



BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY



Apollo

Collaborative genome annotation editing

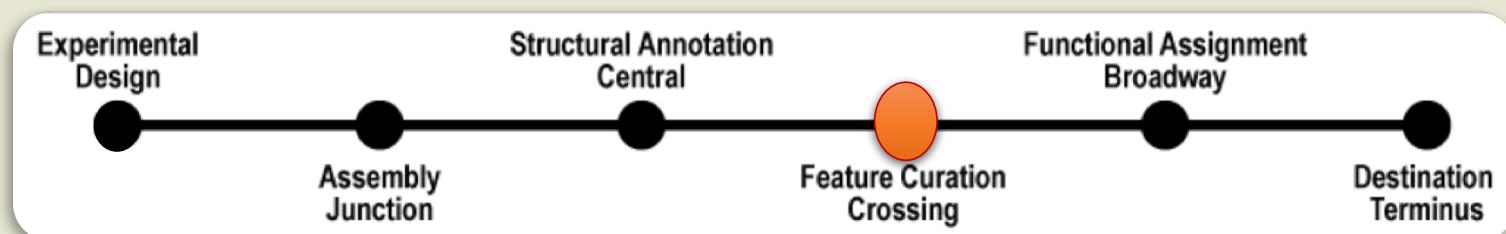
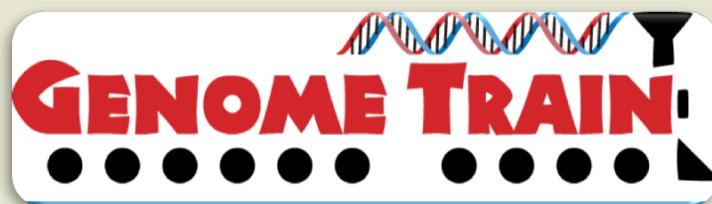
A workshop for the EMBL-ABR Community

Monica Munoz-Torres, PhD | @monimunozto

Phoenix Bioinformatics

A workshop for EMBL-ABR. 02 November, 2017

<http://GenomeArchitect.org>



Reference



editing functionality



Reference



begin with a new gene model



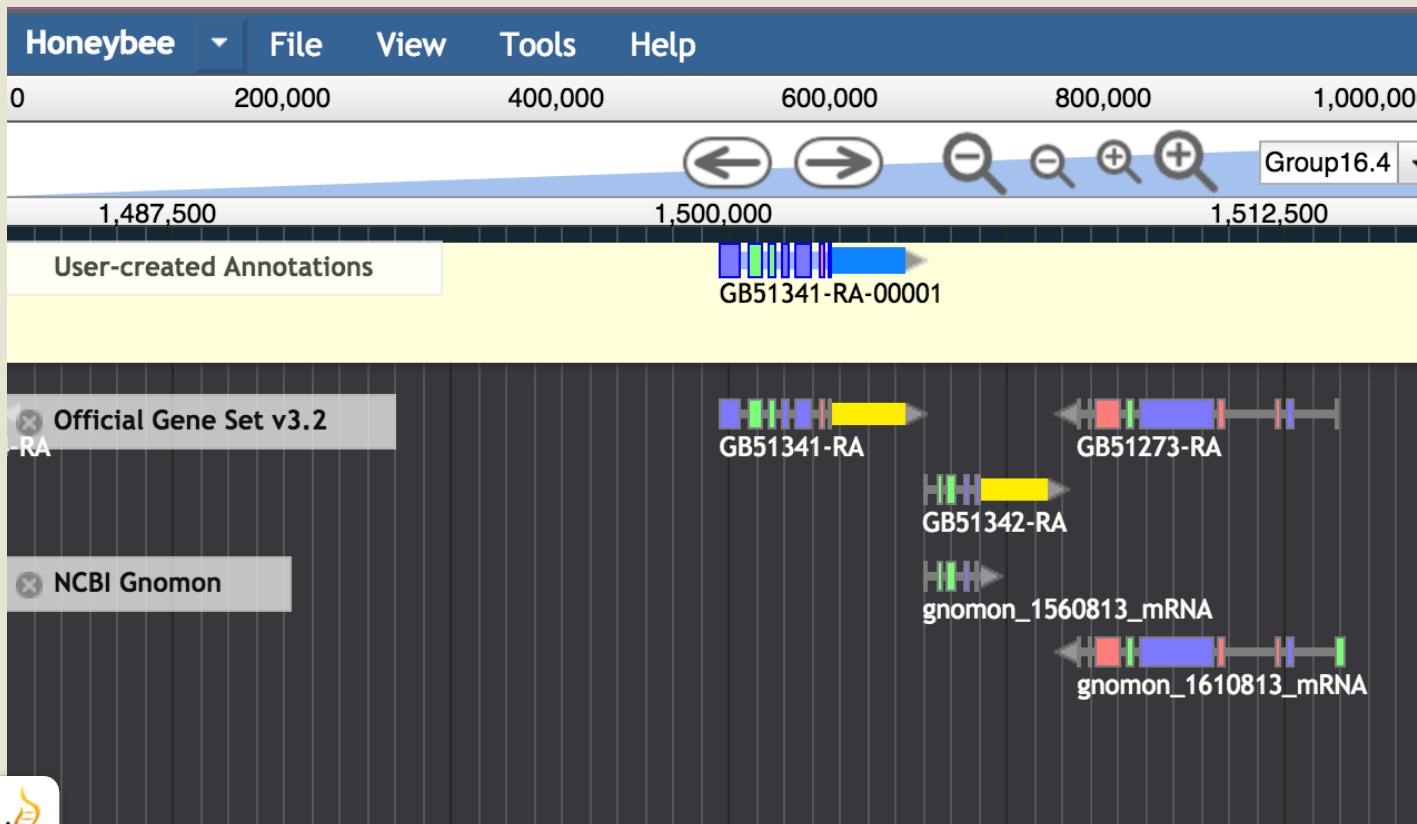
Creating a new annotation



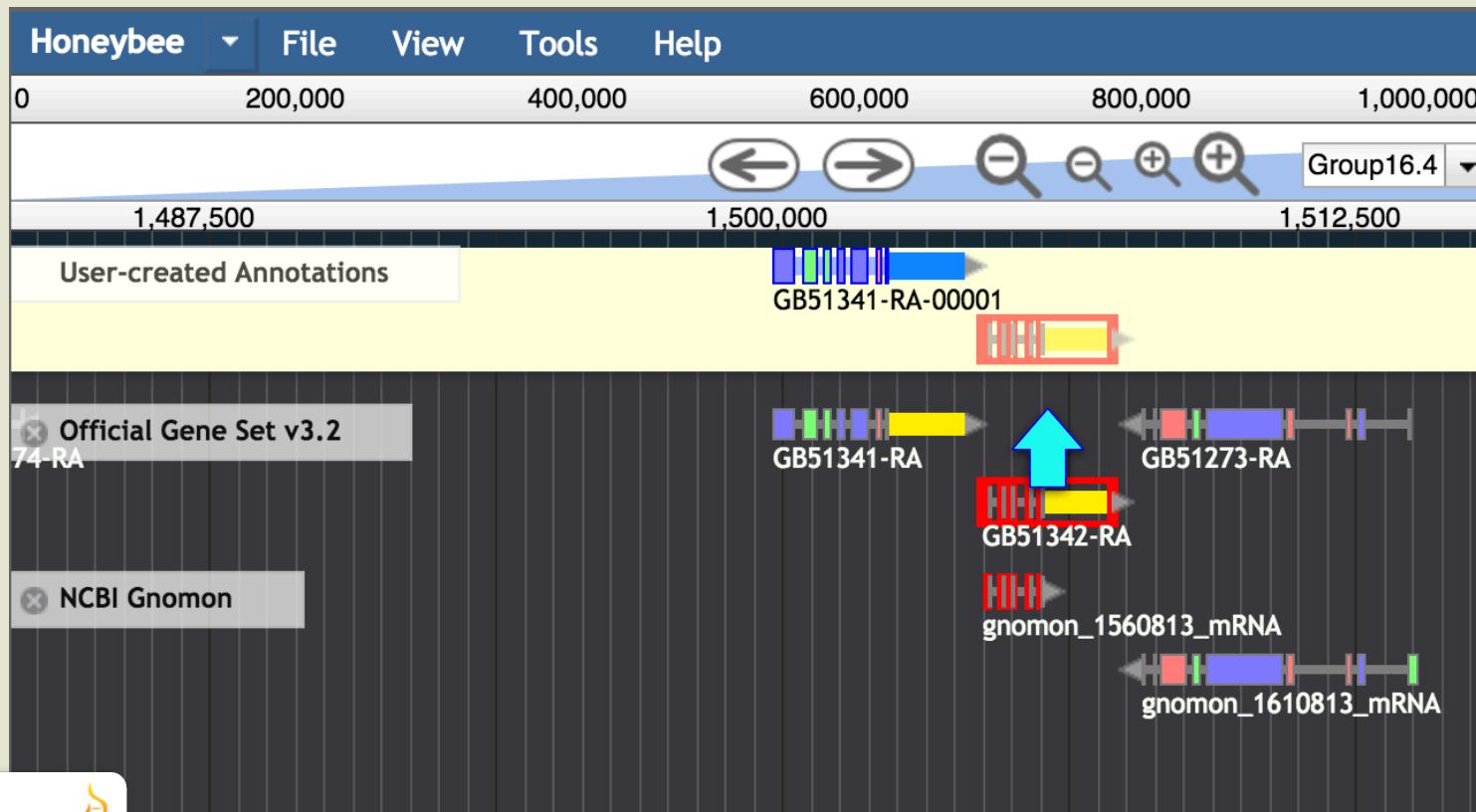
- Choose appropriate evidence from list of “Tracks” on **annotator panel**.
- Select & drag elements from evidence track into the ‘User-created Annotations’ area.
- Hovering over annotation in progress brings up an information pop-up.



