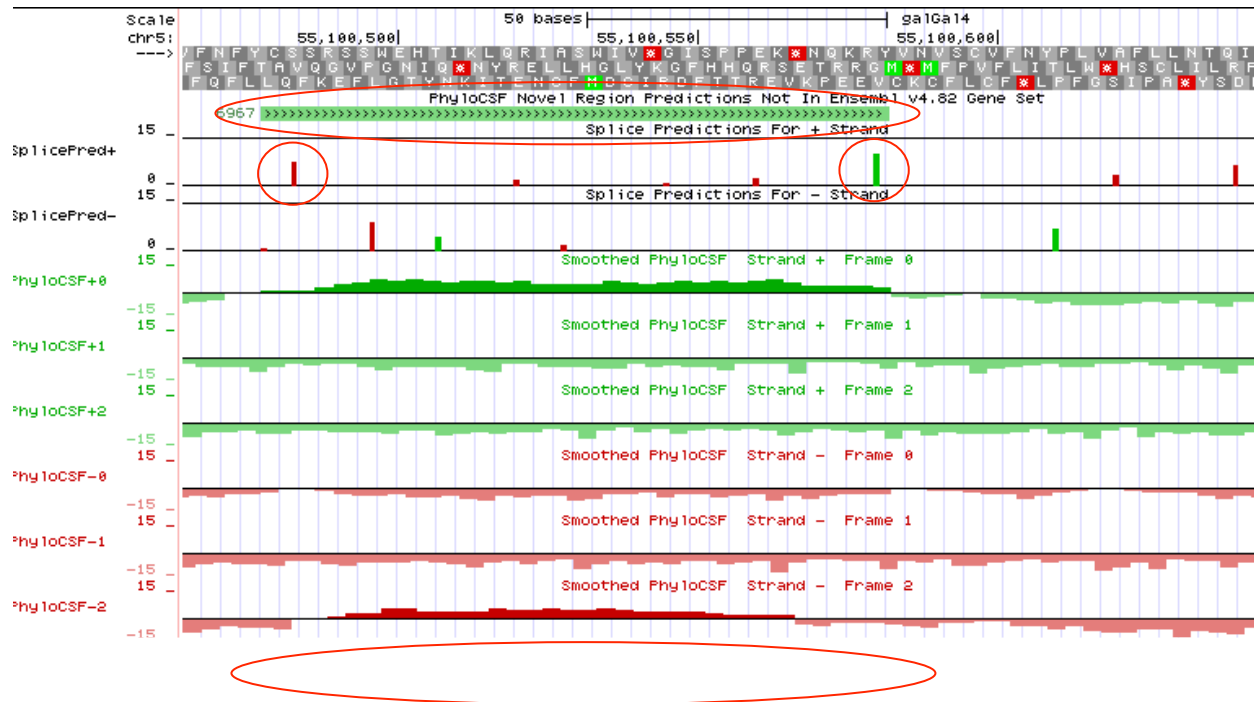
New PhyloCSF Tracks – Splice Predictions



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

RESEARCH

Open Access



Extensive identification and analysis of conserved small ORFs in animals

Sebastian D. Mackowiak¹, Henrik Zauber¹, Chris Bielow^{1,2}, Denise Thiel¹, Kamila Kutz¹, Lorenzo Calviello¹, Guido Mastrobuoni¹, Nikolaus Rajewsky¹, Stefan Kempa¹, Matthias Selbach¹ 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 lincRNAs
 - 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)

```

Human_aa M D R V D R E I A K V E Q Q I L K L K K K Q V K V F A *
Human ATG GAT CGT GTA GAT CGA GAA ATT GCA AAA GTA GAA CAG CAG ATC CTT AAA CTG AAA AAG AAA CAA GTA AAA GTC TTT GCC TAA
Chimp ATG GAT CGT GTA GAT CGA GAA ATT GCA AAA GTA GAA CAG CAG ATC CTT AAA CTG AAA AAG AAA CAA GTA AAA GTC TTT GCC TAA
IntronPred
                                <2
    
```

