HAVANA gene annotation updates

Mouse genomes meeting

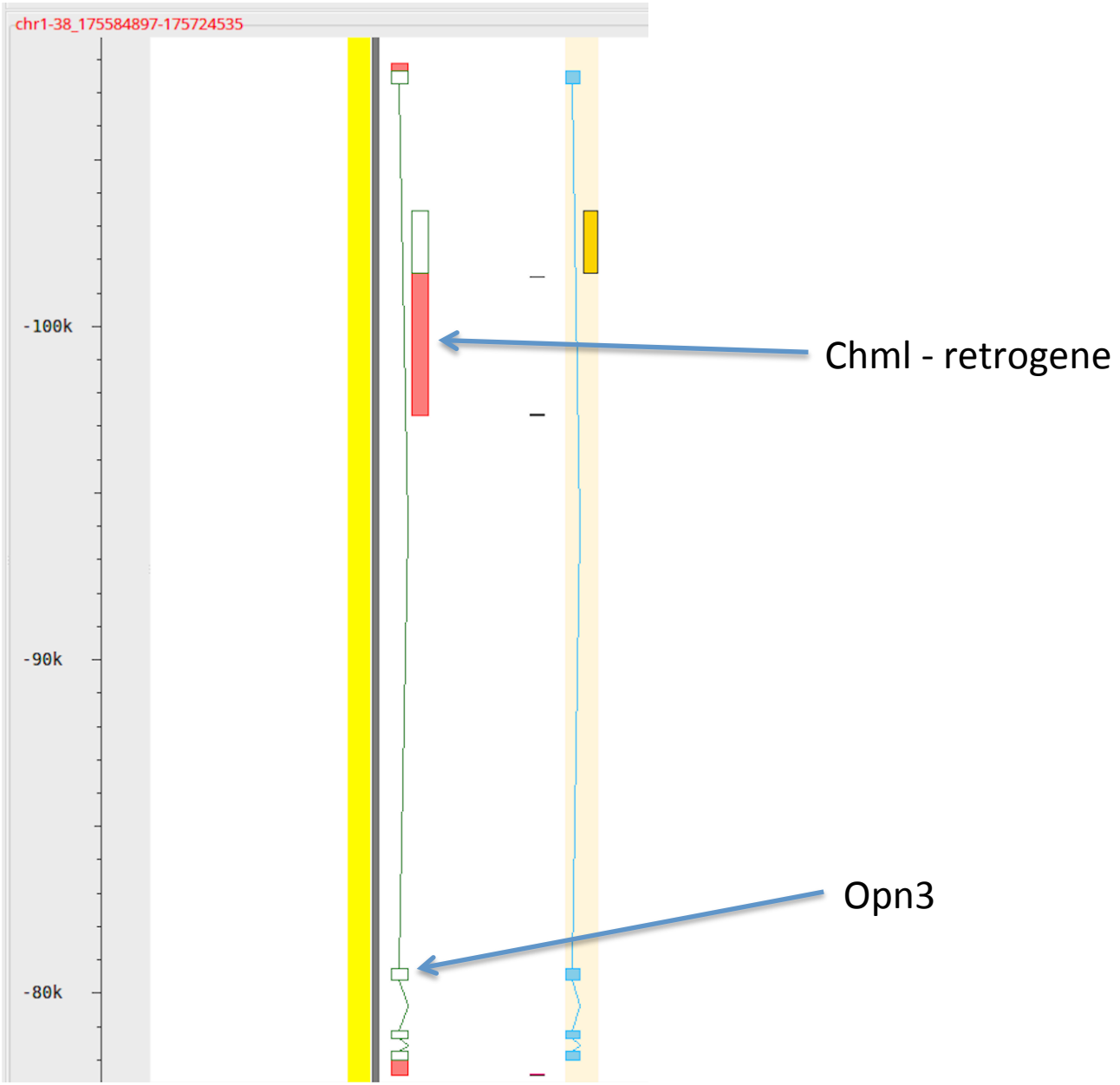
10/12/2015

Charlie Steward

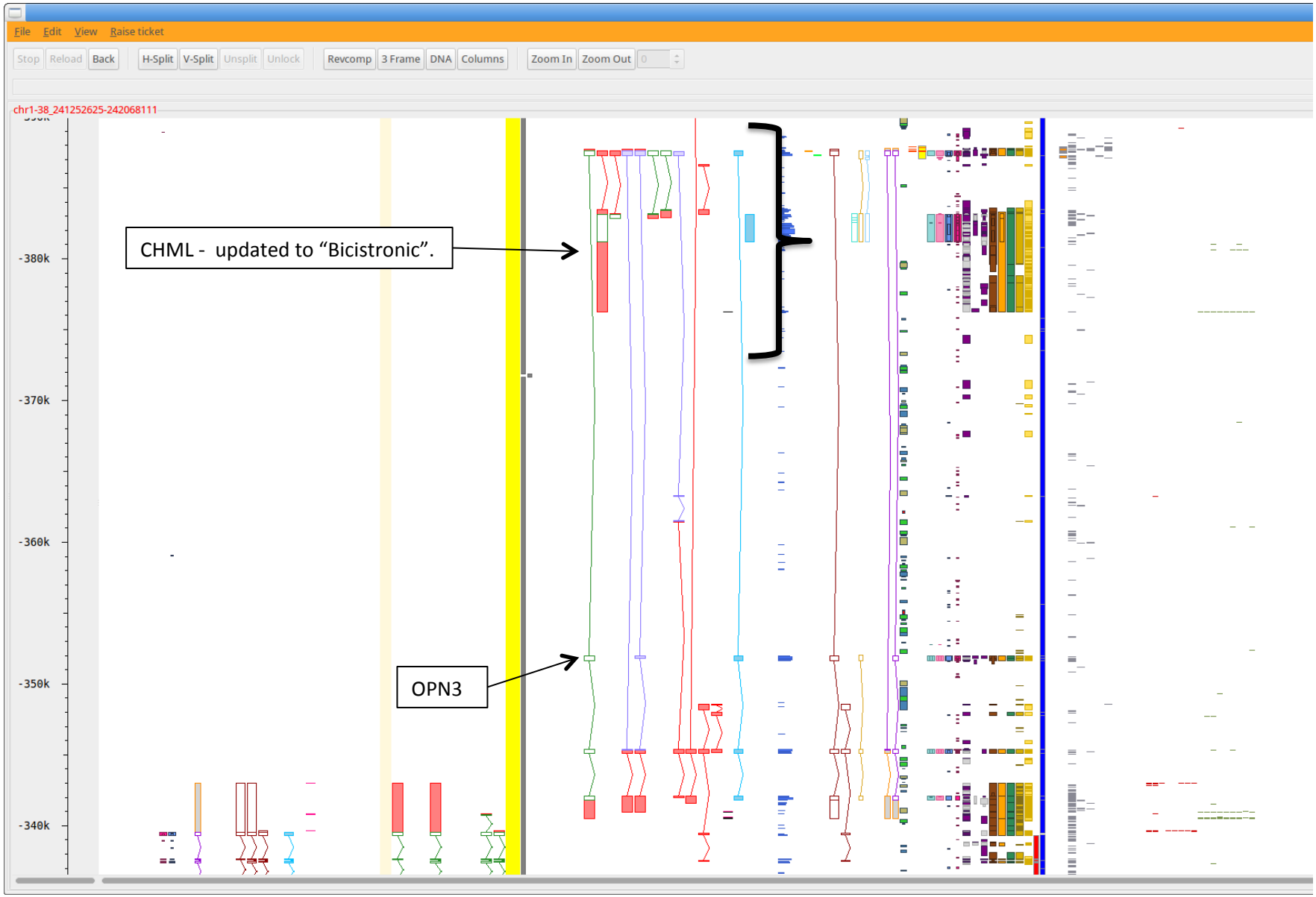
1) Black 6 reference updates

- Examples of what we have found so far from the gencode_novel_unconfident.gff file ...

