



BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY



Apollo

Collaborative genome annotation editing

2018 Winter School – Whole Genome Sequencing, Assembly, and Annotation

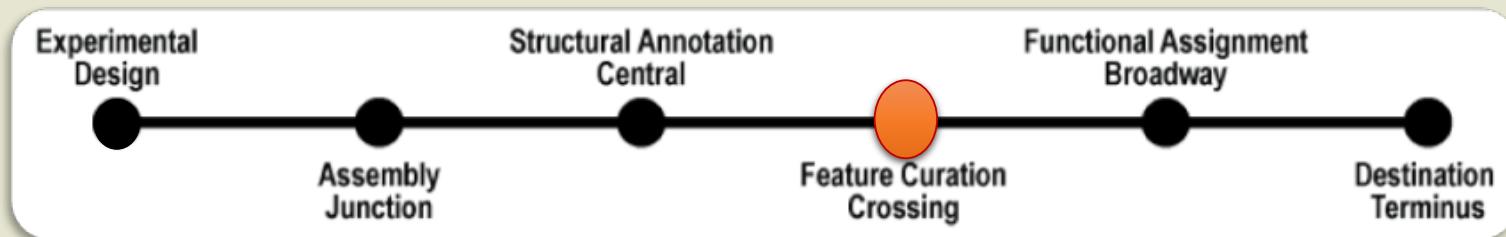
Monica Munoz-Torres, PhD | @monimunozto

Phoenix Bioinformatics – for Lawrence Berkeley National Laboratory

09 February, 2018



<http://GenomeArchitect.org>



Files: bit.ly/wgsaa18-1

Server 1: bit.ly/wgsaa18-s1

Server 2: bit.ly/wgsaa18-s2



Begin at exercise						
Your number	Username	Password	Server	Organism	Exercise	
1	user.one@example.com	userone	1	Honey0	1	
2	user.two@example.com	usertwo	1	Honey0	7	
3	user.three@example.com	userthree	1	Honey1	1	
4	user.four@example.com	userfour	1	Honey1	7	
5	user.five@example.com	userfive	1	Honey2	1	
6	user.six@example.com	usersix	1	Honey2	7	
7	user.seven@example.com	userseven	1	Honey3	1	
8	user.eight@example.com	useight	1	Honey3	7	
9	user.nine@example.com	usernine	1	Honey4	1	
10	user.ten@example.com	userten	1	Honey4	7	
11	user.eleven@example.com	useleven	1	Honey5	1	
12	user.twelve@example.com	usertwelve	1	Honey5	7	
13	user.thirteen@example.com	userthirteen	1	Honey6	1	
14	user.fourteen@example.com	userfourteen	1	Honey6	7	
15	user.fifteen@example.com	userfifteen	1	Honey7	1	
16	user.sixteen@example.com	usersixteen	1	Honey7	7	
17	user.seventeen@example.com	userseventeen	2	Honey0	1	
18	user.eIGHTEEN@example.com	useIGHTEEN	2	Honey0	7	
19	user.nineteen@example.com	usernineteen	2	Honey1	1	
20	user.twenty@example.com	usertwenty	2	Honey1	7	
21	user.twentyone@example.com	usertwentyone	2	Honey2	1	
22	user.twentytwo@example.com	usertwentytwo	2	Honey2	7	
23	user.twentythree@example.com	usertwentythree	2	Honey3	1	
24	user.twentyfour@example.com	usertwentyfour	2	Honey3	7	
25	user.twentyfive@example.com	usertwentyfive	2	Honey4	1	
26	user.twentysix@example.com	usertwentysix	2	Honey4	7	
27	user.twentyseven@example.com	usertwentyseven	2	Honey5	1	
28	user.twentyeight@example.com	usertwentyeight	2	Honey5	7	
29	user.twentynine@example.com	usertwentynine	2	Honey6	1	
30	user.thirty@example.com	userthirty	2	Honey6	7	
31	user.thirtyone@example.com	userthirtyone	2	Honey7	1	
32	user.thirtytwo@example.com	userthirtytwo	2	Honey7	7	

Users	Server	URL
1-16	1	http://bit.ly/wgsaa18-s1
17-32	2	http://bit.ly/wgsaa18-s2

Begin at exercise						
Your number	Username	Password	Server	Organism	Exercise	
1	user.one@example.com	userone	1	Honey0	1	
2	user.two@example.com	usertwo	1	Honey0	7	
3	user.three@example.com	userthree	1	Honey1	1	
4	user.four@example.com	userfour	1	Honey1	7	
5	user.five@example.com	userfive	1	Honey2	1	