TCONS_I2_00016218_chr20:25733316-25733399:+

```

Human_aa M S T R P L Q H F Y S Q R C T R *
Human ATG TCA ACC AGG CCT CTT CAG CAT TTT TAT AGT CAA AGA TGT ACA AGA TAG
Bushbaby ATG GTG GCC CGG CCT CTT CAG CAT TTT TGC AGT TGA GGA TGT TGT AGA TAG
Rat ATG TCA GCC AGG CCT CTT CAA CAT TTT TAG AGC CAA GGA TGC TC- --- ---
Kangaroo_rat ATG TCA GCC AGA CCT CTT CAG CAT TTT TAC AGC CAA GGA TGT TCT AGA TAG
Guinea_pig ATG TCA GCC AAG CCT CTT CAG CAT TTT TGT AGC CAA GGA TAT TCC AGA GAG
Rabbit ATG TCA ACC AGG CCT CTT CAG CAT TTT TAT AGT CAA ACA TGT TCT AGG TGG
Pika ATG TCA GCT AGG CTT CTC CAG GAT TTT TAC AGT TAA ACA TGT TCT ACA TGA
Cow ATG TTG AGC AGG CCT G.. CAC CAT TTT TTT AGT CAA GCA TGT TCT GGA CAG
Cat ATG TCA A-C AGC CTT C.. .....
Megabat GTG TCA ACC AGG CCT C.. CAG CAT TCT TAA AGT CAA GGG TGT TCT AGA CAG
Hedgehog GTG TTC CCT AGT CCT C.. ..... TTT TGC AGT CAG GGA TGT TCT AGA CAT
Shrew ..... TTT TGC AGT CA- AGA TA- TCT AAA CAG
Rock_hyrax ACA TTG ACC TGA TCT CTT CAG CAT TTT TGC AGT CAA GGA TGT TCC AGA CAG
Tenrec ATG TCT ACT AGT CCT TTT CAG CAT TTT TAC AGT CAA TGA TGT ACT AGA CAG
Armadillo ATG TCA ACA AGG CCG CTC CAG CAT TGT TGC AGT CAA GGA TGT TCT AGA CAG
IntronPred
                                1>
    
```