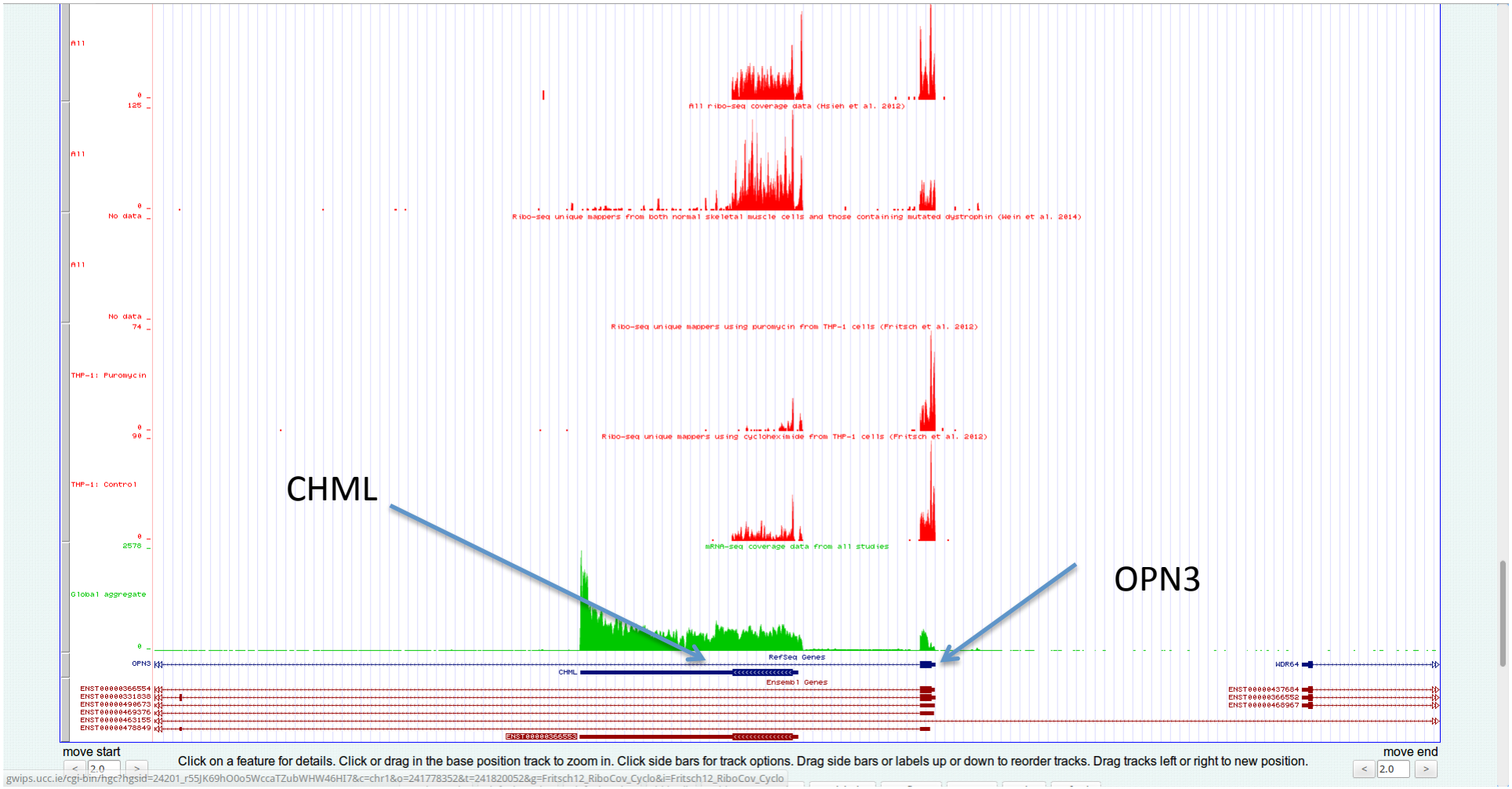
Chml pre-update



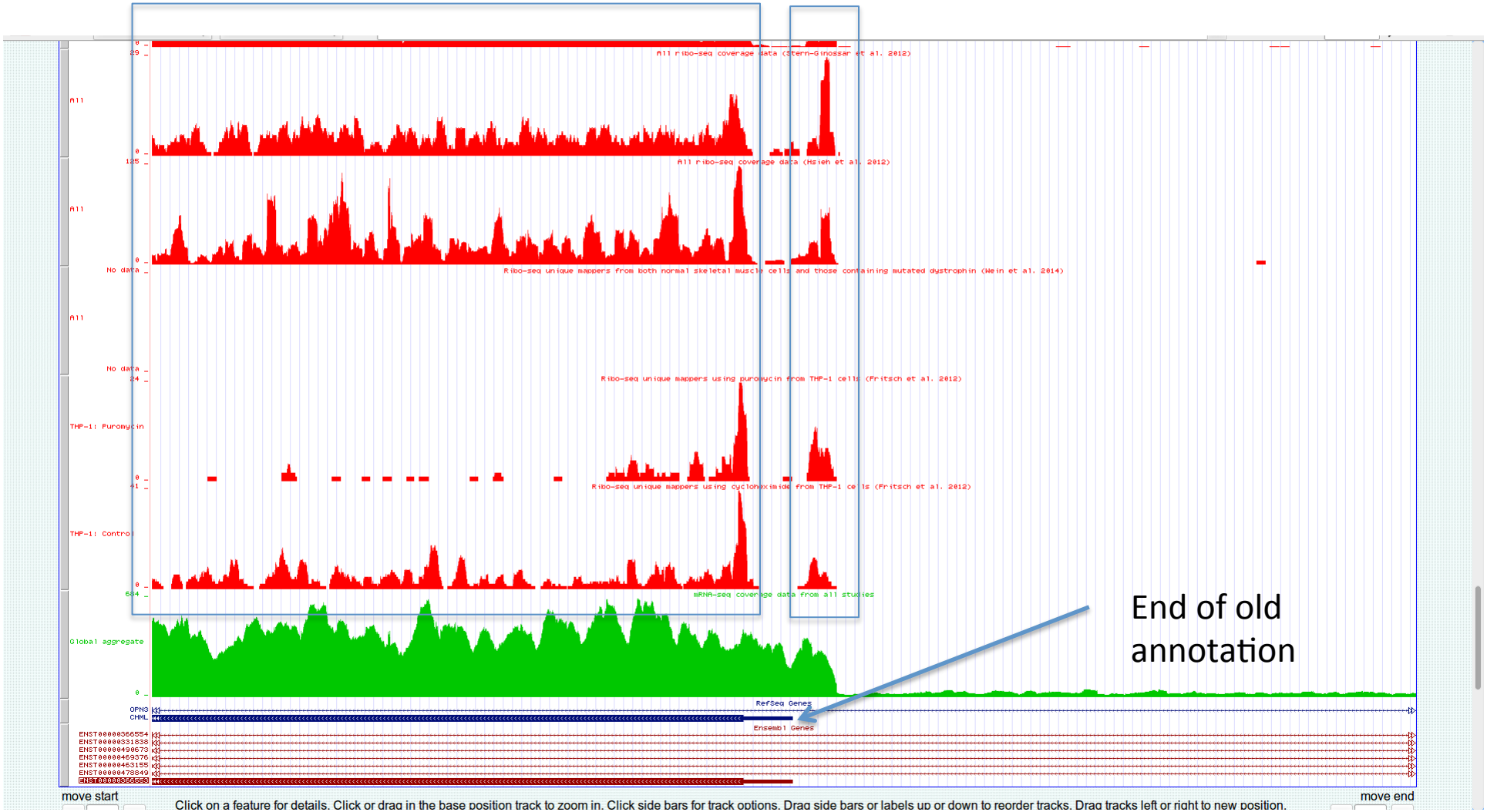
Human chr1 - originally annotated as OPN3 & a single exon gene CHML



