BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY
Berkeley
UNIVERSITY OF CALIFORNIA

Apollo

Collaborative genome annotation editing

A workshop for the Arthropod Genomics Community

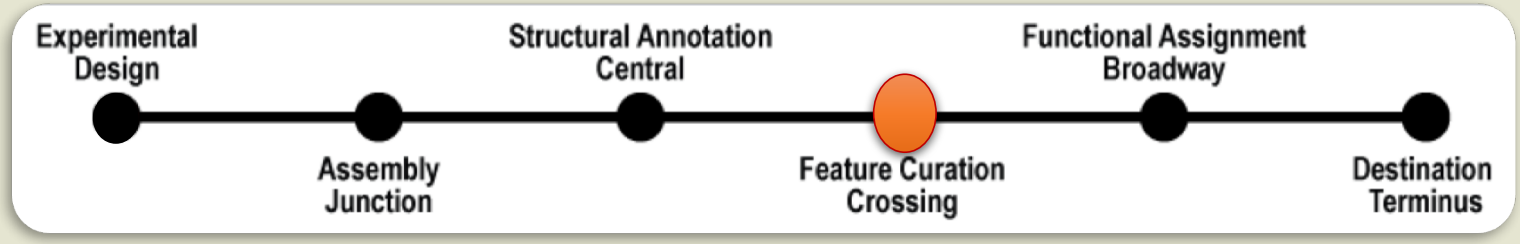
Monica Munoz-Torres, PhD | @monimunozto

Berkeley Bioinformatics Open-Source Projects (BBOP)
Environmental Genomics & Systems Biology Division
Lawrence Berkeley National Laboratory

University of Notre Dame, South Bend, IN. 08 June, 2017

http://GenomeArchitect.org


GENOME TRAIN


Experimental Design
Assembly Junction
Structural Annotation Central
Feature Curation Crossing
Functional Assignment Broadway
Destination Terminus

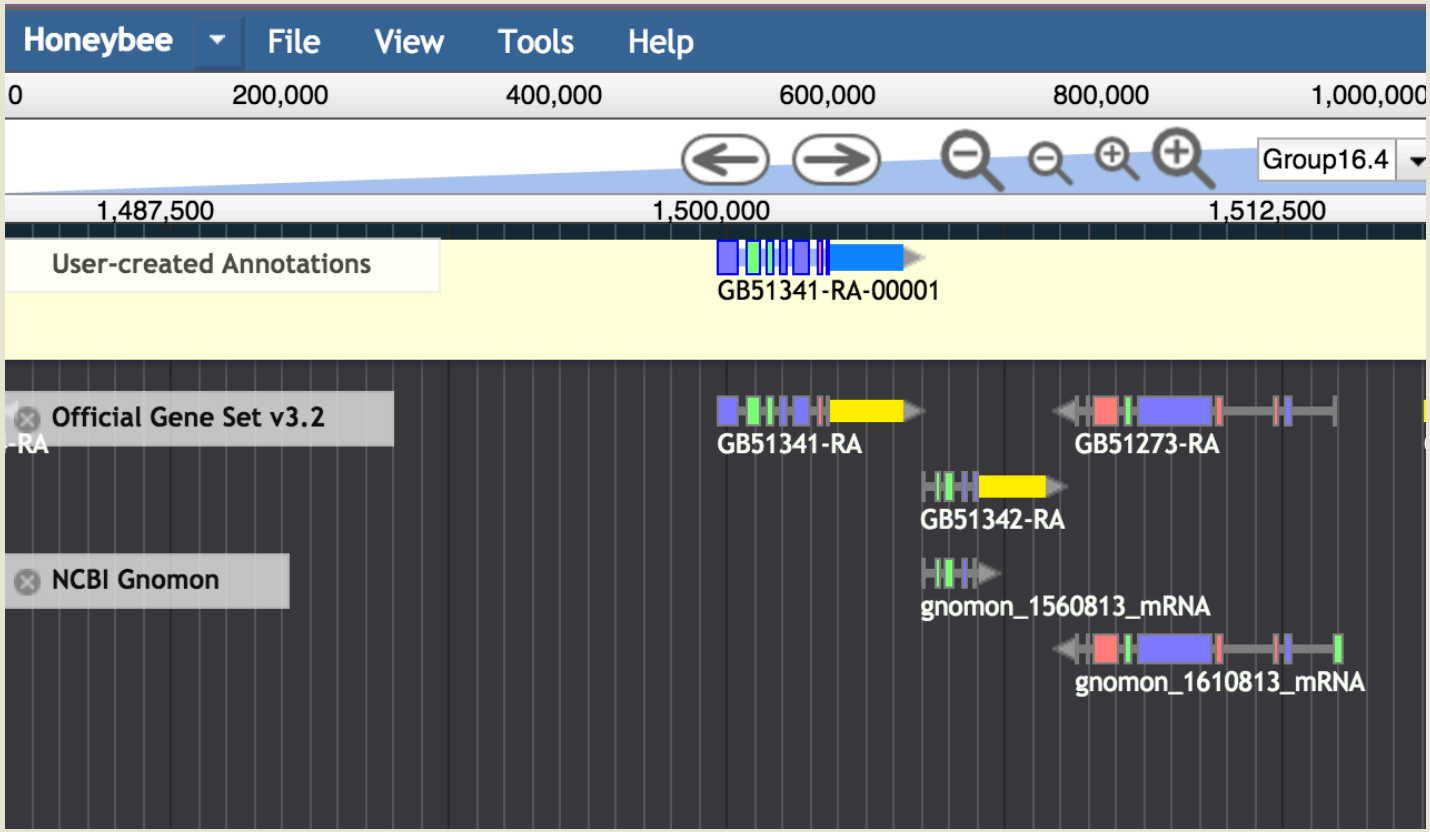
editing functionality

begin with a new gene model

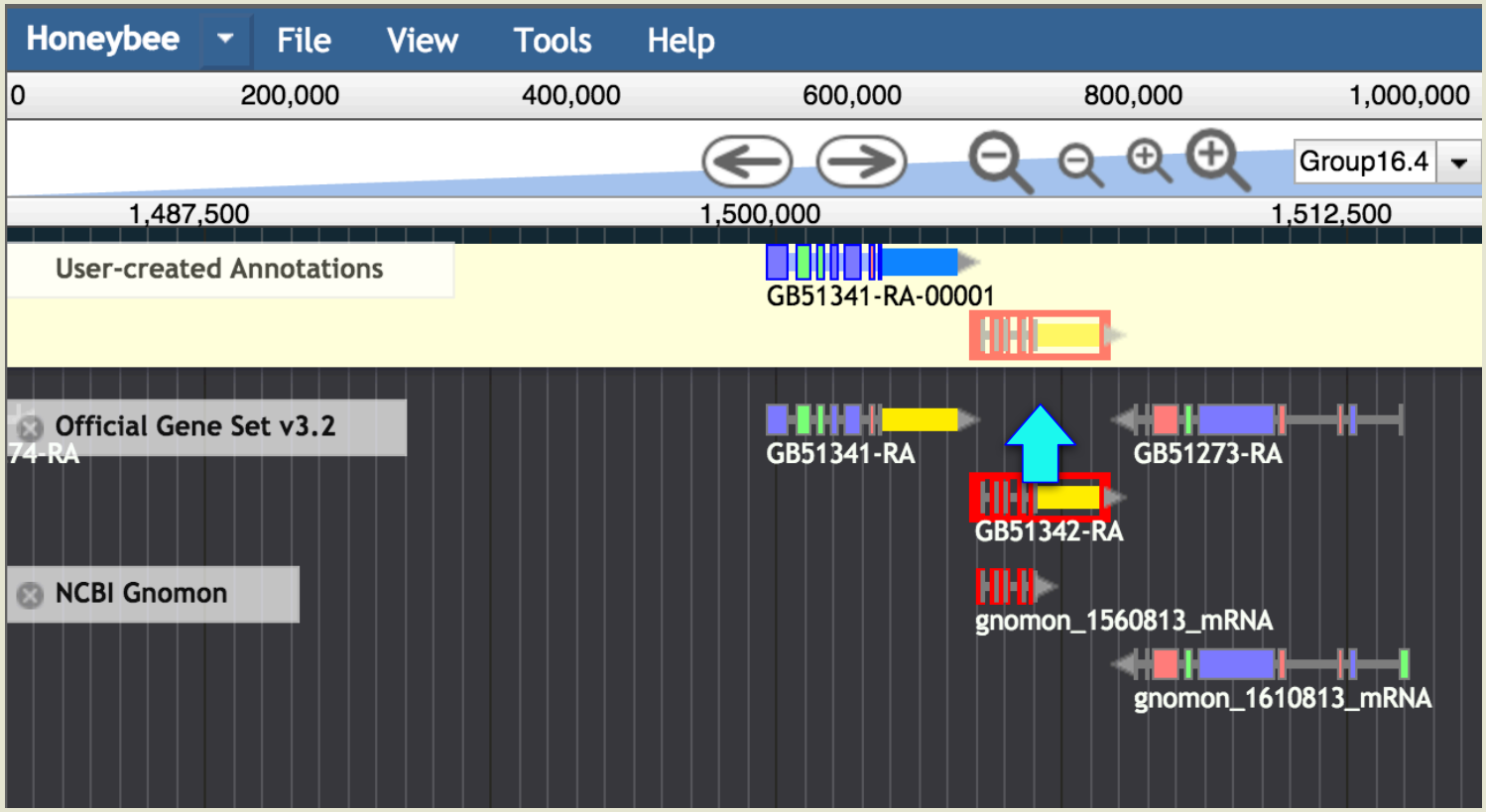
Creating a new annotation

- Choose appropriate evidence from list of “Tracks” on **annotator panel**.
- Select & drag elements from evidence track into the *‘User-created Annotations’* area.
- Hovering over annotation in progress brings up an information pop-up.