Reference



editing functionality



Reference



begin with a new gene model



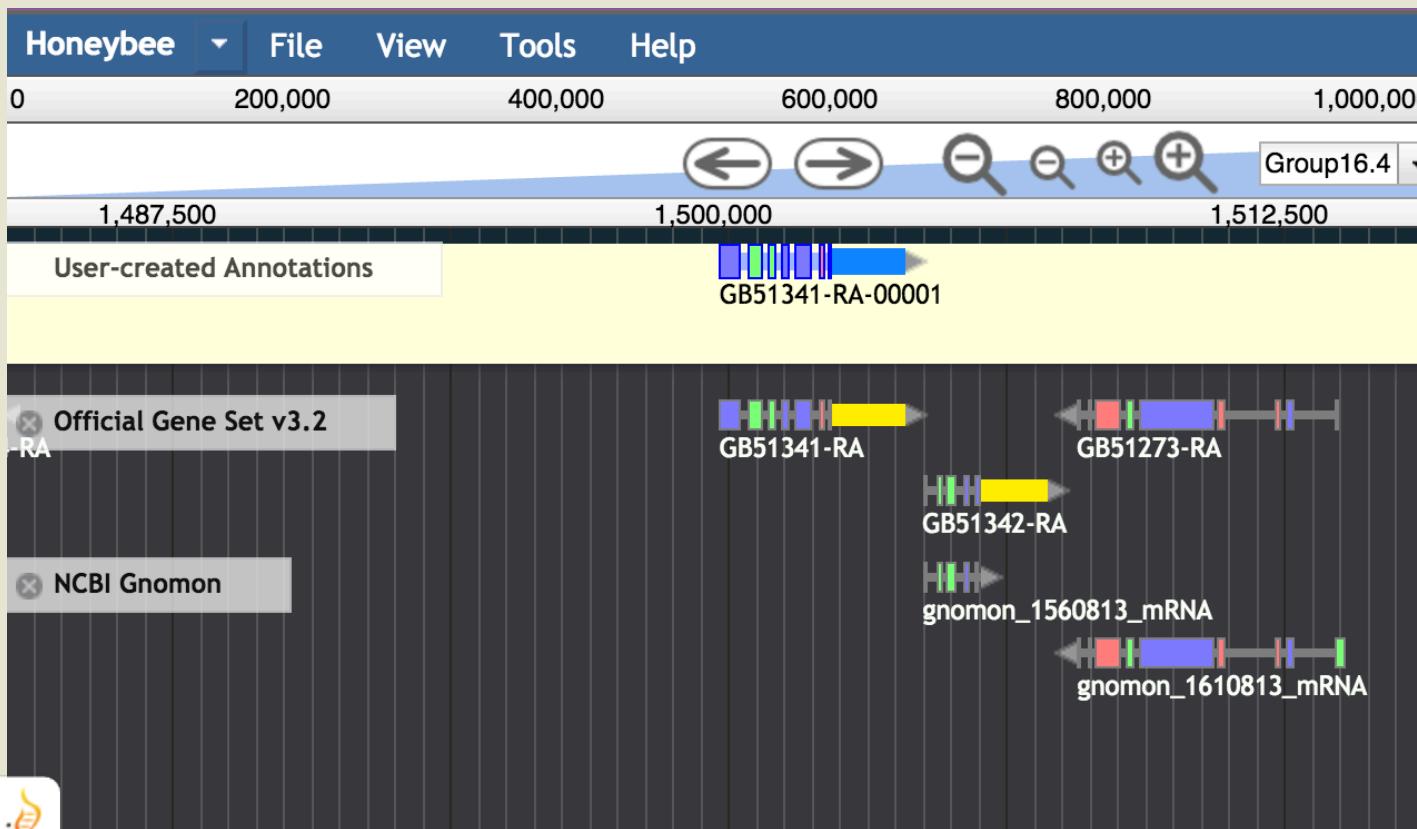
Creating a new annotation



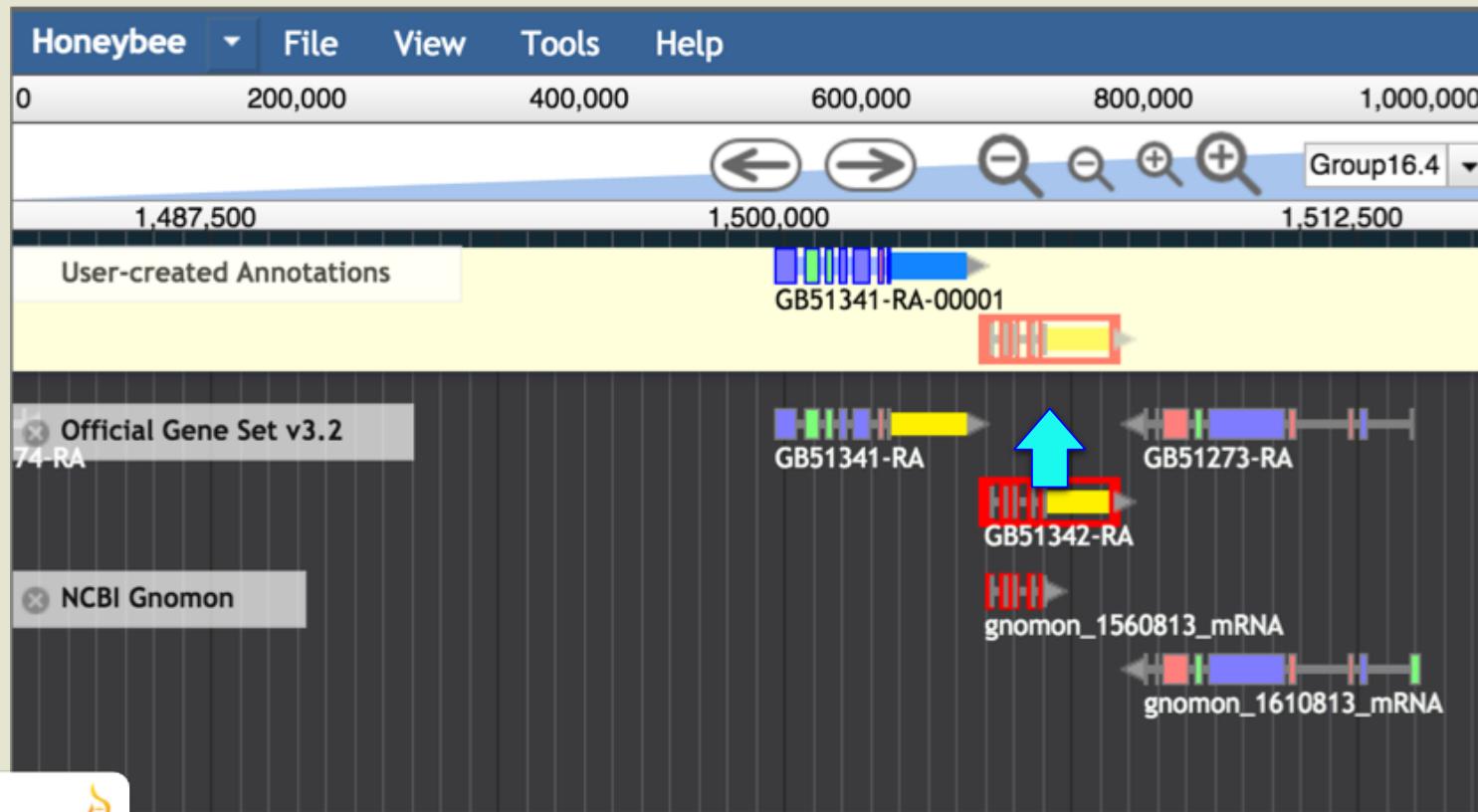
- Choose appropriate evidence from list of “Tracks” on **annotator panel**.
- Select & drag elements from evidence track into the ‘User-created Annotations’ area.
- Hovering over annotation in progress brings up an information pop-up.