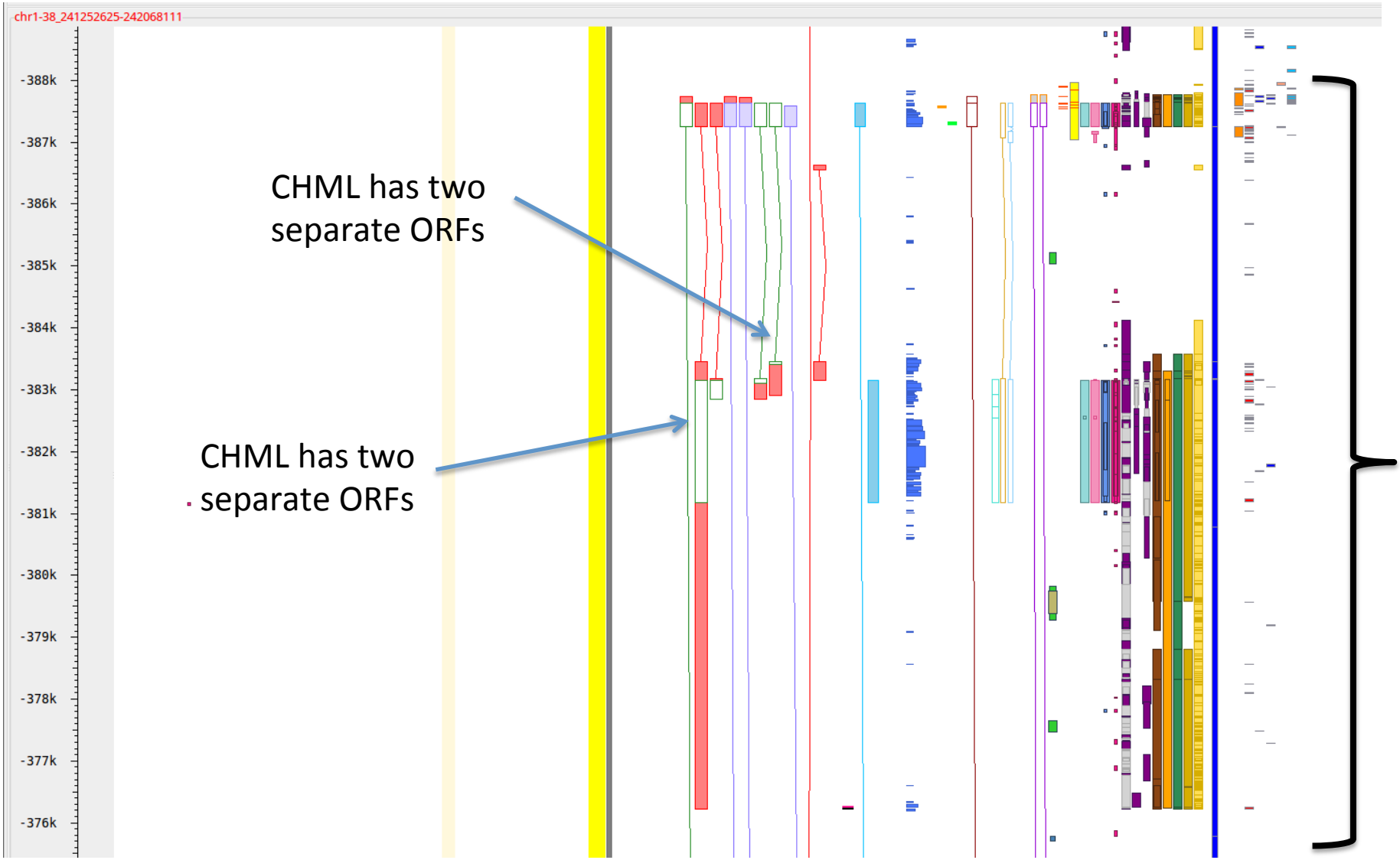
GWIPS Human (zoomed out) displays ribosome profiling data



GWIPS Human (zoomed in). Can clearly see the end of the “upstream ORF”, a dip & then the peak of the downstream (CHML) ORF