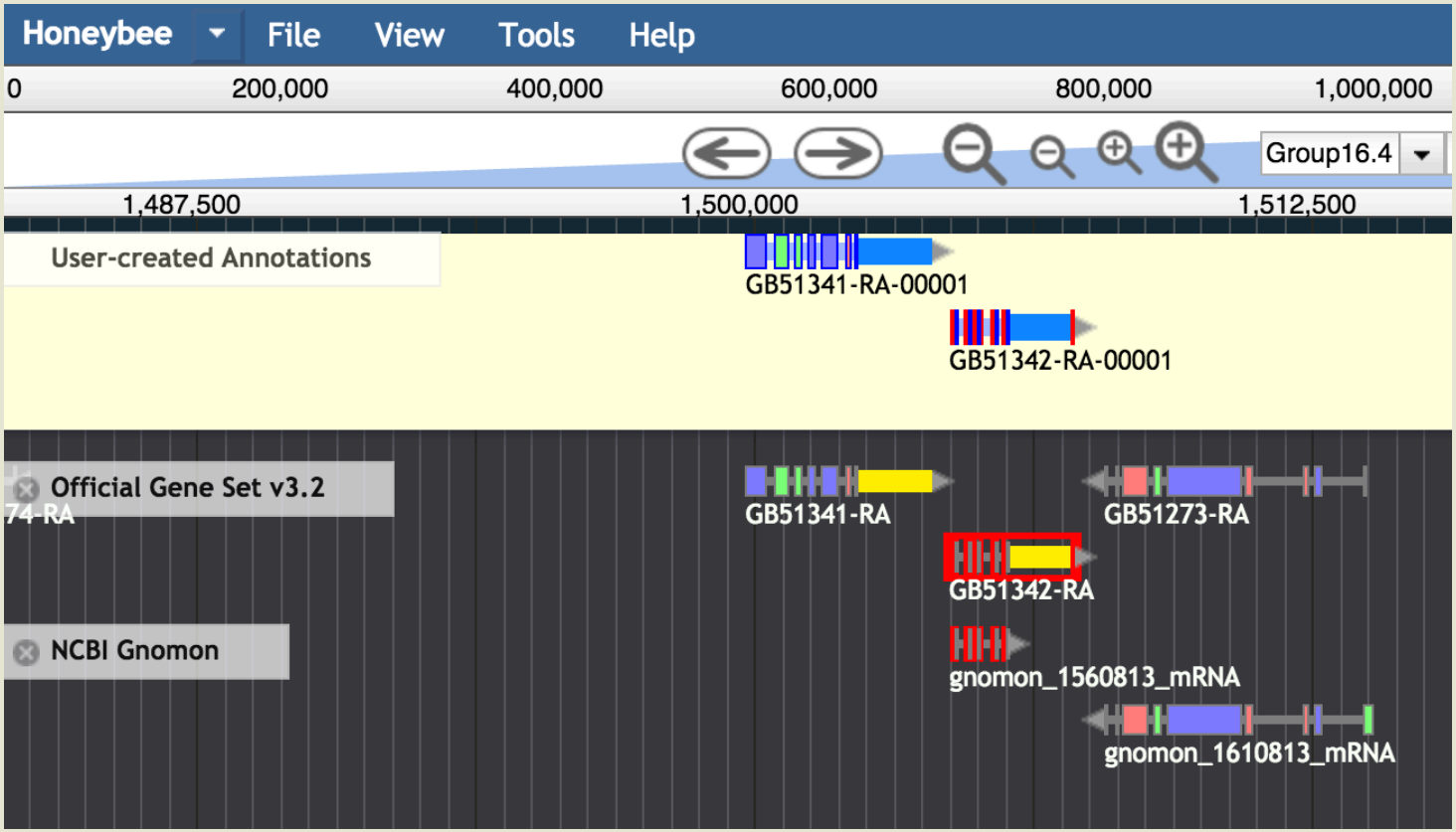
Adding a gene model



Adding a gene model

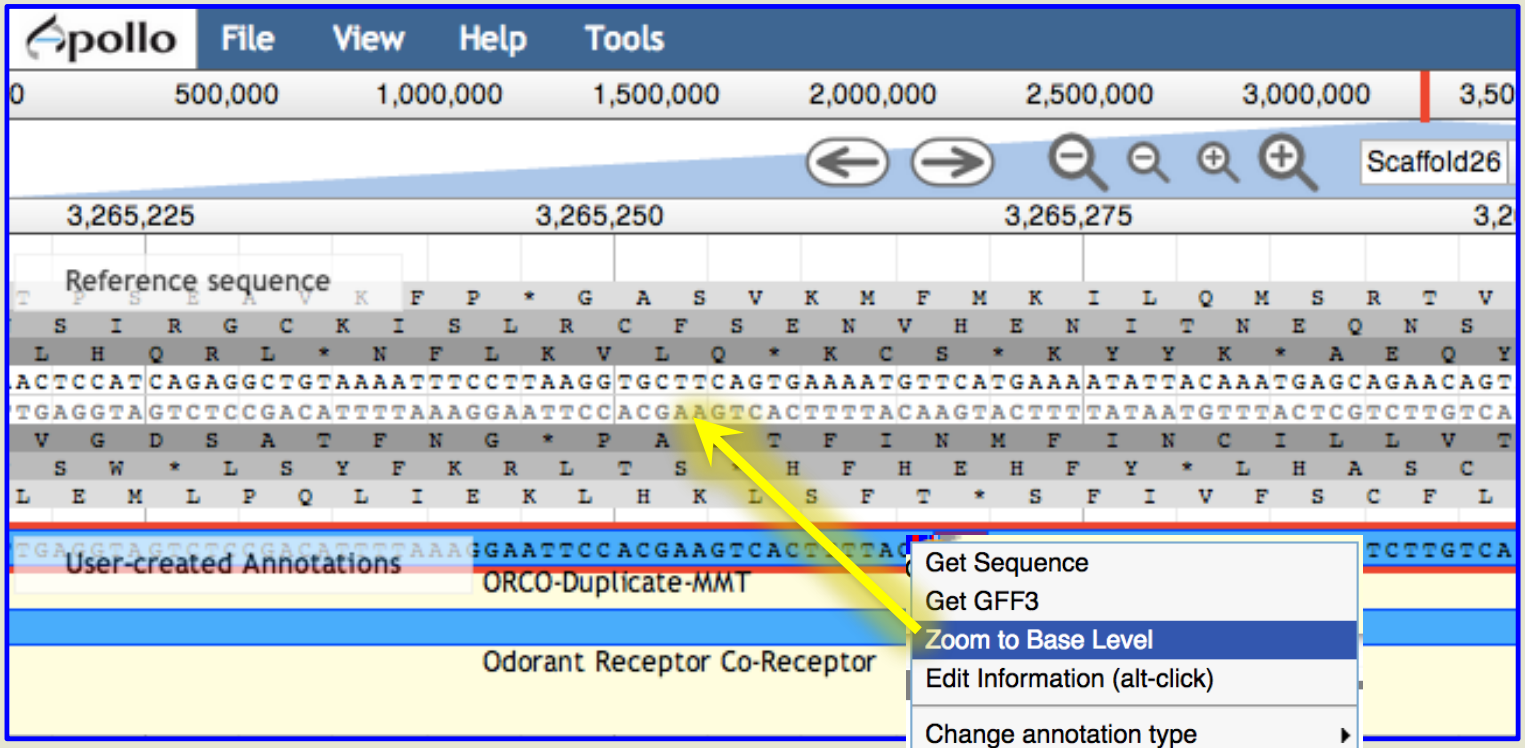


Adding a gene model



the sequence track

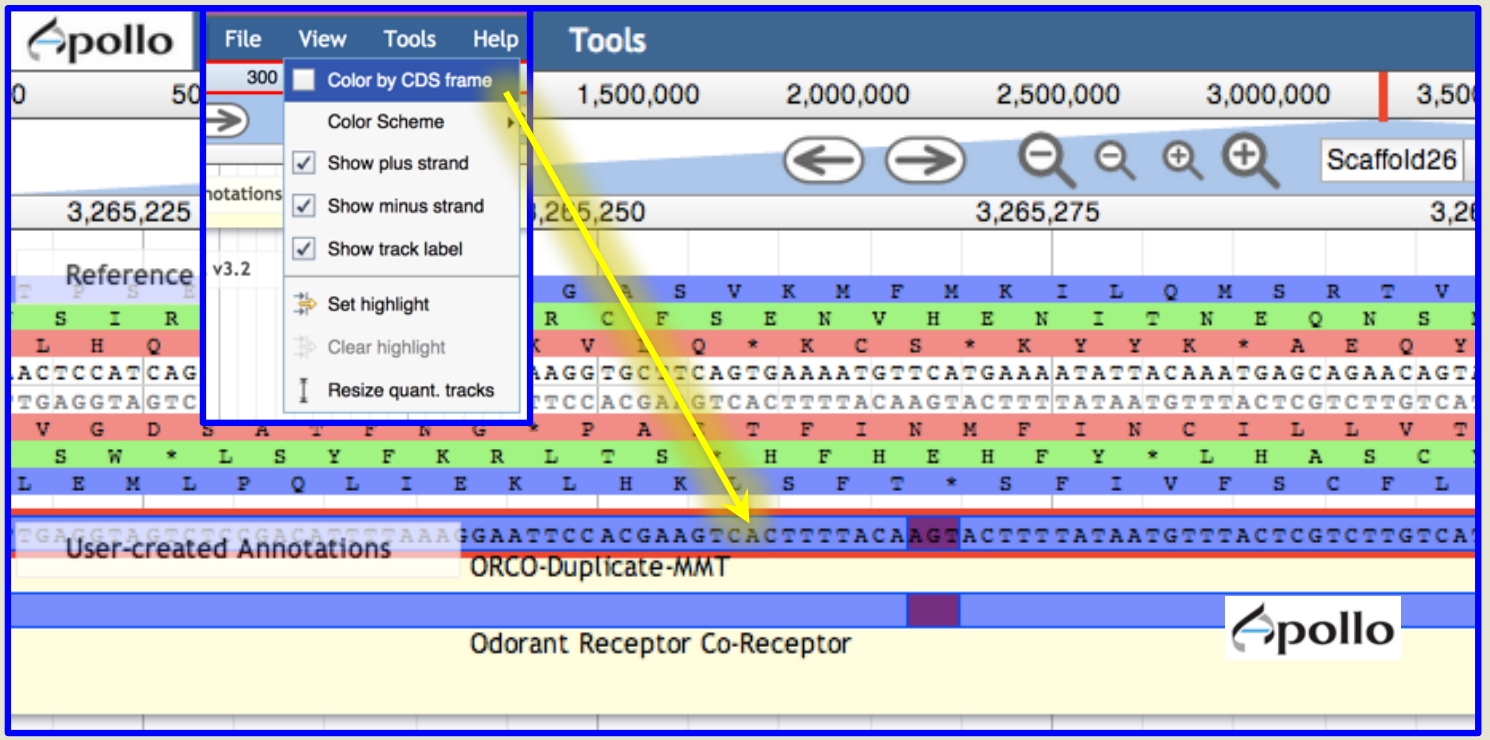
- **'Zoom to base level'** reveals the sequence track.



BERKELEY LAB

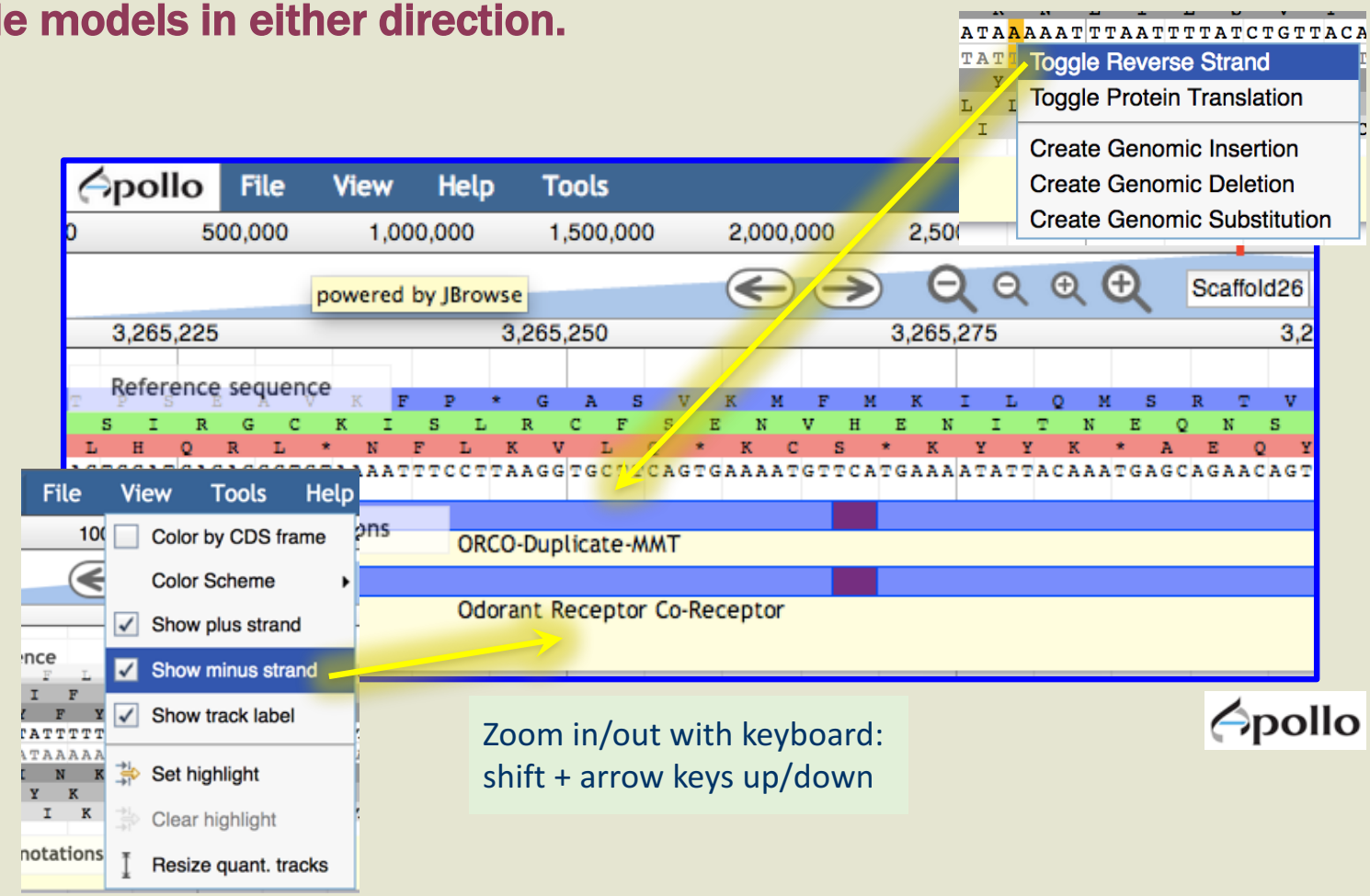
Apollo

Color exons by CDS from the ‘View’ menu.



Toggle reference DNA sequence and translation frames in forward strand.

Also, toggle models in either direction.

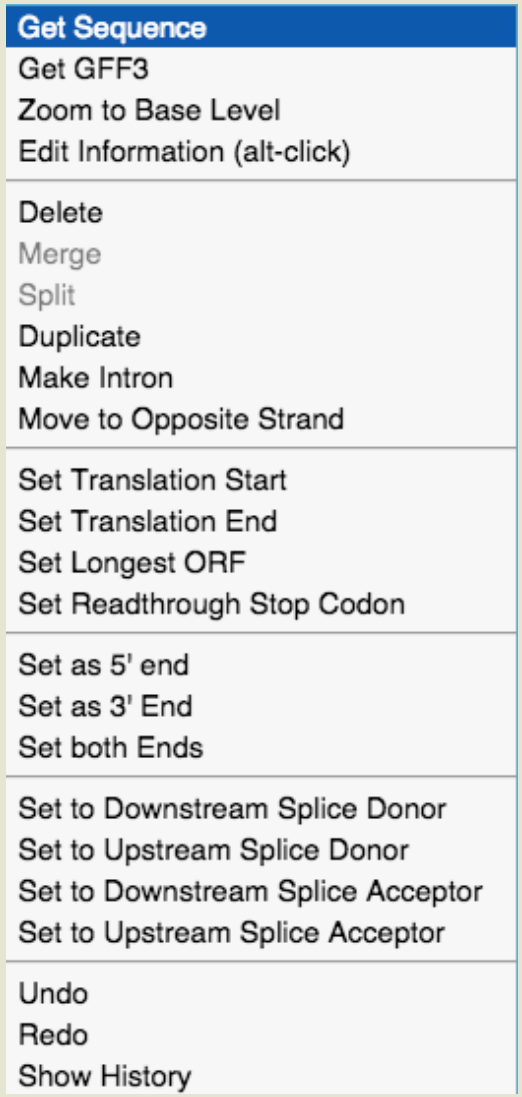


curating simple cases

- “Simple case”:
  - the predicted gene model is correct or nearly correct, and
  - this model is supported by evidence that *completely* or *mostly* agrees with the prediction.
  - evidence that extends beyond the predicted model is assumed to be non-coding sequence.

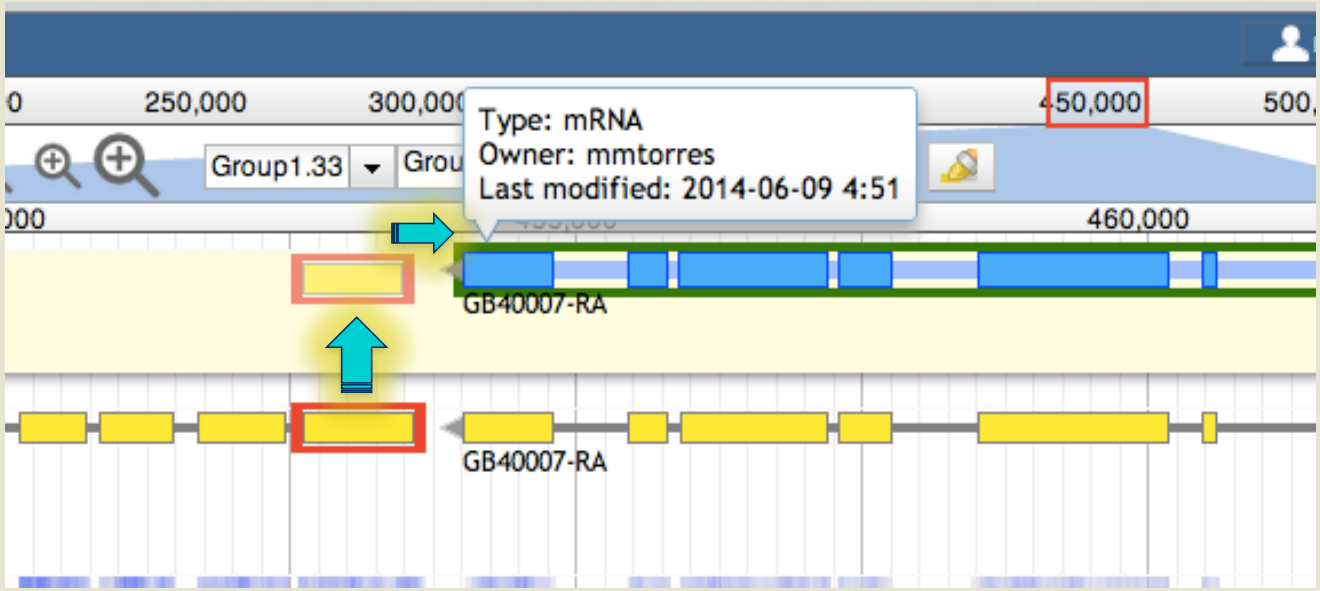
The following are simple modifications.

Editing functionality



SIMPLE CASES

ADDING EXONS



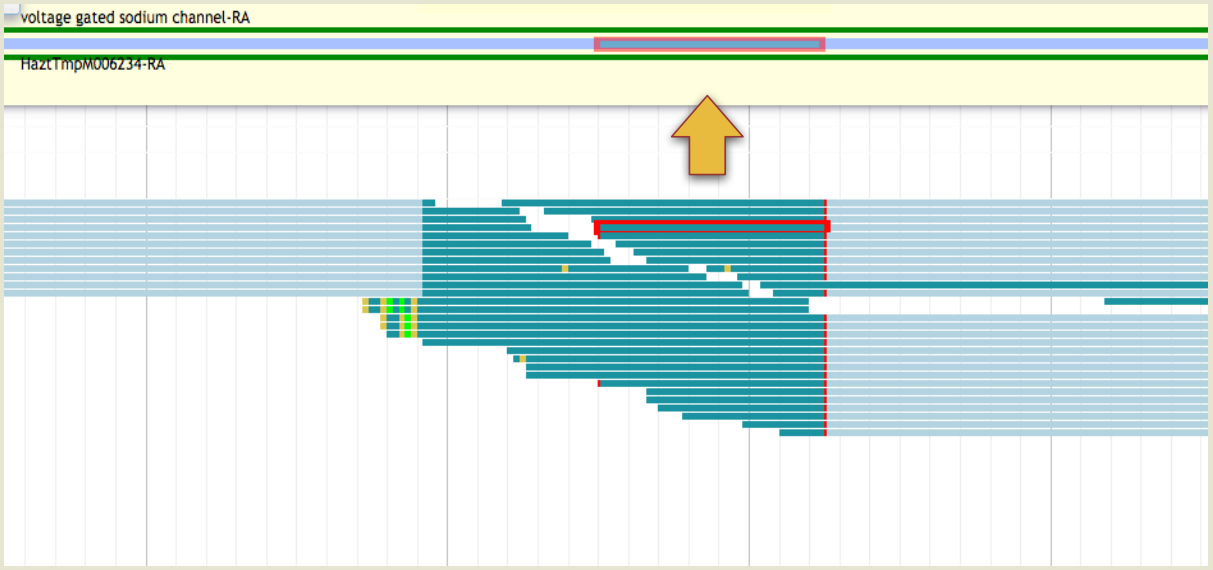
- A confirmation box will warn you if the receiving transcript is not on the same strand as the element from where the '*new*' exon originated.
- Check '***Start***' and '***Stop***' signals after each edit.

SIMPLE CASES

Editing functionality

Example: Adding an exon supported by experimental data

- RNAseq reads show evidence in support of a transcribed product that was not predicted.
- Add exon by dragging up one of the RNAseq reads.

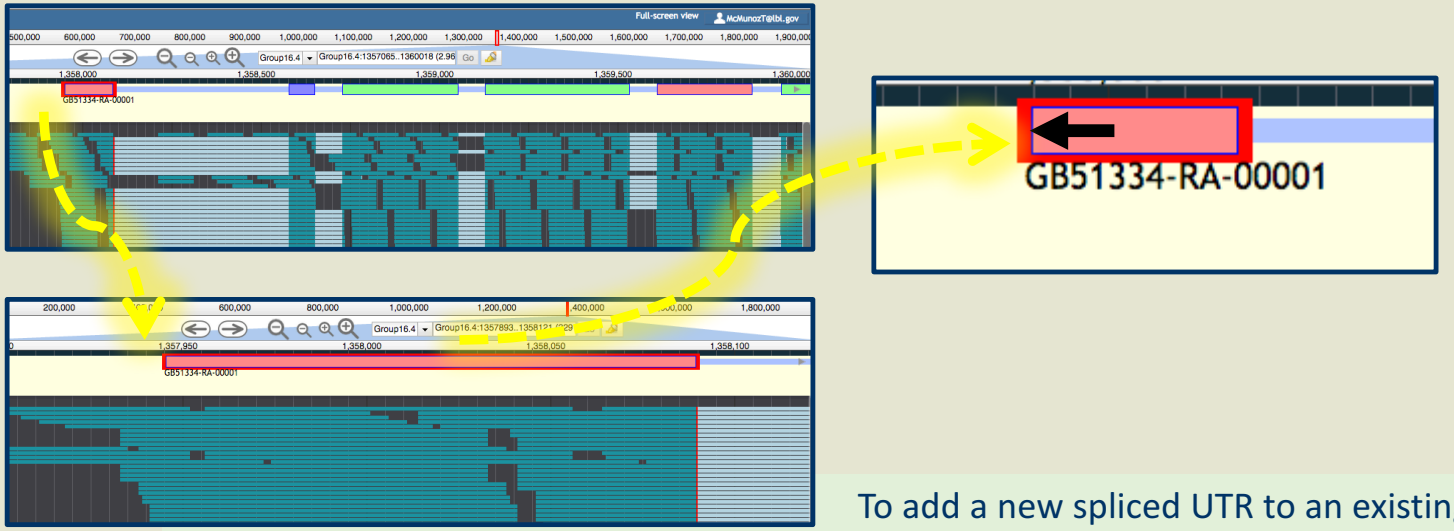


SIMPLE CASES

ADDING UTRs

- If transcript alignment data are available & extend beyond your original annotation, you may extend or add **UTRs**.