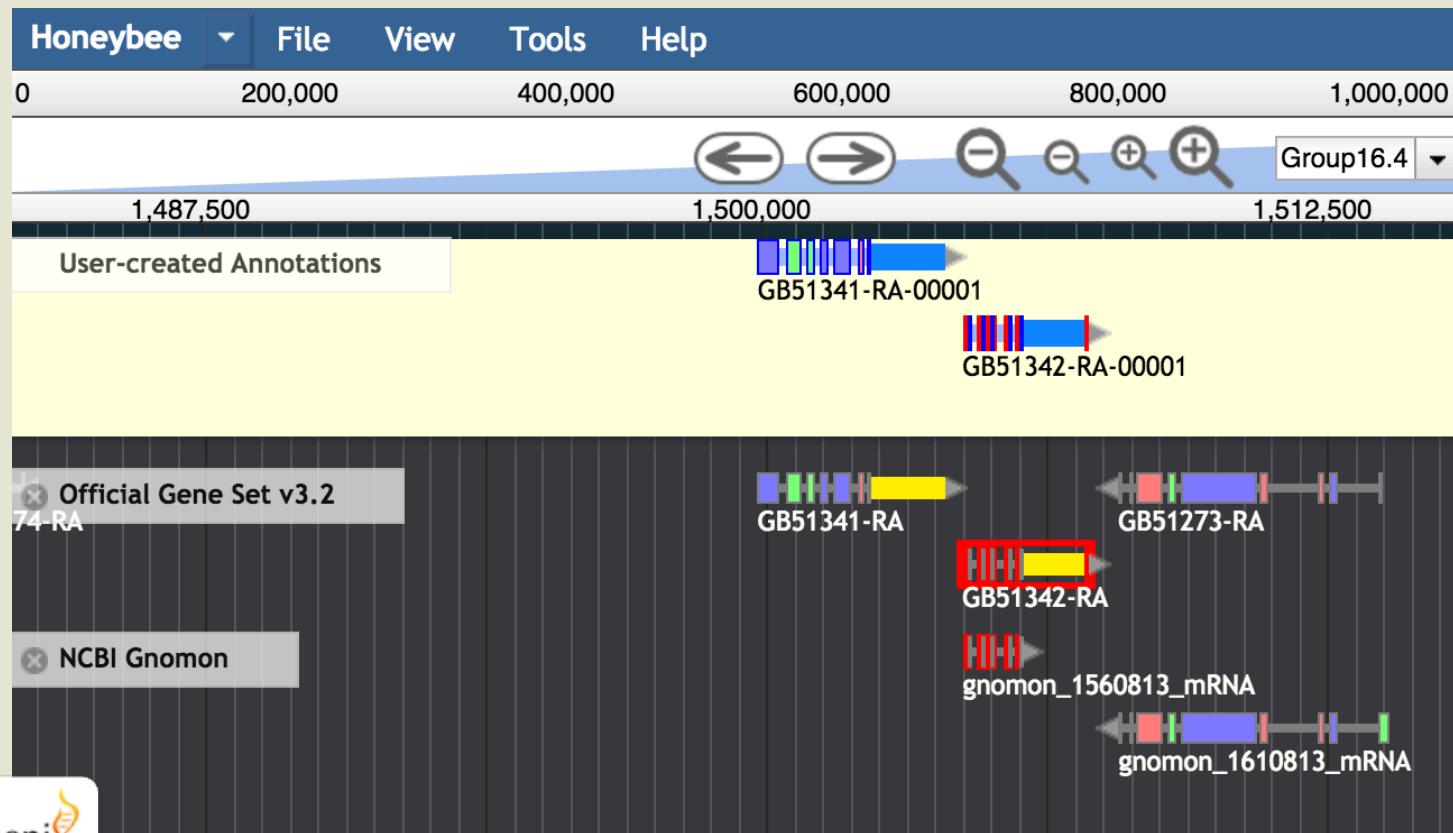
Adding a gene model



Adding a gene model



Adding a gene model



Reference

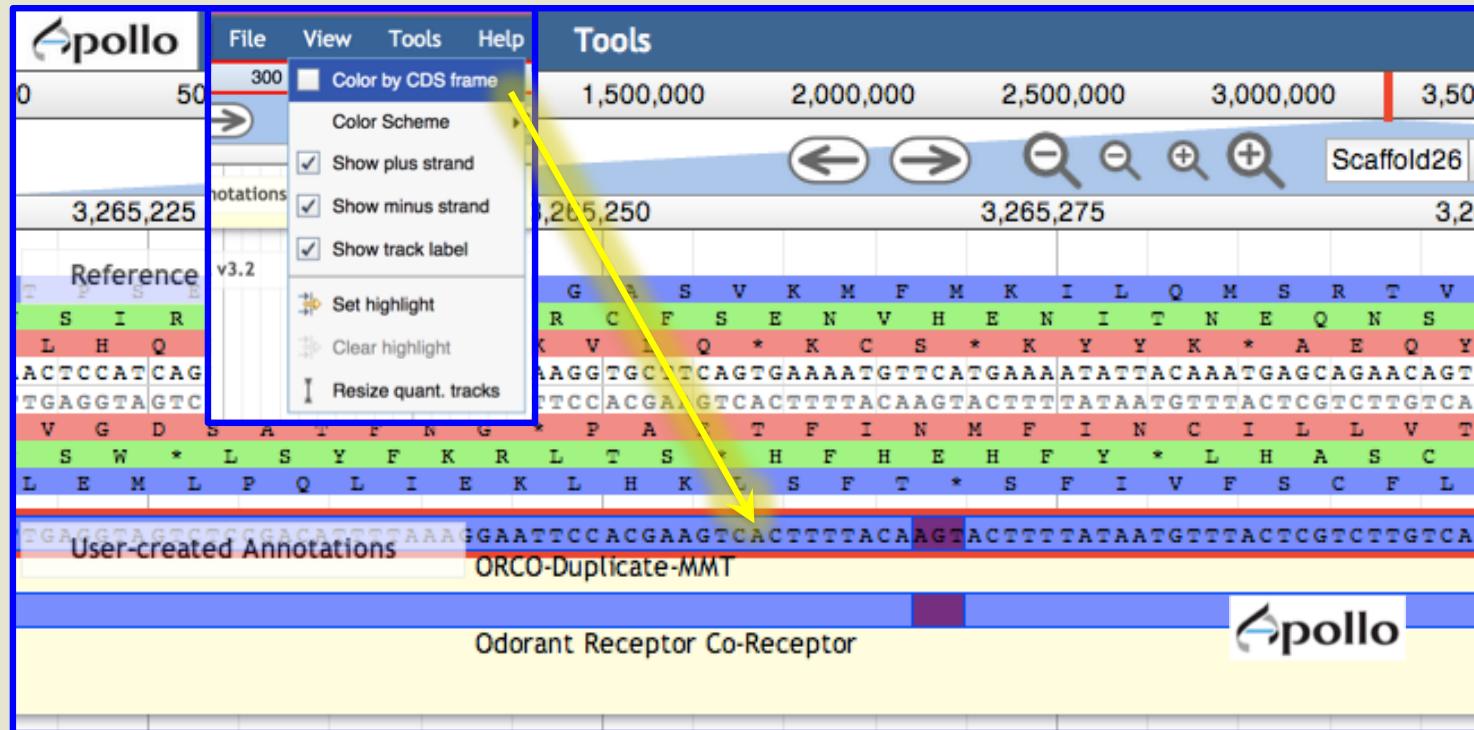
the sequence track



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

The screenshot shows the Apollo software interface. At the top, there's 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. To the right of the scale are several icons: a left arrow, a right arrow, a magnifying glass, and a plus sign. The text "Scaffold26" is displayed next to the icons. Below the scale, four specific genomic coordinates are highlighted: 3,265,225, 3,265,250, 3,265,275, and 3,265,280. The main area displays a reference sequence with amino acid translations above it. A yellow arrow points from the "User-created Annotations" track to a context menu. The menu items are: Get Sequence, Get GFF3, Zoom to Base Level (which is highlighted in blue), Edit Information (alt-click), and Change annotation type. The "User-created Annotations" track contains the text "ORCO-Duplicate-MMT". The "Odorant Receptor Co-Receptor" track is also visible below it. Logos for the Berkeley Lab and Phoenix Bioinformatics are in the bottom corners, along with the Apollo logo.

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 workspace displays a "Reference sequence" with amino acid translations above it. A yellow arrow points from the "View" menu in the bottom-left corner to a submenu where the "Show minus strand" option is highlighted. Another yellow arrow points from the "View" menu to a callout box containing the text "Zoom in/out with keyboard: shift + arrow keys up/down".

At the bottom left, there are logos for the Berkeley Lab and Phoenix Bioinformatics. The Apollo logo is located at the bottom right.

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

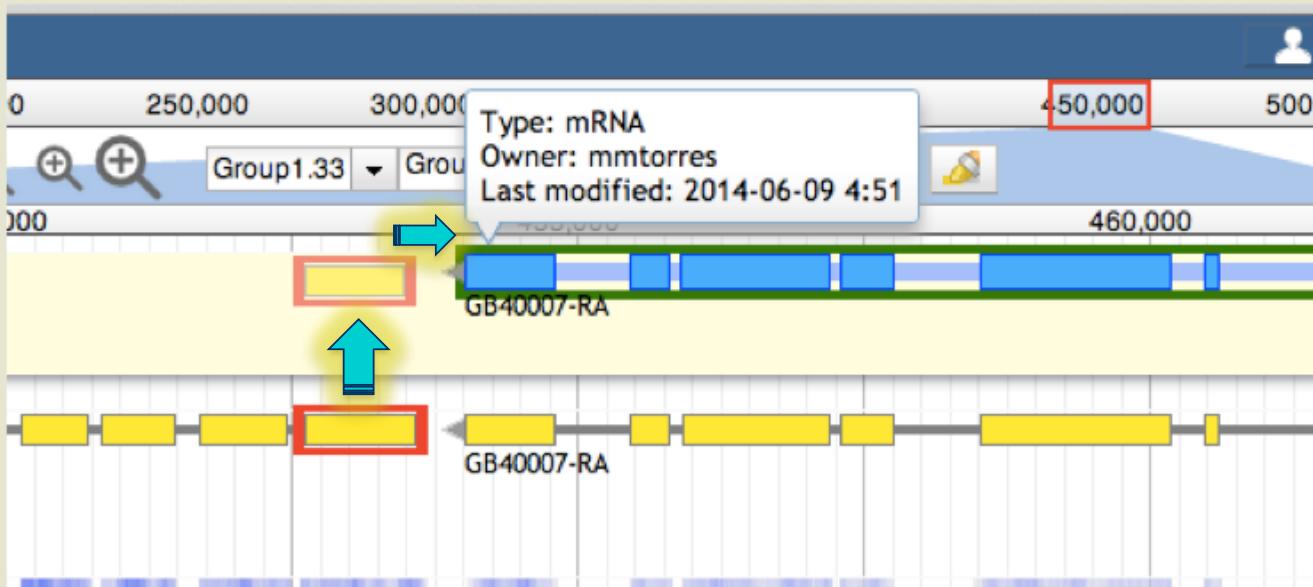
Redo

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.