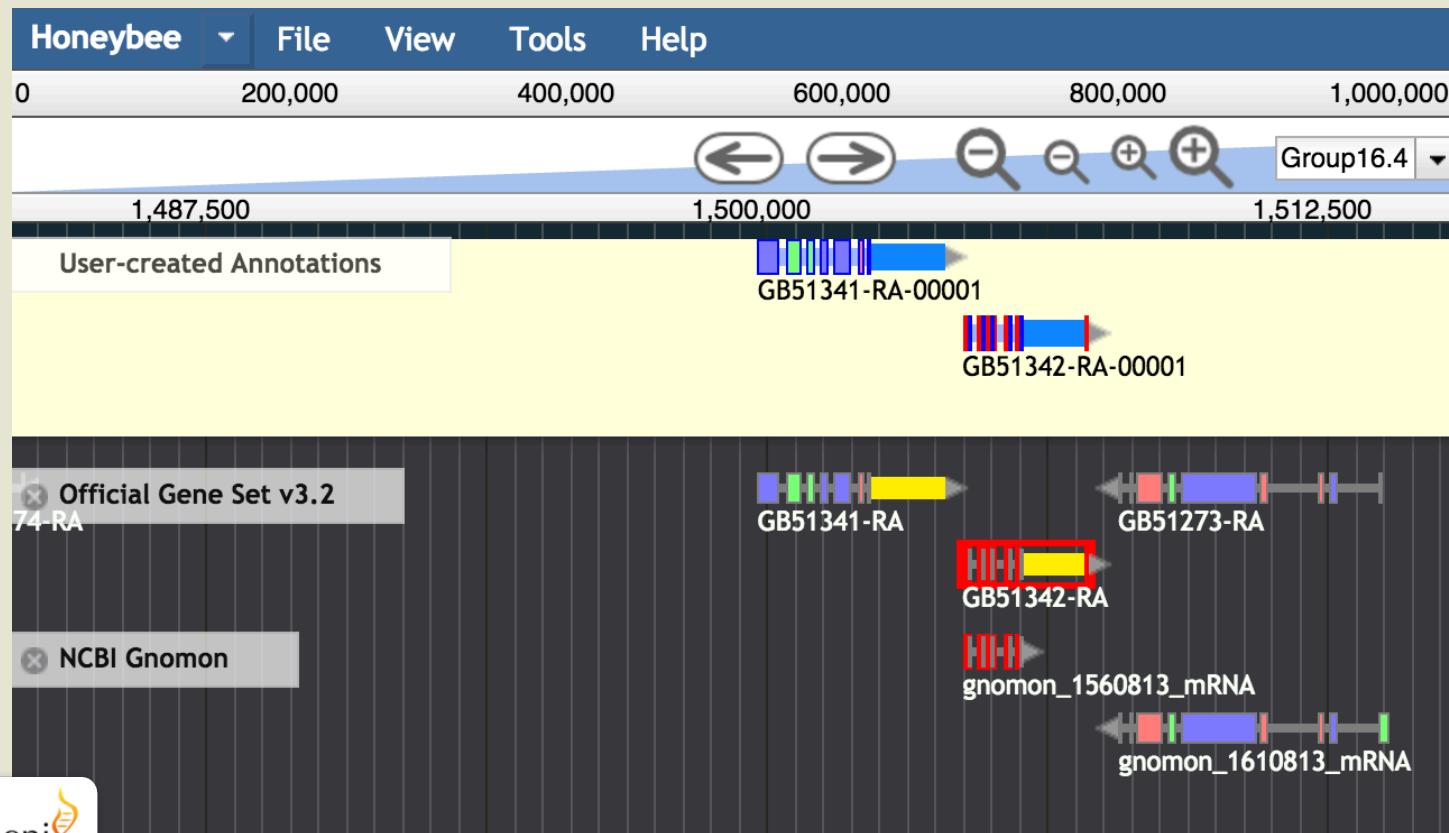
Adding a gene model



Adding a gene model



Adding a gene model

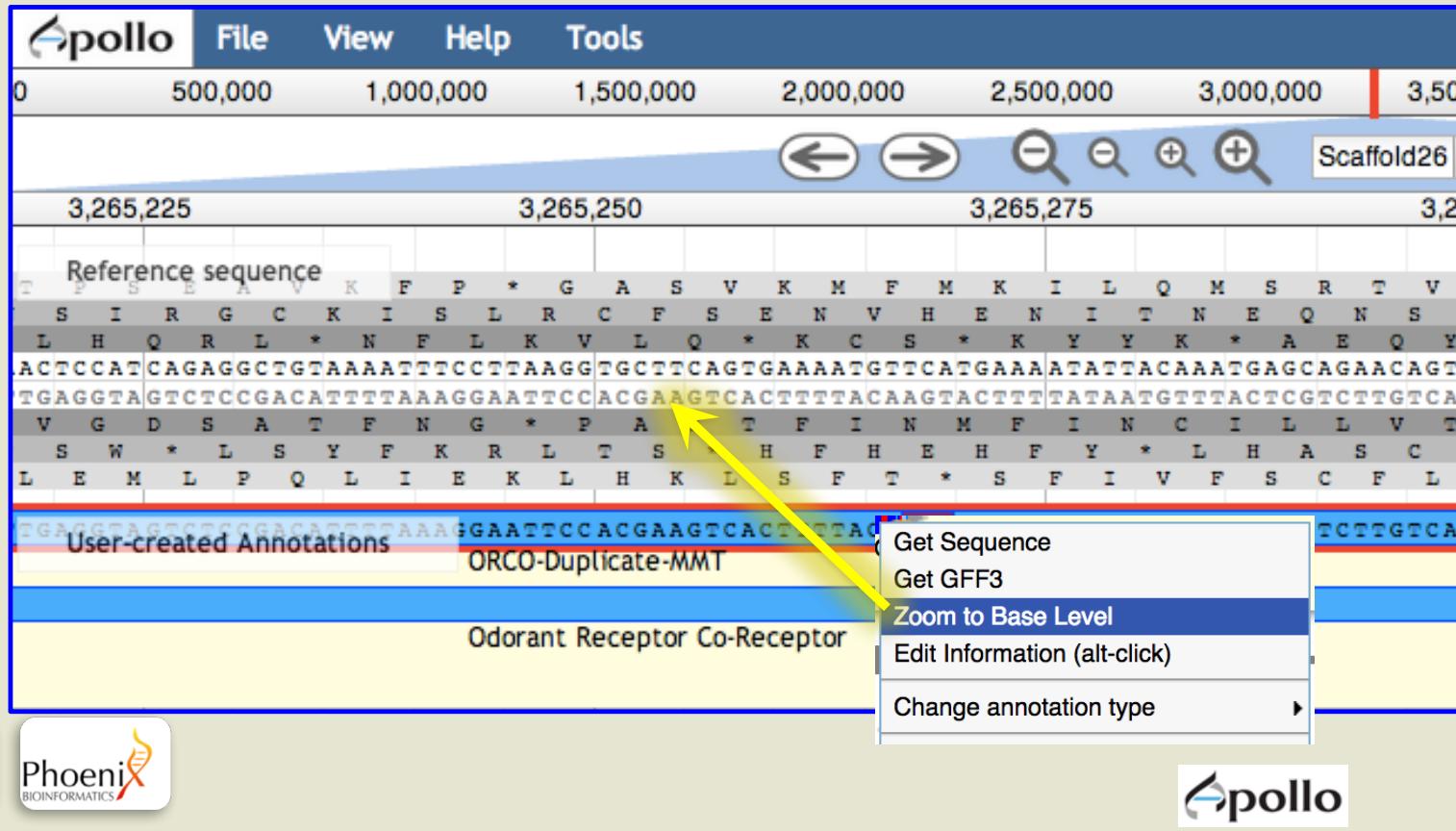


Reference

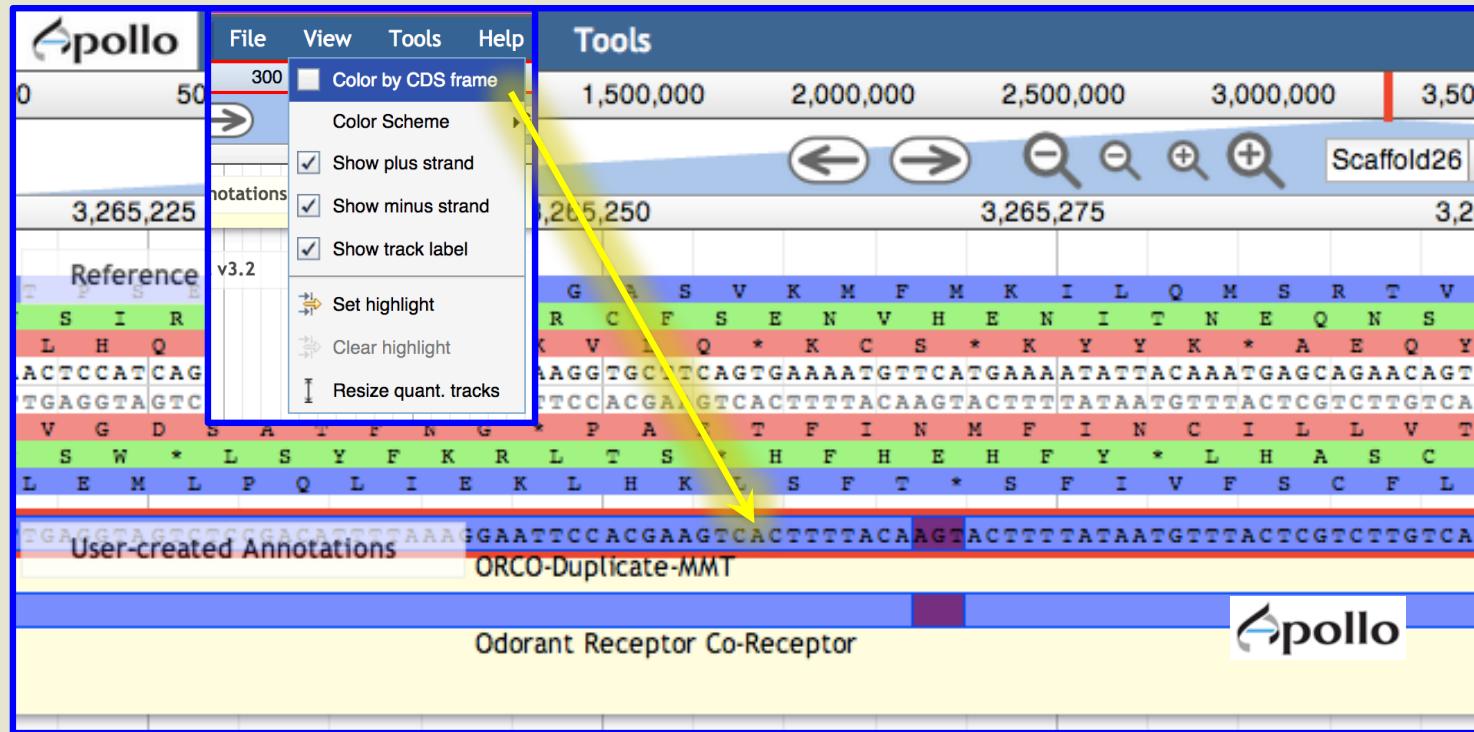
the sequence track



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



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 it. A yellow arrow points from the "View" menu to a submenu where the "Show minus strand" option is highlighted. Another yellow arrow points from the "Tools" menu to a submenu where the "ORCO-Duplicate-MMT" and "Odorant Receptor Co-Receptor" models are listed.

A green callout box at the bottom right provides instructions: "Zoom in/out with keyboard: shift + arrow keys up/down".

Logos for the BERKELEY LAB, Phoenix Bioinformatics, and mutations are visible in the bottom left corner.



