



**BERKELEY LAB**  
LAWRENCE BERKELEY NATIONAL LABORATORY



# Apollo

## Collaborative genome annotation editing

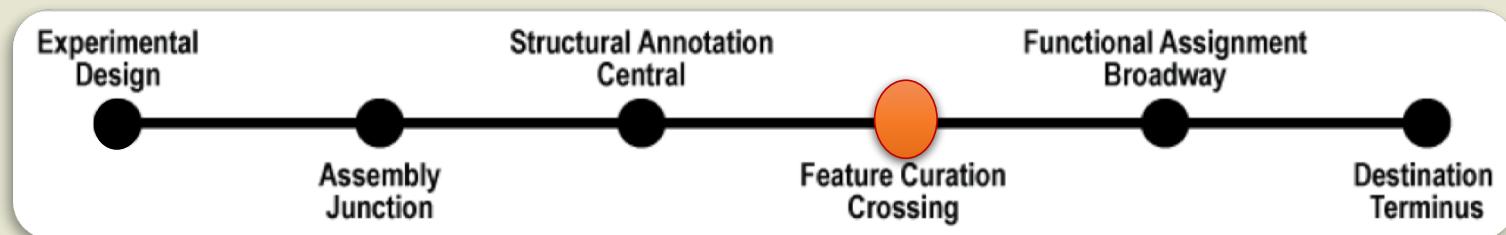
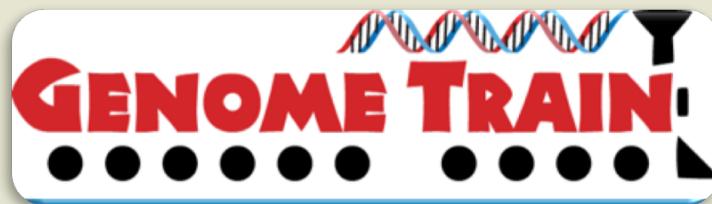
*A workshop for the Arthropod Genomics Community*

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>





editing functionality



begin with a new gene model

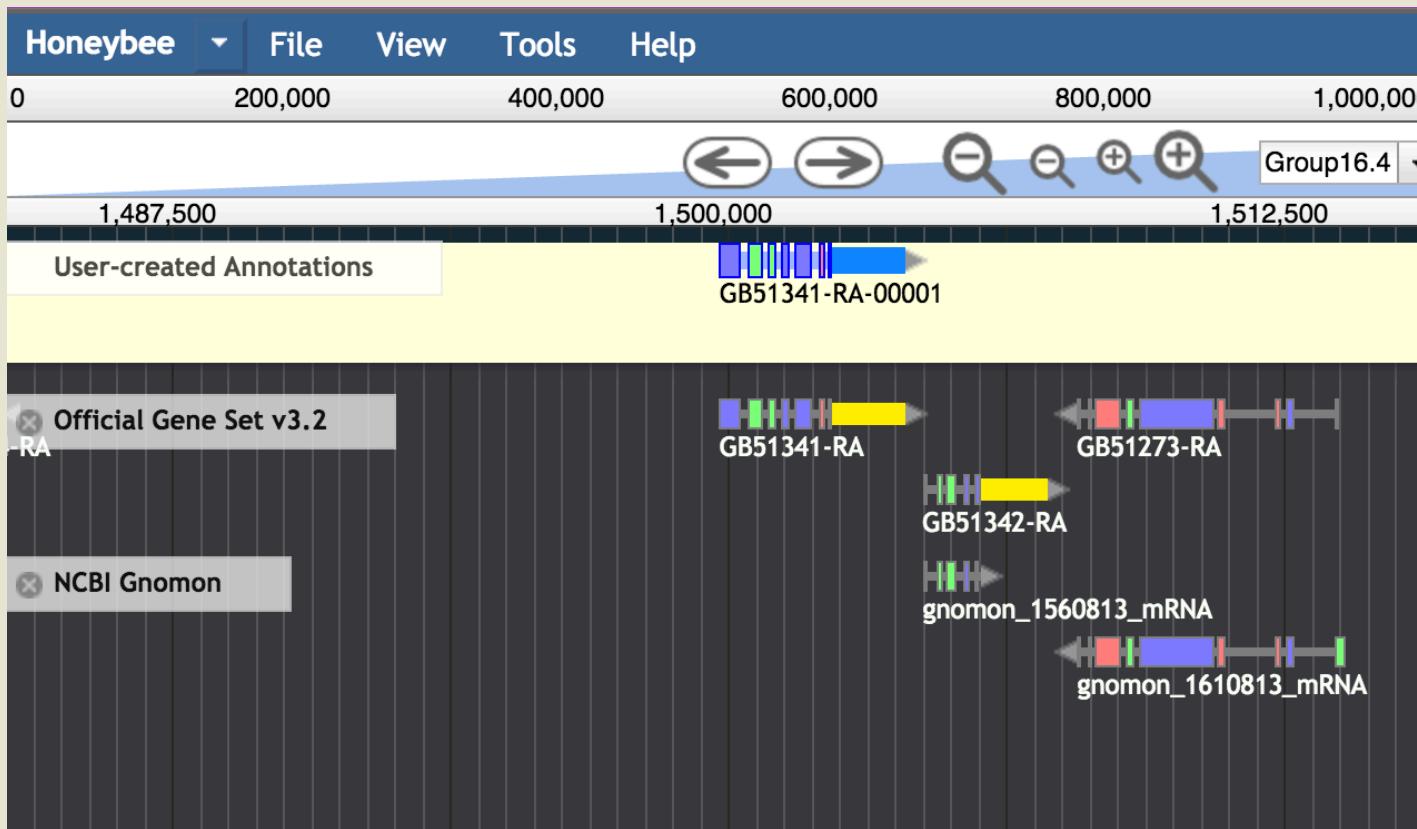
# Creating a new annotation



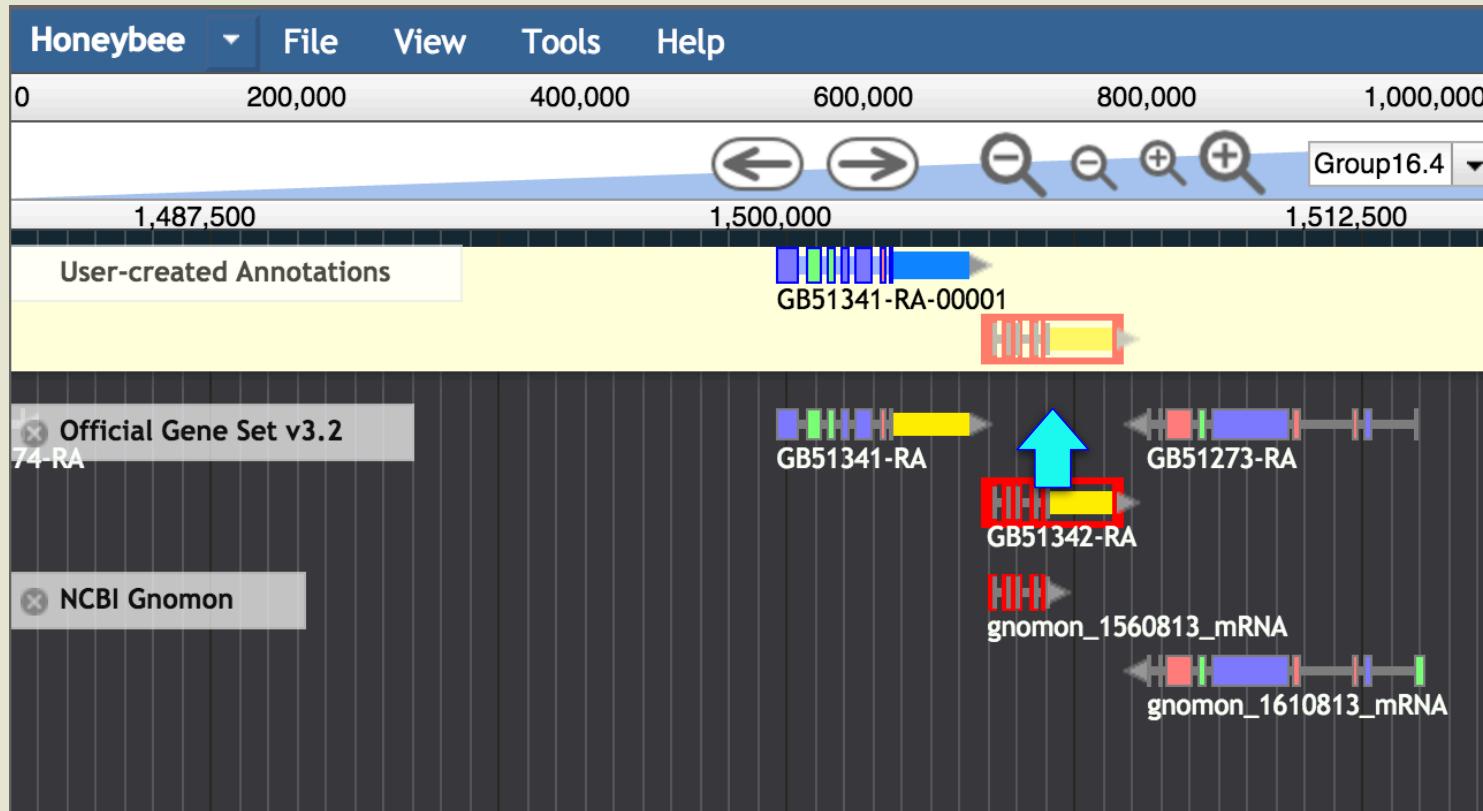
- Choose appropriate evidence from list of “Tracks” on **annotator panel**.
- Select & drag elements from evidence track into the ‘*User-created Annotations*’ area.
- Hovering over annotation in progress brings up an information pop-up.