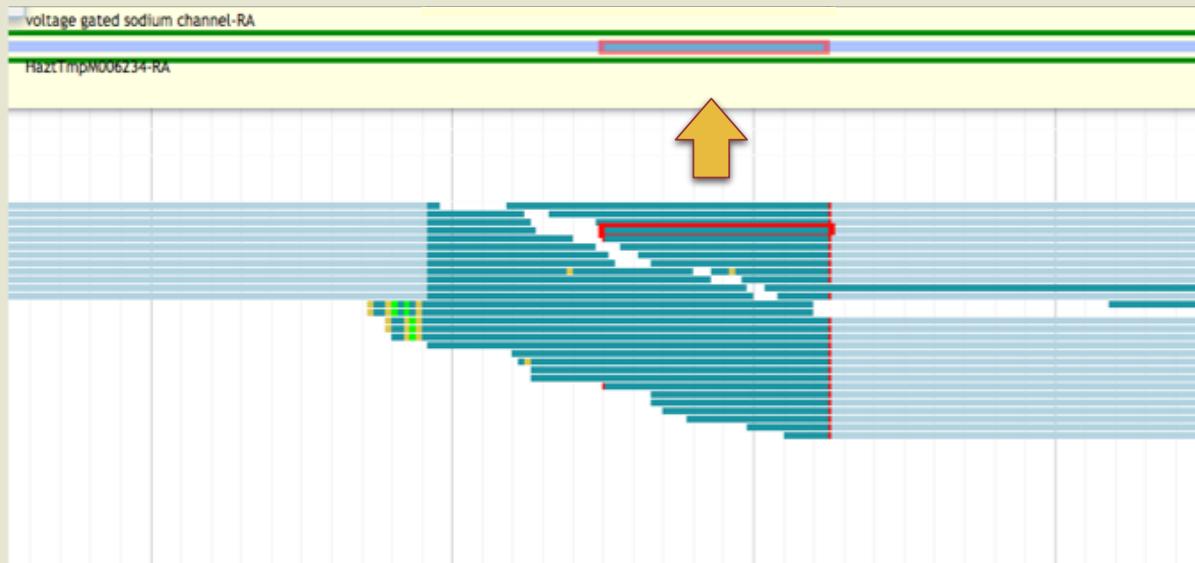


SIMPLE CASES

Editing functionality

Example: Adding an exon supported by experimental data

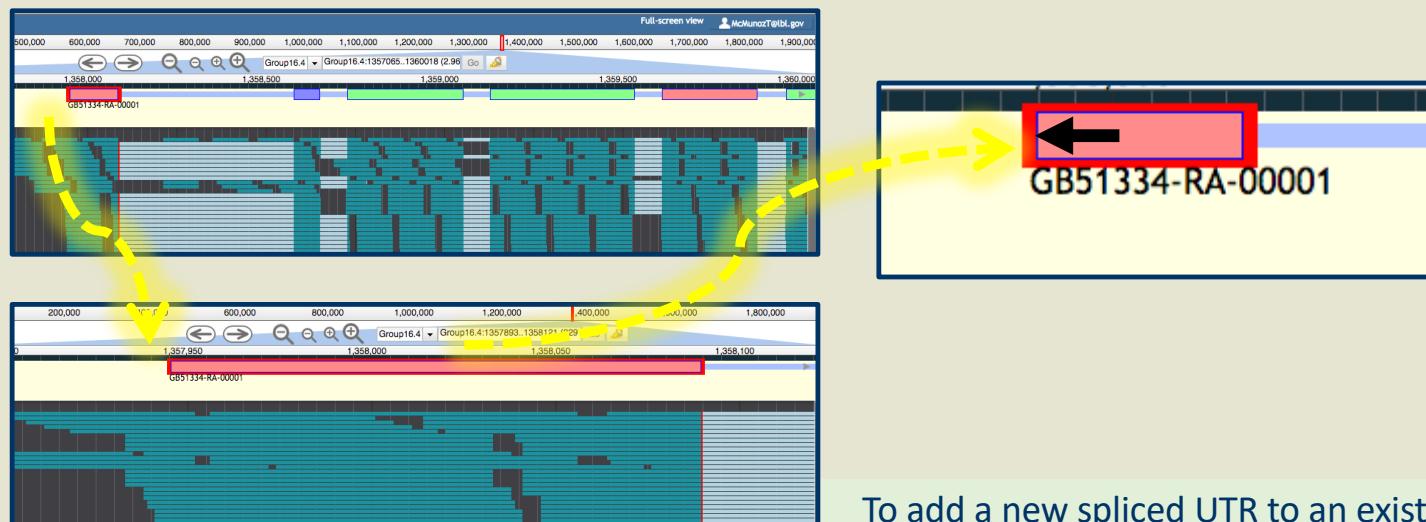
- RNAseq reads show evidence in support of a transcribed product that was not predicted.
- Add exon by dragging up one of the RNAseq reads.



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.

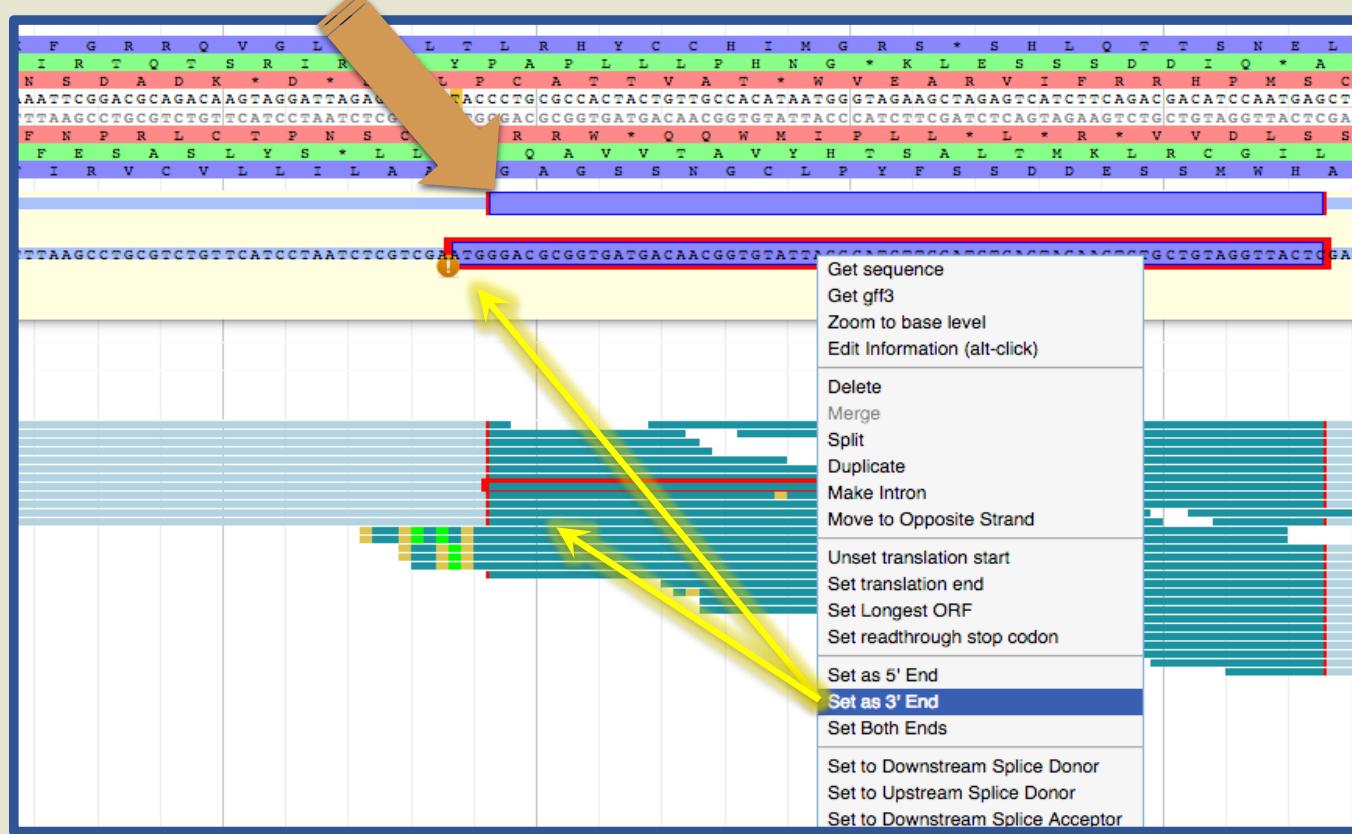


To add a new spliced UTR to an existing annotation also follow the procedure for adding an exon, or to 'Set as X' end'.