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

Mackowiak ORF

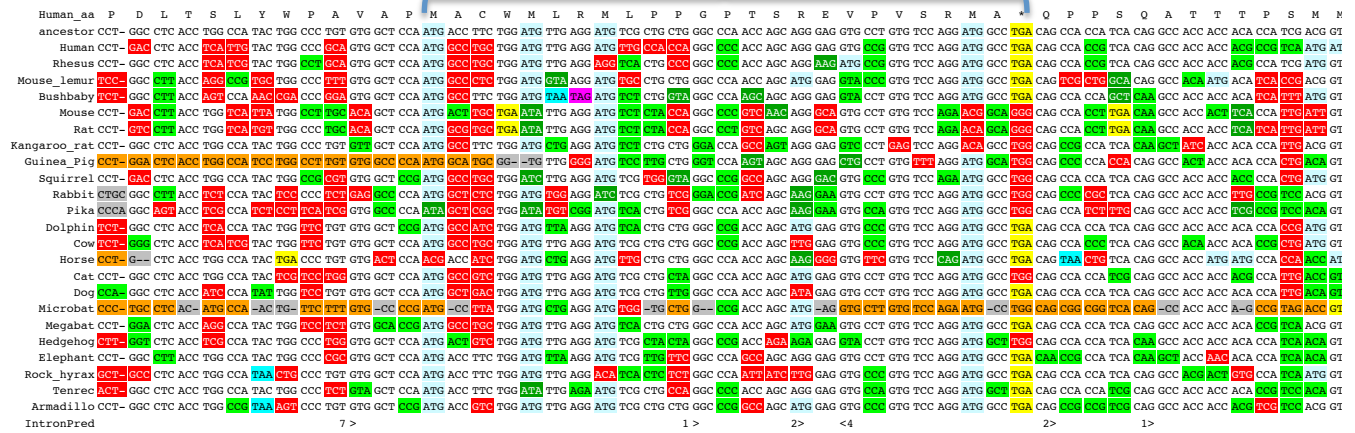
| Species | M | R | Y | L | H | H | H | F | P | P | G | C | L | K | F | Q | N | C | V | V | D | R | C | F | V | L | K | V | T | D | H | G | Y | A | E | P | L | D | T | * | Q | D | P | Q | P | W | P | A | P | E | | | | | | |
|------------|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|-----|
| Human_aa | M | R | Y | L | H | H | H | F | P | P | G | C | L | K | F | Q | N | C | V | V | D | R | C | F | V | L | K | V | T | D | H | G | Y | A | E | P | L | D | T | * | Q | D | P | Q | P | W | P | A | P | E | | | | | | |
| ancest | GTG | CGG | TAT | CTG | CAC | CAT | CGA | CAT | TTC | CCT | CAT | GGC | CGC | CTC | AAG | TCC | CGA | AAC | TGT | GTG | GTG | GAT | GGA | CGC | TTT | GTG | CTG | AAG | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GAG | CTC | CTG | GAT | GCT | CAG | CGG | GCT | CCC | CGC | CCC | CGG | CCA | GCC | CCA | GAA | | | | | |
| Human | ATG | AGG | TAT | CTG | CAC | CAC | CAC | CAT | TTC | CCT | CCT | GGC | TGC | CTC | AAG | TTC | CAA | AAC | TGT | GTG | GTG | GAG | AGA | TGC | TTT | GTG | CTC | AAG | GTC | ACT | GAC | CAT | GGT | TAT | GCA | GAG | CCC | CTG | GAC | ACT | TAG | CAG | GAT | CCC | CAA | CCC | TCG | CCA | GCC | CCA | GAA | | | | | |
| Chimp | GTG | AGC | TAT | CTG | CAC | CAC | CAC | CAT | TTC | CCT | CCT | GGC | TGC | CTC | AAG | TTC | CAA | AAC | TGT | GTG | GTG | GAG | AGA | TGC | TTT | GTG | CTC | AAG | GTC | ACT | GAC | CAT | GGT | TAT | GCA | GAG | CCC | CTG | GAC | ACT | TAG | CAG | GAT | CCC | CAA | CCC | TCG | CCA | GCC | CCG | GAA | | | | | |
| Rhesus | ATG | AGC | TAT | CTG | CAC | CAC | TGA | TGT | TTC | CCT | CCT | GGC | TGC | CTC | AAG | TCC | CAA | AAC | TGT | GTG | GTG | GAG | AGA | CAC | TTT | GTG | CTC | AAG | GTC | ACT | GAC | CAT | GGT | TAT | GCA | GAG | CCC | CTG | GAC | ACT | CAG | TAG | GCT | CCC | CAA | CCC | TCG | CCA | GCC | CCG | GAA | | | | | |
| Tarsier | ATG | CAG | TAT | CTT | TAC | CAT | CGA | CAT | TTC | CCT | CAC | GGC | CAC | CTC | AAG | TCC | CGA | AAC | TGT | GTG | GTG | GAG | TGT | CGC | TTT | GTG | CTC | AAG | GTC | ACT | GAC | CAC | GGC | TAC | GCA | GCA | CTC | CTG | GAA | GCT | CAG | CGG | SCT | CCC | TGA | CCC | TCG | TCA | GCC | CTG | GAA | | | | | |