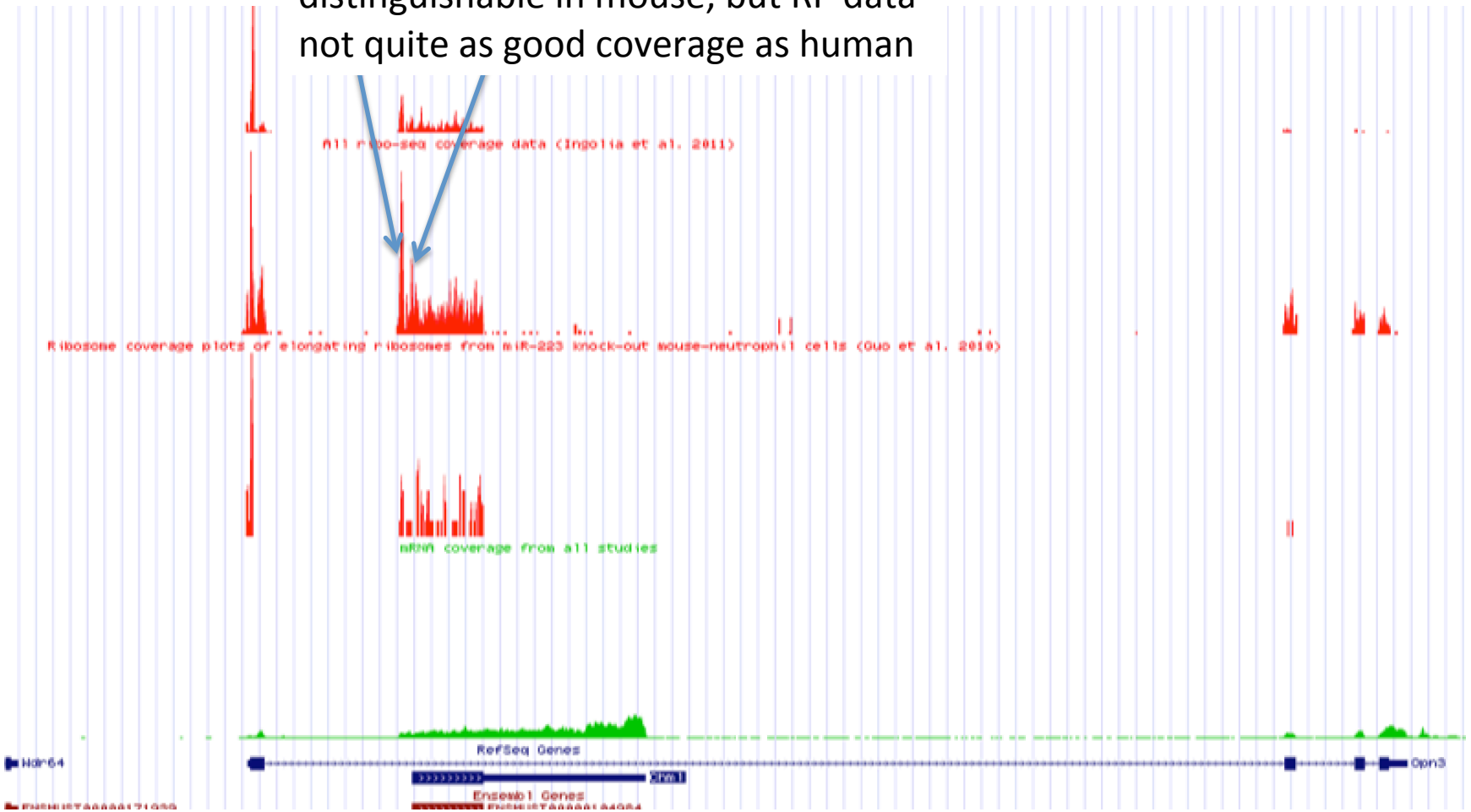


CHML has two different ORFs

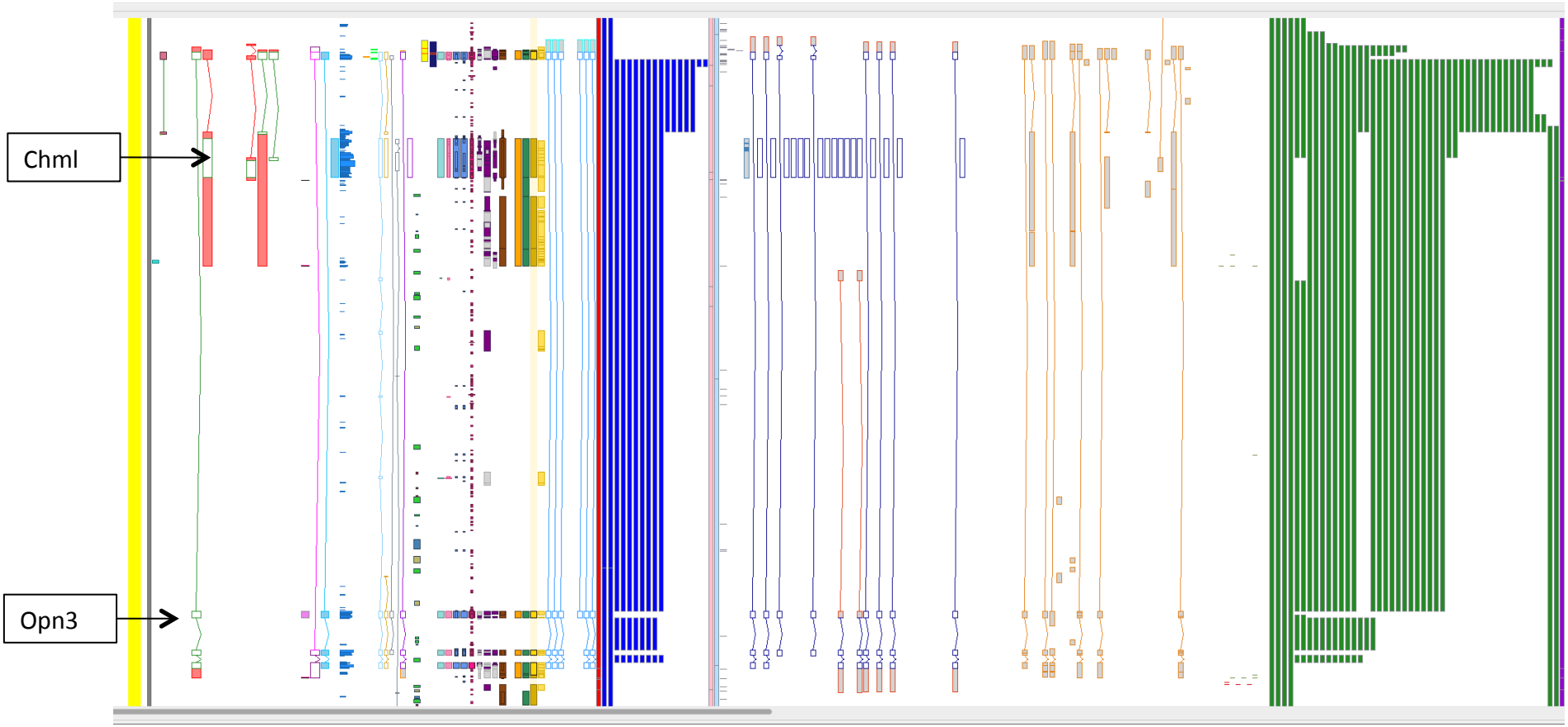


GWIPS data in mouse shows same RP configuration

Two separate ORFs are distinguishable in mouse, but RP data not quite as good coverage as human