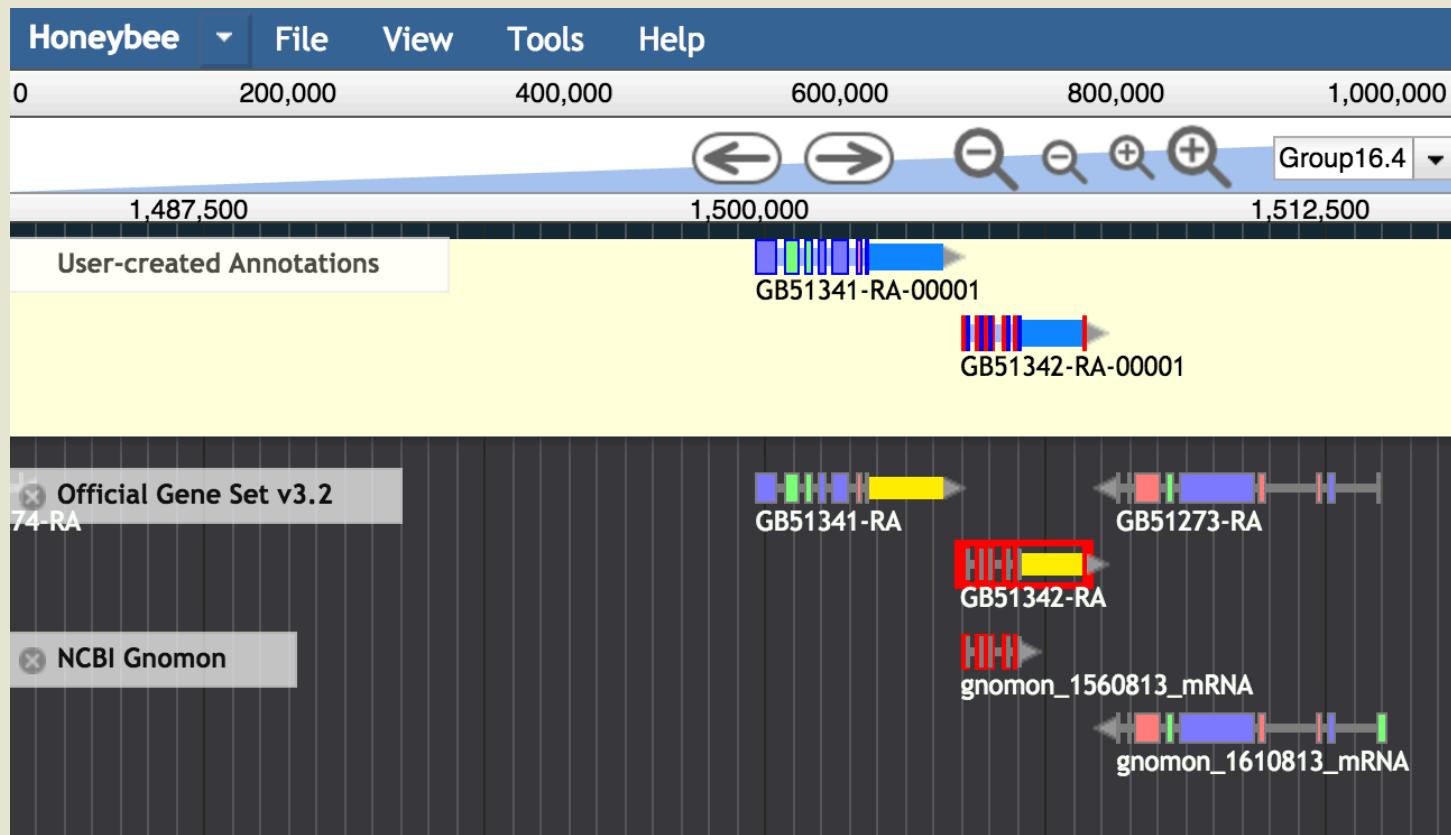
# Adding a gene model



# Adding a gene model



# Adding a gene model

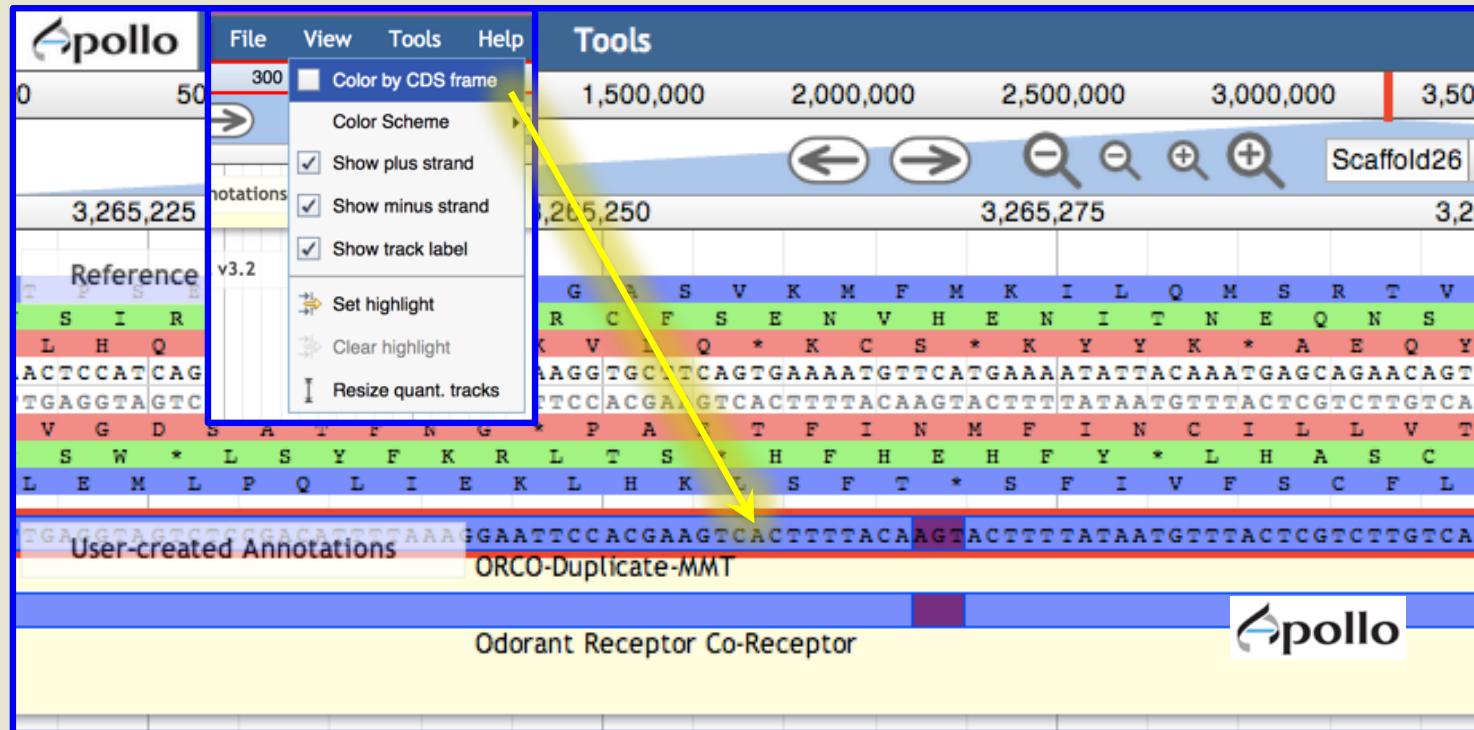


the sequence track

- ‘Zoom to base level’ reveals the sequence track.

The screenshot shows the Apollo software interface. At the top is a menu bar with File, View, Help, and Tools. Below the menu is a horizontal scale from 0 to 3,500,000, with a red vertical marker at 3,500,000. A toolbar below the scale includes arrows for navigation, a magnifying glass for search, and a plus sign for zoom. The main area displays a protein sequence alignment and a DNA sequence track. The DNA sequence is highlighted with a yellow arrow pointing to a specific base pair. A context menu is open over this highlighted area, listing options: Get Sequence, Get GFF3, Zoom to Base Level, Edit Information (alt-click), and Change annotation type. The 'Zoom to Base Level' option is highlighted with a blue background. The bottom of the interface features several tracks: 'User-created Annotations' (ORCO-Duplicate-MMT), 'Odorant Receptor Co-Receptor', and a reference sequence track above it. The Berkeley Lab logo is in the bottom left corner, and the Apollo logo is in the bottom right corner.

## Color exons by CDS from the 'View' menu.



## Toggle reference DNA sequence and translation frames in forward strand.

Also, toggle models in either direction.

The screenshot shows the Apollo genome browser interface. At the top, there is a menu bar with File, View, Help, and Tools. Below the menu is a coordinate track showing positions from 0 to 2,500,000. A yellow arrow points from the "Tools" menu to a context menu that includes options like "Toggle Reverse Strand", "Toggle Protein Translation", "Create Genomic Insertion", "Create Genomic Deletion", and "Create Genomic Substitution".

The main panel displays a "Reference sequence" with amino acid translations above the DNA sequence. A yellow arrow points from the "View" menu in the bottom-left corner to a submenu where the "Show minus strand" option is highlighted. Other options in the submenu include "Color by CDS frame", "Color Scheme", "Show plus strand", "Show track label", "Set highlight", "Clear highlight", and "Resize quant. tracks".

A green callout box in the bottom-right corner provides instructions: "Zoom in/out with keyboard: shift + arrow keys up/down". The Apollo logo is located in the bottom right corner.

**BERKELEY LAB**  
Lawrence Berkeley National Laboratory

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.



SIMPLE CASES

# Editing functionality



SIMPLE CASES

## Get Sequence

Get GFF3  
Zoom to Base Level  
Edit Information (alt-click)