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.



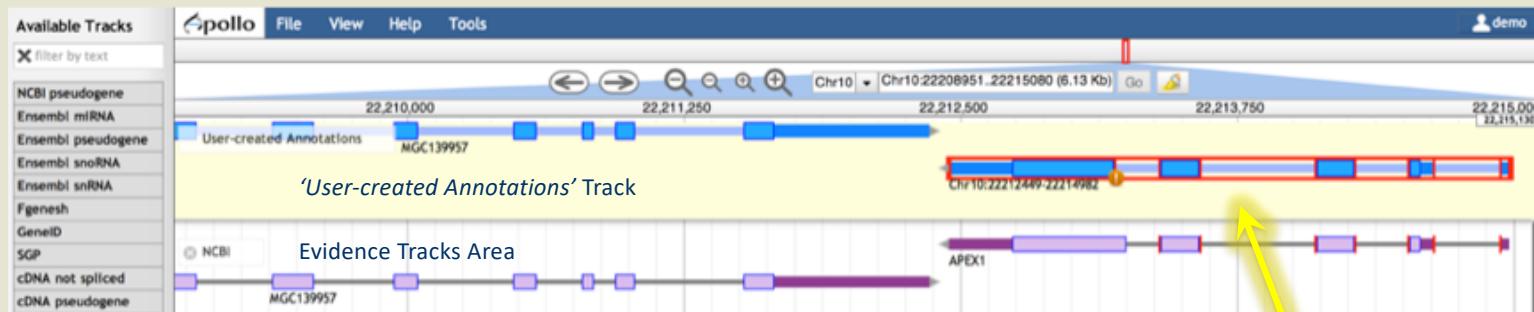
SIMPLE CASES



ORFs - setting & recalculating

Apollo's editing logic (brain):

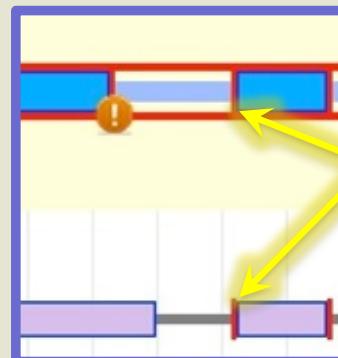
- selects **longest ORF** as CDS
- **recalculates ORF** after each edit, unless set



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



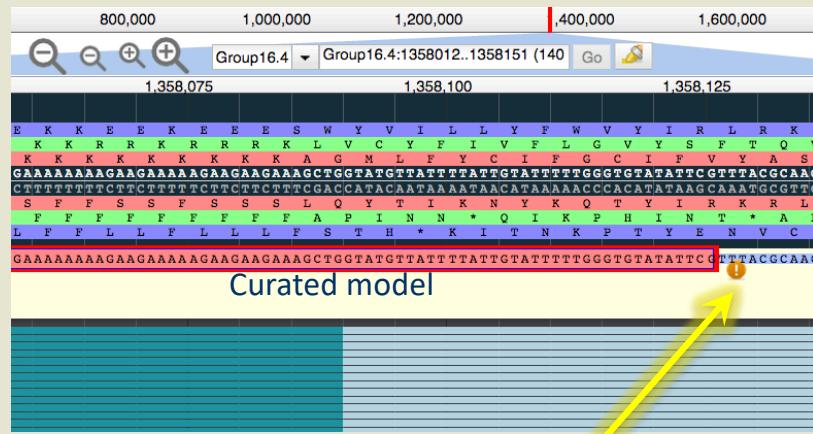
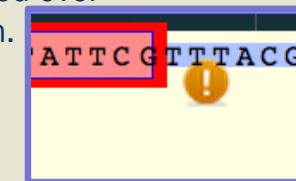
SPLICE SITES

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.



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 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 zoomed-in view of the sequence from 78,975 to 79,000, which is highlighted in red. Another yellow dashed arrow points from the zoomed-in sequence back to the main sequence. A large grey arrow at the bottom right indicates the movement of the edit.

Sequence View:

78,925 78,950 78,975 79,000

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

Contextual Menu:

-0.2-mRNA-1

- 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**
- Set to Upstream Splice Acceptor
- Undo
- Redo
- Show History

Bottom Panel:

SIMPLE CASES

Phoenix Bioinformatics Logo:

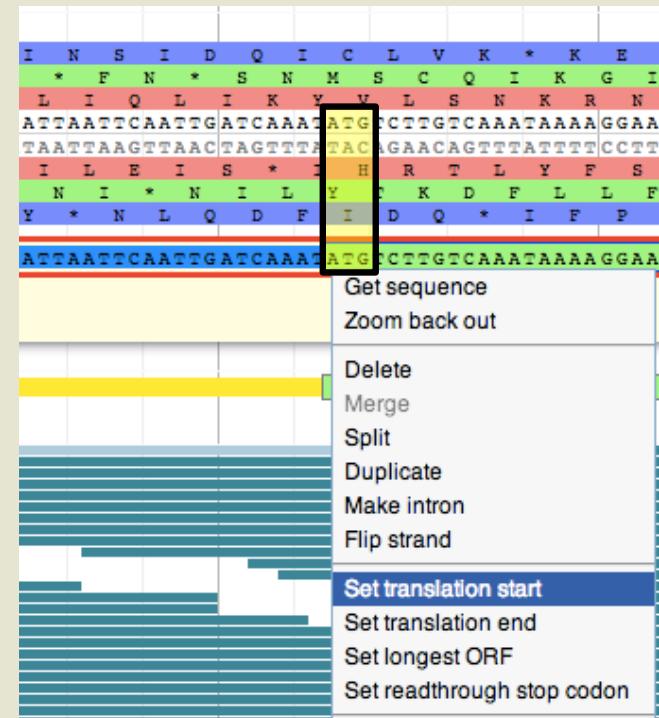


'Start AND 'Stop' SITES

- 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

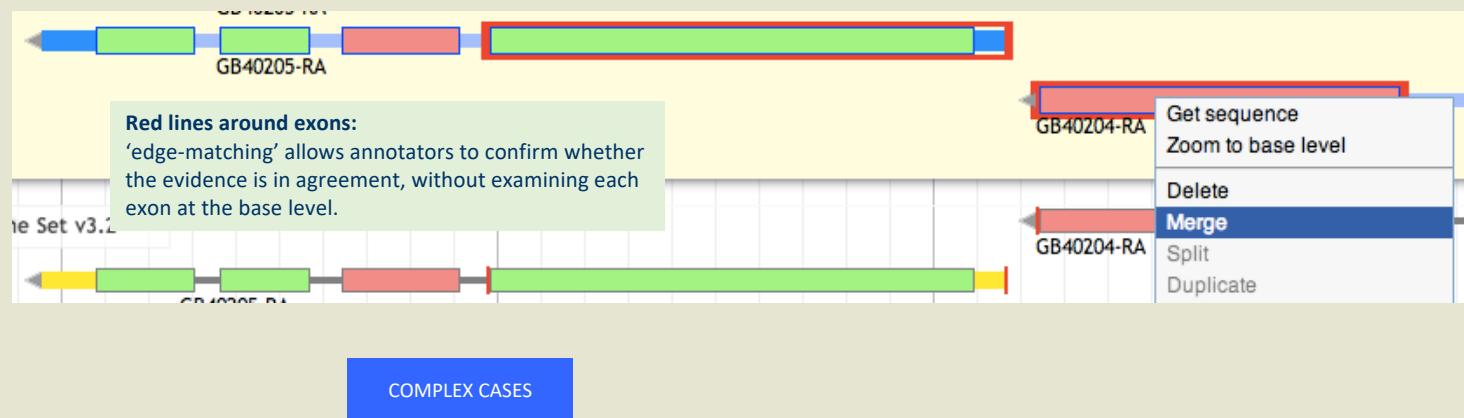


MERGE TWO GENE PREDICTIONS ON THE SAME SCAFFOLD

Evidence may support joining two or more different gene models.