| Mouse | lemur | ATG | CGG | TAT | CTG | CAC | CAT | CGA | GCT | TTC | CCC | CAC | GGC | CGC | CTC | AAG | TCC | CGA | AAC | TGT | GTG | GTG | GAG | GGT | CGC | TTT | GTG | CTC | AAA | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GCA | CTC | CTG | GAG | GCT | CAG | CGG | GCT | CCA | CGA | CCC | CGG | CCA | GCC | CCA | GAA | | | | |
| Bushbaby | ATG | CGG | TAC | CTG | CAC | TGT | CGA | TGT | TTC | CCT | CAT | GGT | CGC | CTC | AAG | TCC | CGA | AAC | TGT | GTG | GTG | GAG | AGC | CAC | TTT | GTG | CTC | AAA | GTC | ACC | GAC | CAT | GGT | TAT | GCA | GAG | GGC | CTC | CTG | GAT | GCT | CAG | CGC | GCT | CCC | CAA | CCC | CAG | CCA | GTC | CCA | GAA | | | | |
| TreeShrew | ATG | AGC | TAT | CTG | CAC | CAT | CGA | CAT | TTC | CCT | CAT | GGC | CGC | CTC | AAG | TCC | GGG | AAC | TGT | GTG | GTG | GAG | STA | GTG | GAT | GGA | CGC | TTG | GTG | CTG | AAG | GTC | ACT | GAC | CAT | GGC | TAT | GCA | GAG | CTC | CTG | GAG | ACC | CAG | TAG | GCT | CCC | TCG | CCC | CGG | CCA | GCC | CCA | GAA | | |
| Mouse | rat | ATG | CTG | CGG | TAT | CTG | CAC | CAT | CGG | CGC | TTC | CCC | CAC | GGG | CGC | CTC | AAG | TCC | AGG | AAC | TGT | GTG | GTG | GAG | ACT | CGC | TTT | GTG | CTC | AAG | ATC | ACT | GAT | CAT | GGC | TAT | GCA | GAG | TTG | CTG | GAG | TCT | CAG | TGT | TCT | TCC | AGG | CCC | CAG | CCA | GCC | CCA | GAA | | | |
| Rat | ATG | CGG | TAT | CTG | CAC | CAT | CGA | CAT | TTC | CCC | CAT | GGG | CGC | CTC | AAG | TCC | AGG | AAC | TGT | GTG | GTG | GAG | ACT | CGC | TTT | GTG | CTC | AAG | ATC | ACT | GAC | CAT | GGT | TAT | GCA | GAG | TTG | CTG | GAG | TCT | CAG | TGG | TCT | TTC | AGG | CCC | CAG | CCG | GCC | CCA | GAA | | | | | |
| Kangaroo | rat | ATG | CGG | TAT | CTG | CAT | CAT | CGA | CAT | TTC | CCC | CAC | GGC | CGT | CTC | AAG | TCT | CGG | AAC | TGT | GTG | GTG | GAT | GGG | DGT | TTG | GTG | CTC | AAG | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GAG | CTC | CTG | GAT | GCT | CAG | CGG | TCT | CCC | TAC | CCC | CAG | CCA | GCT | CCA | GAA | | | | |
| Guinea | Pig | ATG | DGA | TAC | TTG | CAT | CAC | CAC | TTC | CCC | CAT | GGC | CGC | CTC | AAG | TCC | CGG | AAC | TGT | GTG | GAT | GGC | CGC | TTT | GTG | CTC | AAG | GTC | ACT | GAC | TAC | GGC | TAT | GCT | GAG | CTC | CTG | CAT | GGC | CAG | CAG | TGC | SCT | GGG | CCC | CAG | CCA | TCC | CCA | GAA | | | | | | |