Reference

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

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

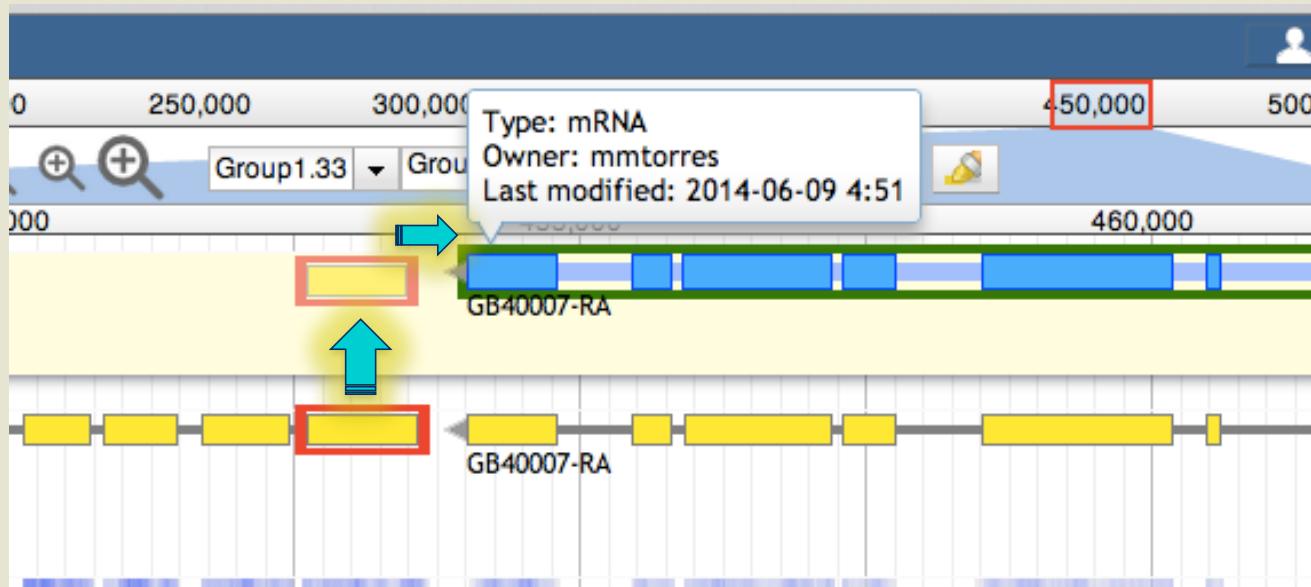
Redo

Show History



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.

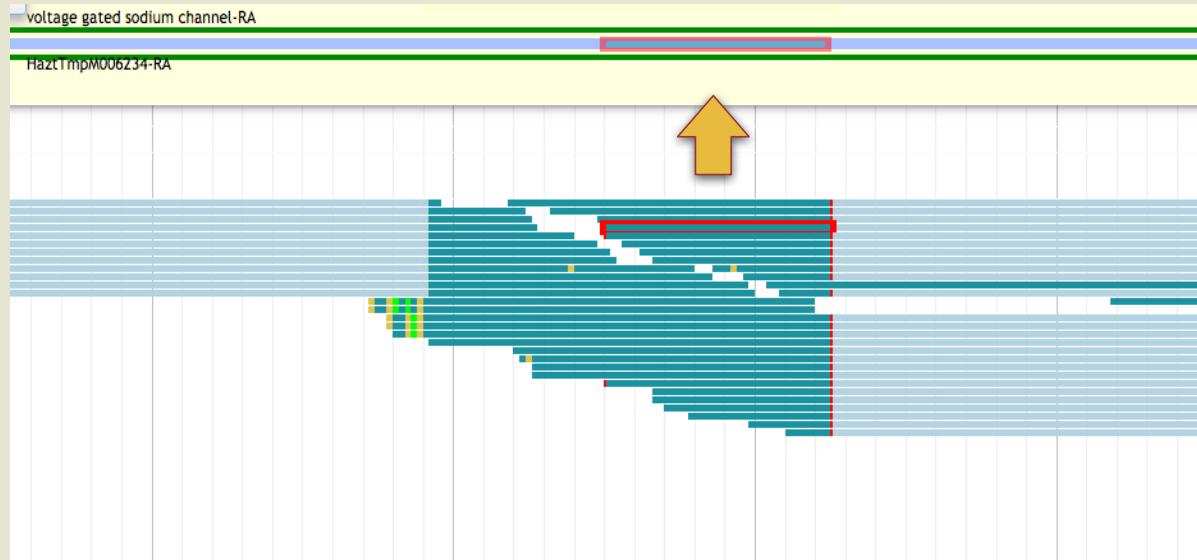


SIMPLE CASES

Editing functionality

Example: Adding an exon supported by experimental data

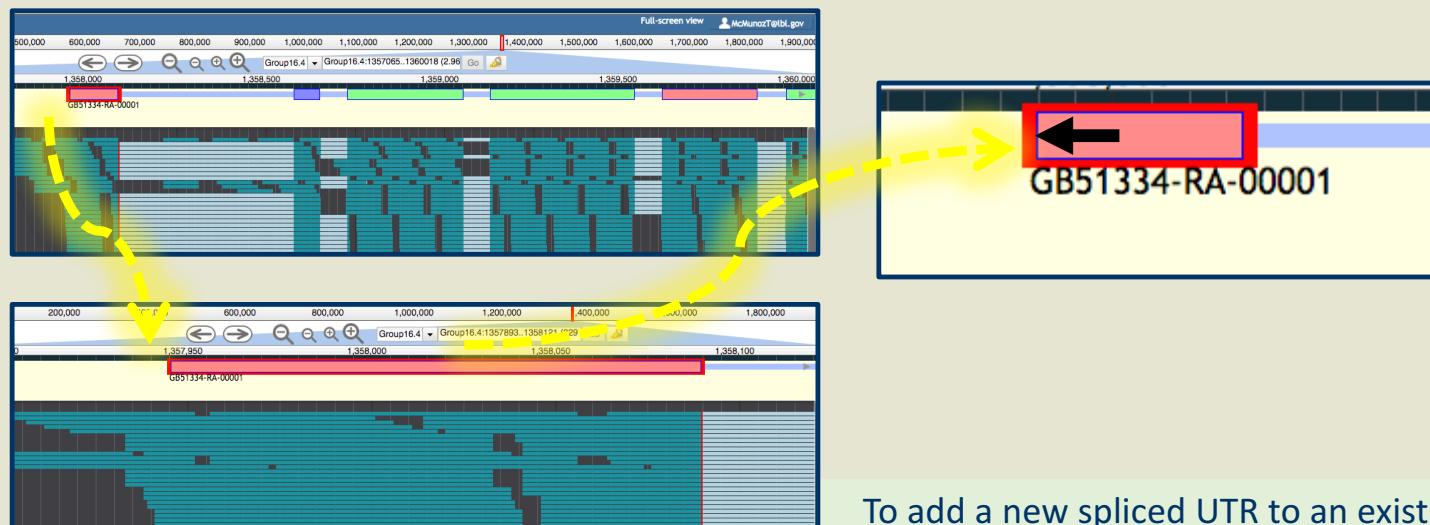
- RNAseq reads show evidence in support of a transcribed product that was not predicted.
- Add exon by dragging up one of the RNAseq reads.



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.



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.



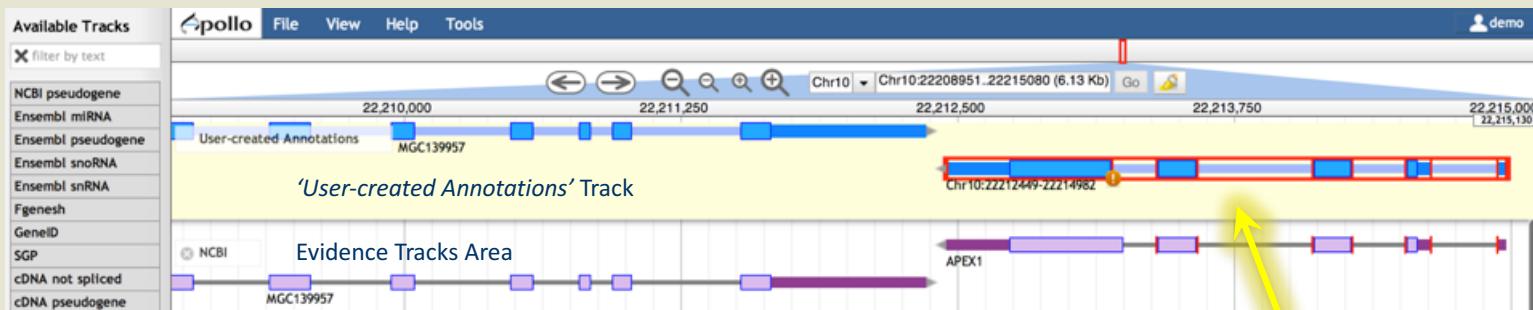
SIMPLE CASES



ORFs - setting & recalculating

Apollo's editing logic (brain):

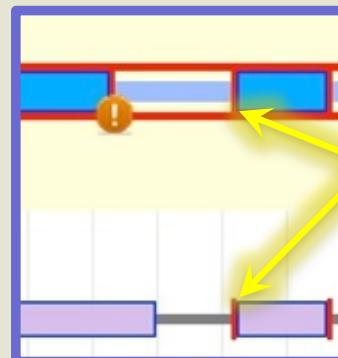
- selects **longest ORF** as CDS
- **recalculates ORF** after each edit, unless set



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



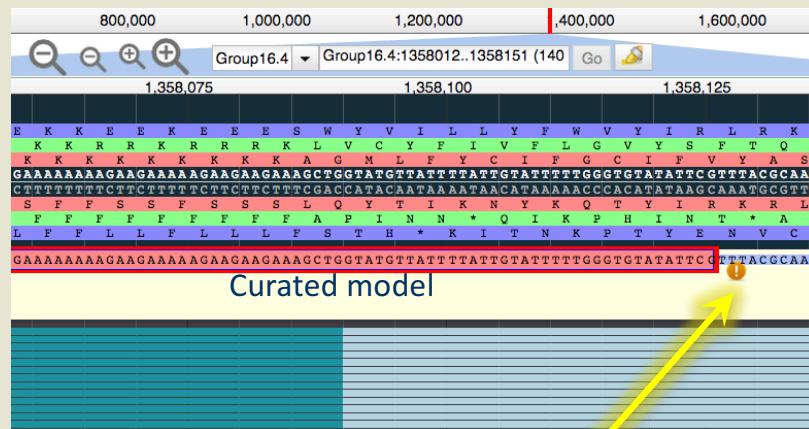
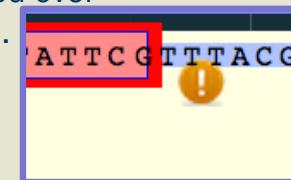
SPLICE SITES

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.



SIMPLE CASES

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.