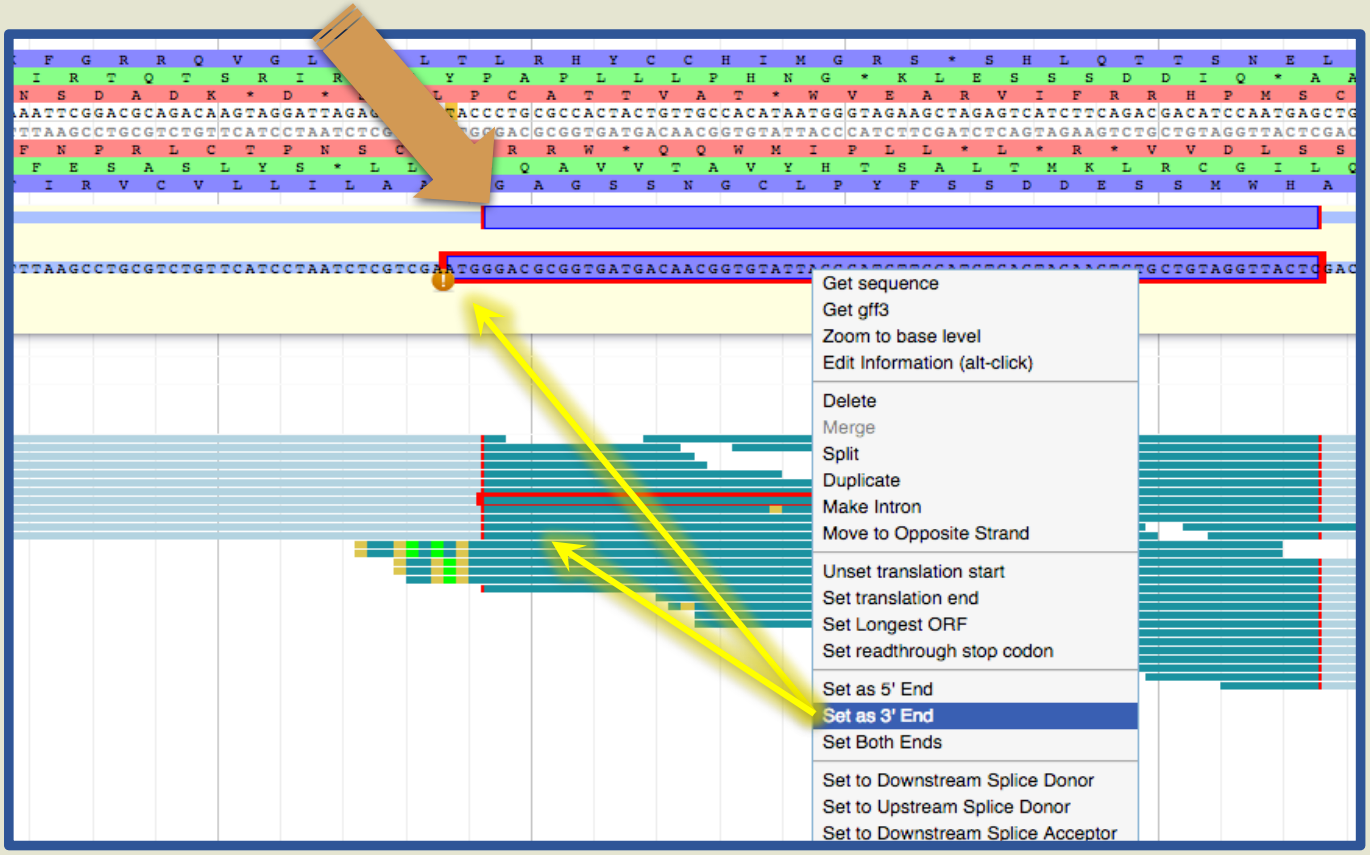
1. Right click at the exon edge and ***'Zoom to base level'***.
2. Place the cursor over the edge of the exon *until it becomes a black arrow* then click and drag the edge of the exon to the new coordinate position that includes the UTR.



To add a new spliced UTR to an existing annotation also follow the procedure for adding an exon, or to '*Set as X' end*'.

SIMPLE CASES

MATCHING EXON BOUNDARY TO EVIDENCE



To modify an exon boundary and match data in the evidence tracks: select both the offending exon and the element with the correct boundary, then right click on the annotation to select '*Set 3' end*' or '*Set 5' end*' as appropriate.

BERKELEY LAB

SIMPLE CASES

CHECK FOR EXON INTEGRITY

1. Two exons from different tracks sharing the same start/end coordinates display a red bar to indicate **matching edges**.
2. Selecting the whole annotation or one exon at a time, use this **edge-matching** function and scroll along the length of the annotation, **verifying exon boundaries against available data**.

   Use square [ ] brackets to scroll from exon to exon.

   User curly { } brackets to scroll from annotation to annotation.
3. Check if cDNA / RNAseq reads lack one or more of the annotated exons or include additional exons.

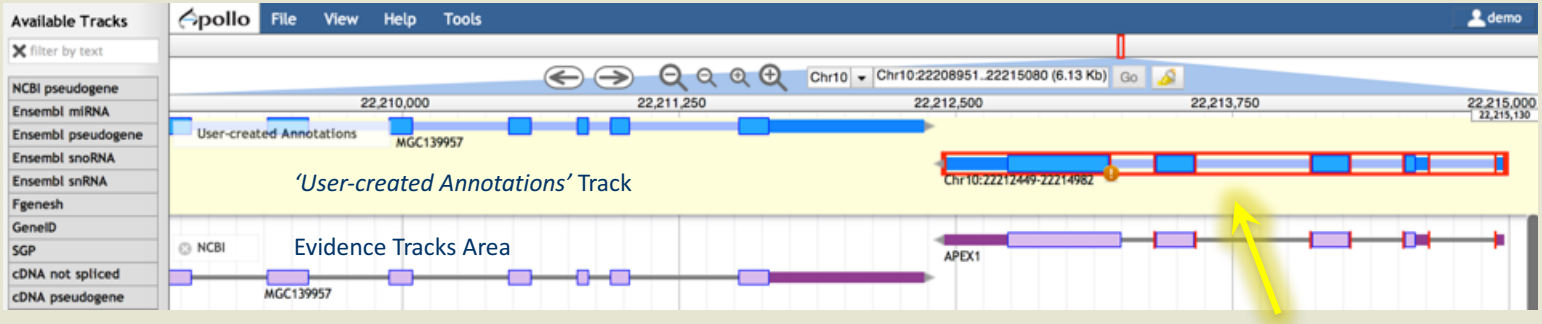
BERKELEY LAB

SIMPLE CASES

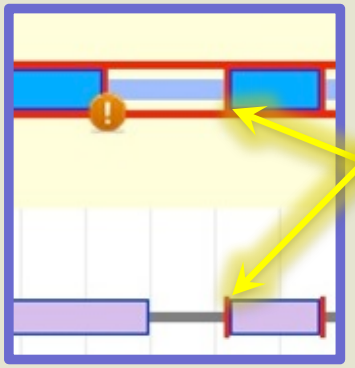
ORFs - setting & recalculating

Apollo's editing logic (brain):

- selects **longest ORF** as CDS
- **recalculates ORF** after each edit, unless set



Double click selects the entire model



Edge-matching

Red lines around exons:
'edge-matching' allows annotators to confirm whether the evidence is in agreement, without examining each exon at the base level.

SIMPLE CASES

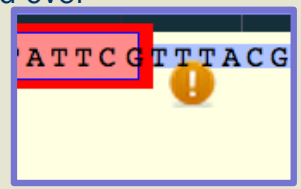
SPLICE SITES

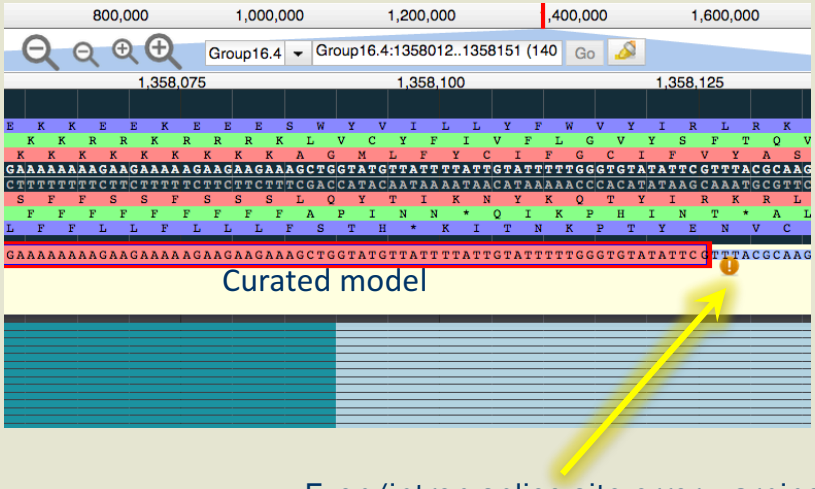
Canonical splice sites:

forward strand
5’-...exon]GT / AG[exon...-3’

reverse strand, not reverse-complemented:
3’-...exon]GA / TG[exon...-5’

Non-canonical splices are indicated with orange circles with a white exclamation point inside, placed over the edge of the offending exon.





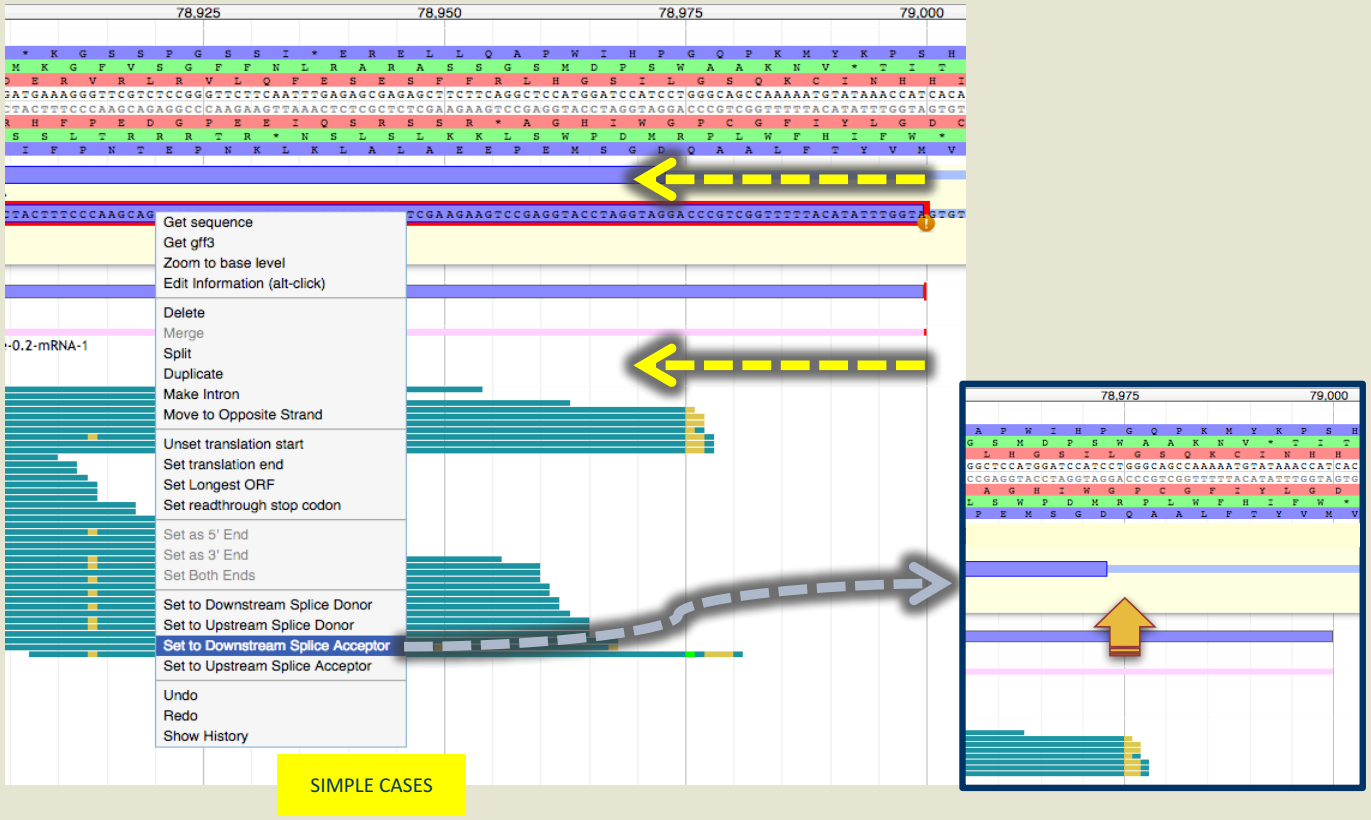
Exon/intron splice site error warning

Zoom to review non-canonical splice site warnings. Although these may not always have to be corrected (e.g. GC donor), they should be flagged with a comment.

SIMPLE CASES

Editing functionality

Example: Adjusting exon boundaries supported by experimental data



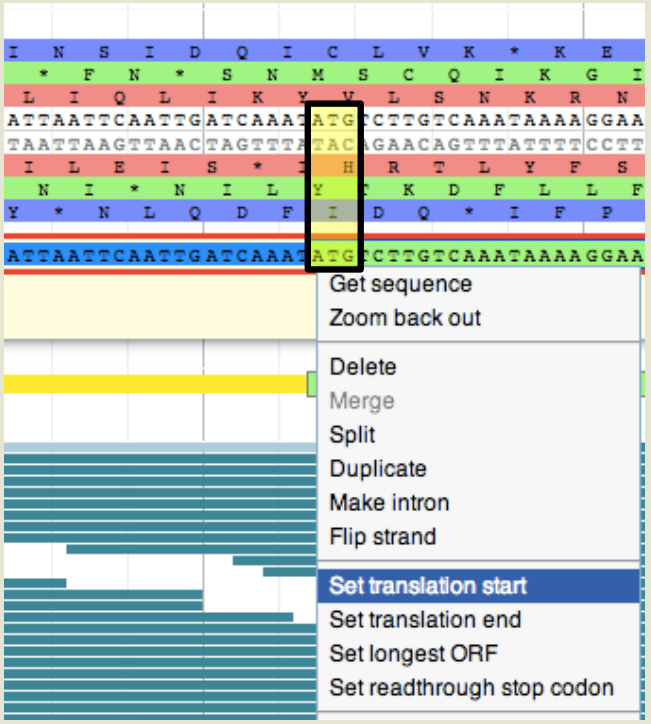
SIMPLE CASES

'*Start*' AND '*Stop*' SITES

- Apollo calculates the longest possible **open reading frame (ORF)** that includes canonical '***Start***' and '***Stop***' signals within the predicted exons.
- If '***Start***' appears to be incorrect, modify it by selecting an in-frame '***Start***' codon further up or downstream, depending on evidence (e.g. proteins, RNAseq).

  It may be present outside the predicted gene model, within a region supported by another evidence track.

  In very rare cases, the actual '***Start***' codon may be non-canonical (non-ATG).



SIMPLE CASES

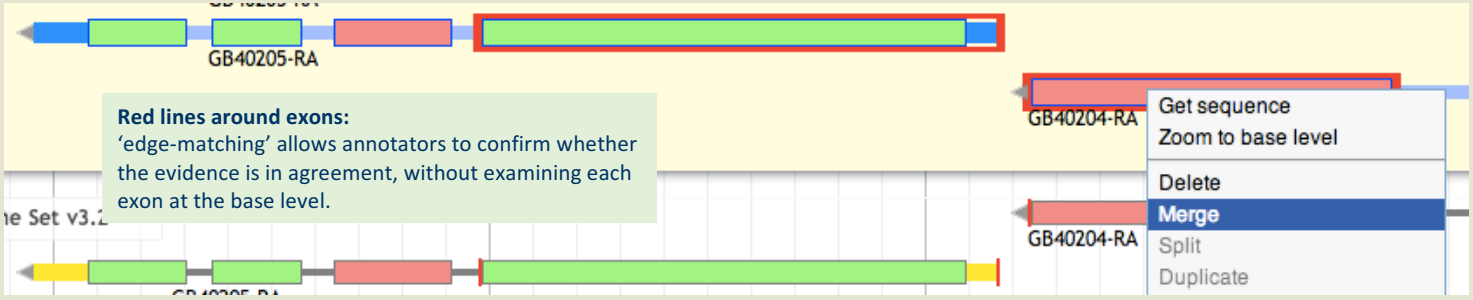
curating complex cases

MERGE TWO GENE PREDICTIONS ON THE SAME SCAFFOLD

Evidence may support joining two or more different gene models.

Warning: protein alignments may have incorrect splice sites and lack non-conserved regions!