Delete

Merge

Split

Duplicate

Make Intron

Move to Opposite Strand

Set Translation Start

Set Translation End

Set Longest ORF

Set Readthrough Stop Codon

Set as 5' end

Set as 3' End

Set both Ends

Set to Downstream Splice Donor

Set to Upstream Splice Donor

Set to Downstream Splice Acceptor

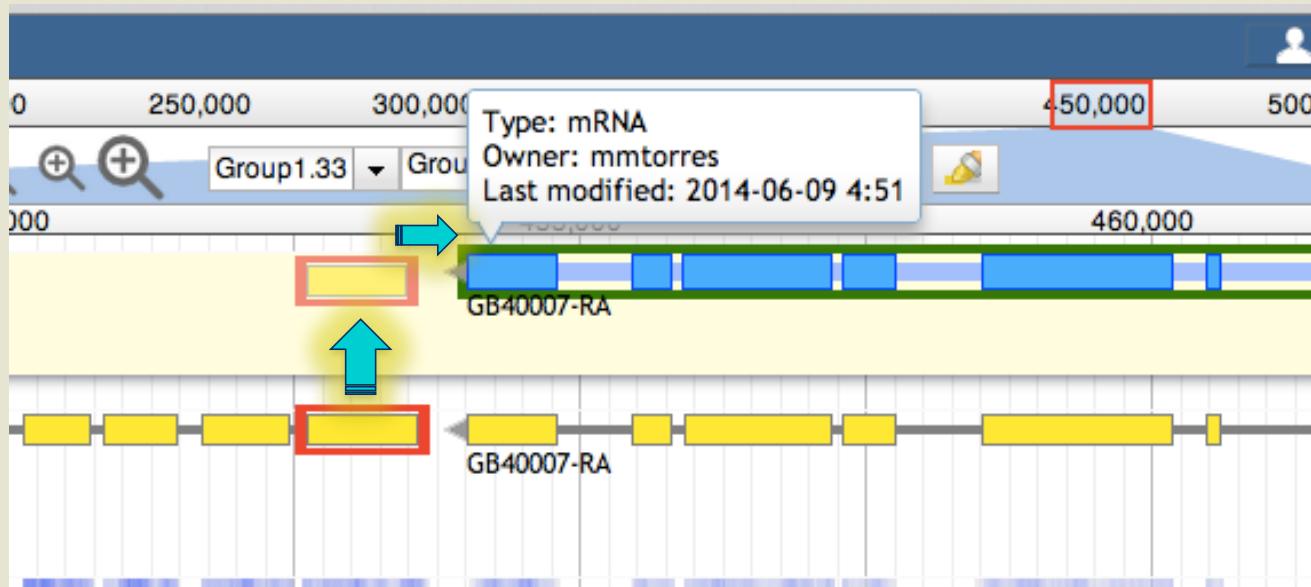
Set to Upstream Splice Acceptor

Undo

Redo

Show History

## 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.

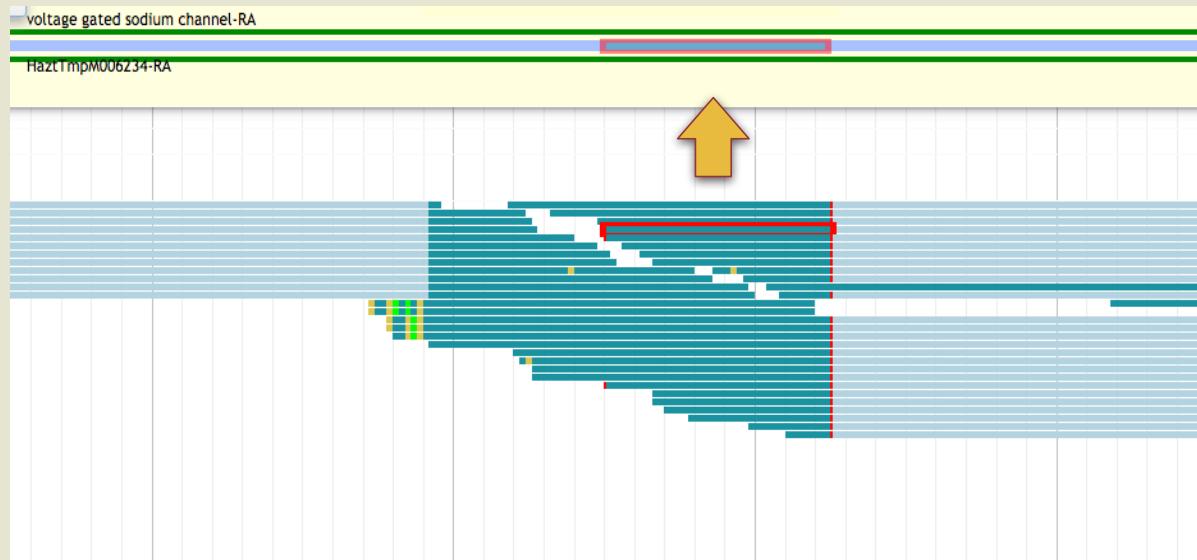


SIMPLE CASES

# Editing functionality

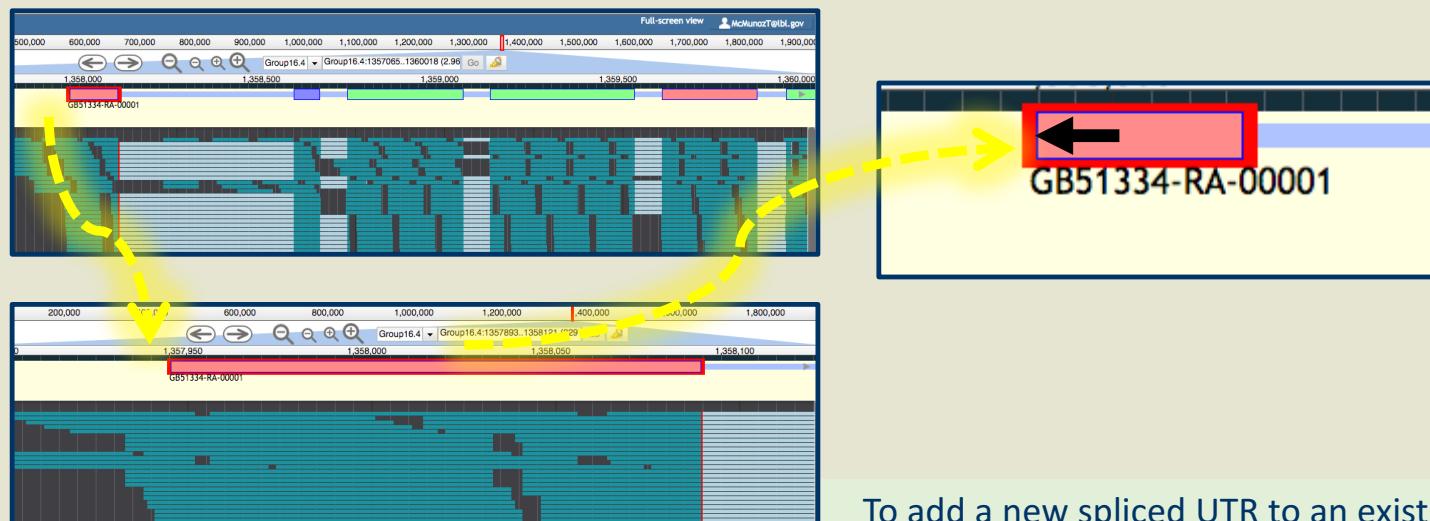
Example: Adding an exon supported by experimental data

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



## 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.



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.

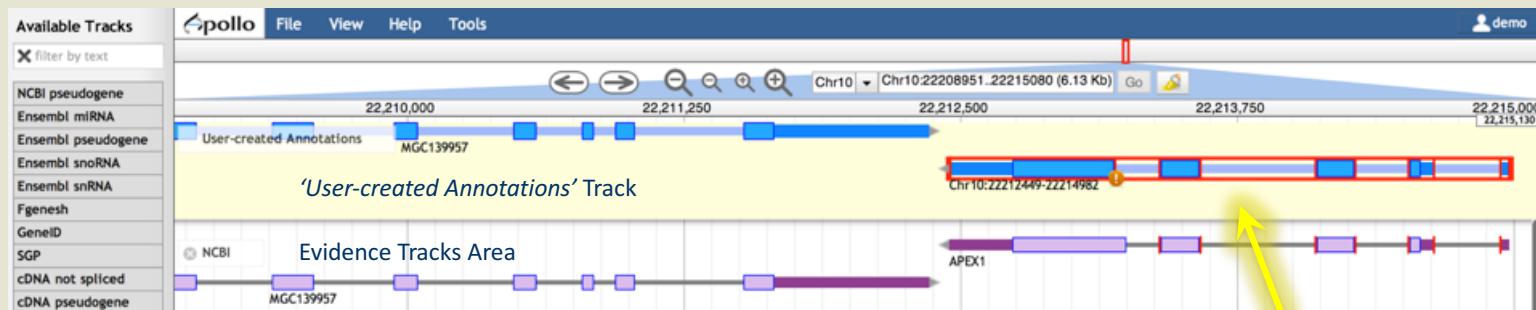


SIMPLE CASES

## ORFs - setting & recalculating

Apollo's editing logic (brain):

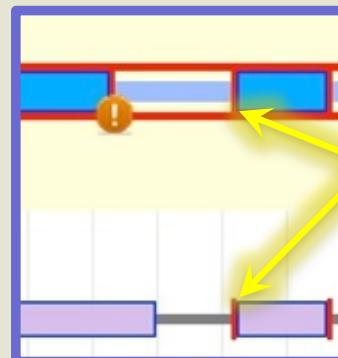
- selects **longest ORF** as CDS
- **recalculates ORF** after each edit, unless set



Double click selects the entire model

Red lines around exons:

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



Edge-matching

SIMPLE CASES



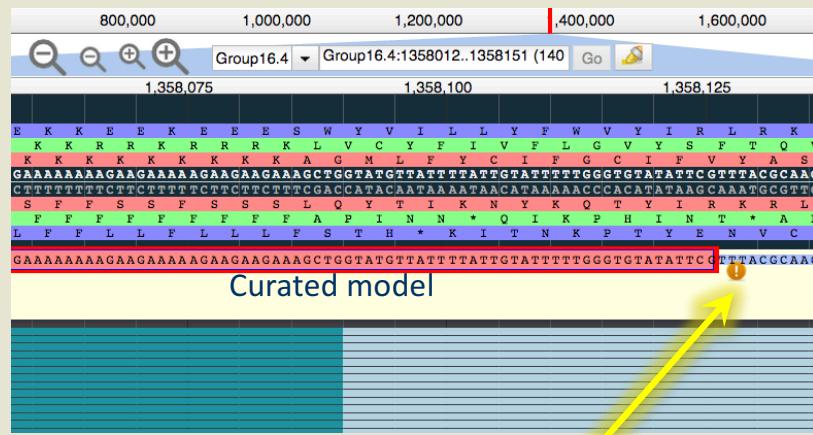
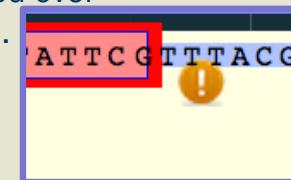
# SPLICE SITES

## 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.



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

The screenshot shows a bioinformatics tool interface for editing mRNA sequences. At the top, a sequence is displayed with coordinates 78,925, 78,950, 78,975, and 79,000. Below the sequence, a red box highlights a segment of the sequence: "TCGAAGAAGTCGAGGTACCTAGGTAGGACCCGTCGGTTTACATATTTGGTAGTGT". A yellow arrow points from this red box to a secondary sequence window.

The main window includes a context menu for the highlighted sequence:

- Get sequence
- Get gff3
- Zoom to base level
- Edit Information (alt-click)

A dropdown menu for the entry "-0.2-mRNA-1" contains the following options:

- Delete
- Merge
- Split
- Duplicate
- Make Intron
- Move to Opposite Strand
- Unset translation start
- Set translation end
- Set Longest ORF
- Set readthrough stop codon
- Set as 5' End
- Set as 3' End
- Set Both Ends
- Set to Downstream Splice Donor
- Set to Upstream Splice Donor
- Set to Downstream Splice Acceptor** (highlighted in blue)
- Set to Upstream Splice Acceptor
- Undo
- Redo
- Show History

A yellow box labeled "SIMPLE CASES" is located at the bottom left of the menu area.

A large grey arrow points from the bottom of the main window towards a smaller secondary window on the right.

The secondary window displays a zoomed-in view of the sequence between coordinates 78,975 and 79,000. It shows a blue bar above the sequence and a red bar below it, with a yellow arrow pointing upwards from the main window towards this secondary window.