Editing functionality

Example: Adjusting exon boundaries supported by experimental data

The screenshot shows the Phoenix Bioinformatics software interface for editing mRNA sequences. The main window displays a sequence from position 78,925 to 79,000. A yellow dashed arrow points from the top sequence to a red highlighted region in the middle panel, which contains the sequence: `TCGAAGAAGTCGAGGTACCTAGGTAGGACCCGTCGGTTTACATATTTGGTAGTGT`. Another yellow dashed arrow points from the top sequence to the bottom panel, which shows the mRNA structure with exons in teal and introns in grey. A large grey arrow points from the bottom panel to a secondary window on the right.

Sequence details:

- 78,925: * K G S P F G S I * E R E L L Q A F N I H F G Q P K M Y K P S H
- 78,950: E R V R L A V L O F E S E S F F R L H G S I L W G S Q K C I N H H I
- 78,975: GATGAAGGGGTTGGTCTCGGGGTTCTCAATTGAGAGCCAGGGCTCTAGGCTCCATGGATCATCTGGCAGGCCAAATAATGATAAACATCACAGCTTCCCRAAGCAGAGGCCCAAGARGTTAAACTCTCGCTCTCGAGAAAGCTGGAGGTACCTAGGTAGGACCCGTCGGTTTACATATTTGGTAGTGT
- 79,000: T H F P D Q C R Y E D C S L T R R R T R * N Q S L S L K K L S W P D M R P L W F H I F N *

Bottom panel menu (for -0.2-mRNA-1):

- Get sequence
- Get gff3
- Zoom to base level
- Edit Information (alt-click)
- 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)
- Set to Upstream Splice Acceptor
- Undo
- Redo
- Show History

Bottom panel label: SIMPLE CASES

Right panel (zoomed view):

- 78,975: A P W I H P G Q P K M Y K P S H
- 79,000: G S M D P S W A A K N V * T I T H
- Sequence: GGCCTCATGATCCATCTGGCAGGCCAAATAATGATAAACATCACAGCTTCCCRAAGCAGAGGCCCAAGARGTTAAACTCTCGCTCTCGAGAAAGCTGGAGGTACCTAGGTAGGACCCGTCGGTTTACATATTTGGTAGTGT
- Structure: Shows a single exon (teal) spanning positions 78,975 to 79,000.

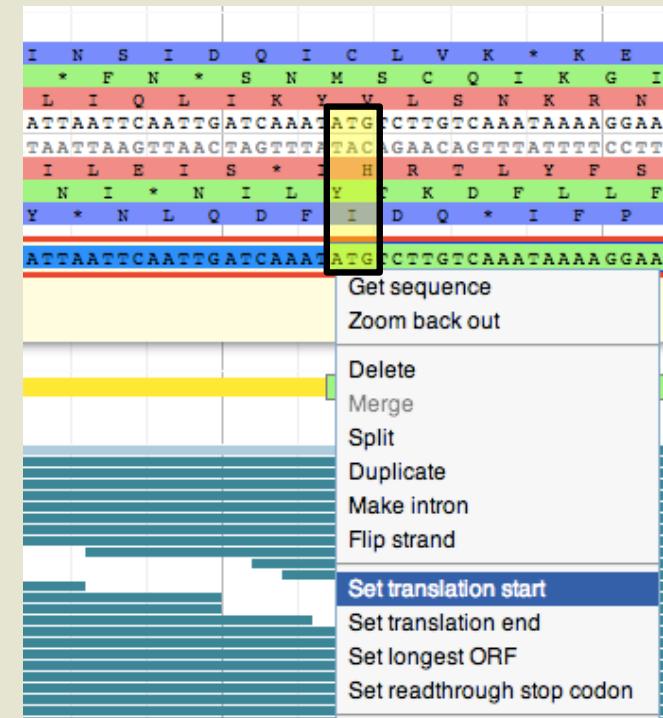
Phoenix Bioinformatics logo:

'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



Reference

curating complex cases

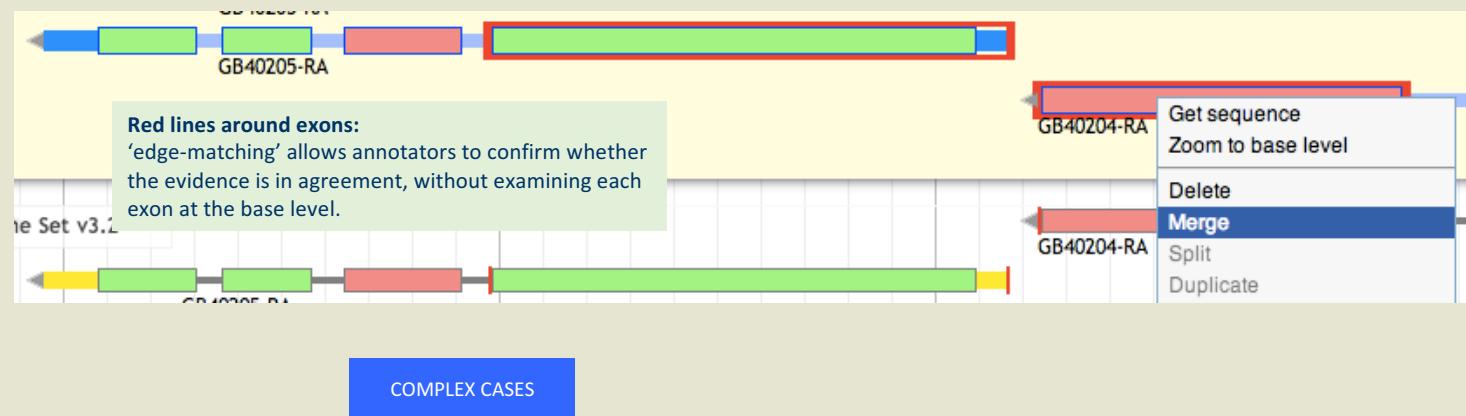


MERGE TWO GENE PREDICTIONS ON THE SAME SCAFFOLD

Evidence may support joining two or more different gene models.

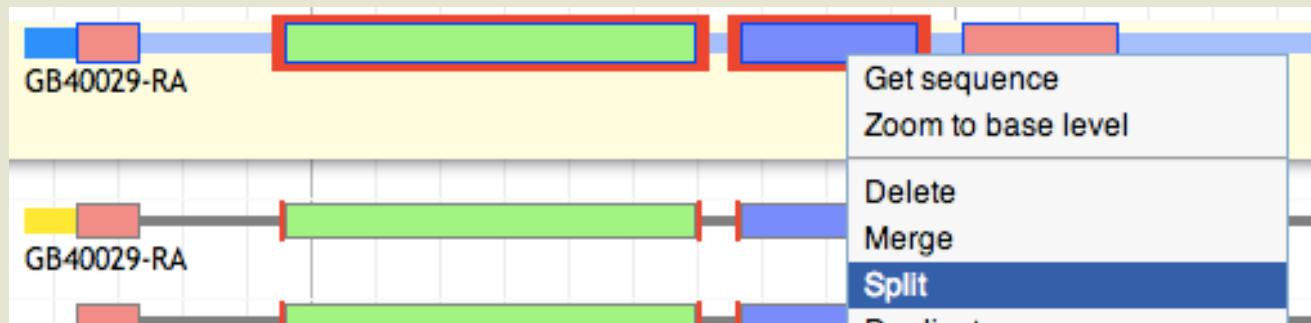
Warning: protein alignments may have incorrect splice sites and lack non-conserved regions!

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 the genome coordinates are set to Chr10:22213112..22213250. A green box highlights the "DNA Track". Below it, a blue box highlights the "User-created Annotations" track. A context menu is open over a specific nucleotide position (highlighted with a red circle), showing options: "Toggle Reverse Strand", "Toggle Protein Translation", "Create Genomic Insertion" (which is selected and highlighted in blue), "Create Genomic Deletion", and "Create Genomic Substitution". To the right of the main track, several floating windows provide additional tools: "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 Berkeley Lab logo. The bottom center has a "COMPLEX CASES" button, and the bottom right corner features the Phoenix Bioinformatics logo.

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

Group1.32 Group1.32:155063..155172 (111 b) Go 🔍

Reference sequence

A R E K L L S D S I S Y M T H K G R I N * T R S L C I F F P F S L L L R
 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 L
 K L G K N F Y P T A * V I * P T K V E L T E R D H F A Y F S L F L F F V
 AAGCTAGGGAAAAAACTTCTATCGACAGCATAAGTTATATGACCCACAAAGGTAGAATTAACTGAACGAGATCACTTGCATATTTCCCTTTCTCTCTTGT

User-created Annotations GB55331-RA-00001

Official Gene Set v3.2

G K G R I N * T R S
 Q R * N * L N E I
 T K V E L T E R D F
 ACAAAAGGTAGAATTAACTGAACGAGATC
 ACAAAAGGTAGAATTAACTGAACGAGATC