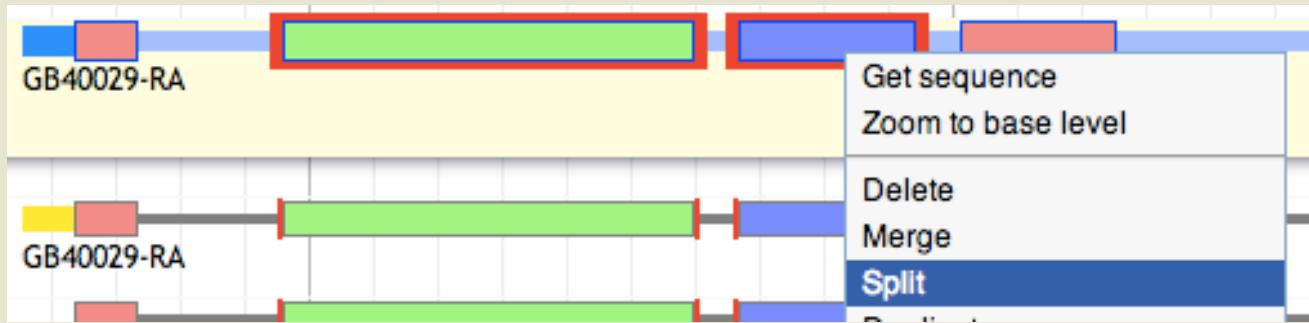
Warning: protein alignments may have incorrect splice sites and lack non-conserved regions!

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), listing 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 annotation 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 bottom right corner features the Phoenix Bioinformatics logo.



COMPLEX CASES



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

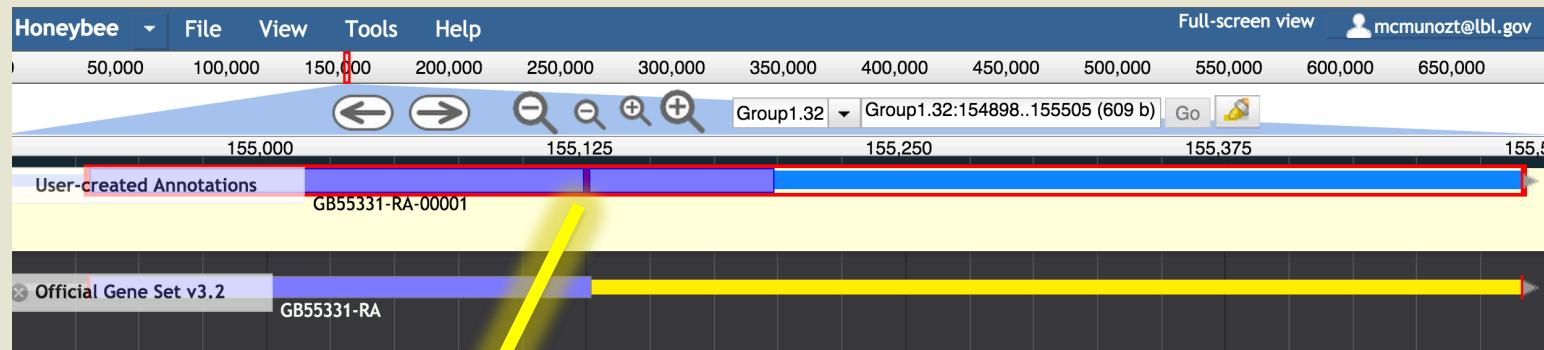
pollo

COMPLEX CASES

BERKELEY LAB

Phoenix Bioinformatics

CORRECTING SELENOCYSTEINE CONTAINING PROTEINS



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 a main window and a modal dialog.

Main Window:

- Top Left:** A dropdown menu labeled "Select mRNA" with options "spe1-RA" and "spe1-RA".
- Top Center:** A search bar containing the word "gene".
- Left Panel:** A table with columns "Name", "Symbol", "Description", "Created", and "Last modified".

Name	DNA mismatch repair protein Msh
Symbol	spe1
Description	DNA mismatch repair protein Msh
Created	2017-03-03
Last modified	2017-03-21
- Middle Left:** A "DBXRefs" section with "DB" and "Accession" fields, and "Add" and "Delete" buttons.
- Middle Right:** A "mRNA" section with fields for "Name", "Symbol", "Description", "Created", and "Last modified".

Name	DNA mismatch repair protein Msh
Symbol	
Description	
Created	2017-03-03
Last modified	2017-03-21
- Bottom Left:** A "Gene Ontology IDs" section listing GO IDs and their corresponding biological processes:
 - GO:0006301: mismatch repair [GO:0006298]
 - GO:0006298: nuclear-transcribed mRNA catabolic process, no-go decay [GO:0070966]
 - GO:0001591: dopamine neurotransmitter receptor activity, coupled via Gi/Go [GO:0001591]
 - GO:0000811: GINS complex [GO:0000811]
- Bottom Right:** A "Comments" section with a text input field containing "Extended 3' UTR using Forager RNAseq reads as -".

Modal Dialog:

A yellow arrow points from the "Enter new DB" field in the mRNA section to the "Gene Ontology IDs" section, indicating a relationship or a copy/paste action.

A yellow arrow points from the "Gene Ontology IDs" section in the main window to the "Comments" section, indicating a relationship or a copy/paste action.

Logos:

- Berkeley Lab:** UC Berkeley National Laboratory logo.
- Phoenix Bioinformatics:** Logo with a DNA helix icon.