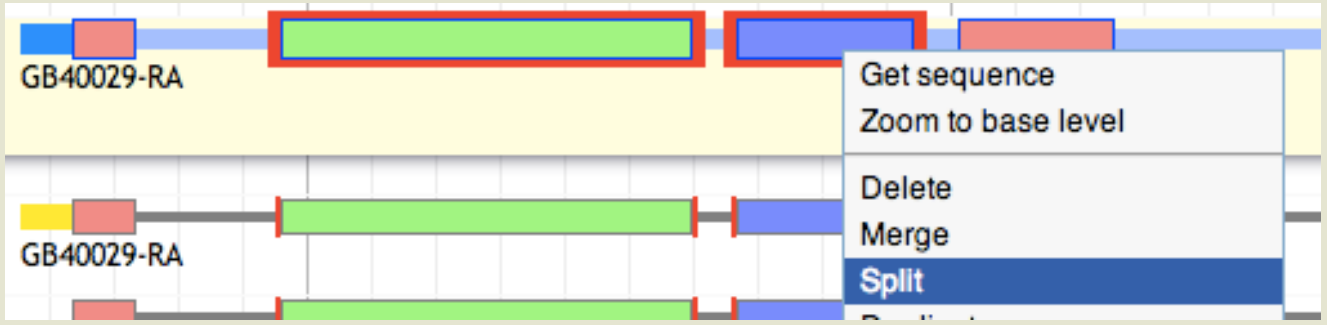
1. In **'*User-created Annotations*'** area shift-click to select an intron from each gene model and right click to select the **'Merge'** option from the menu.
2. Drag supporting evidence tracks over the candidate models to corroborate overlap, or review edge matching and coverage across models.
3. Check the resulting translation by querying a protein database e.g. UniProt, NCBI nr. Add comments to record that this annotation is the result of a merge.



COMPLEX CASES

SPLIT A GENE PREDICTION

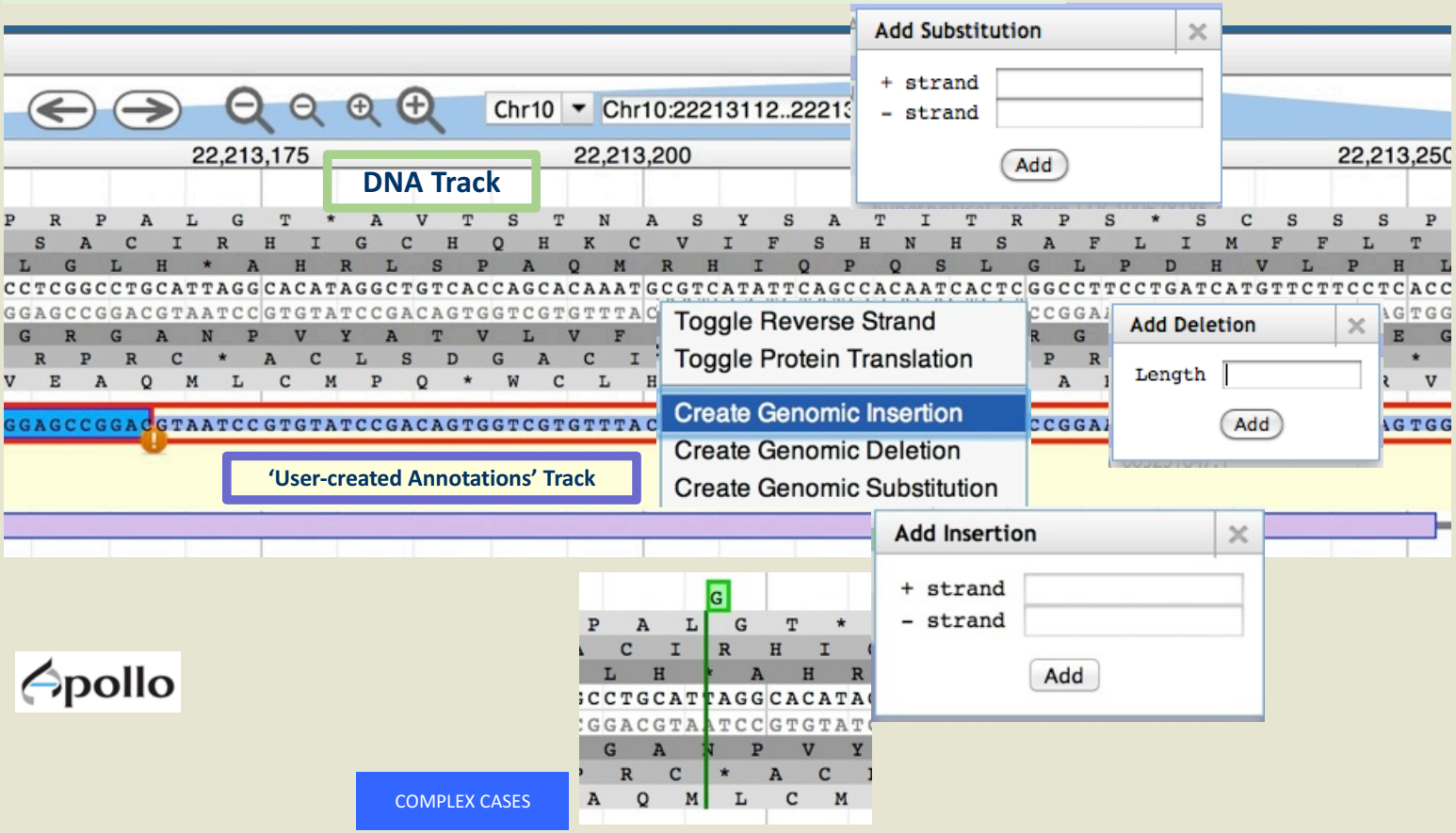
- One or more splits may be recommended when:
  - different segments of the predicted protein align to two or more different gene families
  - predicted protein doesn't align to known proteins over its entire length
  - Transcript data may support a split; BUT - first, verify whether they are alternative transcripts.



COMPLEX CASES

ANNOTATE FRAMESHIFTS AND CORRECT SINGLE-BASE ERRORS

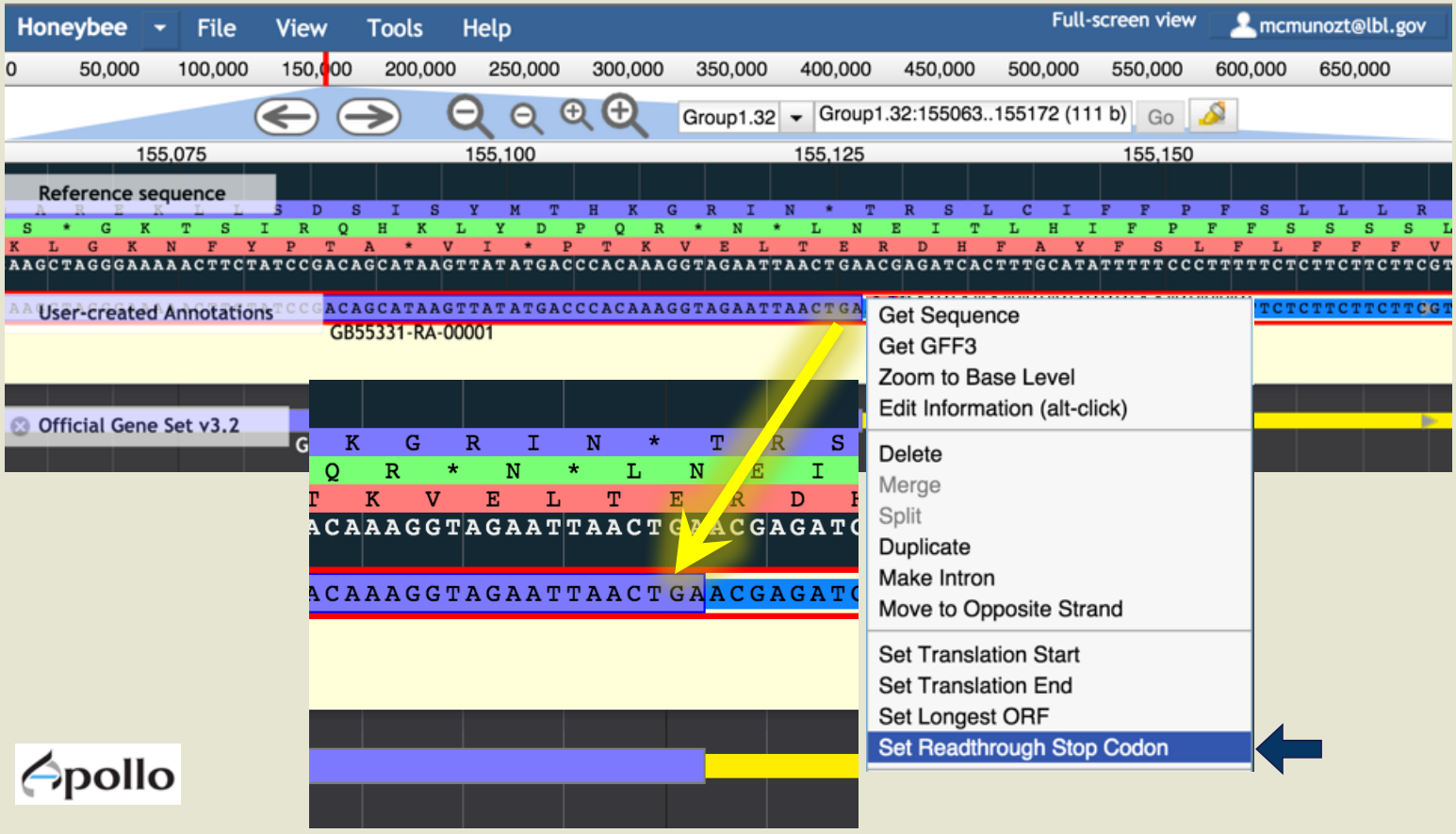
Always remember: when annotating gene models using Apollo, you are looking at a 'frozen' version of the genome assembly and you will not be able to modify the assembly itself.



Apollo

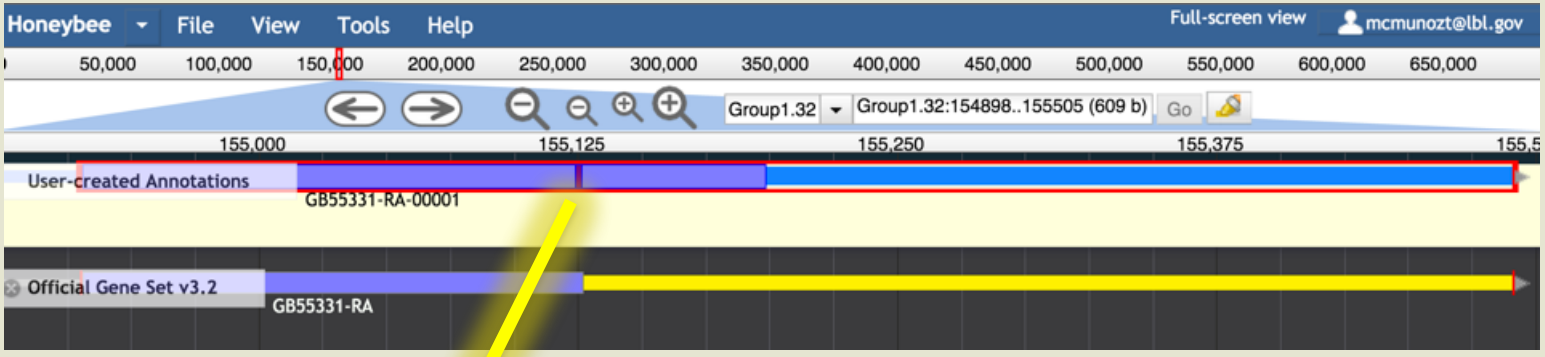
COMPLEX CASES

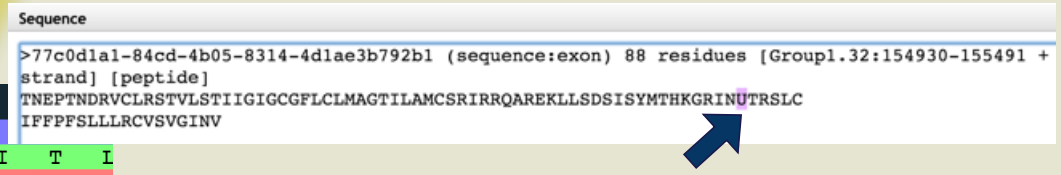
CORRECTING SELENOCYSTEINE CONTAINING PROTEINS

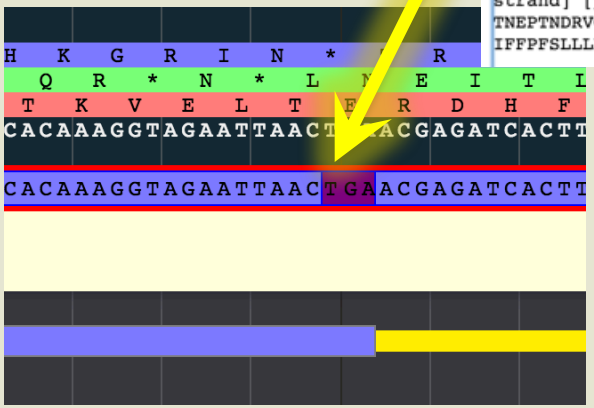


COMPLEX CASES

CORRECTING SELENOCYSTEINE CONTAINING PROTEINS







COMPLEX CASES

BERKELEY LAB

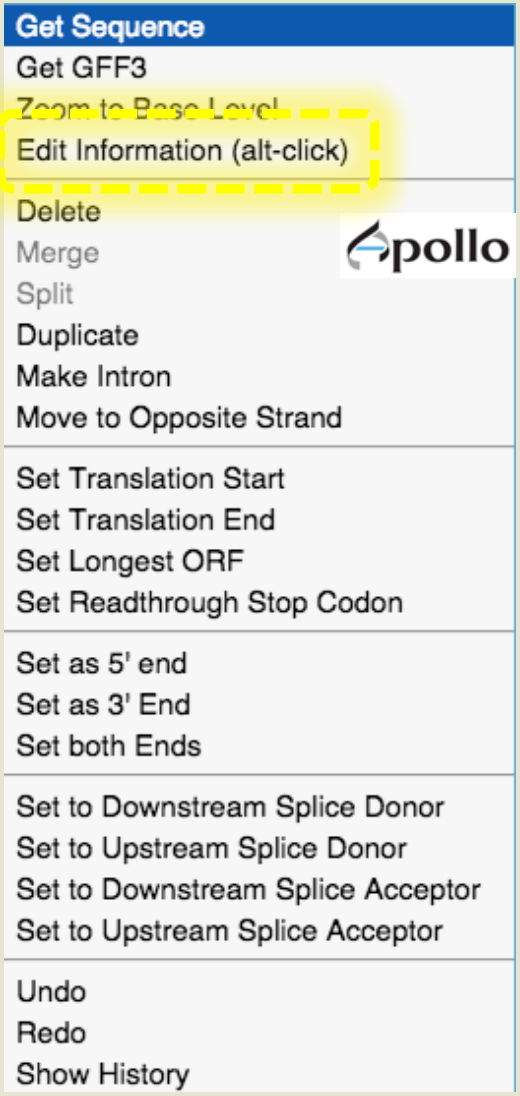
ANNOTATING FRAMESHIFTS, CORRECTING SINGLE-BASE ERRORS & SELENOCYSTEINES

1. Apollo allows annotators to make single base modifications or frameshifts that are reflected in the sequence and structure of any transcripts overlapping the modification. These manipulations do NOT change the underlying genomic sequence. If you determine that you need to make one of these changes, zoom in to the nucleotide level and right click over a single nucleotide on the genomic sequence to access a menu that provides options for creating insertions, deletions or substitutions.
2. The **'*Create Genomic Insertion*'** feature will require you to enter the necessary string of nucleotide residues that will be inserted to the right of the cursor's current location. The **'*Create Genomic Deletion*'** option will require you to enter the length of the deletion, starting with the nucleotide where the cursor is positioned. The **'*Create Genomic Substitution*'** feature asks for the string of nucleotide residues that will replace the ones on the DNA track.
3. Once you have entered the modifications, Apollo will recalculate the corrected transcript and protein sequences, which will appear when you use the right-click menu **'*Get Sequence*'** option. Since the underlying genomic sequence is reflected in all annotations that include the modified region you should alert the curators of your organisms database using the **'*Comments*'** section to report the CDS edits.
4. In special cases such as selenocysteine containing proteins (read-throughs), right-click over the offending/premature **'*Stop*'** signal and choose the **'*Set readthrough stop codon*'** option from the menu.