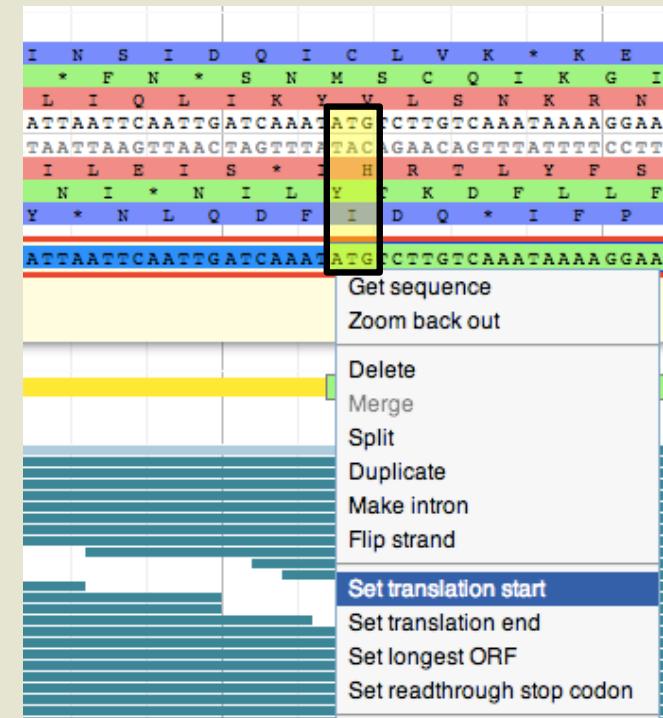


## 'Start' AND 'Stop' SITES

- 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

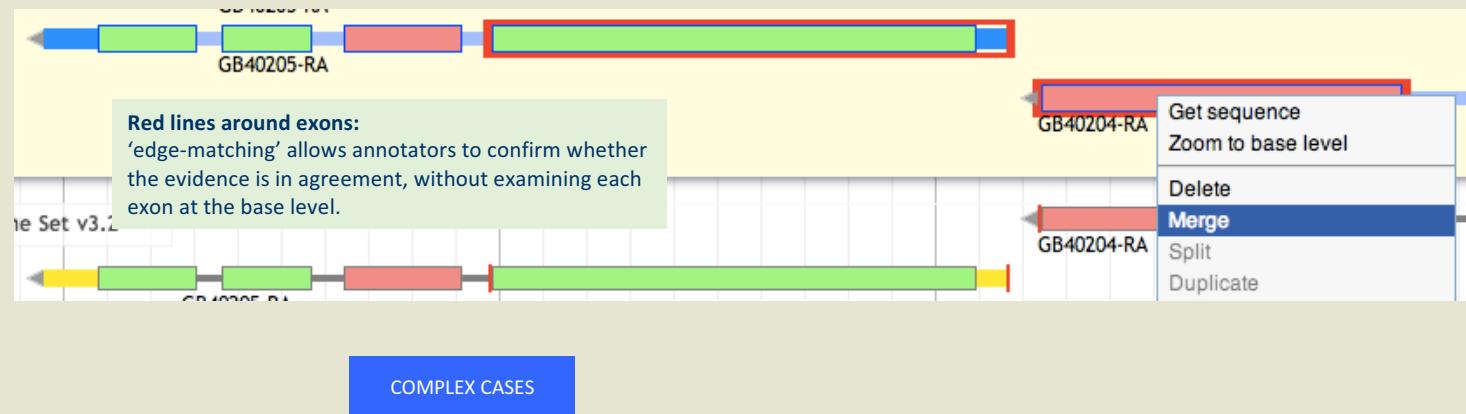
curating complex cases

## MERGE TWO GENE PREDICTIONS ON THE SAME SCAFFOLD

Evidence may support joining two or more different gene models.

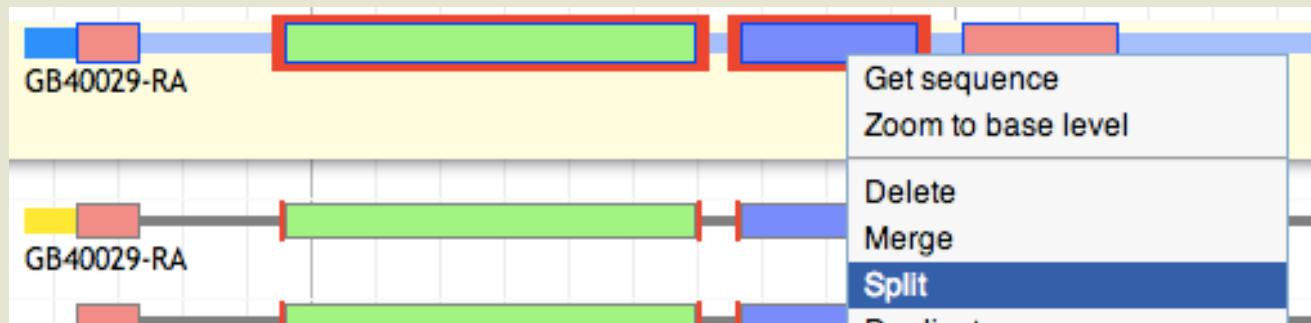
**Warning:** protein alignments may have incorrect splice sites and lack non-conserved regions!

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.



## 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.

The screenshot shows the Apollo genome annotation interface. At the top, there are search and navigation tools, and a header indicating the chromosome (Chr10) and genomic coordinates (Chr10:22213112..22213125). Below the header, a DNA sequence track is shown with amino acid translations above it. A green box highlights the "DNA Track". A blue box highlights the "User-created Annotations" track, which contains a red line and a small orange circle. A context menu is open over the "User-created Annotations" track, listing options: "Toggle Reverse Strand", "Toggle Protein Translation", "Create Genomic Insertion" (which is selected), "Create Genomic Deletion", and "Create Genomic Substitution". To the right of the main window, several floating tool boxes are visible: "Add Substitution" (with "+ strand" and "- strand" fields and an "Add" button), "Add Deletion" (with "Length" field and "Add" button), and "Add Insertion" (with "+ strand" and "- strand" fields and an "Add" button). The bottom left corner features the Apollo logo, and the bottom center has a "COMPLEX CASES" button.



# CORRECTING SELENOCYSTEINE CONTAINING PROTEINS

Honeybee ▾ File View Tools Help Full-screen view mcmunozt@lbl.gov

0 50,000 100,000 150,000 200,000 250,000 300,000 350,000 400,000 450,000 500,000 550,000 600,000 650,000

155,075 155,100 155,125 155,150

Reference sequence

S \* G K T S I R Q H K L Y D P Q R \* N \* L N E I T L H I F P F F S S S S L

AAGCTAGGGAAAAAACTTCTATCCGACAGCATAAGTTATATGACCCACAAAGGTAGAATTAACTGAACGAGATCACTTGCAATTTCCCTTTCTCTCTTCTCGT

User-created Annotations GB55331-RA-00001

Official Gene Set v3.2

Get Sequence  
Get GFF3  
Zoom to Base Level  
Edit Information (alt-click)  
Delete  
Merge  
Split  
Duplicate  
Make Intron  
Move to Opposite Strand  
Set Translation Start  
Set Translation End  
Set Longest ORF  
Set Readthrough Stop Codon

apollo

COMPLEX CASES

The screenshot shows a genomic sequence viewer for Honeybee. At the top, there's a navigation bar with 'Honeybee' dropdown, 'File', 'View', 'Tools', 'Help', 'Full-screen view', and a user email 'mcmunozt@lbl.gov'. Below the navigation is a zoomed-in sequence view with coordinates 155,075 to 155,150. The sequence is color-coded by amino acid: purple for Alanine (A), green for Serine (S), red for Threonine (T), blue for Glutamine (Q), yellow for Cysteine (C), pink for Histidine (H), black for Methionine (M), brown for Isoleucine (I), grey for Leucine (L), and orange for Valine (V). A yellow arrow points to a red 'T' codon at position 155,125, which is part of the sequence 'ACAAAGGTAGAATTAACTGAACGAGATC'. To the right of this codon is a context menu with options like 'Get Sequence', 'Edit Information (alt-click)', and 'Set Readthrough Stop Codon'. A blue arrow points to the 'Set Readthrough Stop Codon' option. On the left, there are sections for 'User-created Annotations' (containing 'GB55331-RA-00001') and 'Official Gene Set v3.2'. The bottom left features the 'apollo' logo and a Berkeley Lab logo. A blue button at the bottom center says 'COMPLEX CASES'.

## CORRECTING SELENOCYSTEINE CONTAINING PROTEINS

Honeybee    File    View    Tools    Help    Full-screen view    mcmunozt@lbl.gov

50,000 100,000 150,000 200,000 250,000 300,000 350,000 400,000 450,000 500,000 550,000 600,000 650,000

155,000 155,125 155,250 155,375 155,500

User-created Annotations    GB55331-RA-00001

Official Gene Set v3.2    GB55331-RA

**Sequence**

>77c0d1a1-84cd-4b05-8314-4d1ae3b792b1 (sequence:exon) 88 residues [Group1.32:154930-155491 + strand] [peptide]

TNEPTNDRVCLRSTVLSTIIGIGCGFLCLMAGTILAMCSRIRQAREKLLSDSISYMTHKGRINUTRSLC

IFFPPFSLLLRCVSGINV

COMPLEX CASES

The screenshot shows a genomic browser interface for the Honeybee genome. At the top, there's a navigation bar with 'Honeybee' (dropdown), 'File', 'View', 'Tools', 'Help', 'Full-screen view', and an email address 'mcmunozt@lbl.gov'. Below the navigation is a coordinate scale from 50,000 to 650,000. Two tracks are visible: 'User-created Annotations' (blue bar) and 'Official Gene Set v3.2' (yellow bar). A yellow arrow points from the sequence viewer below to a specific residue in the peptide sequence above. A blue arrow points from the sequence viewer to the same residue.



# 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

- Get Sequence**
- Get GFF3
- Zoom to Base Level**
- Edit Information (alt-click)**
- Delete
- Merge
- Split
- Duplicate
- Make Intron
- Move to Opposite Strand
- Set Translation Start
- Set Translation End
- Set Longest ORF
- Set Readthrough Stop Codon
- Set as 5' end
- Set as 3' End
- Set both Ends
- Set to Downstream Splice Donor
- Set to Upstream Splice Donor
- Set to Downstream Splice Acceptor
- Set to Upstream Splice Acceptor
- Undo
- Redo
- Show History



# Information Editor

The screenshot illustrates the Information Editor interface, showing two main panels: a left panel for editing gene information and a right panel for mRNA information.

**Left Panel (Gene Information):**

- Select mRNA:** spe1-RA  
spe1-RA
- gene** (highlighted)
- Name:** DNA mismatch repair protein Msh
- Symbol:** spel1
- Description:** DNA mismatch repair protein Msh
- Created:** 2017-03-03
- Last modified:** 2017-03-21
- DBXRefs:** DB Accession

A yellow box highlights the "gene" tab in the top navigation bar.

**Right Panel (mRNA Information):**

- mRNA** (highlighted)
- Name:** DNA mismatch repair protein Msh
- Symbol:**
- Description:**
- Created:** 2017-03-03
- Last modified:** 2017-03-21
- DBXRefs:** DB Accession
- NCBI Gene:** LOC725348
- BeeBase:** GB40028
- Enter new DB:** Enter new accession
- Attributes:** Tag Value

A yellow arrow points from the "DBXRefs" section of the mRNA panel to the "Enter new DB" field.