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

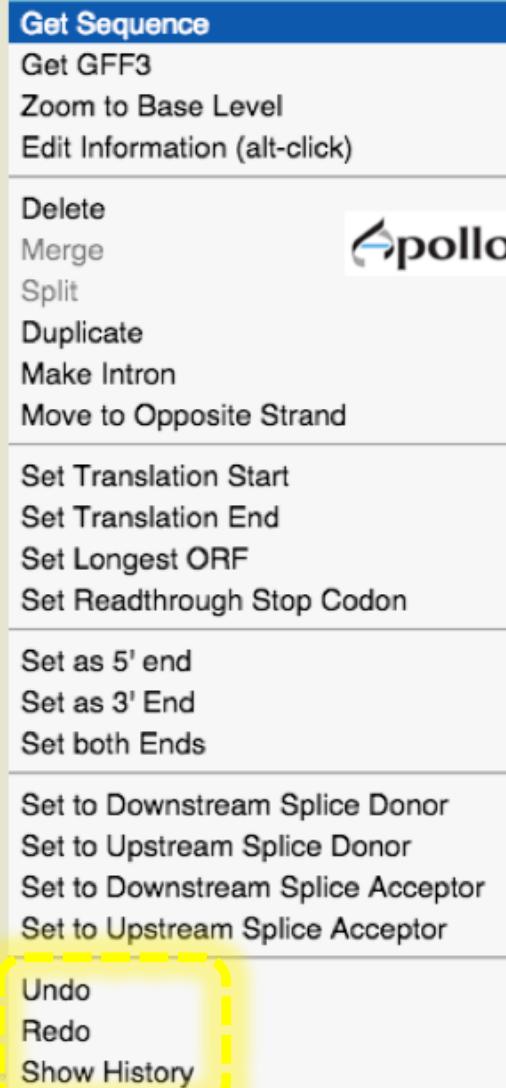
Add Delete



history



Keeping track of each edit



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

pollo

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



40



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



Files: bit.ly/wgsaa18-1

Server 1: bit.ly/wgsaa18-s1

Server 2: bit.ly/wgsaa18-s2



Begin at exercise						
Your number	Username	Password	Server	Organism	Exercise	
1	user.one@example.com	userone	1	Honey0	1	
2	user.two@example.com	usertwo	1	Honey0	7	
3	user.three@example.com	userthree	1	Honey1	1	
4	user.four@example.com	userfour	1	Honey1	7	
5	user.five@example.com	userfive	1	Honey2	1	
6	user.six@example.com	usersix	1	Honey2	7	
7	user.seven@example.com	userseven	1	Honey3	1	
8	user.eight@example.com	useight	1	Honey3	7	
9	user.nine@example.com	usernine	1	Honey4	1	
10	user.ten@example.com	userten	1	Honey4	7	
11	user.eleven@example.com	usereleven	1	Honey5	1	
12	user.twelve@example.com	usertwelve	1	Honey5	7	
13	user.thirteen@example.com	userthirteen	1	Honey6	1	
14	user.fourteen@example.com	userfourteen	1	Honey6	7	
15	user.fifteen@example.com	userfifteen	1	Honey7	1	
16	user.sixteen@example.com	usersixteen	1	Honey7	7	
17	user.seventeen@example.com	userseventeen	2	Honey0	1	
18	user.eIGHTEEN@example.com	useIGHTEEN	2	Honey0	7	
19	user.nineteen@example.com	usernineteen	2	Honey1	1	
20	user.twenty@example.com	usertwenty	2	Honey1	7	
21	user.twentyone@example.com	usertwentyone	2	Honey2	1	
22	user.twentytwo@example.com	usertwentytwo	2	Honey2	7	
23	user.twentythree@example.com	usertwentythree	2	Honey3	1	
24	user.twentyfour@example.com	usertwentyfour	2	Honey3	7	
25	user.twentyfive@example.com	usertwentyfive	2	Honey4	1	
26	user.twentysix@example.com	usertwentysix	2	Honey4	7	
27	user.twentyseven@example.com	usertwentyseven	2	Honey5	1	
28	user.twentyeight@example.com	usertwentyeight	2	Honey5	7	
29	user.twentynine@example.com	usertwentynine	2	Honey6	1	
30	user.thirty@example.com	userthirty	2	Honey6	7	
31	user.thirtyone@example.com	userthirtyone	2	Honey7	1	
32	user.thirtytwo@example.com	userthirtytwo	2	Honey7	7	

Users	Server	URL
1-16	1	http://bit.ly/wgsaa18-s1
17-32	2	http://bit.ly/wgsaa18-s2

Begin at exercise						
Your number	Username	Password	Server	Organism	Exercise	
1	user.one@example.com	userone	1	Honey0	1	
2	user.two@example.com	usertwo	1	Honey0	7	
3	user.three@example.com	userthree	1	Honey1	1	
4	user.four@example.com	userfour	1	Honey1	7	
5	user.five@example.com	userfive	1	Honey2	1	



Apollo on the Web

instructions

- Public Honey bee demo available at:**

genomearchitect.org/demo/

- Username:**

demo@demo.com

- Password:**

demo



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

```
GTTTAAGAGTGTTCGCGCCAATTGTTCGCGCGAGACTGGCCGTGCAGACCGAGCTGTTATAGCCCGTCT  
CCGCTCTGCTCTGCTGATCCATCGATCACCTACGCATCGATCCCTCGTTGATCAACGTGGTCTGAGC  
TGGAGCGTTGAGCGCCGCTATCAGACTGGCGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG  
CTTTAGAATCTTGCCTTGTGACGATATGGCTGGTCCAGCTTGCCTGCCCCGCCATCGCTTAC  
AGCATGGGGTGGGCAGAGCAGATGCTACAGGACCCGCCGCTGAAATTGTTTATGGGCTACCGAAGA  
TCGATCAAAGGATCAGGAATCCATCTCGAACATTCTCCCGCGCATTCATCATCGACGATGGCGAGGA  
GAGGTTCGTCTCGTCAAGCTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGGCAGAAT  
CTTGAAGGAGTTGGCAGCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA  
CACCCGGTGGATTCACTGTTGACATGTTGATATTACGACATTGGTTCTGTTCAAGAGACCTTCA  
TGCTATGGTCAAGGGAAATCAGAAGAGTATTCAACGTGCTACTATGCCATAGTTCCAGGCAGAAATT  
ATCACCCATGGAGAAGTTCATGGTGTGAACTTAATAGAAGCCCATCCG
```

Search all genomic
sequences



The screenshot shows the Honeybee Blat interface. At the top, there's a menu bar with "Honeybee", "File", "View", "Tools", and "Help". A yellow arrow points from the "Tools" menu down to a "Search sequence" dialog box. The main window displays genomic coordinates: 0, 200,000, 462,500, and 465,000. Below these coordinates is a section labeled "User-created Annotations". The "Search sequence" dialog box contains the sequence: GTTTAAGAGTGTTCGCGCCAATTGTTCGCGCGAGACTGGCCGTGCAGACCGAGCTGTTATAGCCCGTCT CCGCTCTGCTCTGCTGATCCATCGATCACCTACGCATCGATCCCTCGTTGATCAACGTGGTCTGAGC TGGAGCGTTGAGCGCCGCTATCAGACTGGCGCAGAGAAAAACTGAATGGAGGCACCGGCAGTTGGACG CTTTAGAATCTTGCCTTGTGACGATATGGCTGGTCCAGCTTGCCTGCCCCATCGCTTAC AGCATGGGGTGGGCAGAGCAGATGCTACAGGACCCGCCGCTGAAATTGTTTATGGGCTACCGAAGA TCGATCAAAGGATCAGGAATCCATCTCGAACATTCTCCCGCGCATTATCATCGACGATGGCGAGGA GAGGTTCGTCTCGTCAAGCTGGATAGCGCCATGATAGGAAACGGCGTTCGTCAAACGGTGGCAGAAT CTTGAAGGAGTTGGCAGCTGTACACAGAGAAAAATGTGATGATCAGTGCAACTCACTCGCACTCCA CACCCGGTGGATTCACTGTTGACATGTTGATATTACGACATTGGTTCTGTTCAAGAGACCTTCA TGCTATGGTCAAGGGAAATCAGAAGAGTATTCAACGTGCTACTATGCCATAGTTCCAGGCAGAAATT ATCACCCATGGAGAAGTTCATGGTGTGAACTTAATAGAAGCCCATCCG

Search all genomic sequences



Blat results

Enter sequence

```
GTAAAGAGTGTTCGCGCCAATTGTTCGCCCGAGACTGGCCGTGCAGACCAAGCTGTTAGCCGCT  
CCGCTCTGCTCTGCTGATCATTACCGCATGATCCTCGTGCAGACCTGGTCATGAGC  
TGAGCGTTGAGCCCGCTATCAGACGGCGAGAGAAAAACTGATGGAGGCCACCGGCACTGGACG  
CTTAGAGATCCCTGGCTGTTGACGATGGCTGGTCAGCTGGGTGCCGGCCATCGCGCTTAC  
AGCATCGGGGGGGGGAGACAGATCCCTTCGACACAGGGCCCGCTGAAATTGTTTATGGCTACCGAAGA  
TCGATCAAAAAGGATCAGGAATCCCTTCGACACAGGGCCCGCTGAAATTGTTTATGGCTACCGAAGA  
GAGGTTCTGCTCTGTCAGCGTGTAGCGCCATGATAGGAAACGGCGTTCGTCACACAGGGTTCGAGAAT  
CTTGAAGAAAGGTTGGACGCTGACACAGAGAAAATGATGATGATCAGTGCACACTCTCGCACTCCA  
CACCCGGGGATTCATGTTGAGTGTGATATTACGACATTCCGGTTTCGTCAGAGAACCTTCGA  
TGCTATGTCAGGAAATCAGAAGAGTATTCAACGTGCTCACTATGCCATAGTTCCAGGCAGAAATTCA  
ATCACCCATGGAGAAAGTTTGTTGTAACATTAATAGAAGGCCATCG
```