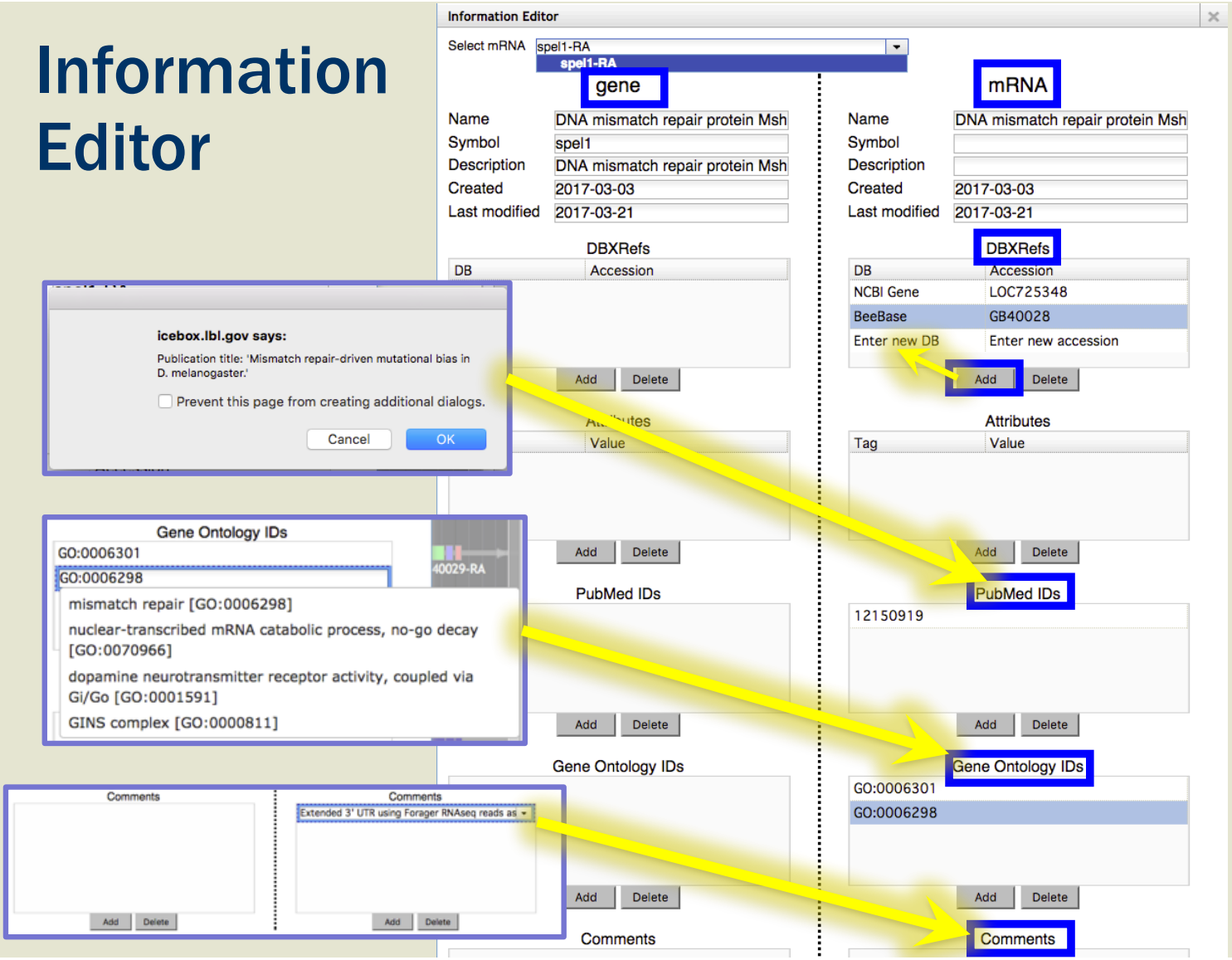
COMPLEX CASES

adding metadata

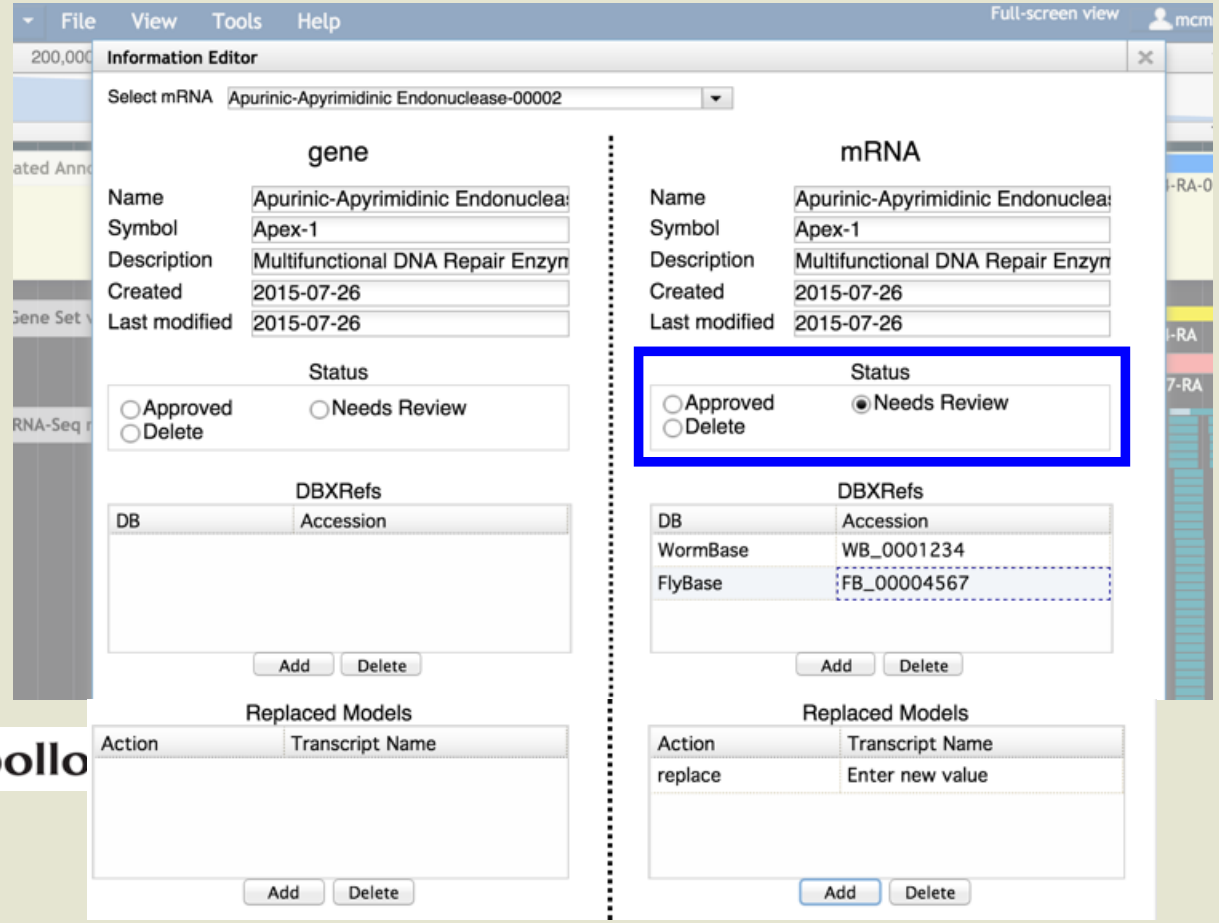
Information Editor



Information Editor



Information Editor

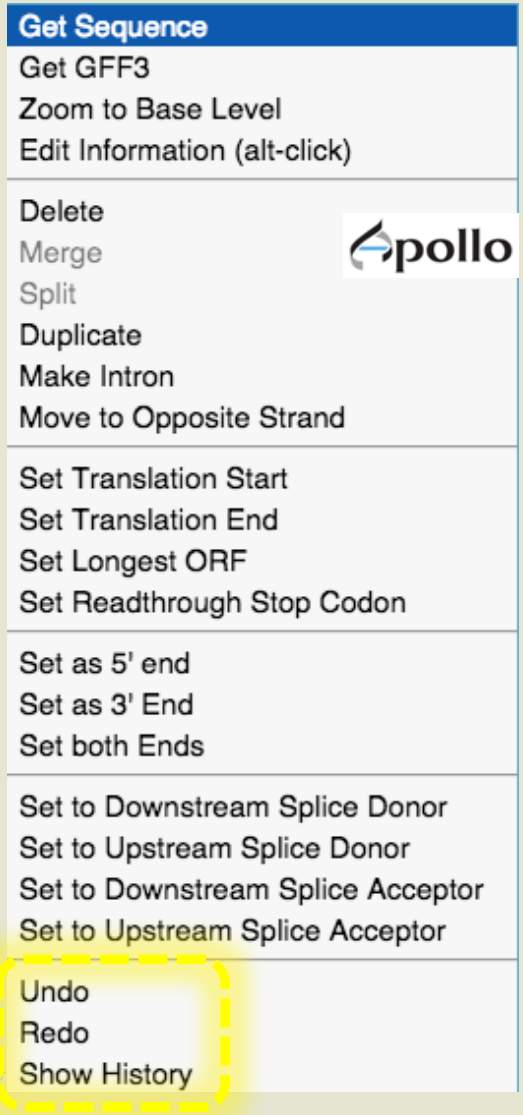


Apollo

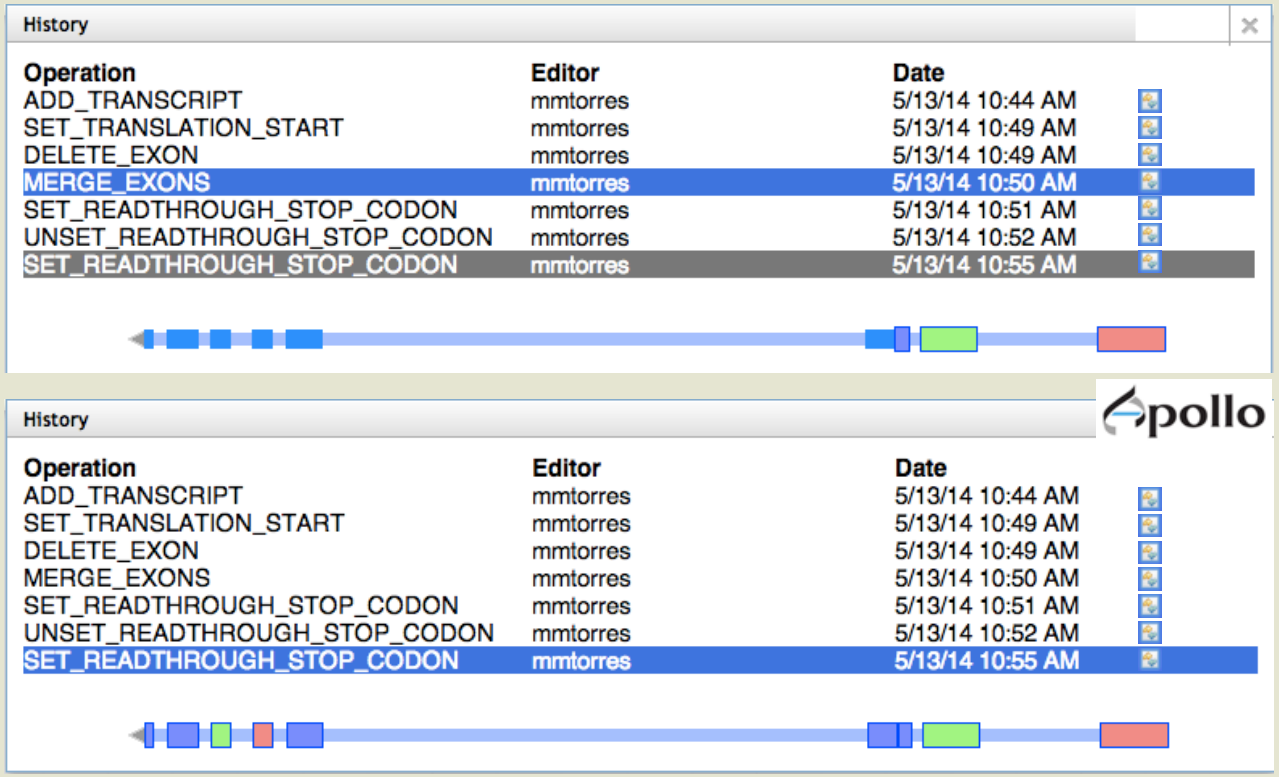
BERKELEY LAB

history

Keeping track of each edit



Annotations, annotation edits, and **History**: are stored in a centralized database.



checklist

COMPLETING THE ANNOTATION

- Follow this checklist until you are satisfied the annotation is the best representation of the underlying biology.
- And remember to…
  - comment to validate your annotation, even if you made no changes to an existing model. Think of comments as your 'vote of confidence'.
  - add a comment to inform the community of unresolved issues you think this model may have.

Always Remember: Apollo curation is a community effort so please use comments to communicate the reasons for your annotation. Your comments will be visible to everyone.

CHECKLIST
for accuracy and integrity

- Check ***'Start'*** and ***'Stop'*** sites.
- Check **splice sites**: most splice sites display these residues **…]5'-GT/AG-3'[…**
- Check if you can annotate **UTRs**, for example using RNA-Seq data:
  - align it against relevant genes/gene family
  - blastp against NCBI's RefSeq or nr
- Check & comment **gaps** in the genome.
- Additional functionality may be necessary:
  - **merge** 2 gene predictions - same scaffold
  - **'merge'** 2 gene predictions - different scaffolds
  - **split** a gene prediction
  - annotate **frameshifts**
  - annotate selenocysteines, correcting single-base and other assembly errors, etc.

- **Add**:
  - Important project information in the form of comments.
  - IDs for this gene model in public or private databases via DBXRefs, e.g. GenBank ID, gene symbol(s), common name(s), synonyms.
  - Comments about the changes you made to each gene model, if any.
  - Any appropriate functional assignments, e.g. via BLAST + HMM (e.g. InterProScan), RNA-Seq or other data of your own, literature searches, etc.

example

Apis mellifera genome data in Apollo

1. Evidence in support of protein coding gene models.

1.1 Consensus Gene Sets:
Official Gene Set v3.2
Official Gene Set v1.0

1.2 Consensus Gene Sets comparison:
OGSv3.2 genes that merge OGSv1.0 and RefSeq genes
OGSv3.2 genes that split OGSv1.0 and RefSeq genes

1.3 Protein Coding Gene Predictions Supported by Biological Evidence:
NCBI Gnomon
Fgenesh++ with RNASeq training data
Fgenesh++ without RNASeq training data
NCBI RefSeq Protein Coding Genes and Low Quality Protein Coding Genes

1.4 *Ab initio* protein coding gene predictions:
Augustus Set 12, Augustus Set 9, Fgenesh, GeneID, N-SCAN, SGP2

1.5 Transcript Sequence Alignment:
NCBI ESTs, *Apis cerana* RNA-Seq, Forager Bee Brain Illumina Contigs, Nurse Bee Brain Illumina Contigs, Forager RNA-Seq reads, Nurse RNA-Seq reads, Abdomen 454 Contigs, Brain and Ovary 454 Contigs, Embryo 454 Contigs, Larvae 454 Contigs, Mixed Antennae 454 Contigs, Ovary 454 Contigs, Testes 454 Contigs, Forager RNA-Seq HeatMap, Forager RNA-Seq XY Plot, Nurse RNA-Seq HeatMap, Nurse RNA-Seq XY Plot

Apis mellifera genome data in Apollo

1. Evidence in support of protein coding gene models (Continued).

1.6 Protein homolog alignment:
Acep_OGSv1.2
Aech_OGSv3.8
Cflo_OGSv3.3
Dmel_r5.42
Hsal_OGSv3.3
Lhum_OGSv1.2
Nvit_OGSv1.2
Nvit_OGSv2.0
Pbar_OGSv1.2
Sinv_OGSv2.2.3
Znev_OGSv2.1
Metazoa_Swissprot

2. Evidence in support of non protein coding gene models

2.1 Non-protein coding gene predictions:
NCBI RefSeq Noncoding RNA
NCBI RefSeq miRNA

2.2 Pseudogene predictions:
NCBI RefSeq Pseudogene

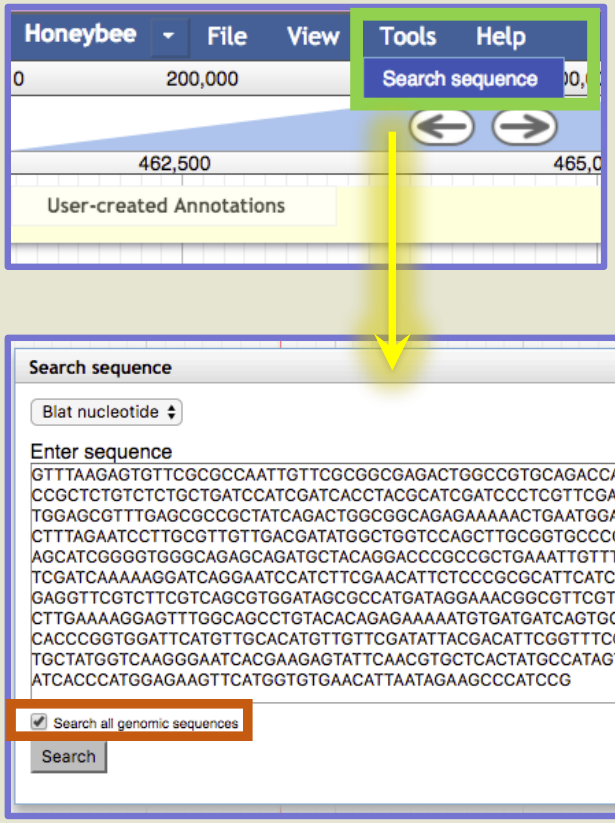
Ceramidase

Ceramidase is an enzyme, which cleaves fatty acids from ceramide, producing sphingosine (SPH), which in turn is phosphorylated by a sphingosine kinase to form sphingosine-1-phosphate (S1P). Ceramide, SPH, and S1P are bioactive lipids that mediate cell proliferation, differentiation, apoptosis, adhesion, and migration.