**Bottom Left Panel (Gene Ontology IDs):**

- Gene Ontology IDs:** GO:0006301, GO:0006298
- 40029-RA** (highlighted)
- Details:**
  - mismatch repair [GO:0006298]
  - nuclear-transcribed mRNA catabolic process, no-go decay [GO:0070966]
  - dopamine neurotransmitter receptor activity, coupled via Gi/Go [GO:0001591]
  - GINS complex [GO:0000811]

A yellow arrow points from the "Gene Ontology IDs" section of the mRNA panel to the "Gene Ontology IDs" section of the gene panel.

**Bottom Right Panel (Comments):**

- Comments:** Extended 3' UTR using Forager RNAseq reads as
- Comments:**
- Comments:**
- Comments:**

A yellow arrow points from the "Comments" section of the mRNA panel to the "Comments" section of the gene panel.

**Logo:** BERKELEY LAB

# Information Editor

File View Tools Help Full-screen view mcm

Select mRNA Apurinic-Apyrimidinic Endonuclease-00002

gene

Name	Apurinic-Apyrimidinic Endonuclea
Symbol	Apex-1
Description	Multifunctional DNA Repair Enzym
Created	2015-07-26
Last modified	2015-07-26

Status

Approved  Needs Review  
 Delete

DBXRefs

DB	Accession
----	-----------

Add Delete

Replaced Models

Action	Transcript Name
replace	Enter new value

pollo

mRNA

Name	Apurinic-Apyrimidinic Endonuclea
Symbol	Apex-1
Description	Multifunctional DNA Repair Enzym
Created	2015-07-26
Last modified	2015-07-26

Status

Approved  Needs Review  
 Delete

DBXRefs

DB	Accession
WormBase	WB_0001234
FlyBase	FB_00004567

Add Delete

Replaced Models

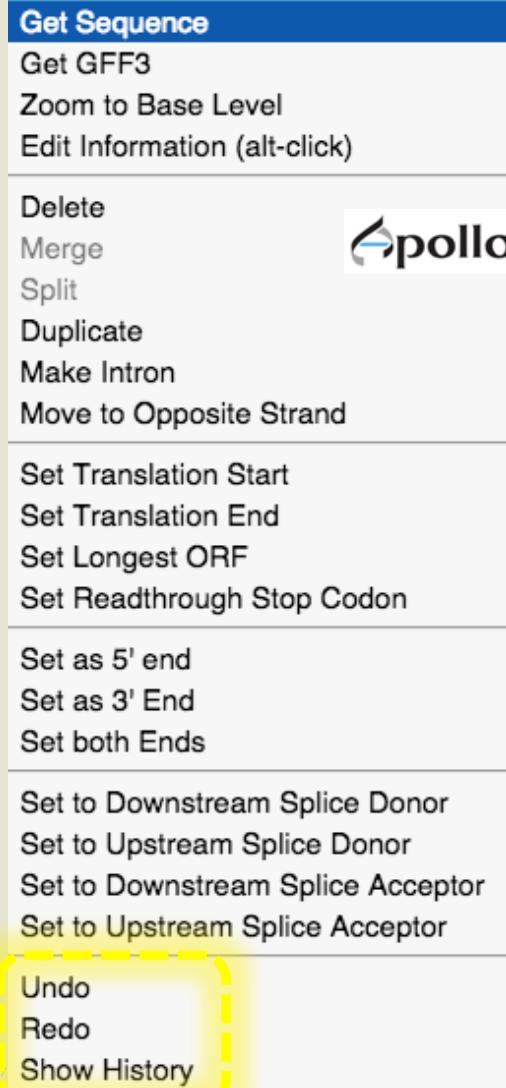
Action	Transcript Name
replace	Enter new value

Add Delete



history

# Keeping track of each edit



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

History		
Operation	Editor	Date
ADD_TRANSCRIPT	mmtorres	5/13/14 10:44 AM
SET_TRANSLATION_START	mmtorres	5/13/14 10:49 AM
DELETE_EXON	mmtorres	5/13/14 10:49 AM
<b>MERGE_EXONS</b>	<b>mmtorres</b>	<b>5/13/14 10:50 AM</b>
SET_READONLY_STOP_CODON	mmtorres	5/13/14 10:51 AM
UNSET_READONLY_STOP_CODON	mmtorres	5/13/14 10:52 AM
<b>SET_READONLY_STOP_CODON</b>	<b>mmtorres</b>	<b>5/13/14 10:55 AM</b>



History		
Operation	Editor	Date
ADD_TRANSCRIPT	mmtorres	5/13/14 10:44 AM
SET_TRANSLATION_START	mmtorres	5/13/14 10:49 AM
DELETE_EXON	mmtorres	5/13/14 10:49 AM
MERGE_EXONS	mmtorres	5/13/14 10:50 AM
SET_READONLY_STOP_CODON	mmtorres	5/13/14 10:51 AM
UNSET_READONLY_STOP_CODON	mmtorres	5/13/14 10:52 AM
<b>SET_READONLY_STOP_CODON</b>	<b>mmtorres</b>	<b>5/13/14 10:55 AM</b>



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, GenID, 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



[GenomeArchitect.org](http://GenomeArchitect.org)

# *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



[GenomeArchitect.org](http://GenomeArchitect.org)

# 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

```
GTTTAAGAGTGTTCGCGCCAATTGTTCGCGCGAGACTGGCGTGCAAGACCGAGCTGTTATAGCCGCGTCT  
CCGCTCTGCTCTGCTGATCCCATCGATCACCTACGCATCGATCCCTCGTTGATCAACGTGGTCAATGAGC  
TGGAGCGTTGAGCGCCGCTATCAGACTGGCGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG  
CTTTAGAATCCTTGCCTGTTGACGATATGGCTGGTCCAGCTTGCCTGGCGCCATCGCTTAC  
AGCATCGGGTGGCAGAGCAGATGCTACAGGACCCGCCGTGAAATTGTTTATGGGCTACCGAAGA  
TCGATCAAAAGGATCAGGAATCCATCTCGAACATTCTCCCGCGATTATCATCGACGATGGCGAGGA  
GAGGTTCGTCTCGTCAGCGTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGGCAGAAT  
CTTGAAGAGGAGTTGGCAGCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA  
CACCCGGTGGATTCATGTTGCACATGTTGATATTACGACATTGGTTCTGTTCAAGAGACCTTCGA  
TGCTATGGTCAAGGGATCAGAAGAGTATTCAACGTGCTACTATGCCATAGTTCCAGGCAGAAATATTC  
ATCACCCATGGAGAAGTTCATGGTGTGAACATTAATAGAAGCCCATCCG
```

Search all genomic  
sequences



The screenshot shows the Honeybee genome browser interface. A yellow arrow points from the "Search sequence" input field in the main window down to the "Search sequence" dialog box.

**Honeybee** ▾ File View Tools Help

0 200,000 Search sequence 200,000

462,500 465,000

User-created Annotations

**Search sequence**

Blat nucleotide ▾

Enter sequence

```
GTTTAAGAGTGTTCGCGCCAATTGTTCGCGCGAGACTGGCGTGCAAGACCGAGCTGTTATAGCCGCGTCT  
CCGCTCTGCTCTGCTGATCCCATCGATCACCTACGCATCGATCCCTCGTTGATCAACGTGGTCAATGAGC  
TGGAGCGTTGAGCGCCGCTATCAGACTGGCGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG  
CTTTAGAATCCTTGCCTGTTGACGATATGGCTGGTCCAGCTTGCCTGGCGCCATCGCTTAC  
AGCATCGGGTGGCAGAGCAGATGCTACAGGACCCGCCGCTGAAATTGTTTATGGGCTACCGAAGA  
TCGATCAAAAGGATCAGGAATCCATCTCGAACATTCTCCCGCGATTATCATCGACGATGGCGAGGA  
GAGGTTCGTCTCGTCAGCGTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGGCAGAAT  
CTTGAAGAGGAGTTGGCAGCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA  
CACCCGGTGGATTCATGTTGCACATGTTGATATTACGACATTGGTTCTGTTCAAGAGACCTTCGA  
TGCTATGGTCAAGGGATCAGAAGAGTATTCAACGTGCTCACTATGCCATAGTTCCAGGCAGAAATATTC  
ATCACCCATGGAGAAGTTCATGGTGTGAACATTAATAGAAGCCCATCCG
```

Search all genomic sequences

Search

# Blat results

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

ID	Start	End	Score	Significance	Identity
Group8.6	1864564	1864709	228	6.6e-60	89.04
Group8.6	1863812	1863918	169	4.8e-42	87.85
Group8.6	1865189	1865302	154	9.8e-38	82.46
GroupUn14..57	103	75	75	7.3e-14	88.24
Group8.6	1871618	1871664	75	7.3e-14	88.24
GroupUn51..1281	1325	71	71	1.3e-12	87.76
Group8.6	1865314	1865354	71	1.1e-12	92.68
Group8.6	1863560	1863582	42	0.00057	95.65
Group1.43	1236401	1236419	37	0.018	100
Group1.17	362426	362443	36	0.05	100
Group1.41	1174204	1174223	36	0.036	95
GroupUn37..494	511	36	36	0.057	100
Group6.38	485127	485144	35	0.065	100



# BIPAA resources - blast

The screenshot shows the BIPAA homepage. At the top, there's a banner with the text "BIPAA Bioinformatics Platform for Agroecosystem Arthropods". Below the banner, there's a navigation bar with several links: "Daktulosphaira vitifoliae", "GO Report", "Blast" (which is highlighted with a blue box), "JBrowse", "Apollo", "Download", and "BIPAA".

This screenshot shows a detailed view of a BIPAA resource page for a polypeptide named DV3012683-PA. The page has a sidebar on the left with links like Overview, Alignments, Analyses, Annotated Terms, GO Annotation, Homology, InterPro, Relationships, and Sequences. The main content area is titled "Overview" and contains a table with the following data:

NAME	DV3012683-RA
UNIQUE NAME	DV3012683-PA
TYPE	polypeptide
ORGANISM	Daktulosphaira vitifoliae (vitiifoliae)
SEQUENCE LENGTH	1700

A yellow arrow points from the "Blast" link on the BIPAA homepage to the "SEQUENCE LENGTH" row in this table.