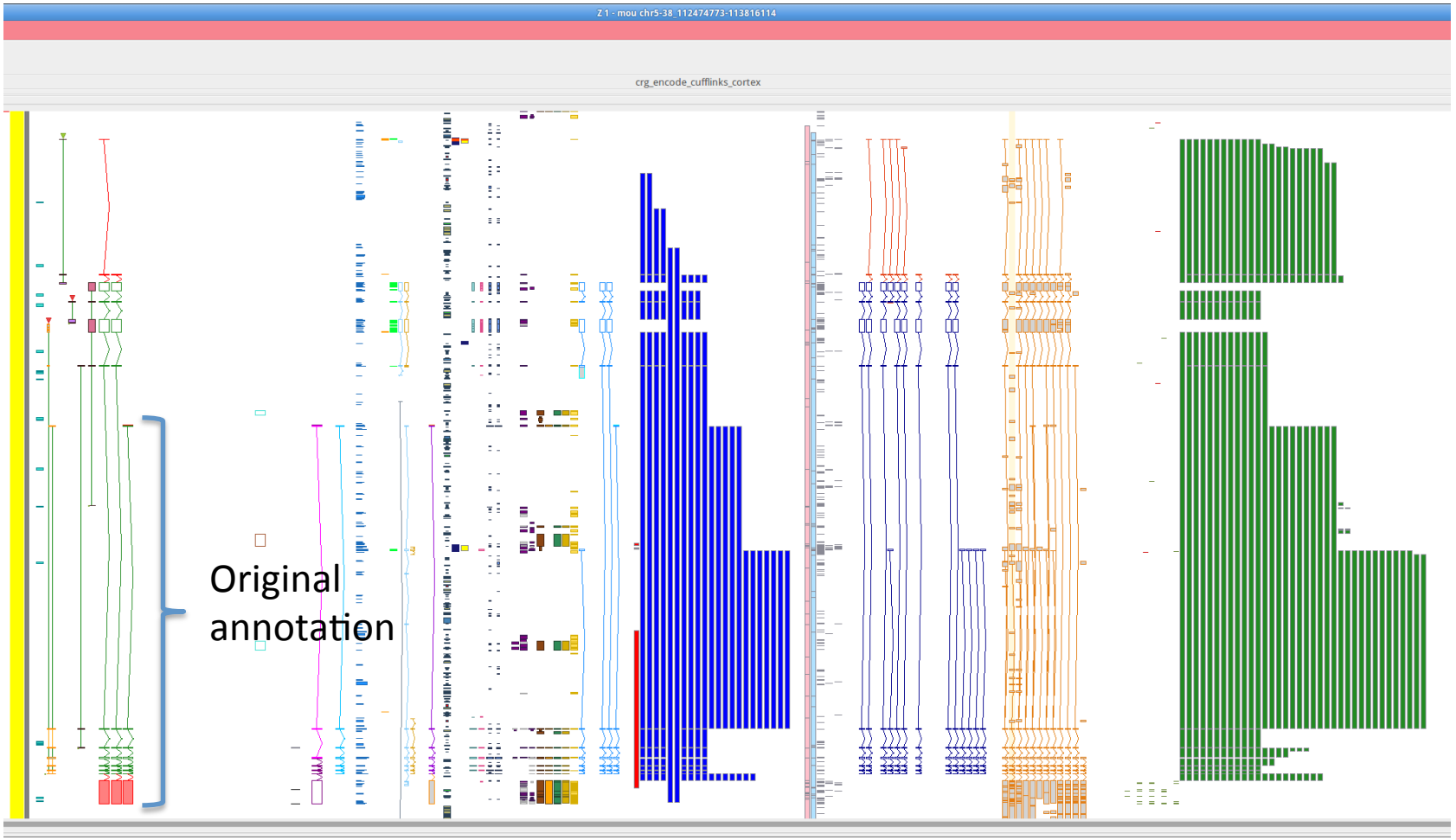
Mouse Opn3 and Chml (bicistronic locus – parent is 15 exon gene on chr X)



The intron of the main variant in Chml is only supported by ENCODE_RNASeq_canonical_intron data but lots of it:

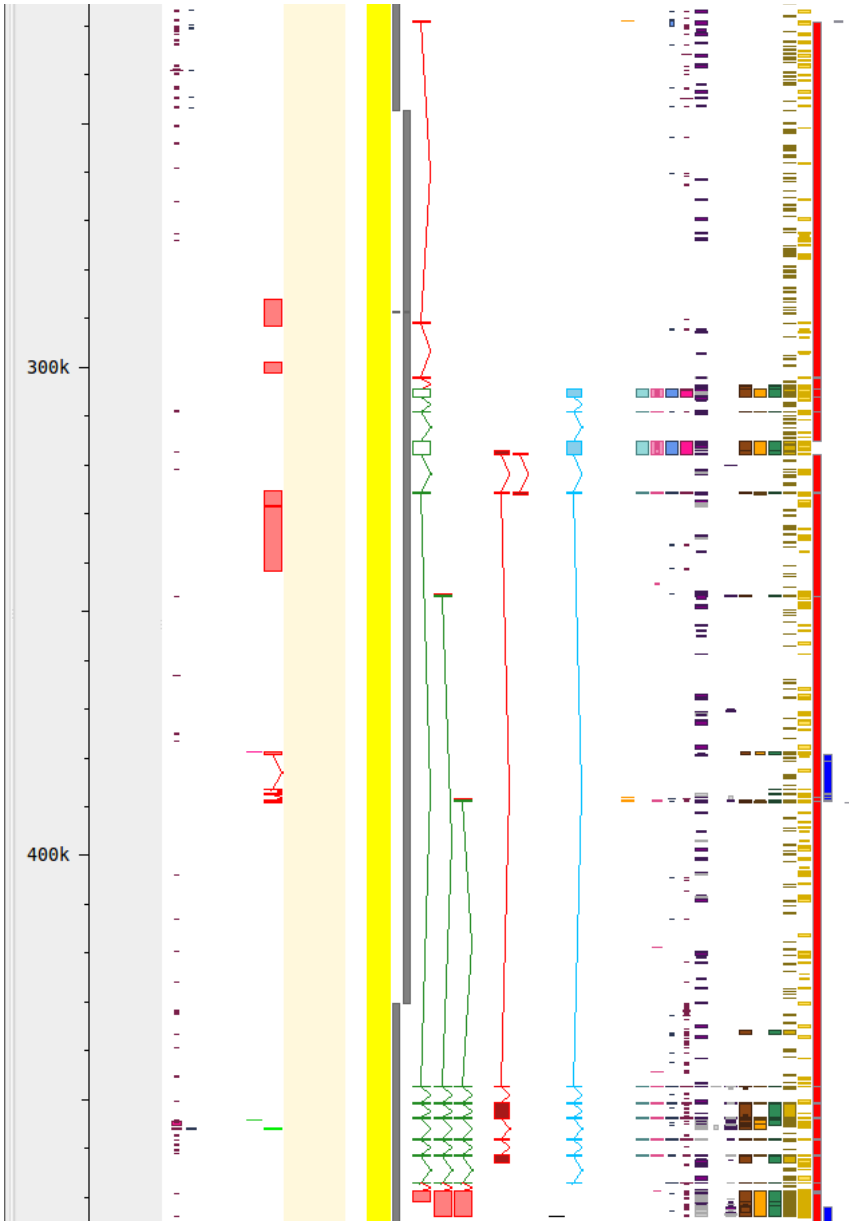
CNS(99), spleen(9), liver(5),kidney(1), frontal lobe(30), cortex(27), thymus(10), hippocampus(2), heart(11), brain(43), testis(11) & cerebellum(39)