It has come to our attention that the honey bee Apis mellifera *ortholog of Ceramidase is fragmented into 2 or more genes in the current gene set (Official Gene Set v3.2).*

Interrogate the genome using Blat

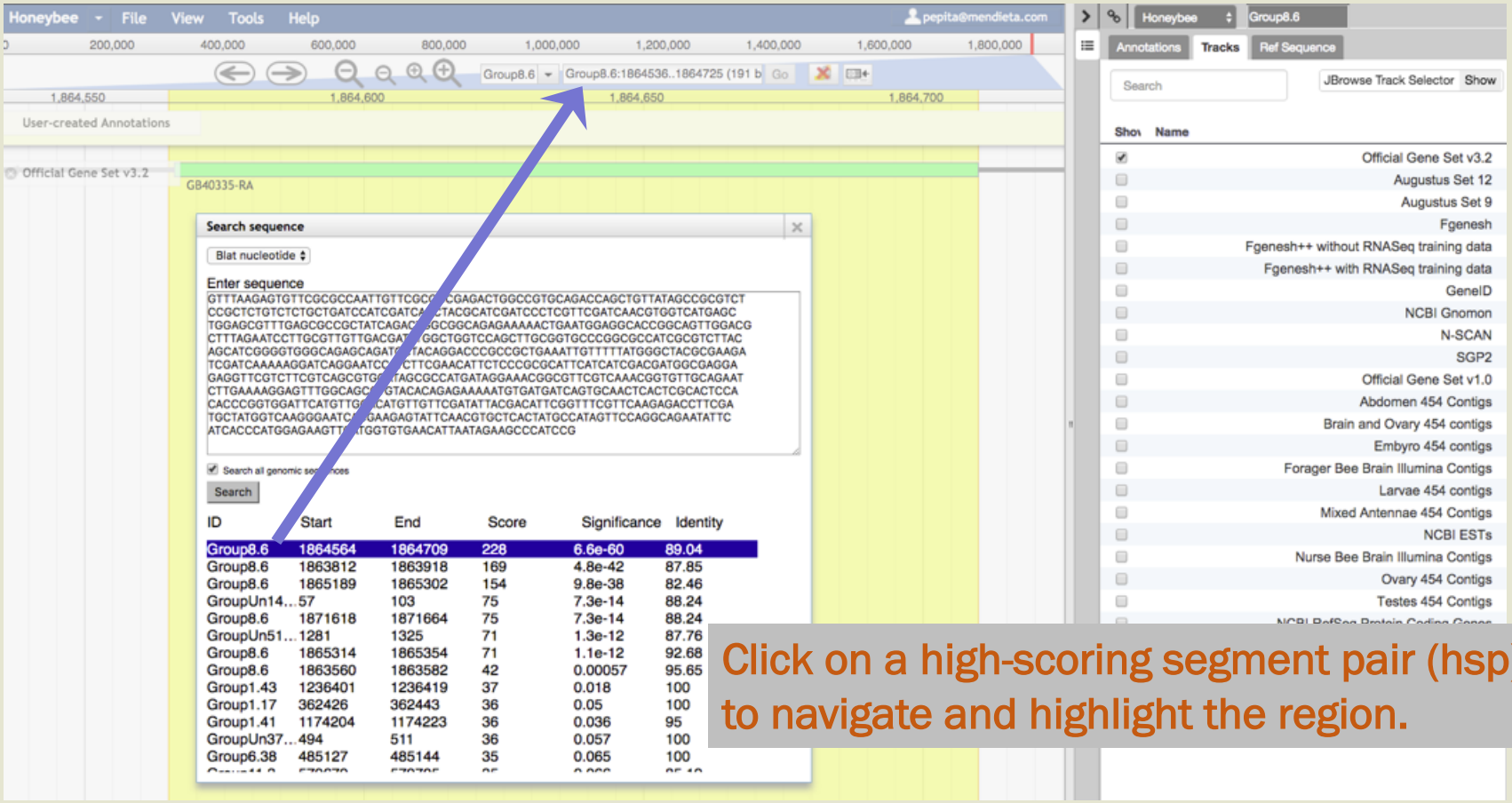
```
--------------------------------------------------------------------------------
Apollo Workshop -
Exercise 5

>B_terrestris_Ceramidase-like
GTTTAAGAGTGTTCGCGCCAATTGTTCGCGGCGAGACTGGCCGTGCAGACCAGCTGTTATAGCCGCGTCT
CCGCTCTGTCTCTGCTGATCCATCGATCACCTACGCATCGATCCCTCGTTCGATCAACGTGGTCATGAGC
TGGAGCGTTTGAGCGCCGCTATCAGACTGGCGGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG
CTTTAGAATCCTTGCGTTGTTGACGATATGGCTGGTCCAGCTTGCGGTGCCCGGCGCCATCGCGTCTTAC
AGCATCGGGGTGGGCAGAGCAGATGCTACAGGACCCGCCGCTGAAATTGTTTTTATGGGCTACGCGAAGA
TCGATCAAAAAGGATCAGGAATCCATCTTCGAACATTCTCCCGCGCATTCATCATCGACGATGGCGAGGA
GAGGTTCGTCTTCGTCAGCGTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGTTGCAGAAT
CTTGAAAAGGAGTTTGGCAGCCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA
CACCCGGTGGATTCATGTTGCACATGTTGTTCGATATTACGACATTCGGTTTCGTTCAAGAGACCTTCGA
TGCTATGGTCAAGGGAATCACGAAGAGTATTCAACGTGCTCACTATGCCATAGTTCCAGGCAGAATATTC
ATCACCCATGGAGAAGTTCATGGTGTGAACATTAATAGAAGCCCATCCG
```



Search all genomic sequences

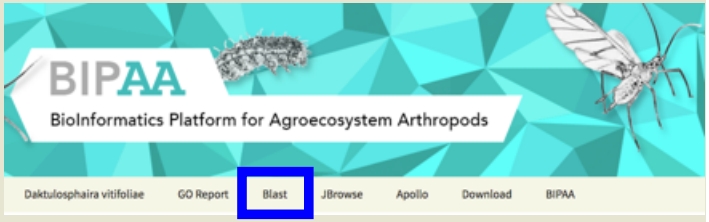
BERKELEY LAB

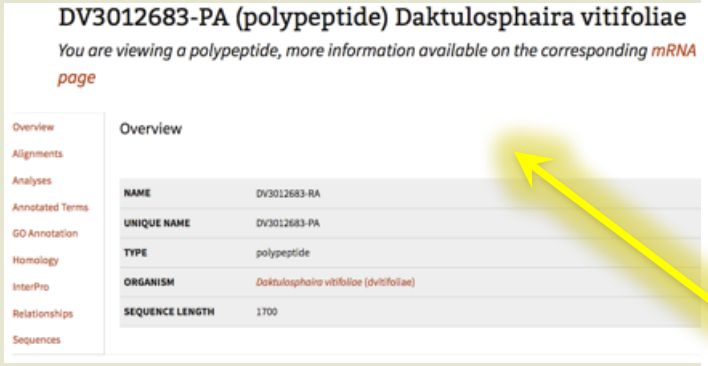
Blat results

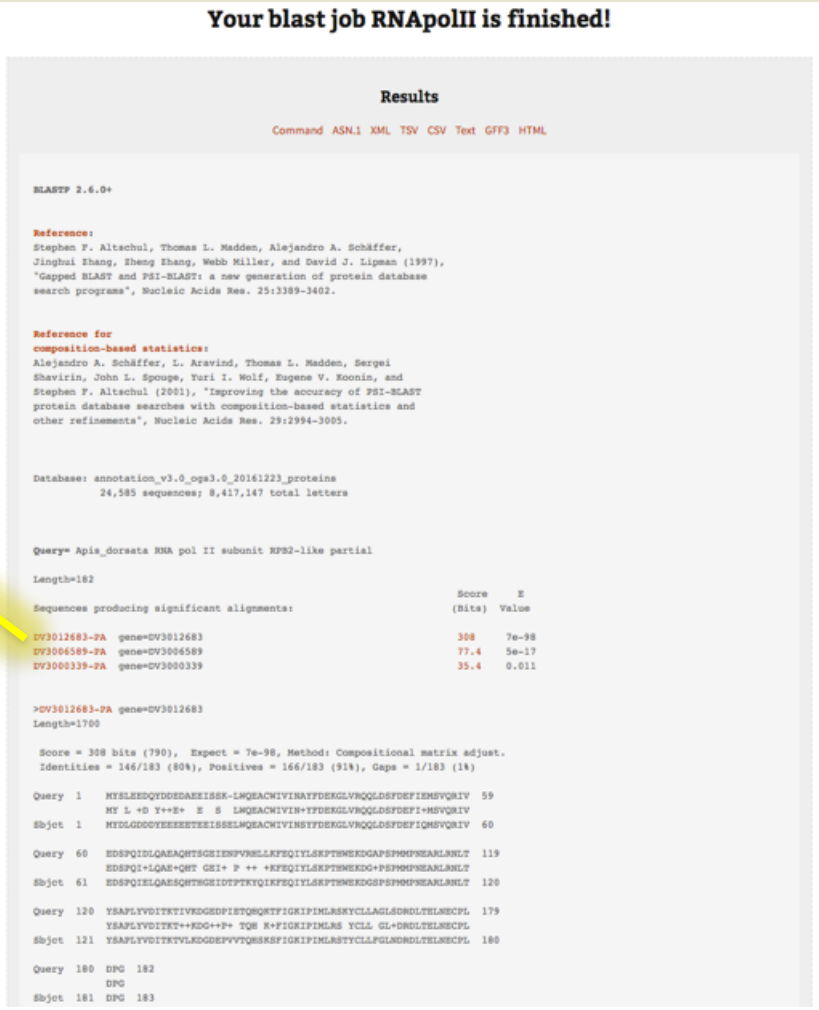


Click on a high-scoring segment pair (hsp) to navigate and highlight the region.

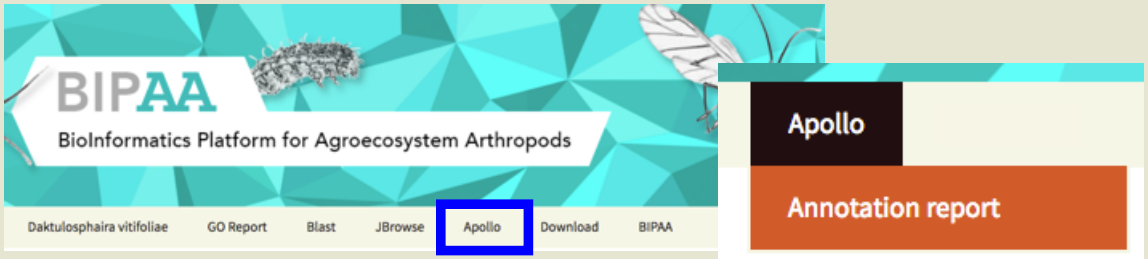
BIPAA resources - blast

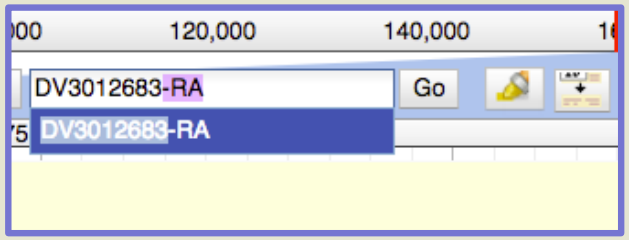




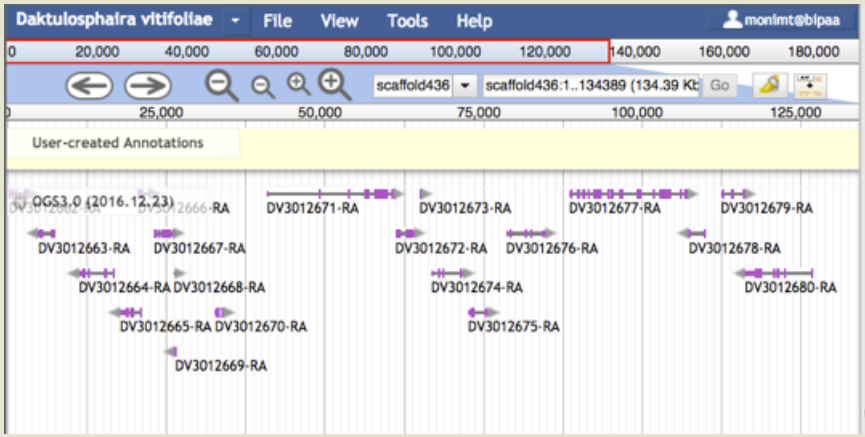


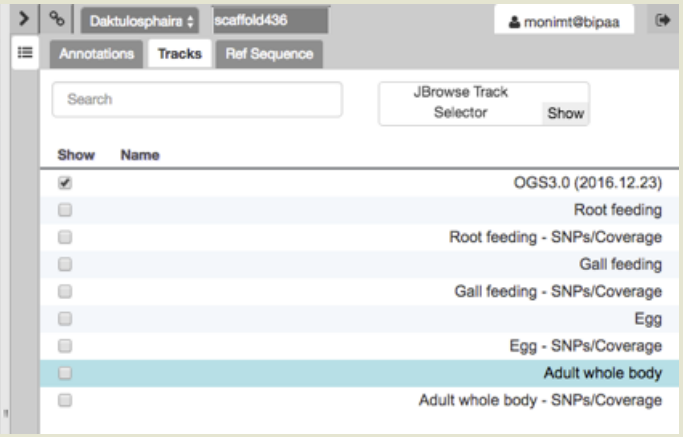
BIPAA resources - Apollo



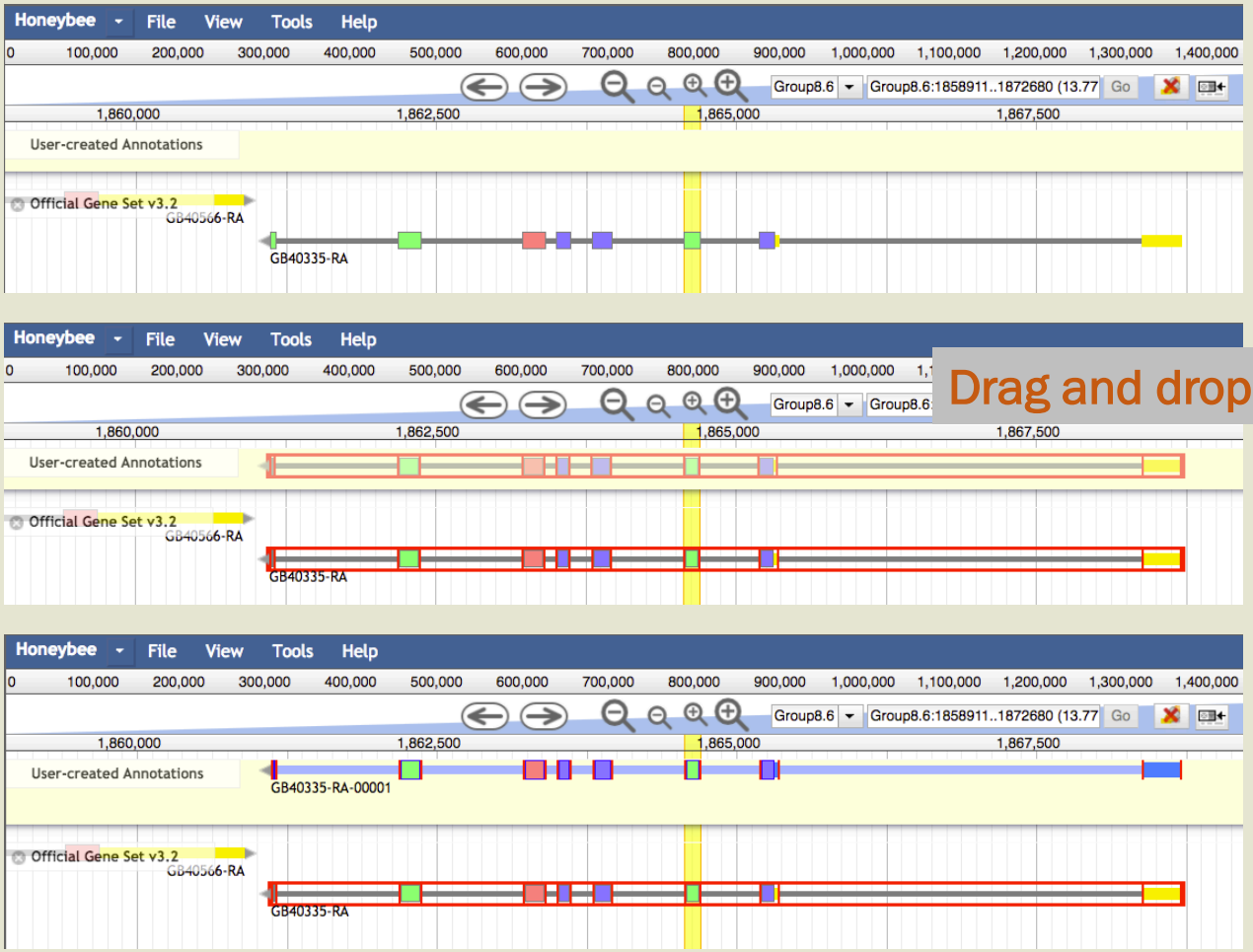


You may find candidate genes from blast results using the 'Search' box with coordinates in main window.