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

pollo

COMPLEX CASES

BERKELEY LAB Lawrence Berkeley National Laboratory

Phoenix BIOINFORMATICS

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

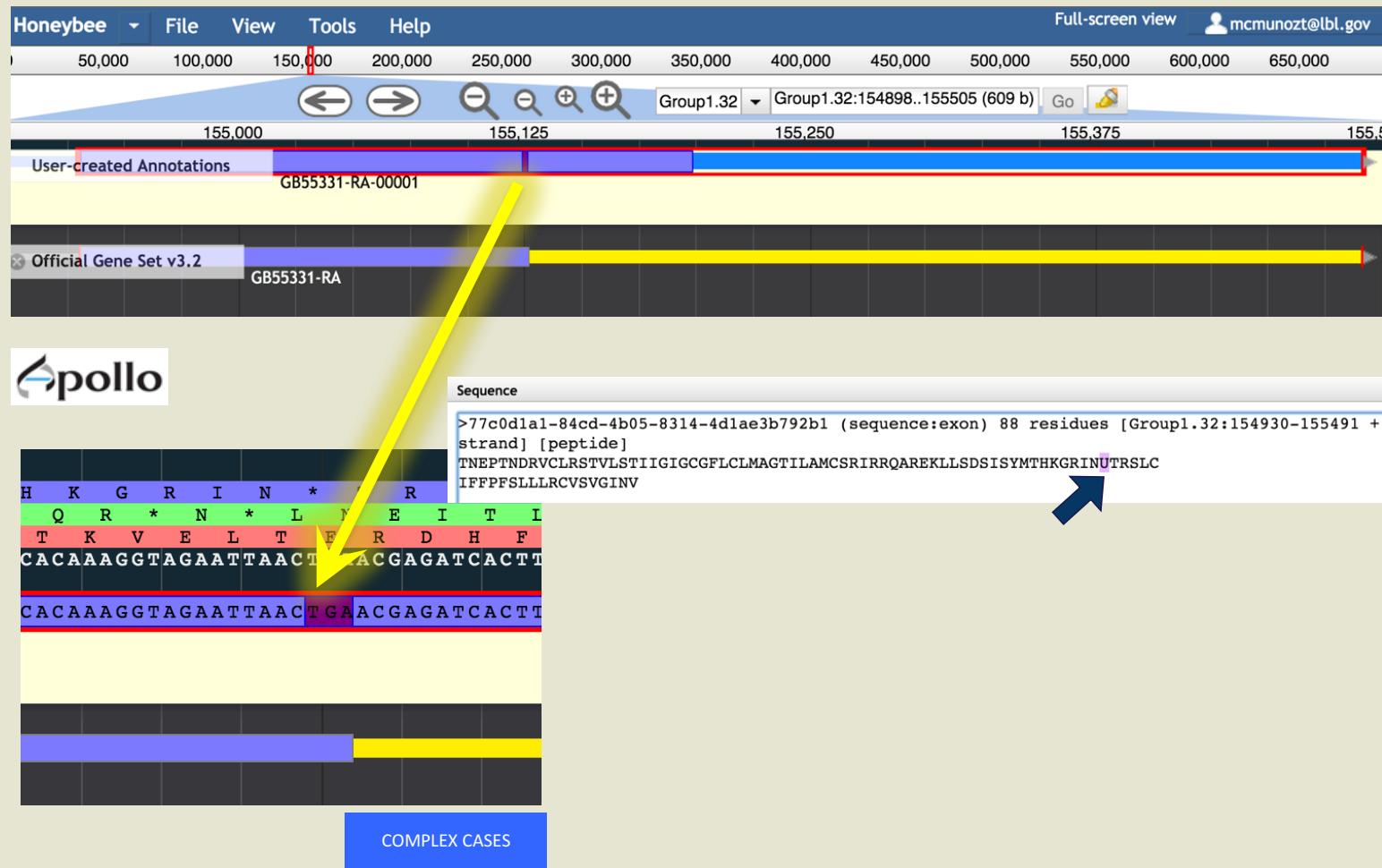
apollo

H K G R I N * R
Q R * N * L N E I T I
T K V E L T F R D H F

CACAAAGGTAGAATTAACT TACGAGATCACTT

CACAAAGGTAGAATTAACT TGAACGAGATCACTT

COMPLEX CASES



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 the creation and linking of gene and mRNA records.

Information Editor Main Window:

- Select mRNA:** spe1-RA
spe1-RA
- gene:** DNA mismatch repair protein Msh
Symbol: spel1
Description: DNA mismatch repair protein Msh
Created: 2017-03-03
Last modified: 2017-03-21
- mRNA:** DNA mismatch repair protein Msh
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

Gene Ontology IDs:

- GO:0006301
- GO:0006298

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]

PubMed IDs:

- 12150919

Comments:

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

Bottom Panel:

- Gene Ontology IDs:** GO:0006301
GO:0006298
- Comments:**

Annotations:

- A yellow arrow points from the "Enter new DB" field in the DBXRefs section of the mRNA record to the "Enter new accession" field in the PubMed IDs section of the mRNA record.
- A yellow arrow points from the "Gene Ontology IDs" section of the mRNA record to the "Comments" section of the mRNA record.
- A yellow arrow points from the "Comments" section of the mRNA record to the "Comments" section of the mRNA record.



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

Add Delete

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

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



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
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
SET_READONLY_STOP_CODON	mmtorres	5/13/14 10:55 AM



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
SET_READONLY_STOP_CODON	mmtorres	5/13/14 10:55 AM



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



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



Follow along



Your number	Email	Password	Server	Organism	Begin at
1	user.one@example.com	userone	1	Honey0	1
2	user.two@example.com	usertwo	2	Honey0	1
3	user.three@example.com	userthree	3	Honey0	1
4	user.four@example.com	userfour	4	Honey0	1
5	user.five@example.com	userfive	5	Honey0	1
6	user.six@example.com	usersix	1	Honey1	7
7	user.seven@example.com	userseven	2	Honey1	7
8	user.eight@example.com	useeight	3	Honey1	7
9	user.nine@example.com	usernine	4	Honey1	7
10	user.ten@example.com	userten	5	Honey1	7
11	user.eleven@example.com	useeleven	1	Honey2	1
12	user.twelve@example.com	usertwelve	2	Honey2	1
13	user.thirteen@example.com	userthirteen	3	Honey2	1
14	user.fourteen@example.com	userfourteen	4	Honey2	1
15	user.fifteen@example.com	userfifteen	5	Honey2	1
16	user.sixteen@example.com	usersixteen	1	Honey3	7
17	user.seventeen@example.com	userseventeen	2	Honey3	7
18	user.eighteen@example.com	useeighteen	3	Honey3	7
19	user.nineteen@example.com	usernineteen	4	Honey3	7
20	user.twenty@example.com	usertwenty	5	Honey3	7
21	user.twentyone@example.com	usertwentyone	1	Honey4	1
22	user.twentytwo@example.com	usertwentytwo	2	Honey4	1
23	user.twentythree@example.com	usertwentythree	3	Honey4	1
24	user.twentyfour@example.com	usertwentyfour	4	Honey4	1
25	user.twentyfive@example.com	usertwentyfive	5	Honey4	1
26	user.twentysix@example.com	usertwentysix	1	Honey5	7
27	user.twentyseven@example.com	usertwentyseven	2	Honey5	7
28	user.twentyeight@example.com	usertwentyeight	3	Honey5	7
29	user.twentynine@example.com	usertwentynine	4	Honey5	7
30	user.twentynine@example.com	usertwentynine	5	Honey5	7

Access Apollo

Your number	Email	Password	Server	Organism	Begin at
1	user.one@example.com	userone	1	Honey0	1
2	user.two@example.com	usertwo	2	Honey0	1

←	user.twentyfour@example.com	usertwentyfour	↑	Honey4	↑
25	user.twentyfive@example.com	usertwentyfive	5	Honey4	1
26	user.twentysix@example.com	usertwentysix	1	Honey5	7
27	user.twentyseven@example.com	usertwentyseven	2	Honey5	7
28	user.twentyeight@example.com	usertwentyeight	3	Honey5	7
29	user.twentynine@example.com	usertwentynine	4	Honey5	7
30	user.twentynine@example.com	usertwentynine	5	Honey5	7



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  
CCGCTCTGCTCTGCTGATCCATCGATCACCTACGCATCGATCCCTCGTTGATCAACGTGGTCAATGAGC  
TGGAGCGTTGAGCGCCGCTATCAGACTGGCGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG  
CTTTAGAATCCTTGCCTGTTGACGATATGGCTGGTCCAGCTTGCCTGGCGCCATCGCTTAC  
AGCATCGGGTGGCAGAGCAGATGCTACAGGACCCGCCGTGAAATTGTTTATGGCTACCGAAGA  
TCGATCAAAAGGATCAGGAATCCATCTCGAACATTCTCCCGCGATTATCATCGACGATGGCGAGGA  
GAGGTTCGTCTCGTCAGCGTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGGCAGAAT  
CTTGAAGAGGAGTTGGCAGCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA  
CACCCGGTGGATTCAATGTTGCACATGTTGATATTACGACATTGGTTCTGTTCAAGAGACCTTCGA  
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  
CCGCTCTGCTCTGCTGATCCATCGATCACCTACGCATCGATCCCTCGTTGATCAACGTGGTCAATGAGC  
TGGAGCGTTGAGCGCCGCTATCAGACTGGCGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG  
CTTTAGAATCCTTGCCTGTTGACGATATGGCTGGTCCAGCTTGCCTGGCGCCATCGCTTAC  
AGCATCGGGTGGCAGAGCAGATGCTACAGGACCCGCCGCTGAAATTGTTTATGGCTACCGAAGA  
TCGATCAAAAGGATCAGGAATCCATCTCGAACATTCTCCCGCGATTATCATCGACGATGGCGAGGA  
GAGGTTCGTCTCGTCAGCGTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGGCAGAAT  
CTTGAAGAGGAGTTGGCAGCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA  
CACCCGGTGGATTCAATGTTGCACATGTTGATATTACGACATTGGTTCTGTTCAAGAGACCTTCGA  
TGCTATGGTCAAGGGATCAGAAGAGTATTCAACGTGCTCACTATGCCATAGTTCCAGGCAGAAATATTC  
ATCACCCATGGAGAAGTTCATGGTGTGAACATTAATAGAAGCCCATCCG
```

Search all genomic sequences

Search



Blat results

Honeybee ▾ File View Tools Help pepita@mendiesta.com

Group8.6 ▾ Group8.6:1864536..1864725 (191 b) Go X

1,864,550 1,864,600 1,864,650 1,864,700

User-created Annotations

Official Gene Set v3.2 GB40335-RA

Search sequence

Blat nucleotide ▾

Enter sequence