Mouse chr5 – jg7343

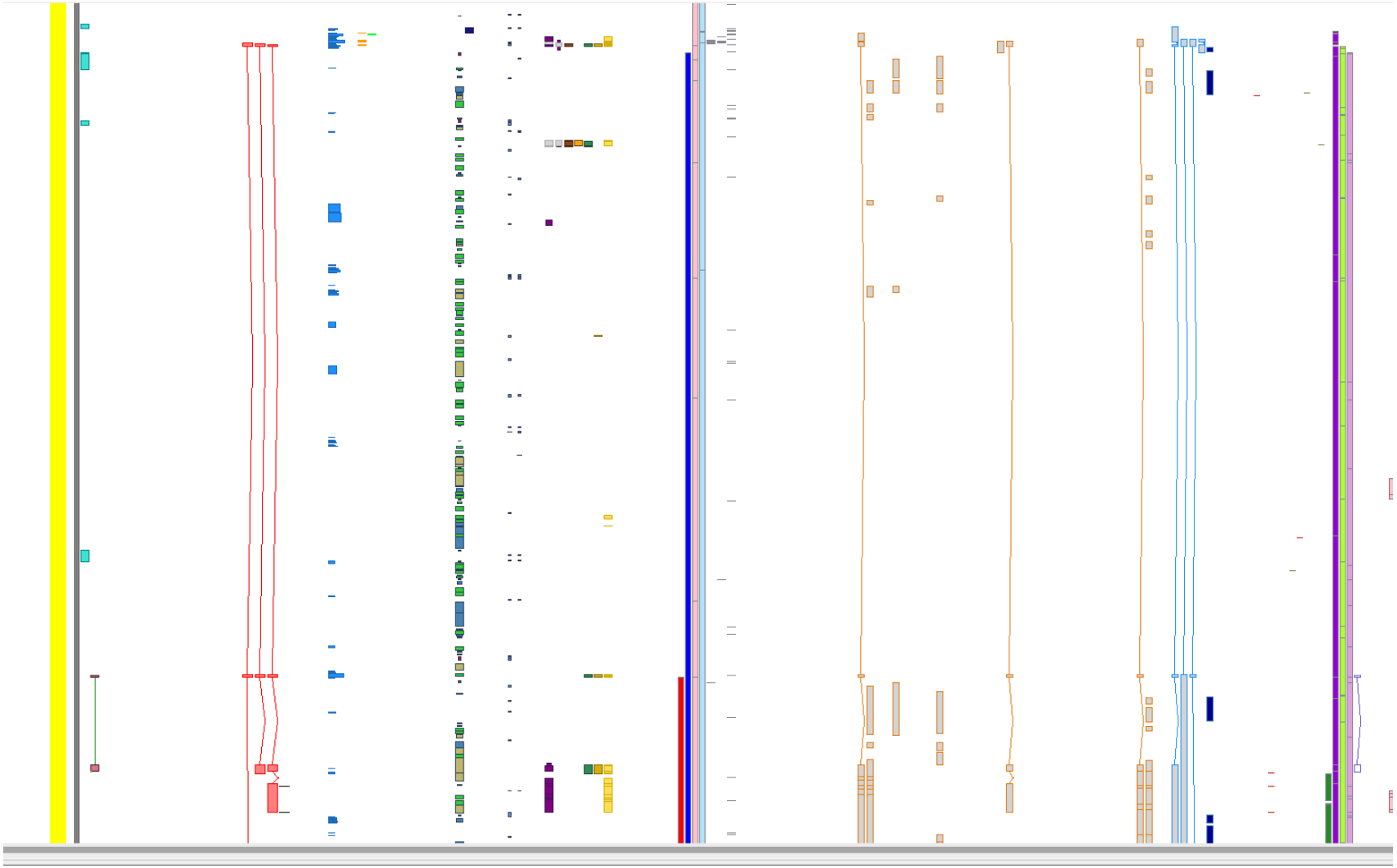


Annotated Novel_CDS based on ESTs & human protein
Supported by ENCODE_RNASeq_canonical_intron data in wholebrain, testis, spleen, liver, cortex, CNS & cerebellum.

Human ortholog of jg7343 - KIAA1671



Mouse chr 7 – jg6157



Annotated lincRNA locus based on mRNA & EST evidence. Good cage tags, CpG island & polyA features. There are potential ORFs, but they are not conserved, yet the ATG is conserved back to Armadillo

Findings…

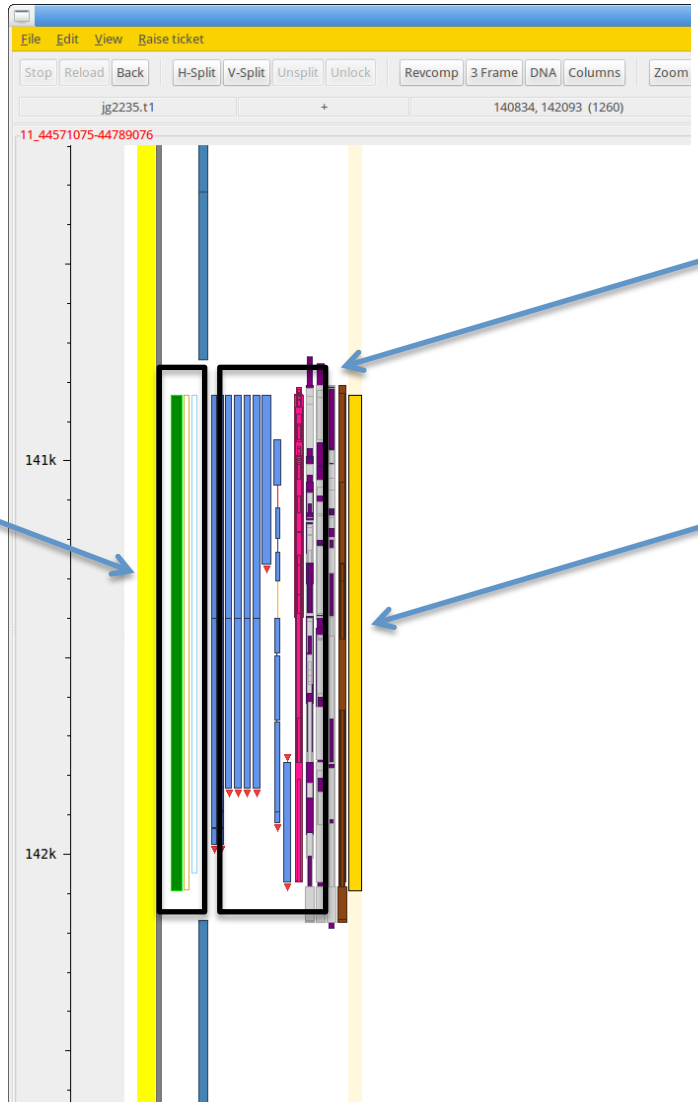
- 84 queries left to do out of a total of 171 “unconfident” predictions
- @50% of new features being added - mix of anything - novel exons, 5' UTR from coding genes etc.
- Canonical intron data is confirming a lot of these predictions.
- Lots of lncRNAs contain potential ORFs that are not conserved

2) Single exon genes

- Examples of what we have found so far. The following examples are from chr 11 NOD …