| Squirrel | ATG | CGG | TAT | CTG | CAT | CAC | CAA | CAT | TTC | CCT | CAC | GGC | CGC | CTC | AAG | TCC | CGG | AAC | TGT | GTG | GTG | GAT | GGA | CGC | TTT | GTG | CTC | AAG | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GAG | CTC | CTG | GAT | SCT | CAG | CAG | TCT | G-- | -- | -- | CCG | CCA | GCC | CCA | GAA | | | | | |
| Pika | ATG | CGG | TAT | CTG | CAC | CAT | CGA | CAC | TTC | CCT | CAT | GGC | CGC | CTC | AAG | TCC | AGA | AAC | TGT | GTG | GTG | GAT | GGG | DGT | TTT | GTG | CTC | AAG | GTC | ACT | GAT | CAT | GGC | TAT | GCA | GAG | CTC | CTG | GAT | GCT | CAG | CAG | GCT | CCC | CGG | CCC | CAG | CCA | GCC | CCA | GAA | | | | | |
| Alpaca | GTG | TCG | TAT | CTG | CAC | CAC | CAG | CAT | TTC | CCT | CAT | GGC | TGC | CTC | AAG | CCC | TGA | AAC | TGT | GTG | CTG | GAT | GGA | TCG | TTT | GTG | CTA | AAG | TTG | ACT | GAC | TAC | GGT | TAT | GCA | GAG | CTC | CTG | GAG | ACT | CAG | CAG | GCT | CCC | TCG | CCC | CAG | CCA | GCC | CCA | GAA | | | | | |
| Dolphin | GTG | TCG | TAT | CTG | CAC | CAT | CGG | CAT | TTC | CCT | CAT | GGC | TGC | CTC | AAG | TCC | TGA | AAC | TGT | GTG | GTG | SCT | GGA | TCG | TTT | GTG | CTG | AAG | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GAG | CTC | CTG | GAG | ACT | CAG | TCG | GCT | CCC | CGC | CCC | CGG | CCA | GCC | CCG | GAA | | | | | |
| Cow | GTG | CGG | TAT | CTA | CAC | CAT | CAG | CGT | TTC | CCT | CAT | GGC | CGC | CTC | AAG | TCC | CAA | AAC | TGT | GTG | GTG | GGT | GGG | TCG | TTT | GTG | CTG | AAG | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GAG | CTC | CTG | GAT | GCT | CAG | AGG | GCT | CCC | CGC | CCC | CAG | CCA | GCC | CCA | GAA | | | | | |
| Horse | GGG | CGG | TAT | CTG | CAC | CAT | CGA | CAT | TTC | CCT | CAC | GGC | CGC | CTC | AAG | TCC | CGA | AAC | CGT | G-- | GTG | GAT | GGA | CGC | TTT | CTA | CTG | AAG | GTC | ACT | GAC | CAT | GGT | TAT | GCA | GAG | CTC | CTG | GAG | GCT | CAG | GGG | GCT | CCC | CGC | TCG | TCG | CTG | GCC | CCA | GAA | | | | | |
| Cat | ATG | CGG | TAT | CTG | CAC | CAT | CGA | CAT | TTC | CCT | CAC | GGG | CGC | CTC | AAG | TCC | CGA | AAC | TGT | GTG | GTG | GAT | GGG | CGC | TTT | CTT | CTC | AAG | GTC | ACT | GAC | CAT | GGT | TAT | GCA | GAG | CTC | CTG | GAG | GCT | CAG | CGG | GCT | CCC | CGC | CCC | CGG | CCA | GCC | CCA | GAA | | | | | |