```
GTAAAGTGTGCGCCAATTGTCGCCAGAGCTGGCCGTGCAGACCGAGCTGTTAGCCGCGCTCTCGCTGCTGATCATCGATCAGCTACCGCATCGATCCCTCGTGTAGTAAACGTTGGTCATGAGCTGGAGCGTTGAGCCGCCCTATCAGACGGGGAGAGAAAACCTGAATGGAGGCCACGGCAGTTGGACGCTTTAGAATCCCTGGTGTGACGATGGCTGTCAGCTGGGTGCCCGCCCATCGTCTTACAGATCGGGTGGGAGAGACATCGACAGACCCCGCCCTGAAATTGTTTATGGCTAGCGGAAGAAGCATCGAAAAGGATCAGGAATCTCGACGAGAGACATTCATCATCGACCATGGCGAGAGAGGTTCGTCCTGTCAGCGTGTAGGGCCATGATAGGAAACGGCGTTCGTCAACCGGTGTTGAGAATCTTGAAGAAGGAGTTGGCAGCTGACACAGAGAAAATGATGATCAGTGCACACTACTCGCACTCTAACCGGGTGGATTCATGTTGATCGATATTACGACATTCGGTTTCGTTCAAGAGACCTTCGAACCCGGTGGATTCATGTTGATCGATATTACGACATTCGGTTTCGTTCAAGAGACCTTCGAACCCCATGGAGAGTTTGTTGAACATTAATAGAAGGCCATCCG
```

Search all genomic sequences Search

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	7.3e-14	88.24	
Group8.6	1871618	1871664	75	7.3e-14	88.24
GroupUn51..1281	1325	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	0.057	100	
Group6.38	485127	485144	35	0.065	100
GroupUn14..57	570070	570076	35	0.066	95.10

Annotations Tracks Ref Sequence

Search JBrowse Track Selector Show

Show Name

- Official Gene Set v3.2
- Augustus Set 12
- Augustus Set 9
- Fgenesh
- Fgenesh++ without RNASeq training data
- Fgenesh++ with RNASeq training data
- GenelD
- NCBI Gnomon
- N-SCAN
- SGP2
- Official Gene Set v1.0
- Abdomen 454 Contigs
- Brain and Ovary 454 contigs
- Embryo 454 contigs
- Forager Bee Brain Illumina Contigs
- Larvae 454 contigs
- Mixed Antennae 454 Contigs
- NCBI ESTs
- Nurse Bee Brain Illumina Contigs
- Ovary 454 Contigs
- Testes 454 Contigs
- NCBI RefSeq Protein Coding Genes

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

BERKELEY LAB Lawrence Berkeley National Laboratory

Phoenix BIOINFORMATICS

BIPAA resources - blast

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

DV3012683-PA (polypeptide) Daktulosphaira vitifoliae
You are viewing a polypeptide, more information available on the corresponding mRNA page

Overview

NAME	DV3012683-RA
UNIQUE NAME	DV3012683-PA
TYPE	polypeptide
ORGANISM	<i>Daktulosphaira vitifoliae</i> (dvitifoliae)
SEQUENCE LENGTH	1700

A yellow arrow points from the "Blast" link on the BIPAA homepage to the "Blast" link on this specific gene page.

Your blast job RNApolII is finished!

Results

Command ASN.1 XML TSV CSV Text GFF3 HTML

BLASTP 2.6.0+

Reference:
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Reference for composition-based statistics:
Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", *Nucleic Acids Res.* 29:2994-3005.

Database: annotation_v3.0_ogs3.0_20161223_proteins
24,585 sequences; 8,417,147 total letters

Query= *Apis_dorsata* RNA pol II subunit RPB2-like partial
Length=182

Sequences producing significant alignments:	Score (Bits)	E Value