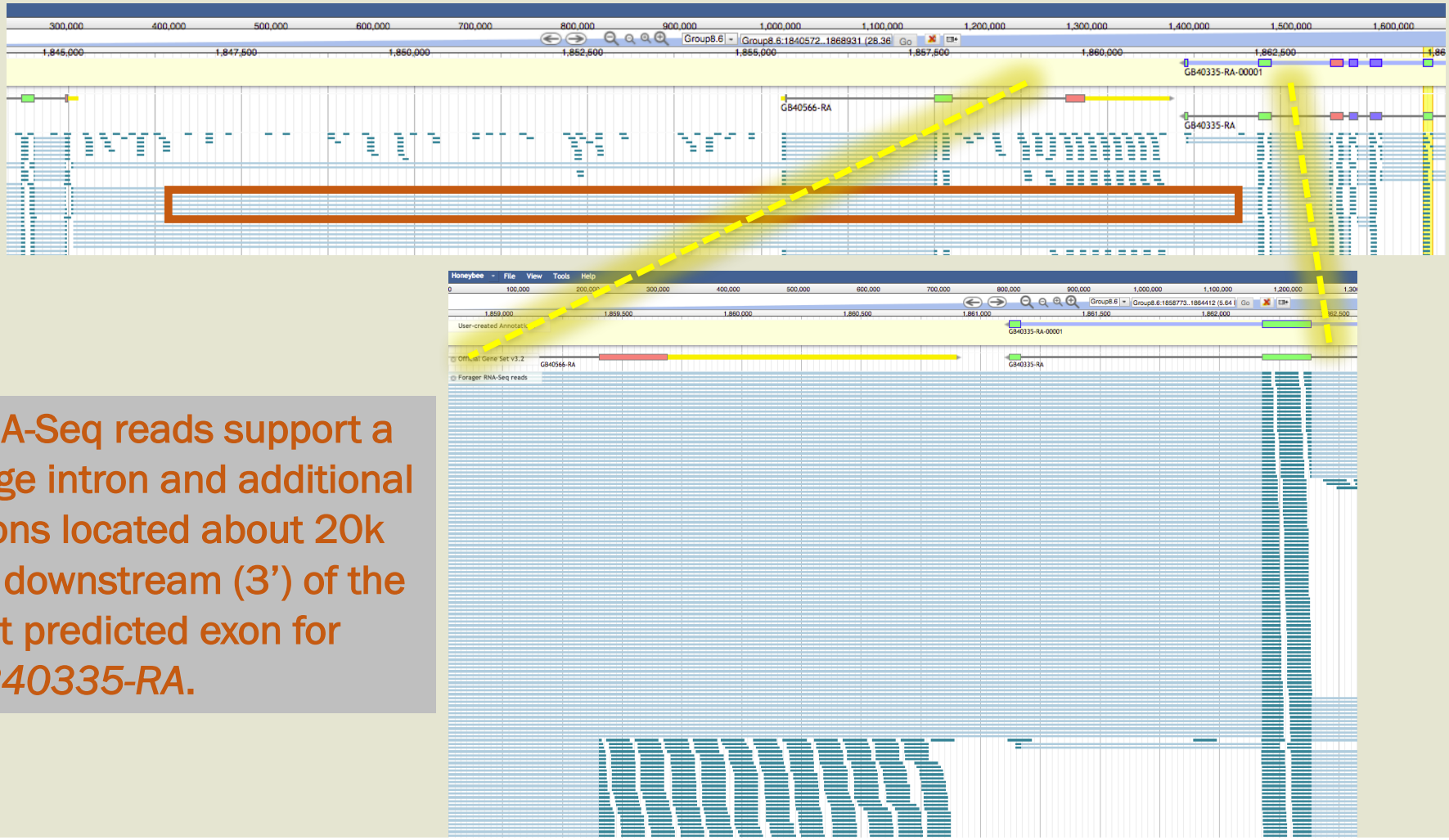


Create a new annotation



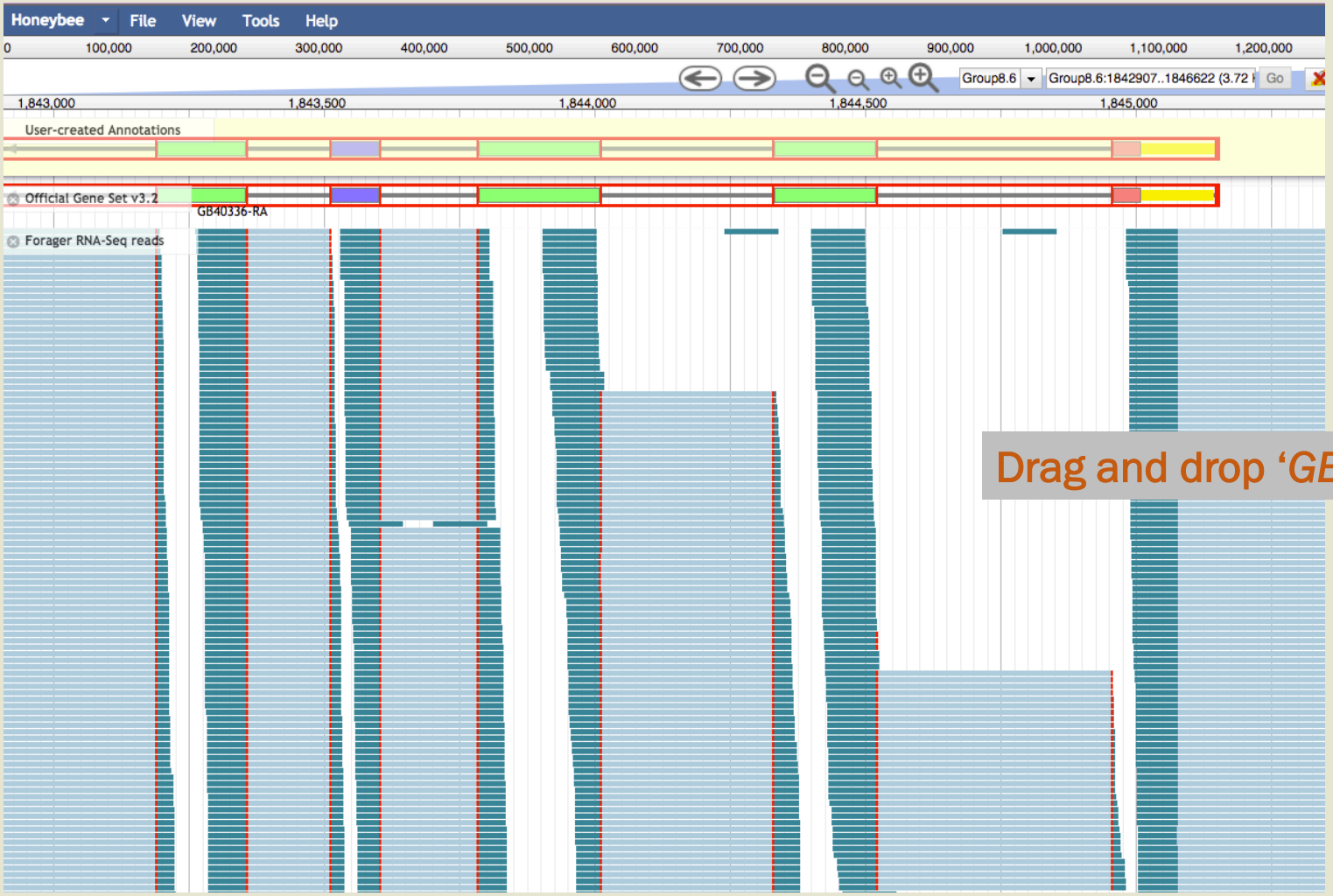
Drag and drop '*GB40335-RA*'

Transcriptomic data support a longer gene



RNA-Seq reads support a large intron and additional exons located about 20k bp downstream (3') of the last predicted exon for *GB40335-RA*.

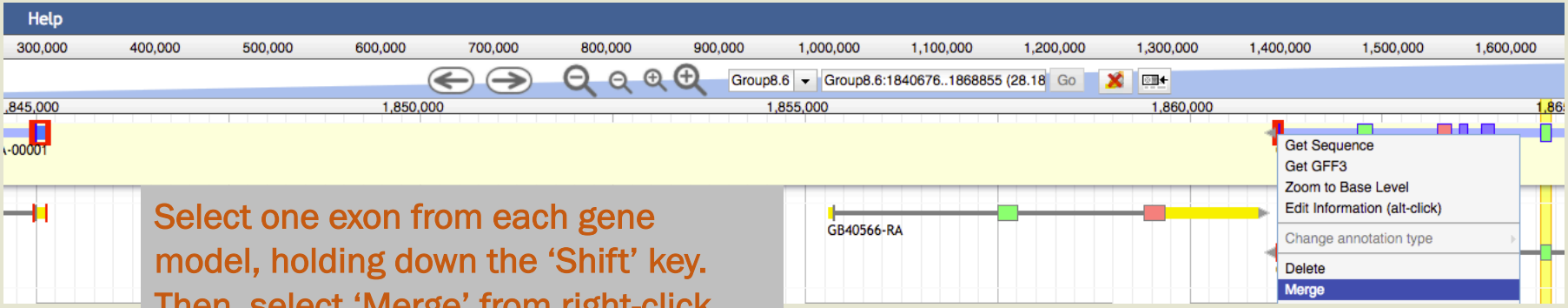
Transcriptomic data support a longer gene



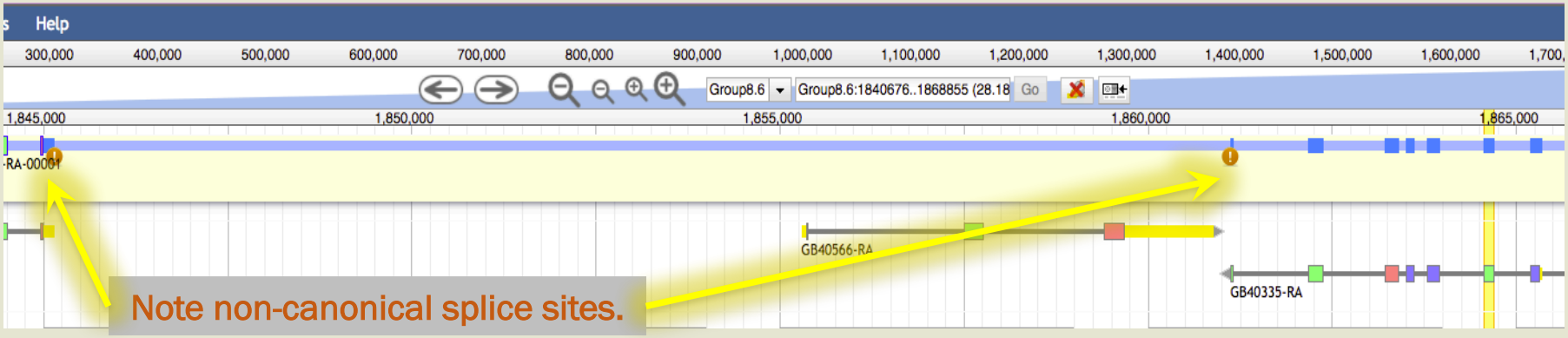
Drag and drop '*GB40336-RA*'

BERKELEY LAB

Merge transcripts

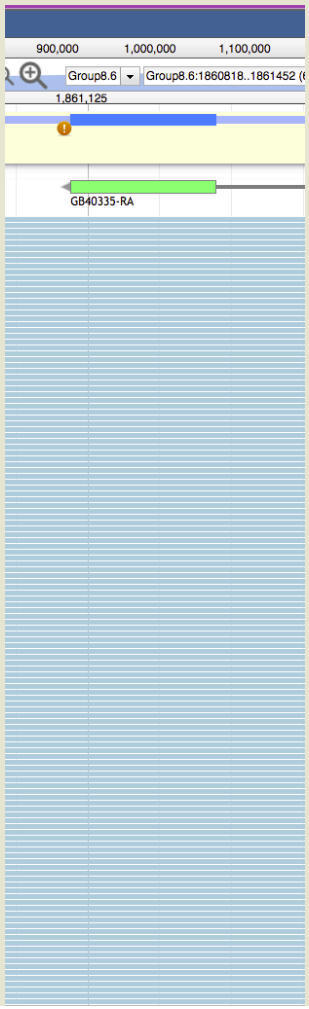


Select one exon from each gene model, holding down the 'Shift' key. Then, select 'Merge' from right-click menu to bring gene models together.

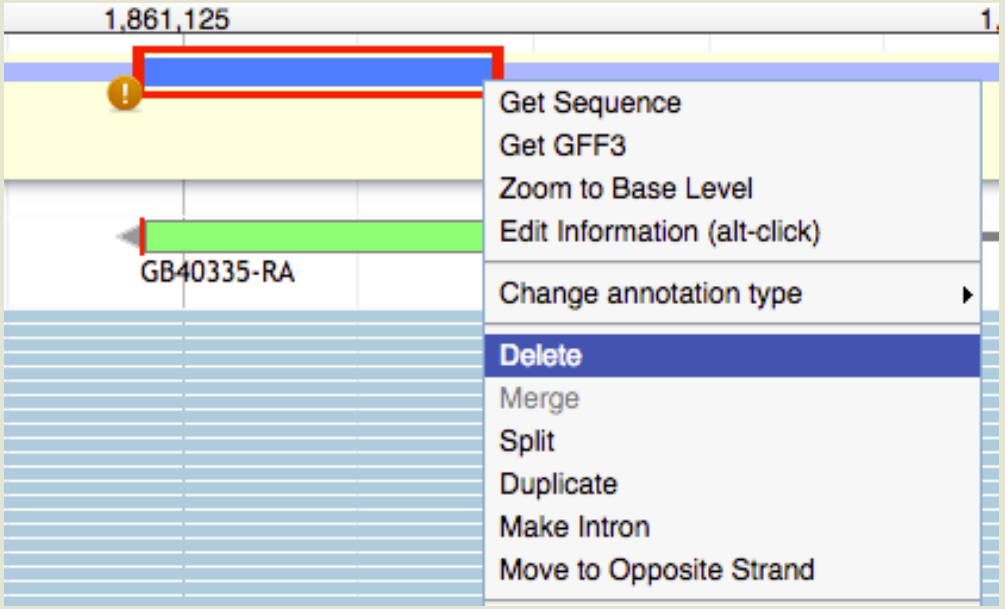


Note non-canonical splice sites.

Exon not supported by RNA-Seq data

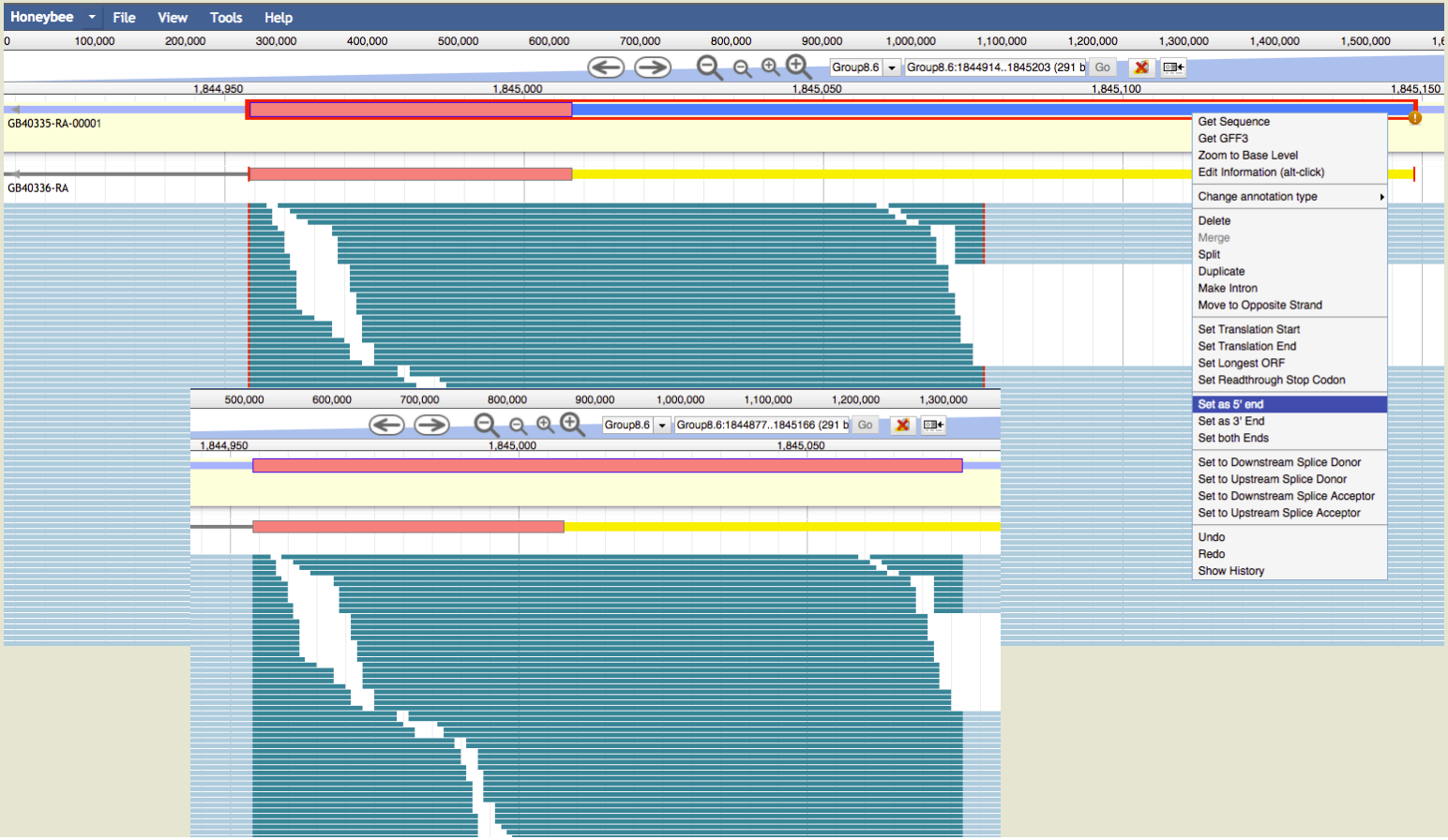


At the end of *GB40335-RA*, select last exon and right-click to choose the 'Delete' option.