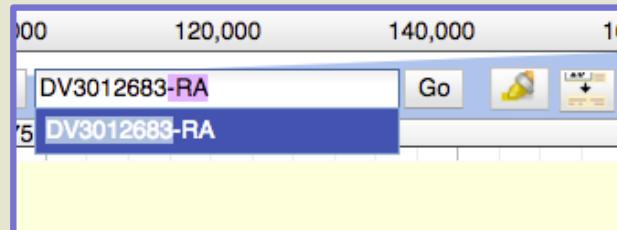
The screenshot shows the results of a BLAST search. At the top, it says "Your blast job RNApolII is finished!". Below that, there's a "Results" section with a "BLASTP 2.6.0+" header and a "Command ASN.1 XML TSV CSV Text GFF3 HTML" menu. The results show a reference section with citations for the BLAST algorithm and its improvements. The database used is "annotation\_v3.0\_ogs3.0\_20161223\_proteins" with 24,585 sequences and 8,417,147 total letters. The query sequence is "Apis\_dorsata RNA pol II subunit RPB2-like partial" with a length of 182. The results table lists three significant alignments:

Score (Bits)	E Value
308	7e-98
77.4	5e-17
35.4	0.011

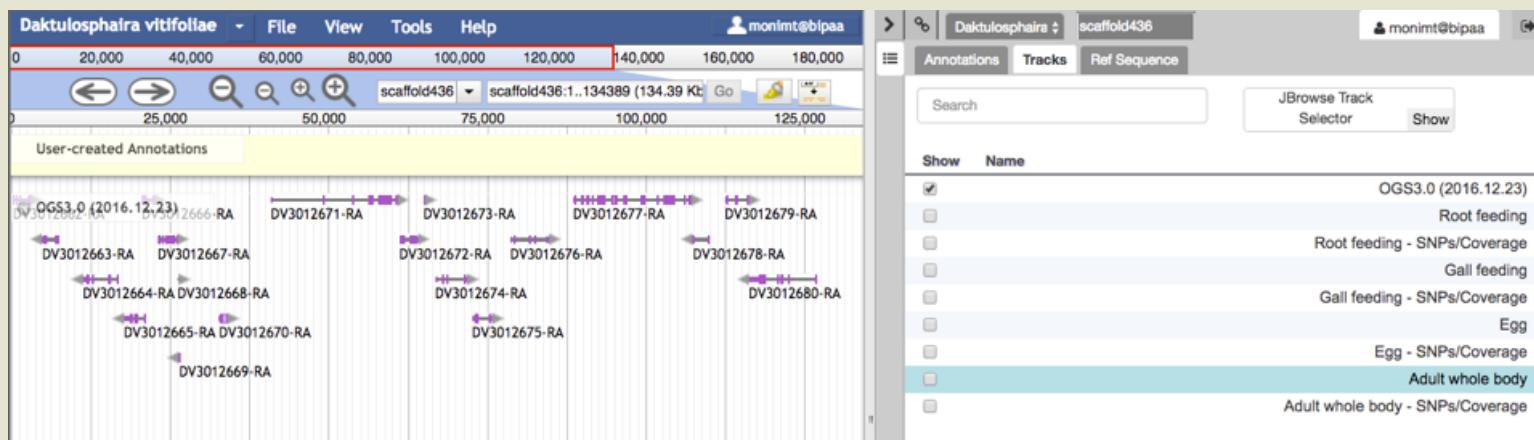
Below the table, the sequence alignment for the top hit is shown, starting with >DV3012683-PA gene=CV3012683. The alignment shows the query sequence followed by the subject sequence with gaps indicated by dashes and identities by matching letters.



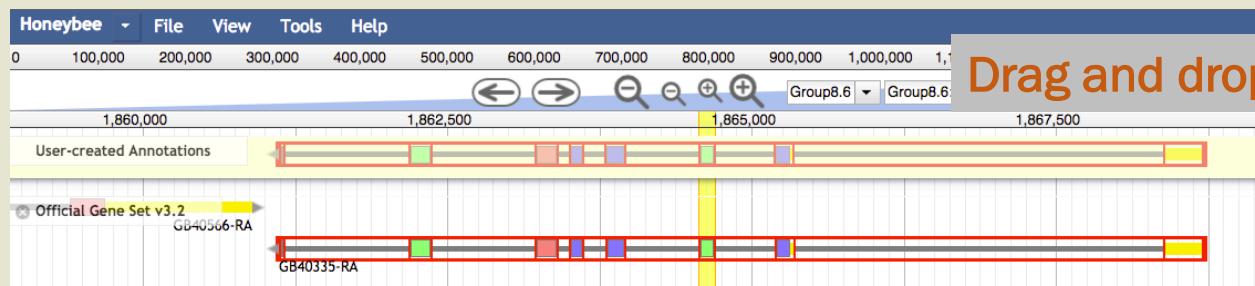
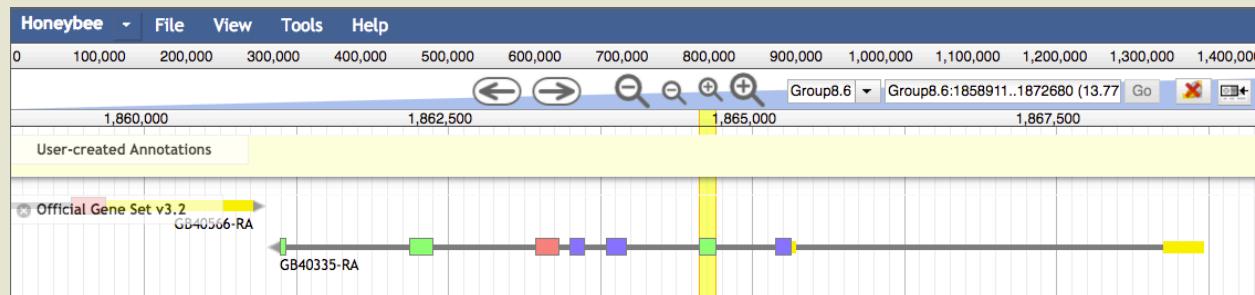
# BIPAA resources - Apollo



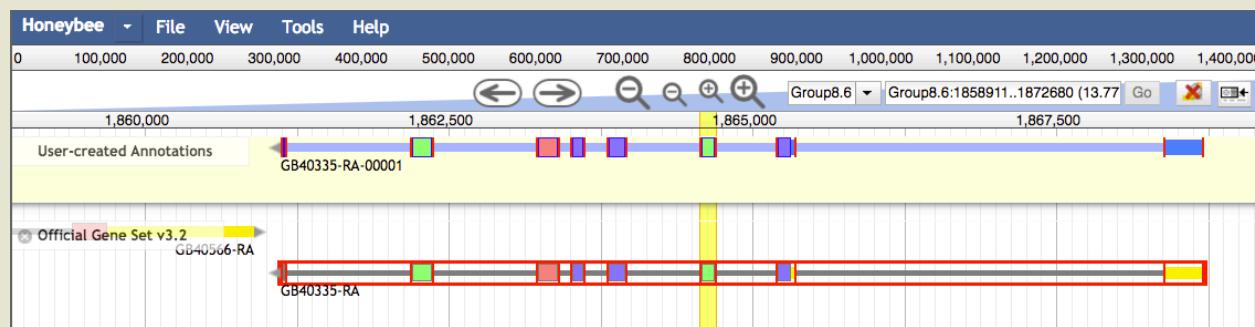
You may find candidate genes from blast results using the 'Search' box with coordinates in main window.