| Dog | ATG | AGC | TAT | CTG | CAC | CAT | CGG | CAT | TTC | CCT | CAC | GGC | CGC | CTC | AAG | TCC | CGA | AAC | TGT | GTG | GTG | GAT | GGA | CGC | TTT | CTT | CTC | AAG | GTC | ACT | GAC | CAT | GGT | TAT | GCA | GAG | CTC | CTG | GAT | GCT | CAG | CGG | GCT | CCC | CGC | CCC | CGG | CCA | GCC | CCA | GAA | | | | | |
| Megabat | GTG | CGG | TAT | CTG | CAC | CAT | CAA | CAT | TTC | CCT | CAT | GGC | CGC | CTC | AAG | TCC | CGA | AAC | TGT | GTG | GTG | GAT | GGC | CGC | TTT | GTG | CTA | AAG | GTC | ACC | GAC | CAT | GGT | TAT | GCA | GAG | CTC | CTG | GAG | ACT | CAG | CGG | GCT | CCC | CGC | CCC | CGG | CCA | GTC | CCA | GAA | | | | | |
| Hedgehog | ATG | AGC | TAC | TTA | CAC | CAC | CGA | CAC | TTC | CCT | CAC | GGC | CGC | CTC | AAG | TCC | GGG | AAC | TGT | GTG | CTA | GAG | AGA | CGC | TTG | CTA | CTG | AAG | ATC | ACT | GAC | CAC | GGC | TAC | TCA | GCA | CTC | CTG | GAT | SCT | CAG | CGG | GCT | CCC | AGG | CCC | CAG | ACA | GCC | CCG | AAG | | | | | |
| Shrew | ATG | CGT | TAC | CTG | CAC | CAT | CGC | CAC | TTC | CCT | CAT | GGC | CGC | CTC | AAG | TCC | AGG | AAC | TGT | GTG | GTG | GAG | GGG | CGC | TTT | CTC | CTG | AAG | GTC | ACT | GAC | TAC | GGC | TAC | GGG | GGG | CTC | CTG | GAT | GCT | CAG | GGT | GCT | CCC | CGC | SCT | TCG | TCG | GCC | CCG | GAA | | | | | |
| Elephant | GTG | CGG | TAT | CTG | CAC | CAT | CGT | CAT | TTC | CCT | CAT | GGC | CGC | CTC | AAG | TCC | CGG | AAC | AGG | ATG | GTG | GAT | GGA | CGC | TTT | GTG | CTG | AAG | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GAG | CTC | CTG | GAT | GCT | CAA | CGG | GCT | CCC | TCG | CCC | CGG | CCA | TCC | CCA | GAA | | | | | |
| Rock_hyrax | ATG | CAG | TAT | CTG | CAT | CAC | CGT | CAT | TTC | CCT | CAC | GGG | CGC | CTC | CAG | TCA | GGG | AAC | TGT | GTG | GTG | GAG | GGG | CAC | TTT | GTG | CTG | AAG | GTC | ACT | GAC | CAC | GGT | TAT | GCA | GAG | CTC | CTG | GAT | GCT | CTC | TGG | GCT | CCC | TCG | CTC | CAA | CCA | GCC | CCA | GAG | | | | | |
| IntronPred | < 7 | | | | | | | | | | | 4 > | | | | | | | | | | | 5 > | | | | | | | | | | | 4 > < 3 | | | | | | | | | | | 4 > | | | | | | | | | | | 0 > |