jg2235.t1 - processed pseudogene

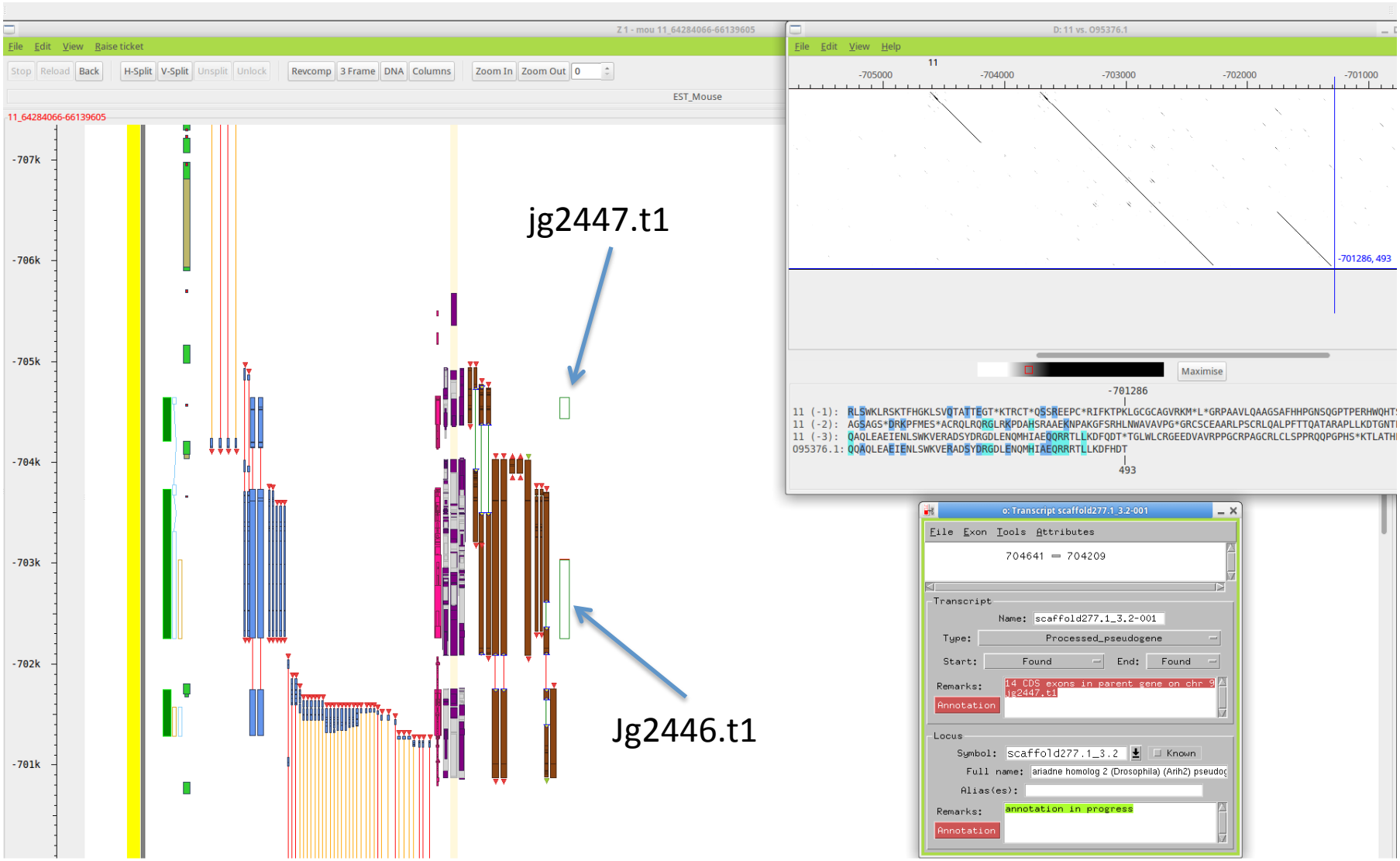
Annotated processed pseudogene based on parent gene homologies