DV3012683-PA gene=DV3012683	308	7e-98
DV3006589-PA gene=DV3006589	77.4	5e-17
DV3000339-PA gene=DV3000339	35.4	0.011

>DV3012683-PA gene=DV3012683
Length=1700

Score = 308 bits (790, Expect = 7e-98, Method: Compositional matrix adjust.
Identities = 146/183 (80%), Positives = 166/183 (91%), Gaps = 1/183 (1%)

Query 1 NYSLLEQQDDDEAIEISK-LNQEAChIVINAYFDDEKLVRQLQLDSPDEFITMSQRIV 59
Sbjct 1 NYI L +D T++E+ E S LNQEAChIVIN+YFDEKLVRQLQLDSPDEFIT+MSVQRIV 60

Query 60 EDSPIQIDQARAQWTSGEIIENPVRELLKFQIYLISKPTINWEDQAPSPMMWNEARLNL 119
Sbjct 61 EDSPIQIDQARAQWTSGEIIENPVRELLKFQIYLISKPTINWEDQAPSPMMWNEARLNL 120

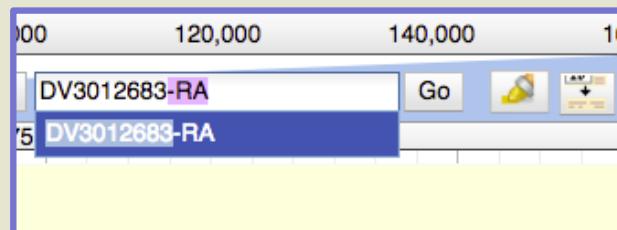
Query 120 YSAPLYVDTIKTIVKDGEDPIETQHQRKTFIGRIPIMLRSKYCLLAGLSDRDLTELNECPL 179
Sbjct 121 YSAPLYVDTIKTIVKDGEDPIETQHQRKTFIGRIPIMLRSYCL GL+DRDLTELNECPL 180

Query 180 DPG 182
Sbjct 181 DPG 183

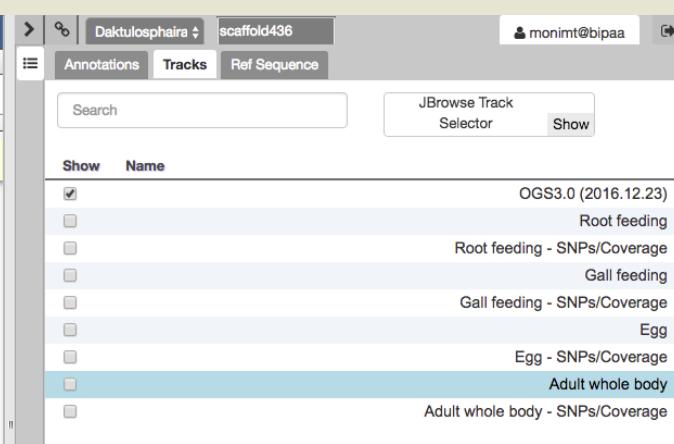
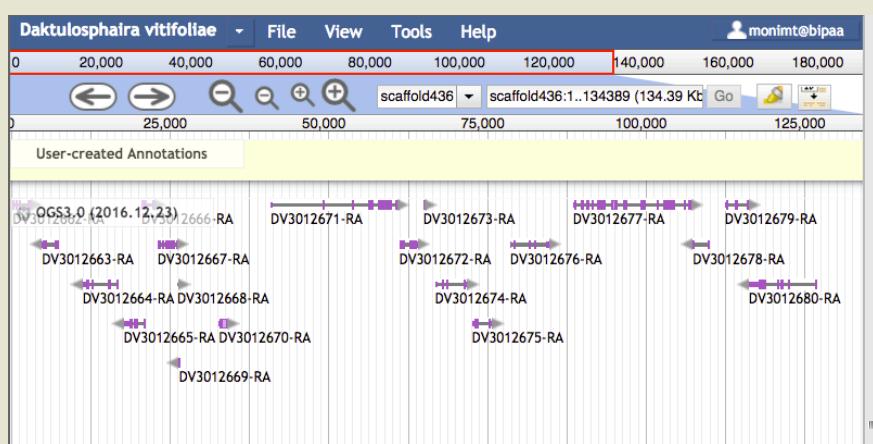
BERKELEY LAB
Lawrence Berkeley National Laboratory

Phoenix Bioinformatics

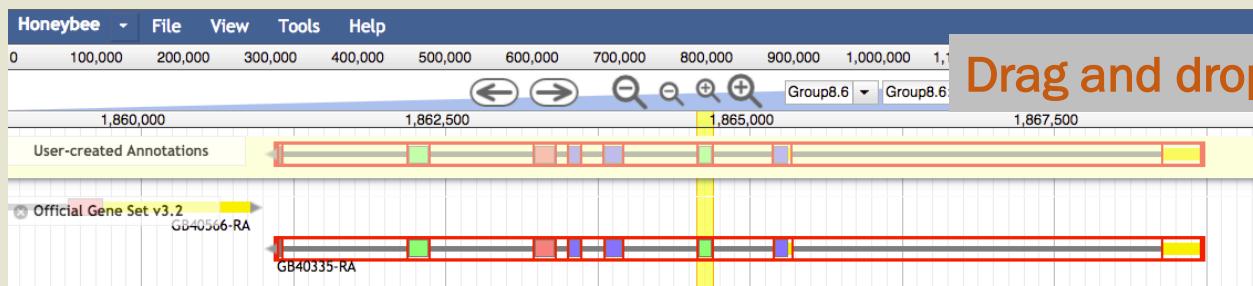
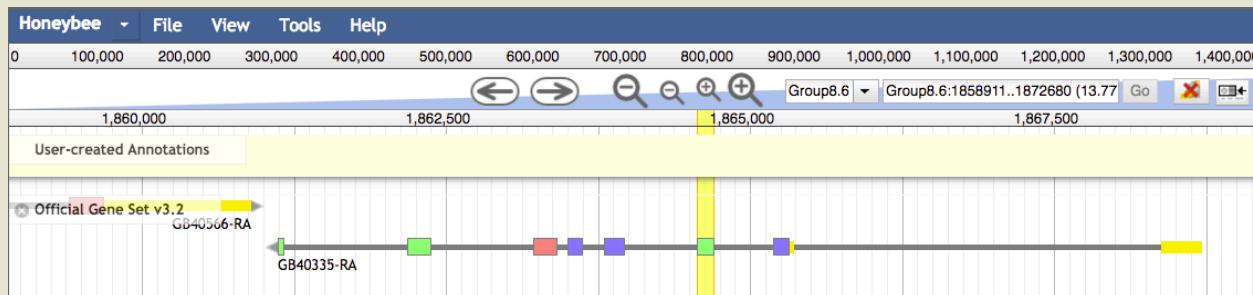
BIPAA resources - Apollo



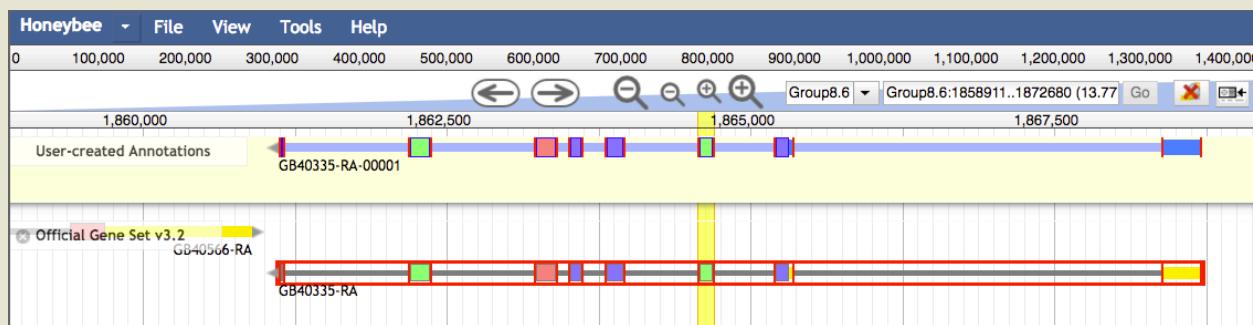
You may find candidate genes from blast results using the 'Search' box with coordinates in main window.



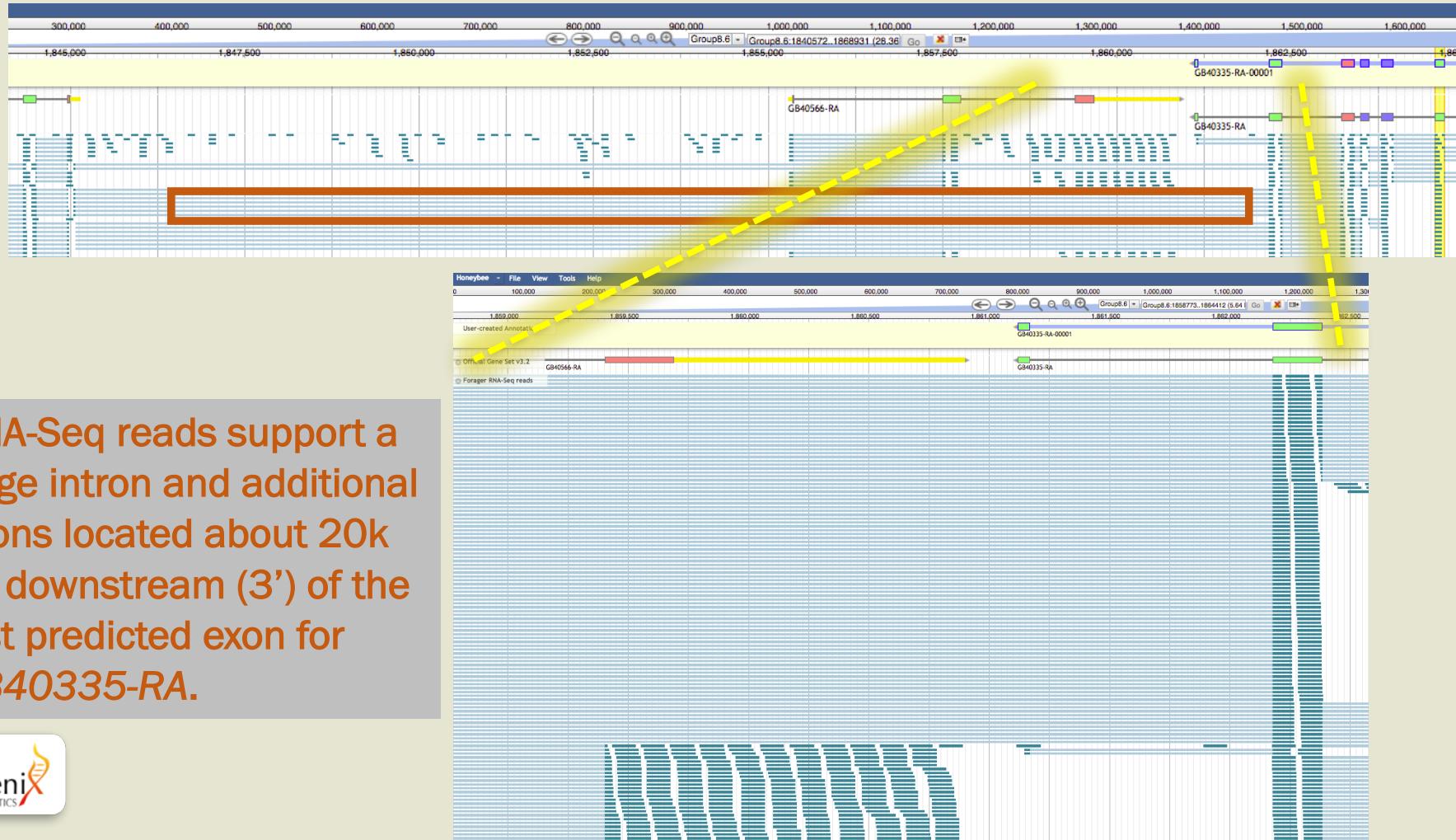
Create a new annotation



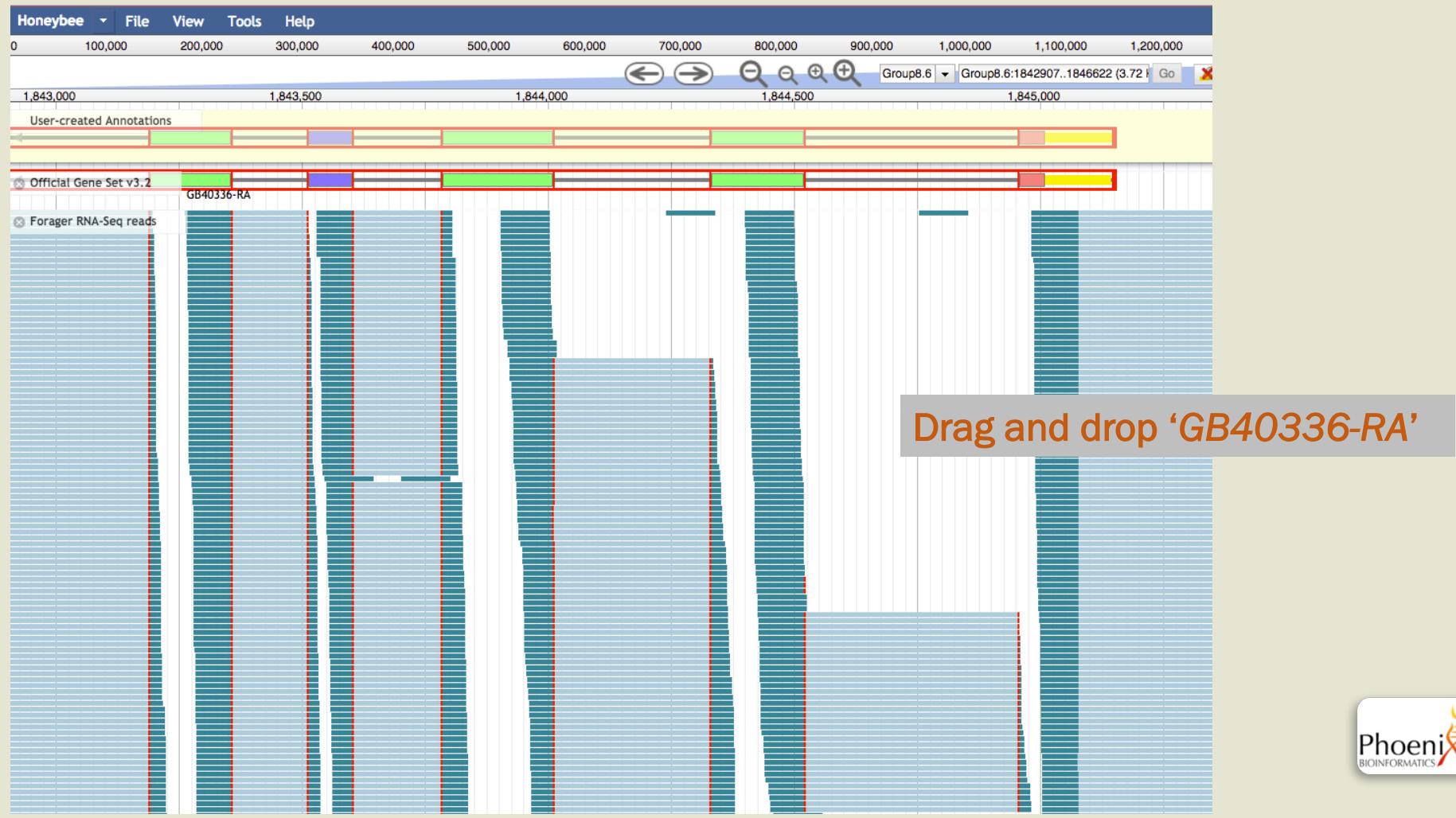
Drag and drop 'GB40335-RA'



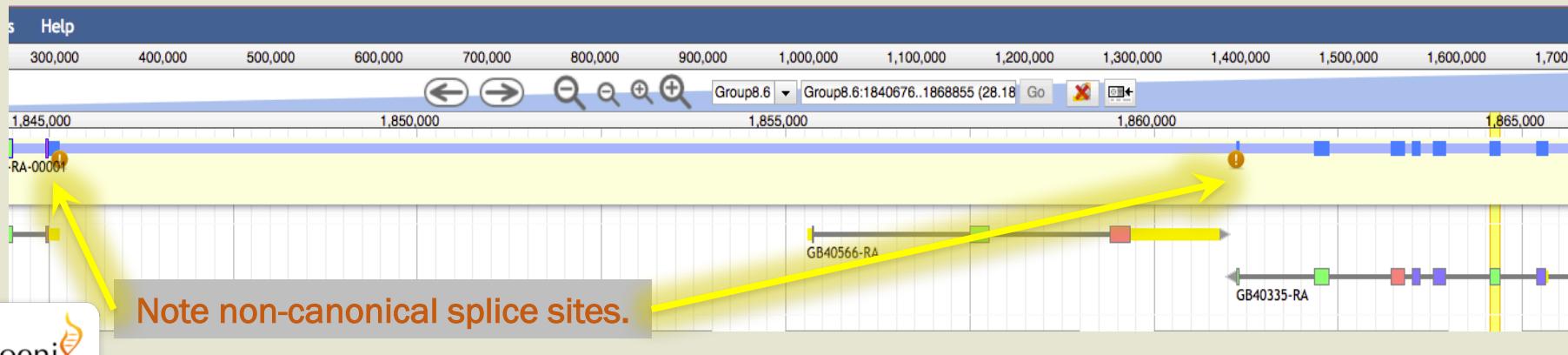
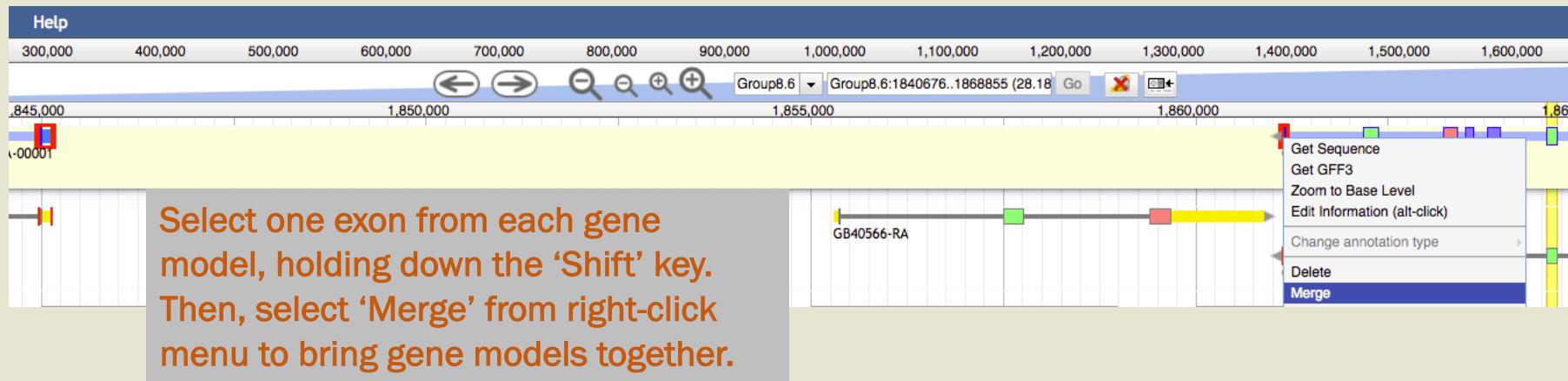
Transcriptomic data support a longer gene



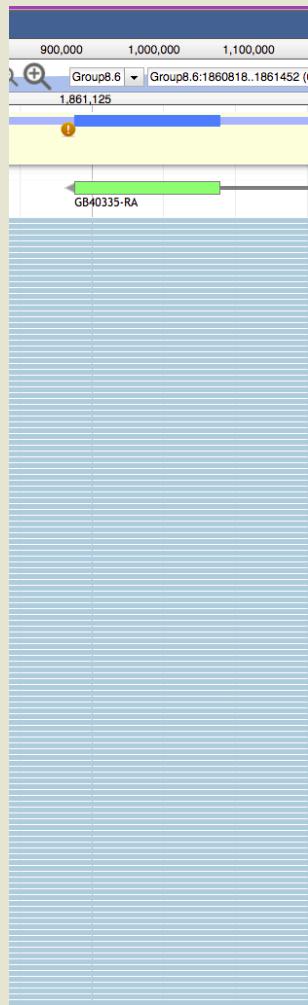
Transcriptomic data support a longer gene



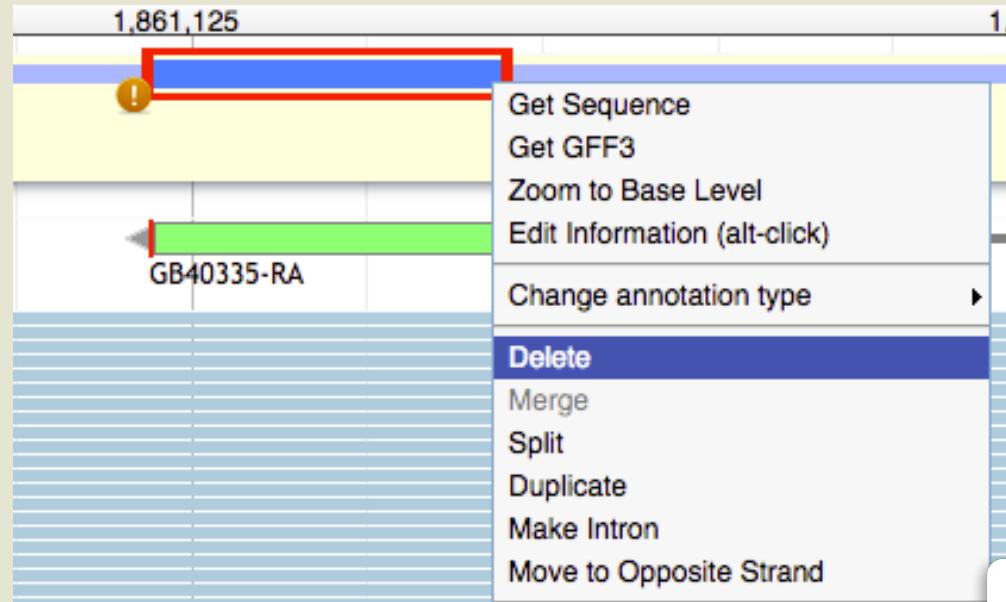
Merge transcripts



Exon not supported by RNA-Seq data

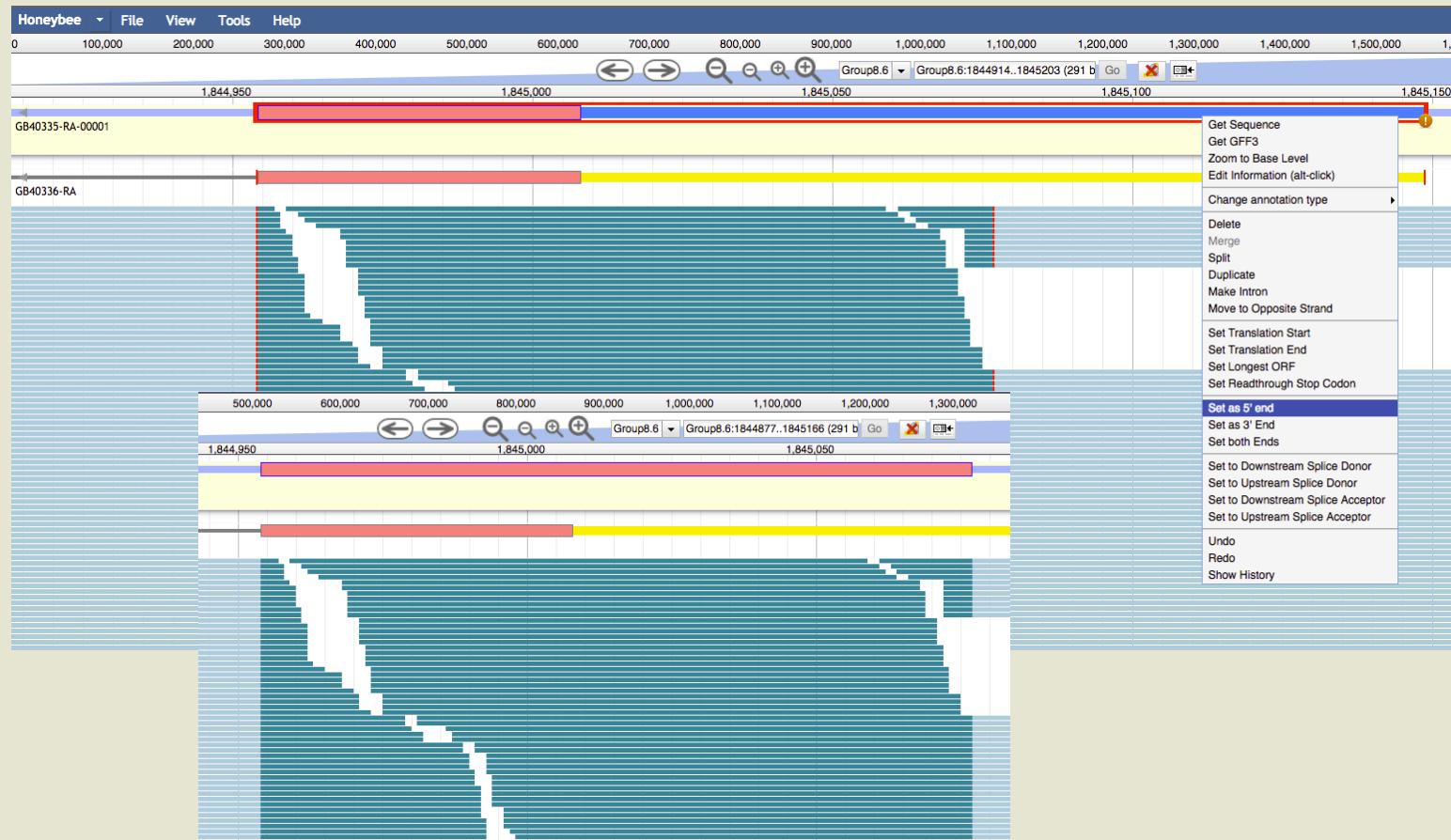


At the end of *GB40335-RA*, select last exon and right-click to choose the 'Delete' option.



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

The screenshot illustrates a workflow for retrieving and comparing a peptide sequence. It consists of three main panels:

- Top Panel:** A genomic browser interface showing two genomic tracks. The top track is labeled "GB40335-RA-00001" and the bottom track is "GB40336-RA". A yellow arrow points from the "Get Sequence" option in a context menu (which is open over the top track) down to the sequence alignment panel.
- Middle Panel:** A "Sequence" viewer window. The sequence is identified as >fccb9943-c0dc-4bbc-b43b-74f6dfe33dfe (sequence:mRNA) 715 residues [Group8.6:1841036-1868721 - strand] [peptide]. The sequence itself is a long string of amino acids.
- Bottom Panel:** The NCBI BLAST suite interface. The query sequence is pasted into the "Enter Query Sequence" field. The search parameters are set to search against the "Non-redundant protein sequences (nr)" database. The algorithm selected is "blast (protein-protein BLAST)".