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

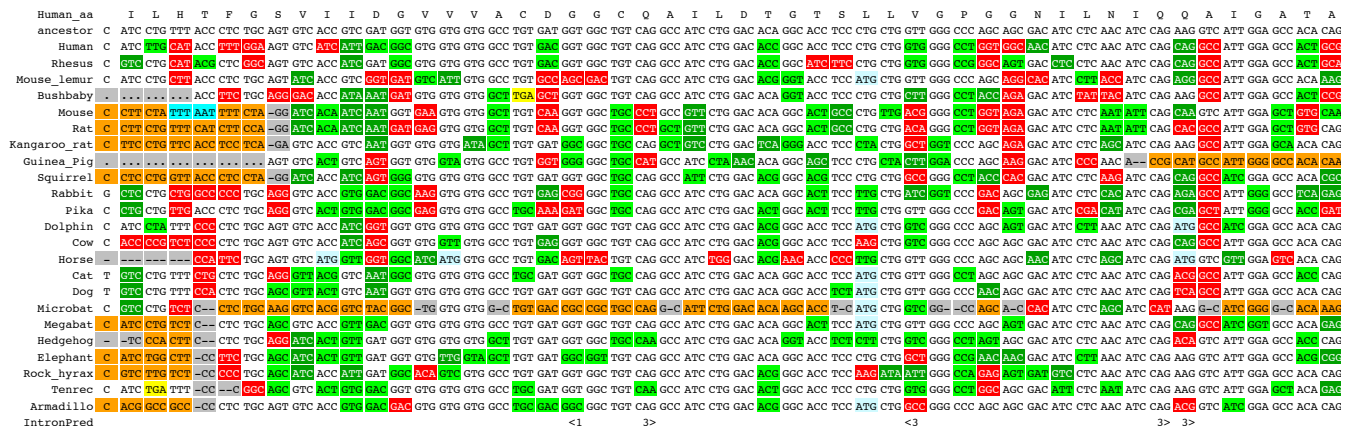
Example: antisense

Mackowiak ORF



Mackowiak ORF

ENST00000457402_chrl:111032091-111032165:-



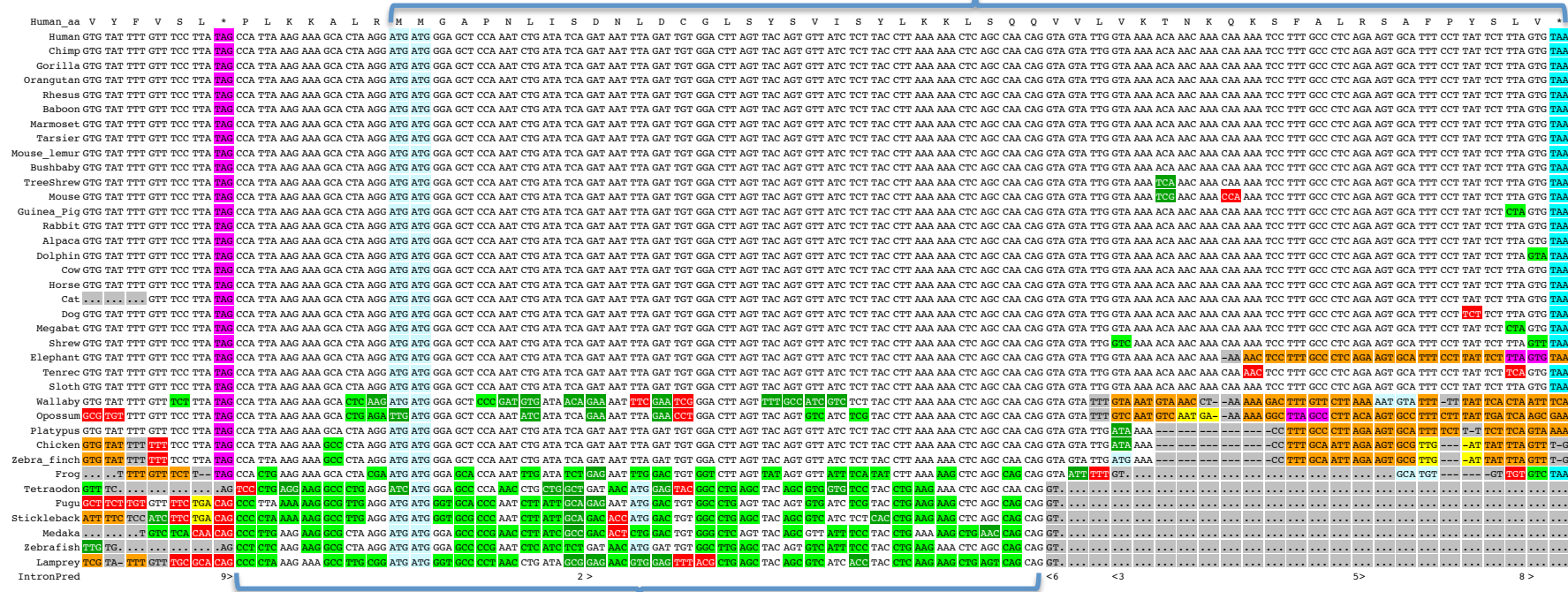
Antisense

ENST00000457402_chrl: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

Mackowiak ORF



ENST00000493637_chrx1:134704382-134704543+

Actual exon

- 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

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?