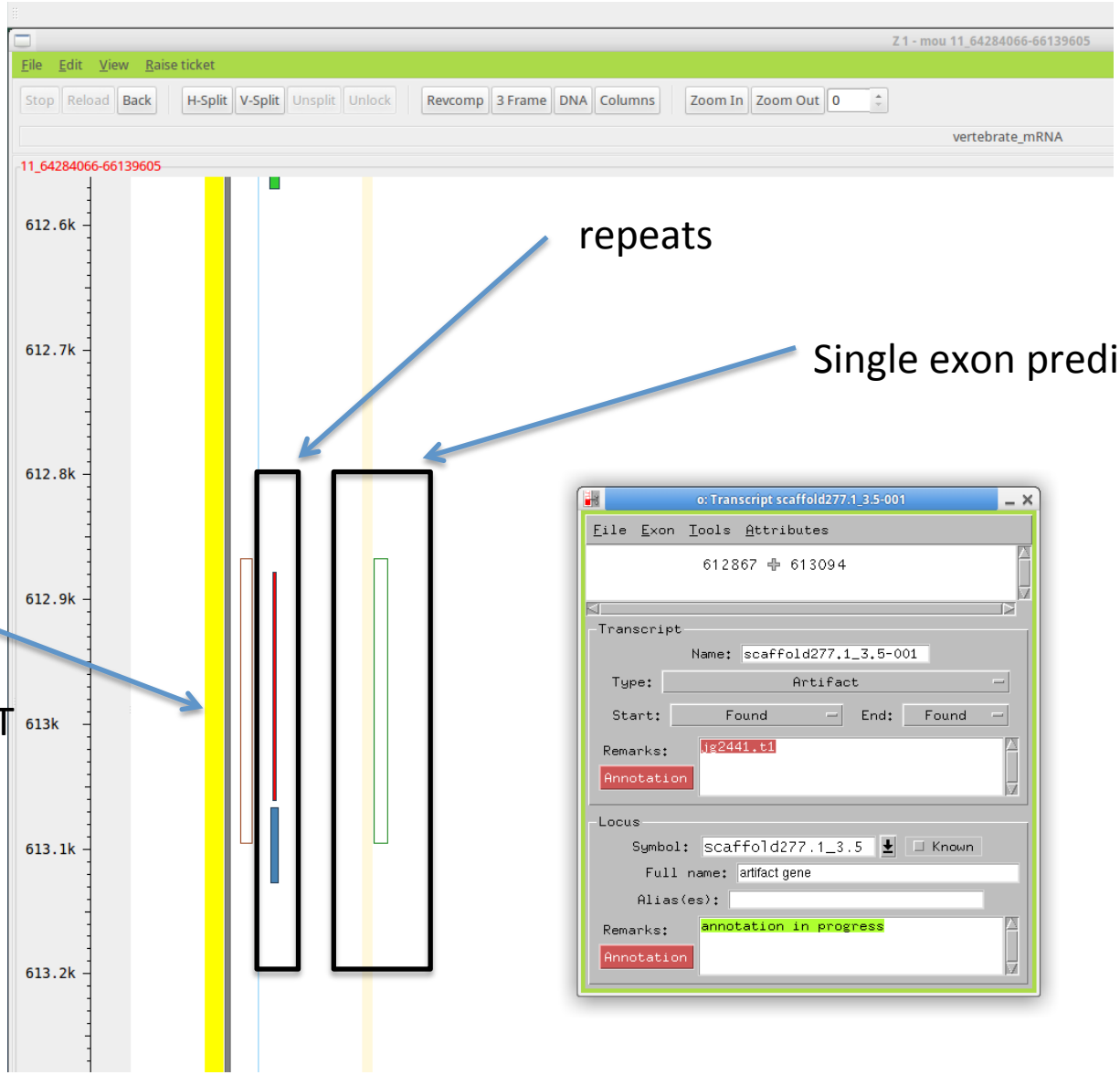
Protein homologies

Single exon prediction

jg2446.t1 and jg2447.t1 – three Arih2 pseudogenes

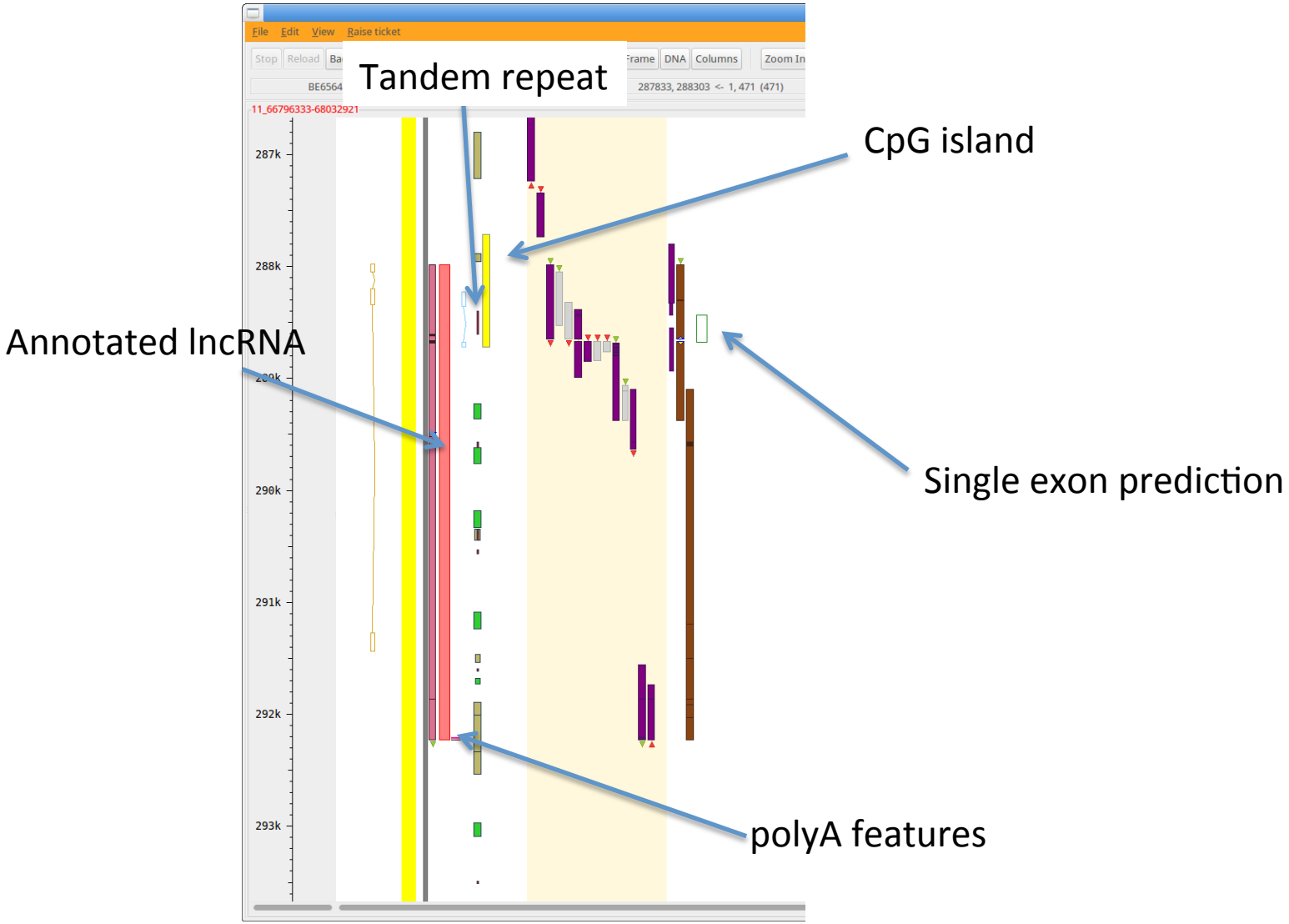


jg2441.t1 – artifact?



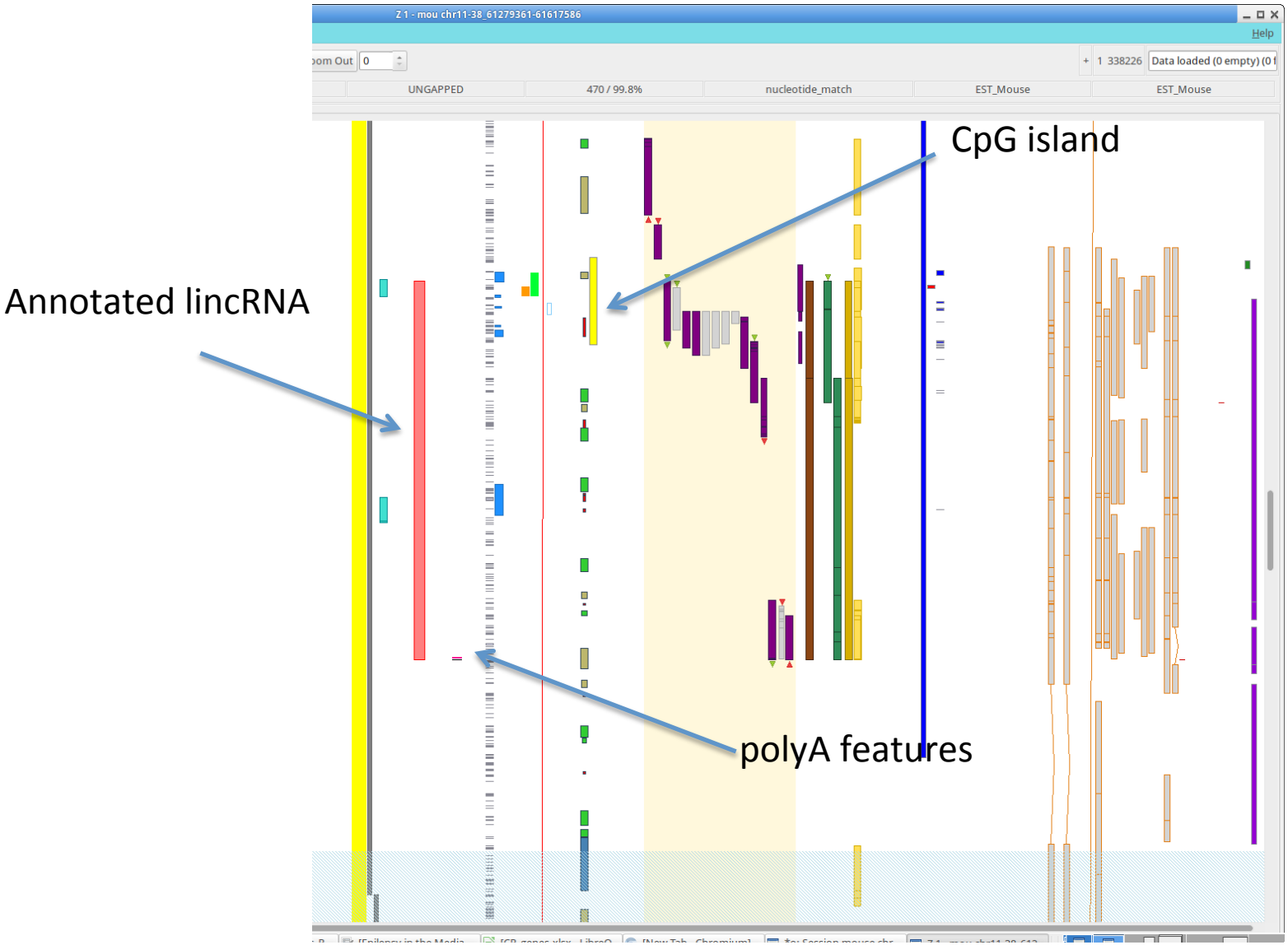
Annotated artifact, based on prediction only matching trf and no matches from BLAST

jg2488.t1 – annotated as a lincRNA in NOD (based on B6 transcript alignment and other evidence)



jg2488.t1 – annotated in B6 reference

(OTTMUSG00000060875) this was the only addition in B6 from the single exon list



Findings…

- Almost all of the predictions are pseudogenes
- Checked around 20 of the “single exons” in B6 and all are conserved
- Are there any pseudogene predictions for the mouse strains that we can use to filter single exon predictions?
- Is it worth filtering on repeat sequences as well or could we lose lncRNAs?