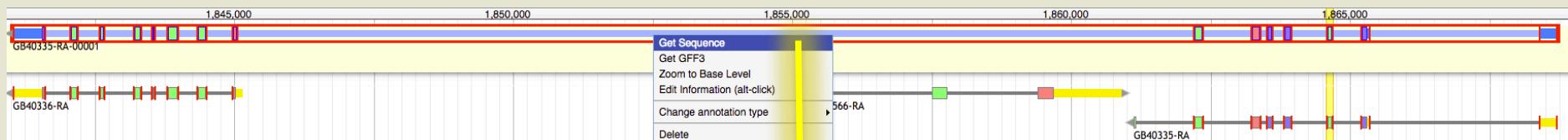
Legend (Visible in Middle Panel):

- Peptide sequence (radio button)
- cDNA sequence (radio button)
- CDS sequence (radio button)
- Genomic sequence (radio button)
- Genomic sequence +/- 500 bases (checkbox)

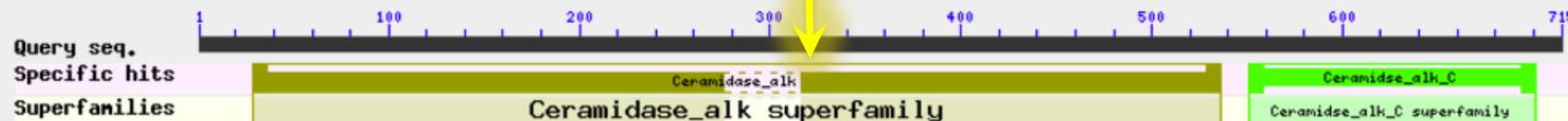
Logos:

- BERKELEY LAB (Lawrence Berkeley National Laboratory)
- Phoenix BIOINFORMATICS

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.

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

DBXRefs

DB	Accession

Add **Delete**

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

Add **Delete**

Don't forget!

Nice to have



Add metadata in ‘Information Editor’

Information Editor

Select mRNA GB40335-RA-00001

gene		mRNA	
Name	neutral ceramidase	Name	neutral ceramidase-00001
Symbol	CCase	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	DBXrefs	DB
Accession		Accession	
Add	Delete	Add	Delete
Attributes		Attributes	
Tag	Value	Tag	Value
Add	Delete	Add	Delete
PubMed IDs		PubMed IDs	
Add	Delete	Add	Delete
Gene Ontology IDs		Gene Ontology IDs	
Add	Delete	Add	Delete
Comments		Comments	
Add	Delete	Add	Delete

Comments
Product of merging GB40335-RA and GB40336-RA
Supporting evidence from Forager RNA-seq reads

Comments

Comments
Product of merging GB40335-RA and GB40336-RA
Supporting evidence from Forager RNA-seq reads

icebox.lbl.gov says:

Publication title: Insights into social insects from the genome of the honeybee *Apis mellifera*.

Cancel OK

NCBI Gene	LOC409628
BeeBase	GB40336
Add	Delete

Attributes

Tag	Value
-----	-------

Junctional annotations using GO IDs. Do not use this field to list publications containing

IDs

17073008	PubMed IDs
Add	Delete

PubMed Identifiers

Gene Ontology terms

Gene Ontology IDs

GO:0017040	ceramidase activity [GO:0017040]
	nuclear-transcribed mRNA catabolic process, dopamine neurotransmitter receptor activity, GINS complex [GO:0000811]

Comments

GO:0017040	ceramide catabolic process [GO:0046514]
	nuclear-transcribed mRNA catabolic process, dopamine neurotransmitter receptor activity, GINS complex [GO:0000811]



Files

<http://bit.ly/apollo-emblabr-exercises1>

<http://bit.ly/apollo-emblabr-exercises2>



Public demo instances



Apollo on the Web

instructions

- Public Honey bee demo available at:

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



Eric Yao



Deepak Unni

JBrowse. Ian Holmes' Lab
University of California, Berkeley

Christine Elsik's Lab,
University of Missouri



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)

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