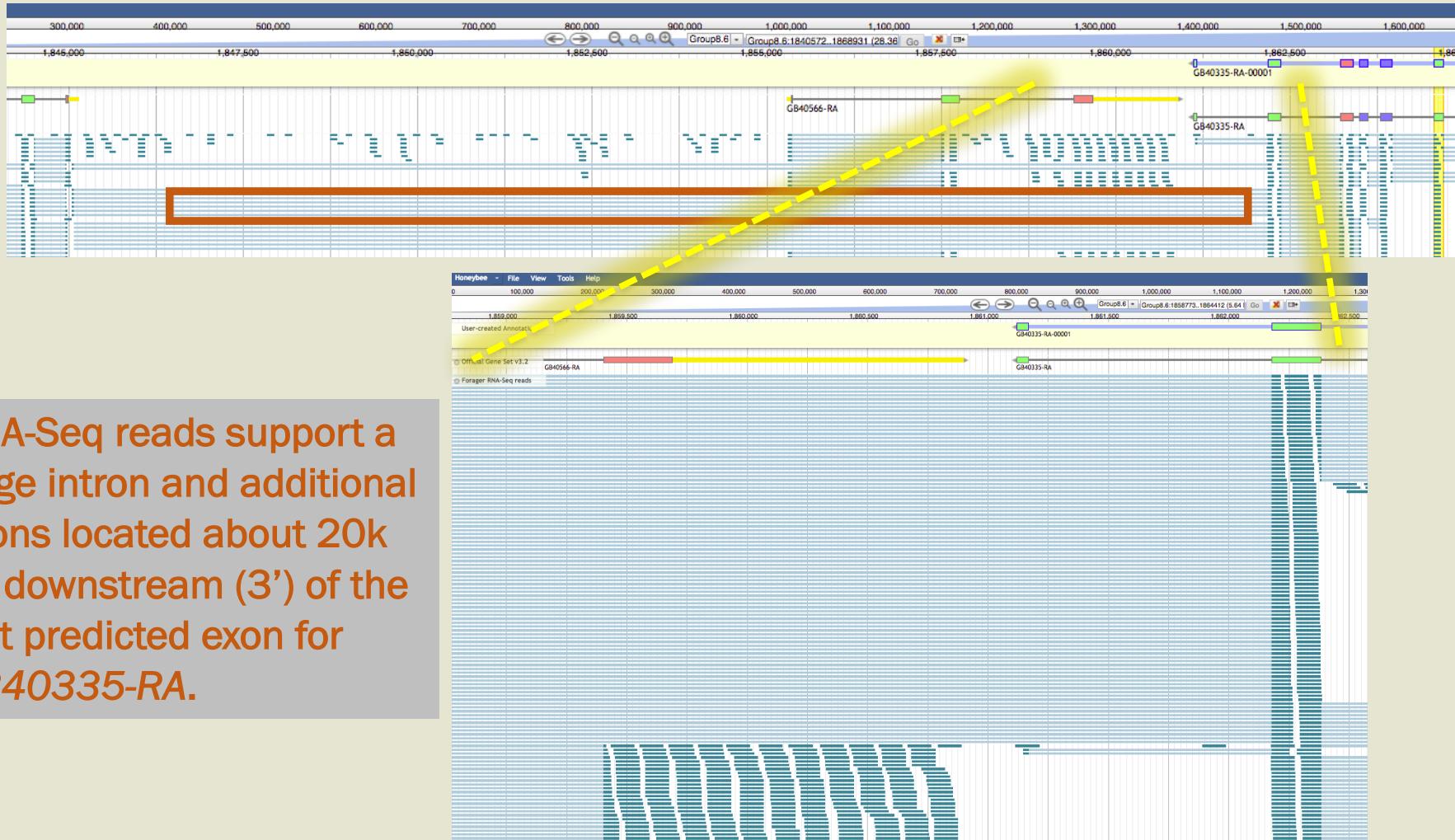
# Create a new annotation



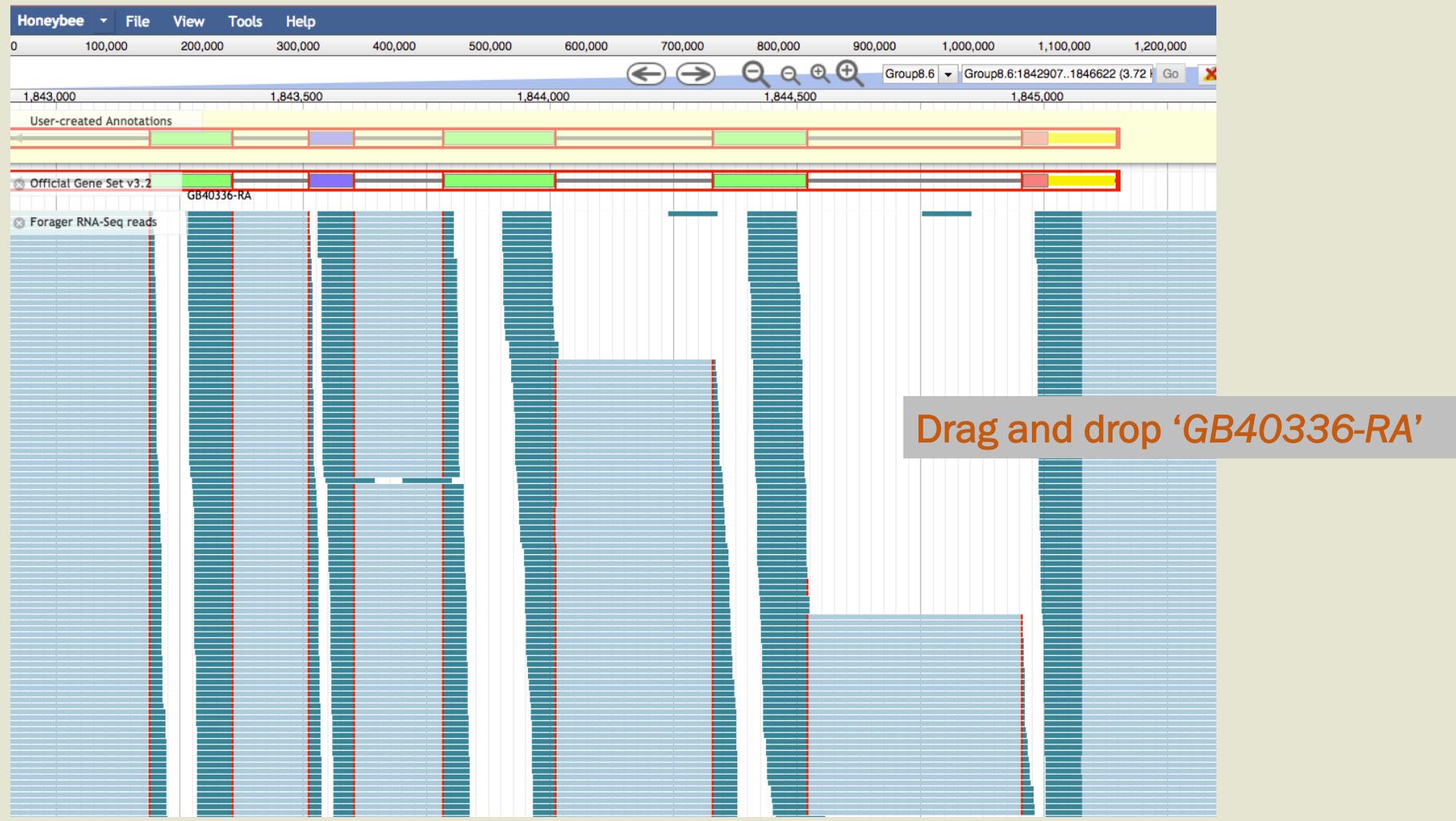
Drag and drop 'GB40335-RA'



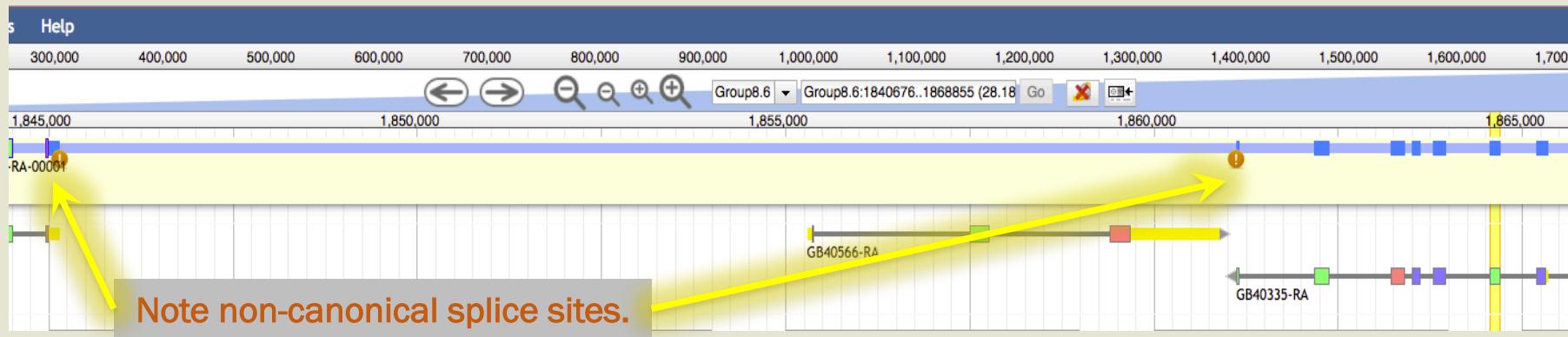
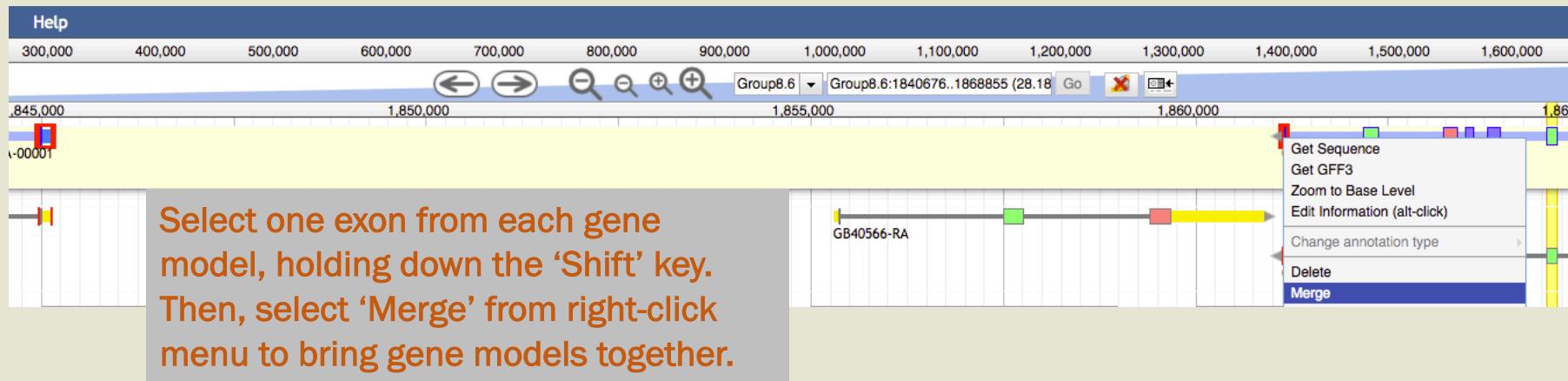
# Transcriptomic data support a longer gene



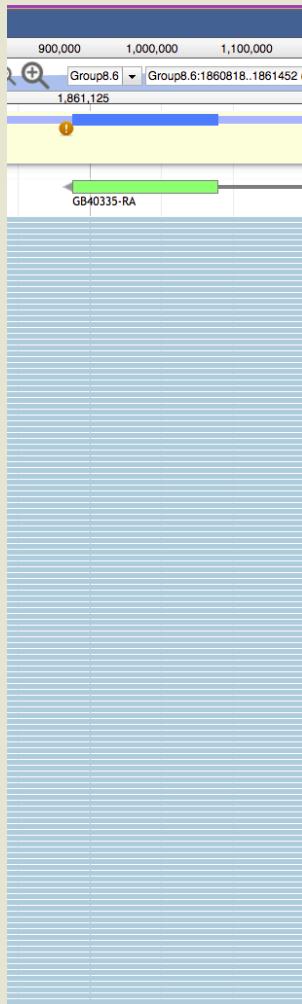
# Transcriptomic data support a longer gene