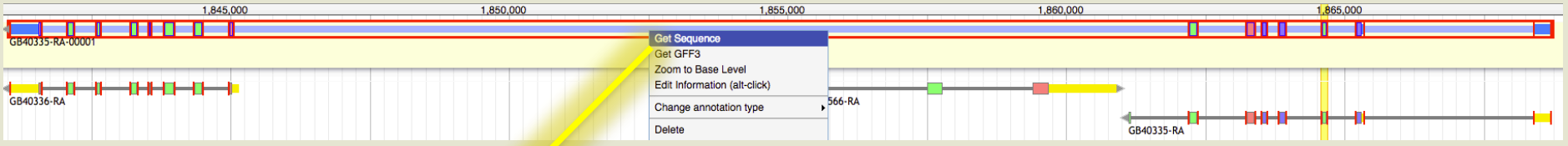


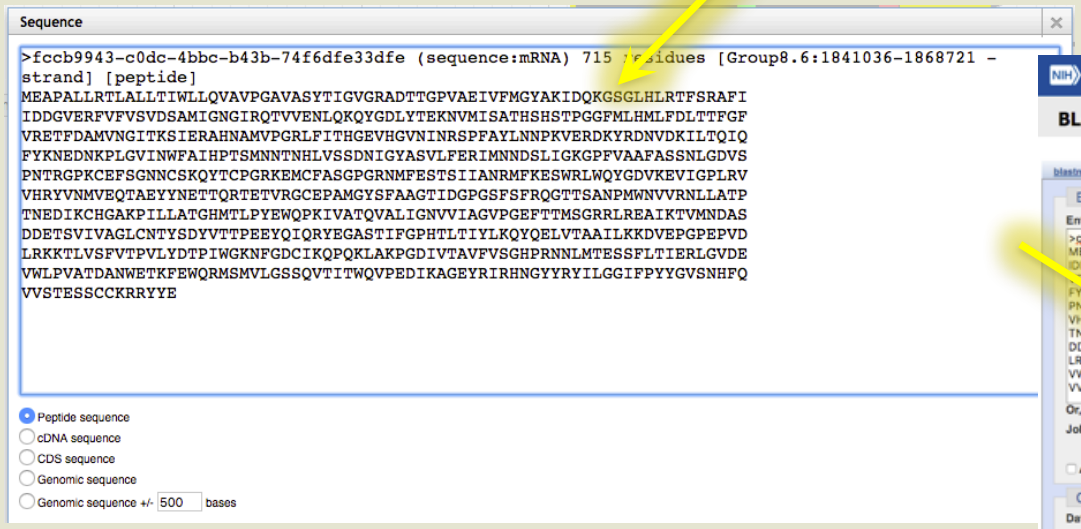
Fix remaining non-canonical splice site

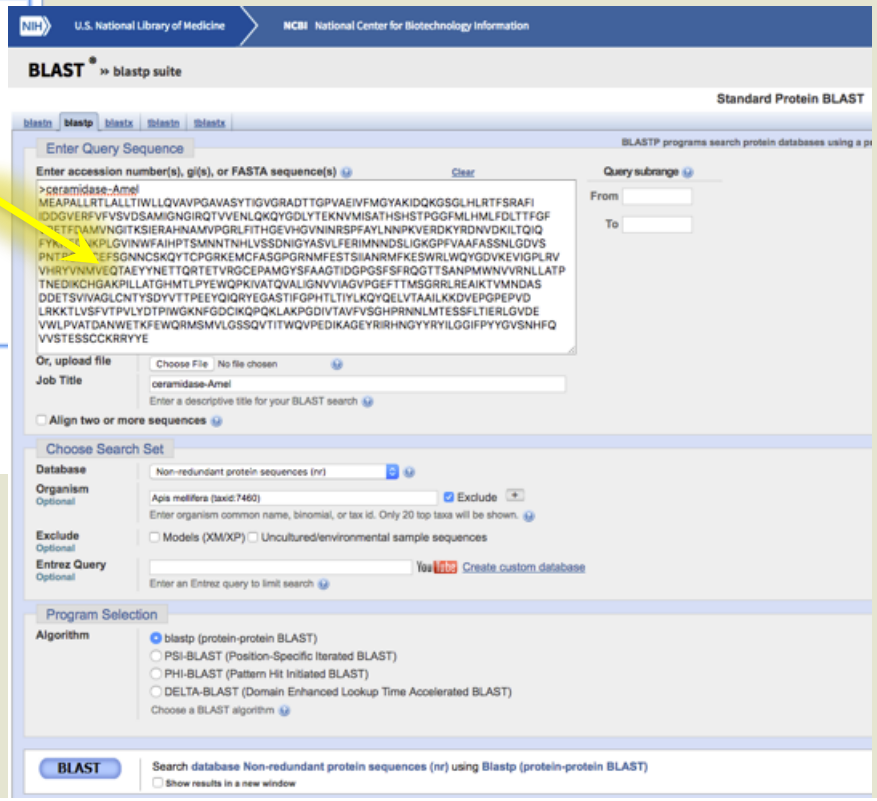
Now, on the other offending exon (was first exon of *GB40336-RA*), use RNA-seq reads - or use '*Set Downstream Splice Acceptor*', or drag the intron/exon boundary manually - to use a canonical splice site.



Retrieve resulting peptide, compare to public databases

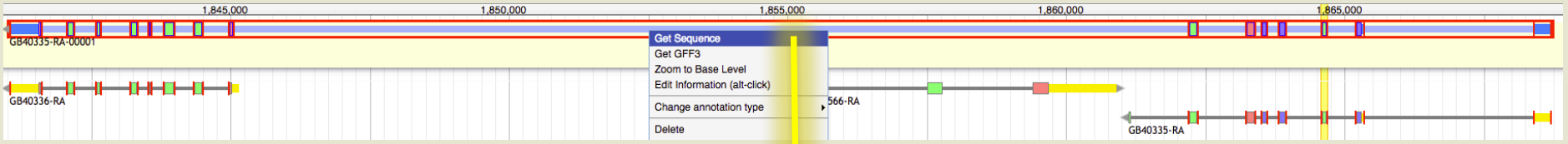


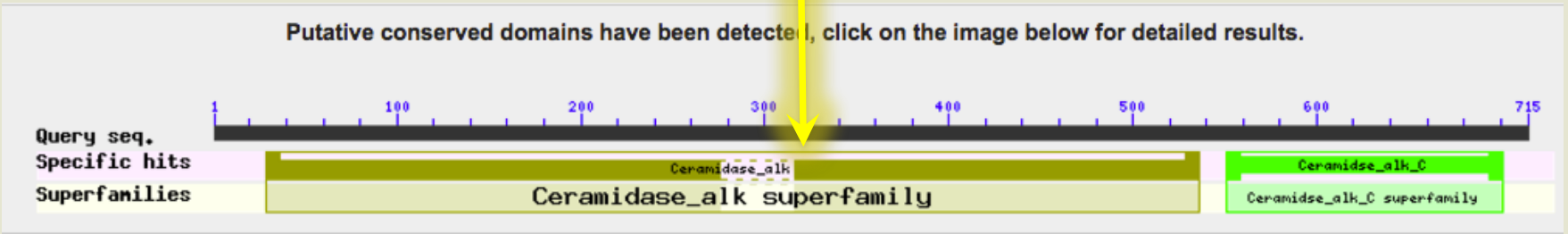




BERKELEY LAB

Results from NCBI blastp vs nr





Sequences producing significant alignments:

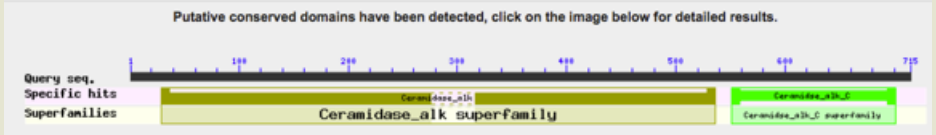
Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: neutral ceramidase [Apis cerana]	1471	1471	100%	0.0	98%	XP_016908167.1
PREDICTED: neutral ceramidase-like isoform X1 [Apis dorsata]	1470	1470	100%	0.0	98%	XP_006612924.1
PREDICTED: neutral ceramidase [Apis florea]	1439	1439	100%	0.0	96%	XP_003691475.1
PREDICTED: neutral ceramidase [Bombus terrestris]	1328	1328	100%	0.0	87%	XP_003397164.1
PREDICTED: neutral ceramidase [Bombus impatiens]	1324	1324	100%	0.0	86%	XP_003489963.1
PREDICTED: neutral ceramidase isoform X1 [Eufriesea mexicana]	1301	1301	100%	0.0	85%	XP_017756753.1
PREDICTED: neutral ceramidase [Ceratina calcarata]	1267	1267	100%	0.0	83%	XP_017893250.1
PREDICTED: neutral ceramidase isoform X2 [Megachile rotundata]	1263	1263	98%	0.0	83%	XP_003703614.1
PREDICTED: neutral ceramidase isoform X1 [Megachile rotundata]	1253	1253	98%	0.0	82%	XP_012141148.1

BERKELEY LAB

Add metadata in '*Information Editor*'



List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	Ceramidase_alk	pfam04734	Neutral/alkaline non-lysosomal ceramidase, N-terminal; This family represents N-terminal ...	29-536	0e+00
[+]	Ceramidse_alk_C	pfam17048	Neutral/alkaline non-lysosomal ceramidase, C-terminal; This family represents C-terminal ...	551-701	4.77e-76

Don't forget!

Information Editor

Select mRNA: GB40335-RA-00001

gene		mRNA	
Name	neutral ceramidase	Name	neutral ceramidase-00001
Symbol	CDase	Symbol	
Description	Enzyme, cleaves fatty acids from (	Description	
Created	2017-03-22	Created	2017-03-22
Last modified	2017-03-22	Last modified	2017-03-22

DBXRefs

DB	Accession

Add Delete

DBXRefs

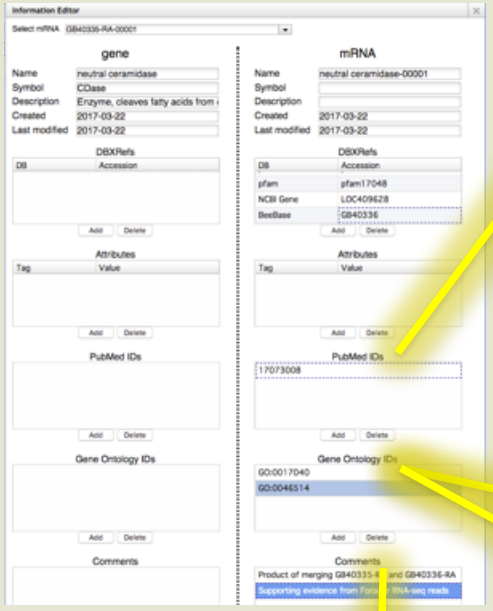
DB	Accession
pfam	pfam17048
NCBI Gene	LOC409628
BeeBase	GB40336

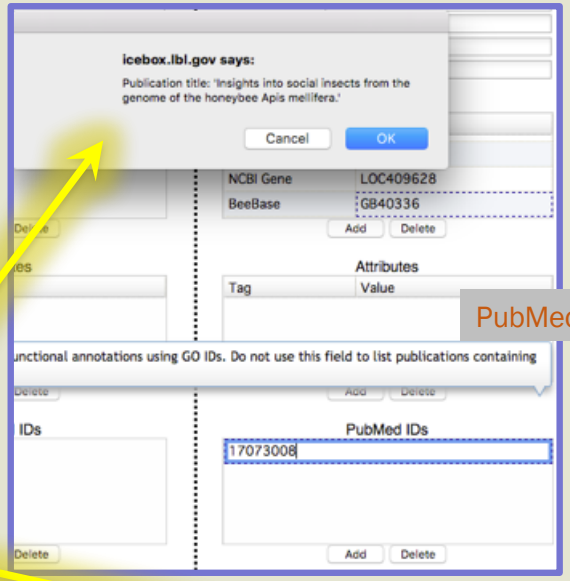
Add Delete

Nice to have

BERKELEY LAB

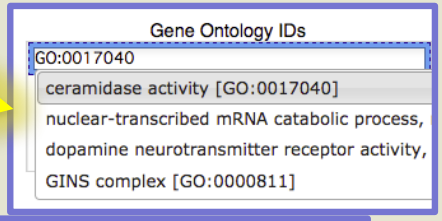
Add metadata in '*Information Editor*'

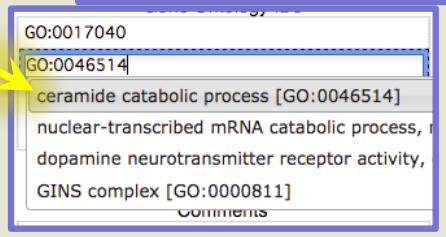




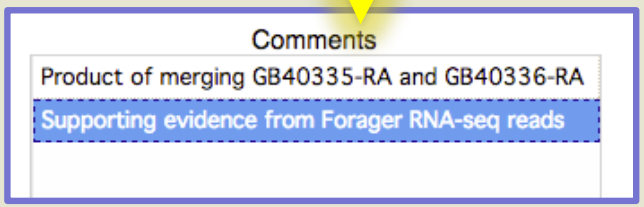
PubMed Identifiers

Gene Ontology terms





Comments



BERKELEY LAB

Public demo instances

APOLLO ON THE WEB
instructions

- Public Honey bee demo available at:

  genomearchitect.org/demo/

- Username:

  demo@demo.com

- Password:

  demo

Apollo

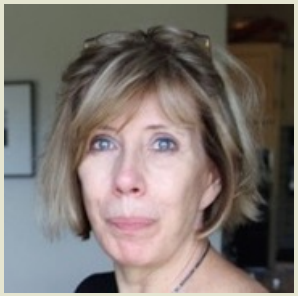
APOLLO demonstration

Demonstration video available at
http://bit.ly/apollo-video1

Apollo

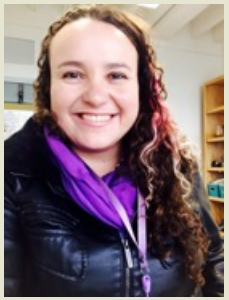
Apollo Development

BBOP



Suzi Lewis
Principal Investigator

Nathan Dunn
Technical Lead



Moni Munoz-Torres
Project Manager

JBrowse. Ian Holmes' Lab
University of California, Berkeley

Eric Yao

Christine Elsik's Lab,
University of Missouri

Deepak Unni

Thank You.

Berkeley Bioinformatics Open-Source Projects,
Environmental Genomics & Systems Biology,
Lawrence Berkeley National Laboratory

Suzanna Lewis & Chris Mungall

Seth Carbon (GO - Noctua / AmiGO)

Eric Douglas (GO / Monarch Initiative)

Nathan Dunn (Apollo)

Monica Munoz-Torres (Apollo / GO)

Collaborators

- **Ian Holmes, Eric Yao,** UC Berkeley (JBrowse)
- **Chris Elsik**, **Deepak Unni**, U of Missouri (Apollo)
- **Paul Thomas,** USC (Noctua)
- Monica Poelchau, USDA/NAL (Apollo)
- Gene Ontology Consortium (GOC)
- i5k Community

Funding

- Work for GOC is supported by NIH grant 5U41HG002273-14 from NHGRI.
- Apollo is supported by NIH grants 5R01GM080203 from NIGMS, and 5R01HG004483 from NHGRI.
- BBOP is also supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

berkeleybop.org

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY

Berkeley UNIVERSITY OF CALIFORNIA

Apollo