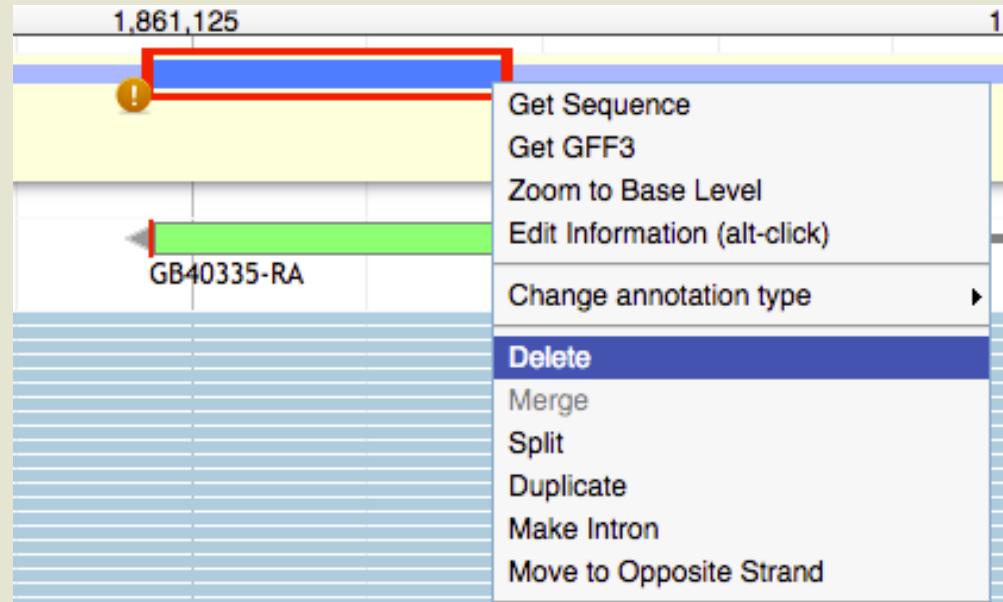
# Merge transcripts



# Exon not supported by RNA-Seq data

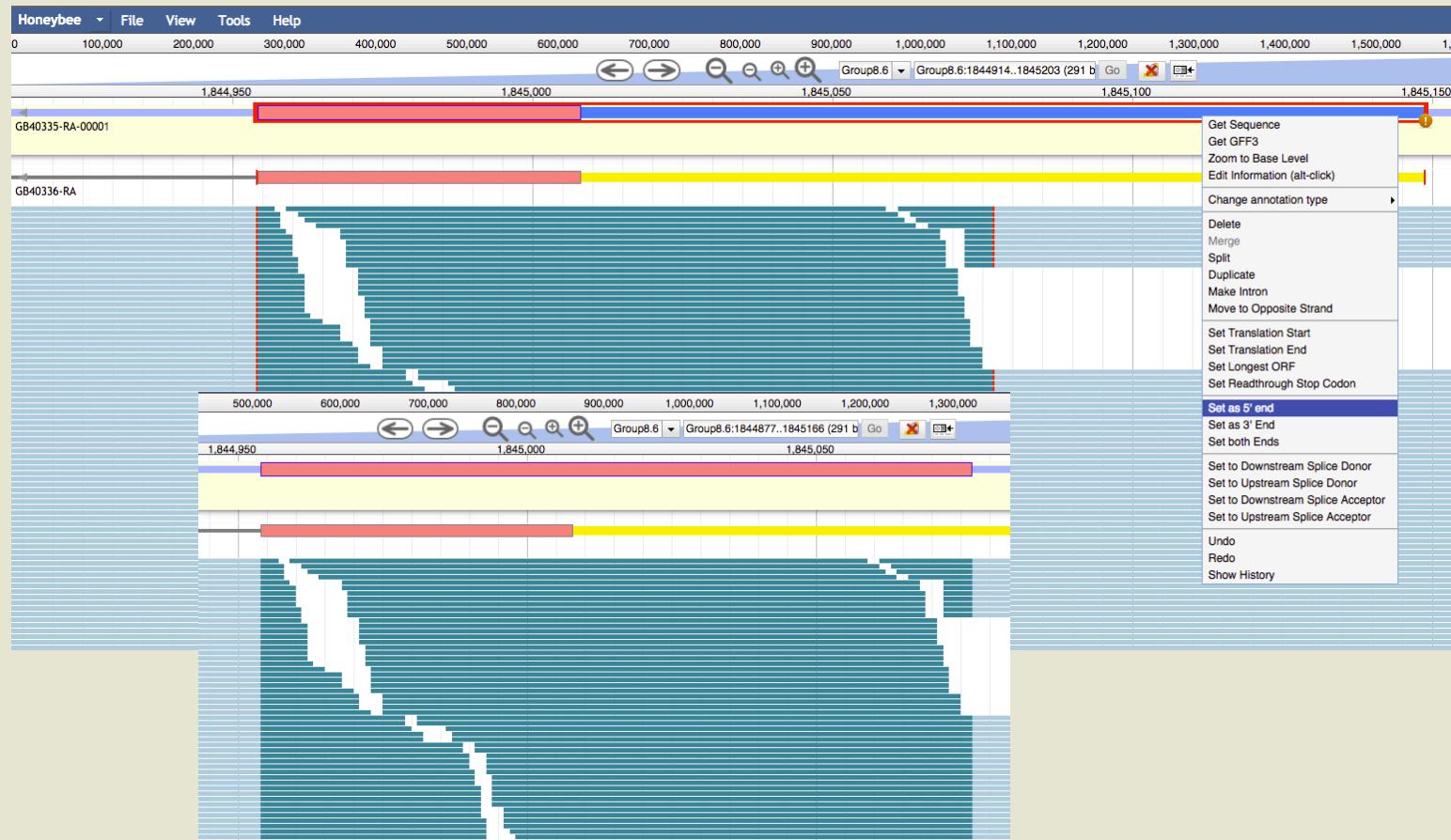


At the end of *GB40335-RA*, select last exon and right-click to choose the 'Delete' option.

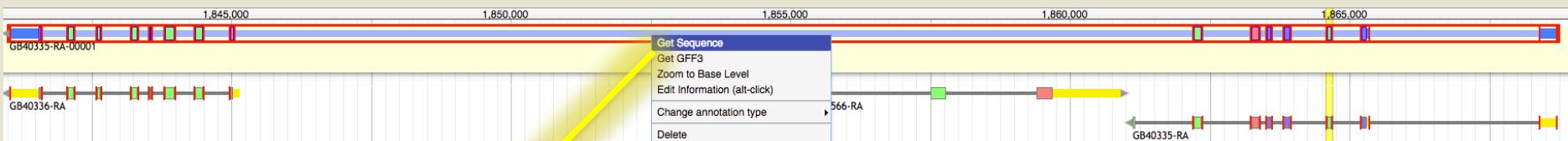


# Fix remaining non-canonical splice site

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



**Sequence**

```
>fccb9943-c0dc-4bbc-b43b-74f6dfe33dfe (sequence:mRNA) 715 residues [Group8.6:1841036-1868721 - strand] [peptide]
MEAPALLRTLALLTILQAVAVPGAVASYTIGVGRADTTGPVAEIVFMGYAKIDQKGSGHLRLRTSRAFI
IDDGVERFVFVSVDSAMIGNGIRQTVVENLQKQYGDLYTEKVMISATHSHSTPGGMFLHMLFDLTTFGF
VRETDAMVNQTSIERAHNAMPGRLFITHGEVHGVIINRSPFAYLNNPKVERDKYRDNDKILITQI
FYKNEDNPKPLGVINWPAIHPTSMNTNHLVSSDNIGYASVLFERIMMNDSLIGKGPVAAFASSNLGDVS
PNTRGPKEFSGNCNSKQYTCGPRKEMCFASGPGRMFESTSIIANRMFKESWRWLWQYGDVKVEIGPLRV
VHRYVNVMEQTAEYYNETTQRTEETVRGCEPAMGYSFAACTIDCPGSFSFRQGTTSANPMWNVVRNLLATP
TNEDIKCHGAKPILLATGHMTPYEWQPKIVATQVALIGNVVIAGVPGFEFTTMSGRRLEAIKTVMDAS
DDETSVIVAGLCNTYSDYVITPEEYQIQRYECASTIFGPHLTILYKLQYQELVTAAILKKDVEPGPEPVD
LRKKTLSVTFVTVPVLYDTPIWGKNGDCIKQPQKLAKPQDIVTAVFVSGHPRNNLMTESSLFTIERLGVD
VWLPVATDANWETKFWQRMSMVLGSSQVTTWQVPEDIKAGEYRIRHNGYYRYILGGIFPYVGVSNHFQ
VYSTESSCCCKRRIYYE
```

Peptide sequence  
 cDNA sequence  
 CDS sequence  
 Genomic sequence  
 Genomic sequence +/- 500 bases

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

**BLAST® » blastp suite**

Standard Protein BLAST

Enter Query Sequence  
 Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange  
 From [?](#)  To [?](#)

Or, upload file  No file chosen [?](#)

Job Title  ceramidase-Amel [?](#)

Align two or more sequences

Choose Search Set

Database  Non-redundant protein sequences (nr) [?](#)

Organism  Apis mellifera (taxid:7460)  Exclude [?](#)

Exclude  Models (XM/XP)  Uncultured/environmental sample sequences

Entrez Query  [Create custom database](#)

Program Selection

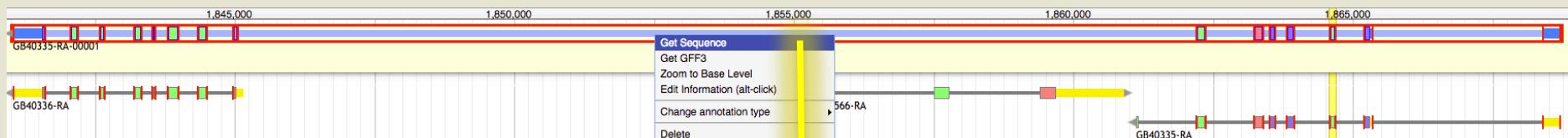
Algorithm  blastp (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

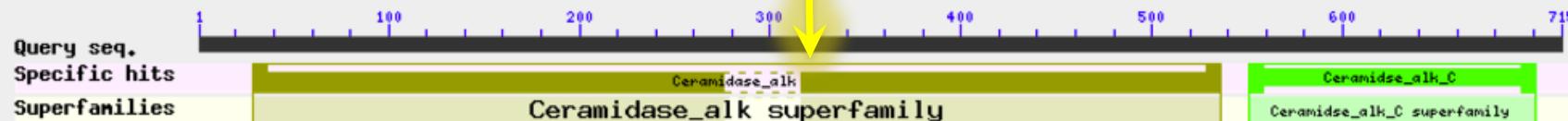
**BLAST** [Search database Non-redundant protein sequences \(nr\) using Blastp \(protein-protein BLAST\)](#)  Show results in a new window



# Results from NCBI blastp vs nr



Putative conserved domains have been detected, click on the image below for detailed results.



Sequences producing significant alignments:

Select: All None Selected:0

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



# Add metadata in ‘Information Editor’

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq., Specific hits Superfamilies

Ceramidase\_alk Ceramidase\_alk superfamily Ceramidase\_alk\_C Ceramidase\_alk\_C superfamily

**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
[+] Ceramidase_alk_C	pfam17048	Neutral/alkaline non-lysosomal ceramidase, C-terminal; This family represents C-terminal ...	551-701	4.77e-76

**Information Editor**

Select mRNA: GB40335-RA-00001

**gene**

Name	neutral ceramidase
Symbol	CDase
Description	Enzyme, cleaves fatty acids from ...
Created	2017-03-22
Last modified	2017-03-22

**mRNA**

Name	neutral ceramidase-00001
Symbol	
Description	
Created	2017-03-22
Last modified	2017-03-22

**DBXRefs**

DB	Accession
pfam	pfam17048
NCBI Gene	LOC409628
BeeBase	GB40336

**Don't forget!**

**Nice to have**



# Add metadata in ‘Information Editor’

The screenshot illustrates the 'Information Editor' interface for adding metadata to a mRNA entry. The main window shows a tree view of metadata categories: gene, mRNA, DBXrefs, Attributes, PubMed IDs, Gene Ontology IDs, and Comments.

**Comments:** A yellow arrow points from the 'Comments' section in the main window to a detailed view of the merged comments for GB40335-RA and GB40336-RA, which mention merging and supporting evidence from Forager RNA-seq reads.

**Gene Ontology terms:** Two boxes show expanded lists of GO IDs and their associated terms. The top box (GO:0017040) includes 'ceramidase activity [GO:0017040]' and 'nuclear-transcribed mRNA catabolic process, dopamine neurotransmitter receptor activity, GINS complex [GO:0000811]'. The bottom box (GO:0046514) includes 'ceramide catabolic process [GO:0046514]', 'nuclear-transcribed mRNA catabolic process, dopamine neurotransmitter receptor activity, GINS complex [GO:0000811]', and a 'Comments' section.

**PubMed Identifiers:** A yellow arrow points from the 'PubMed IDs' section in the main window to a detailed view of the ID 17073008, which is linked to a publication titled 'Insights into social insects from the genome of the honeybee *Apis mellifera*'.

**Publication:** A modal dialog box titled 'icebox.lbl.gov says:' displays the publication information: 'Publication title: 'Insights into social insects from the genome of the honeybee *Apis mellifera*.'

**BERKELEY LAB** logo is visible in the bottom left corner.

# Public demo instances

# **APOLLO ON THE WEB**

## **instructions**

---

- Public Honey bee demo available at:

[genomearchitect.org/demo/](http://genomearchitect.org/demo/)

- Username:

demo@demo.com

- Password:

demo



# Apollo demonstration

---

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



# 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)

[berkeleybop.org](http://berkeleybop.org)



## 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

Berkeley  
UNIVERSITY OF CALIFORNIA