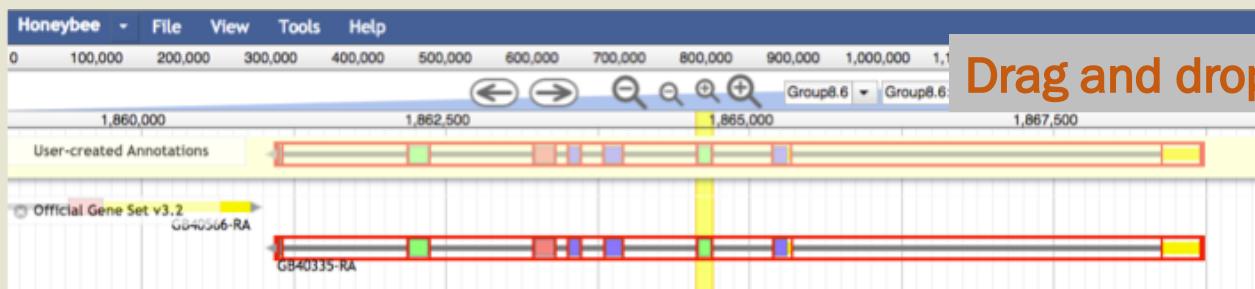
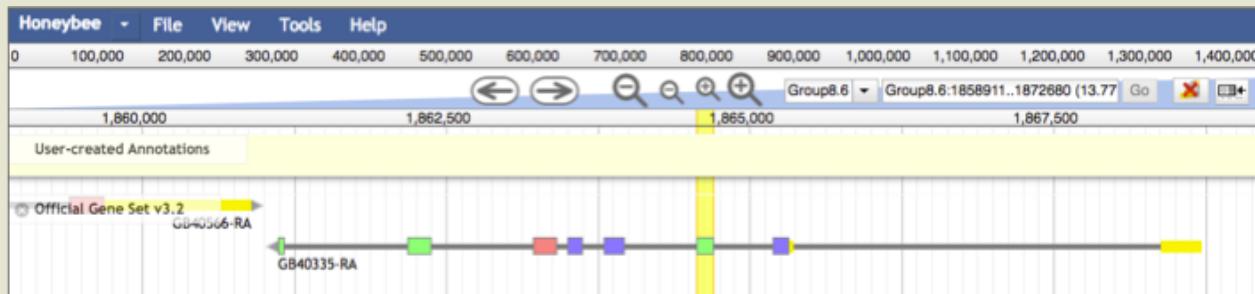
Search

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

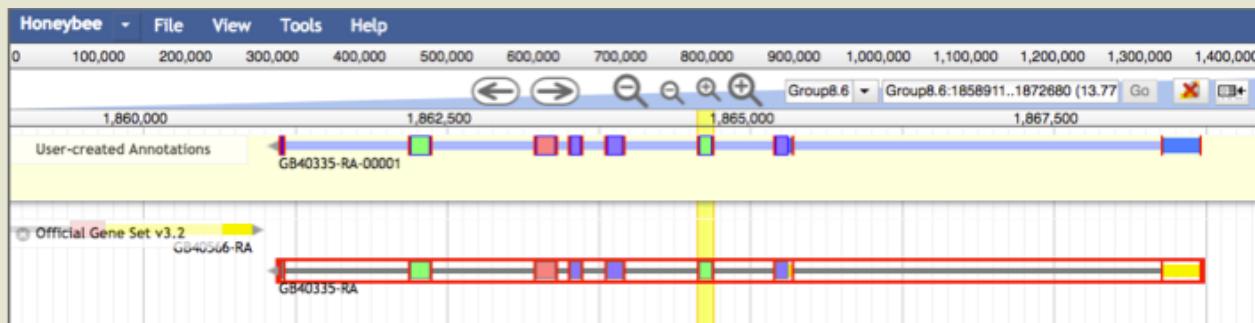
Click on a high-scoring segment pair (hsp) to navigate and highlight the region.



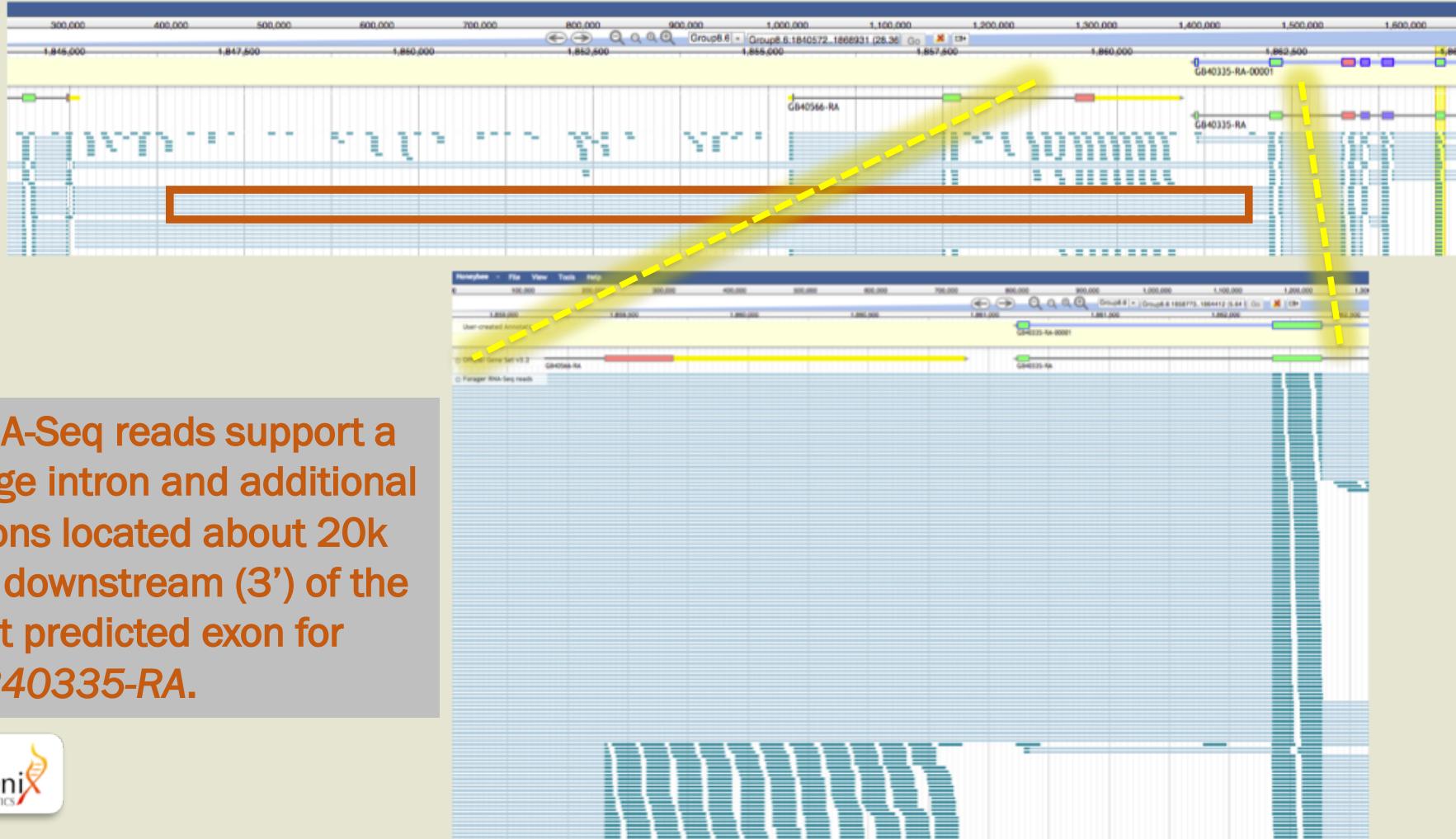
Create a new annotation



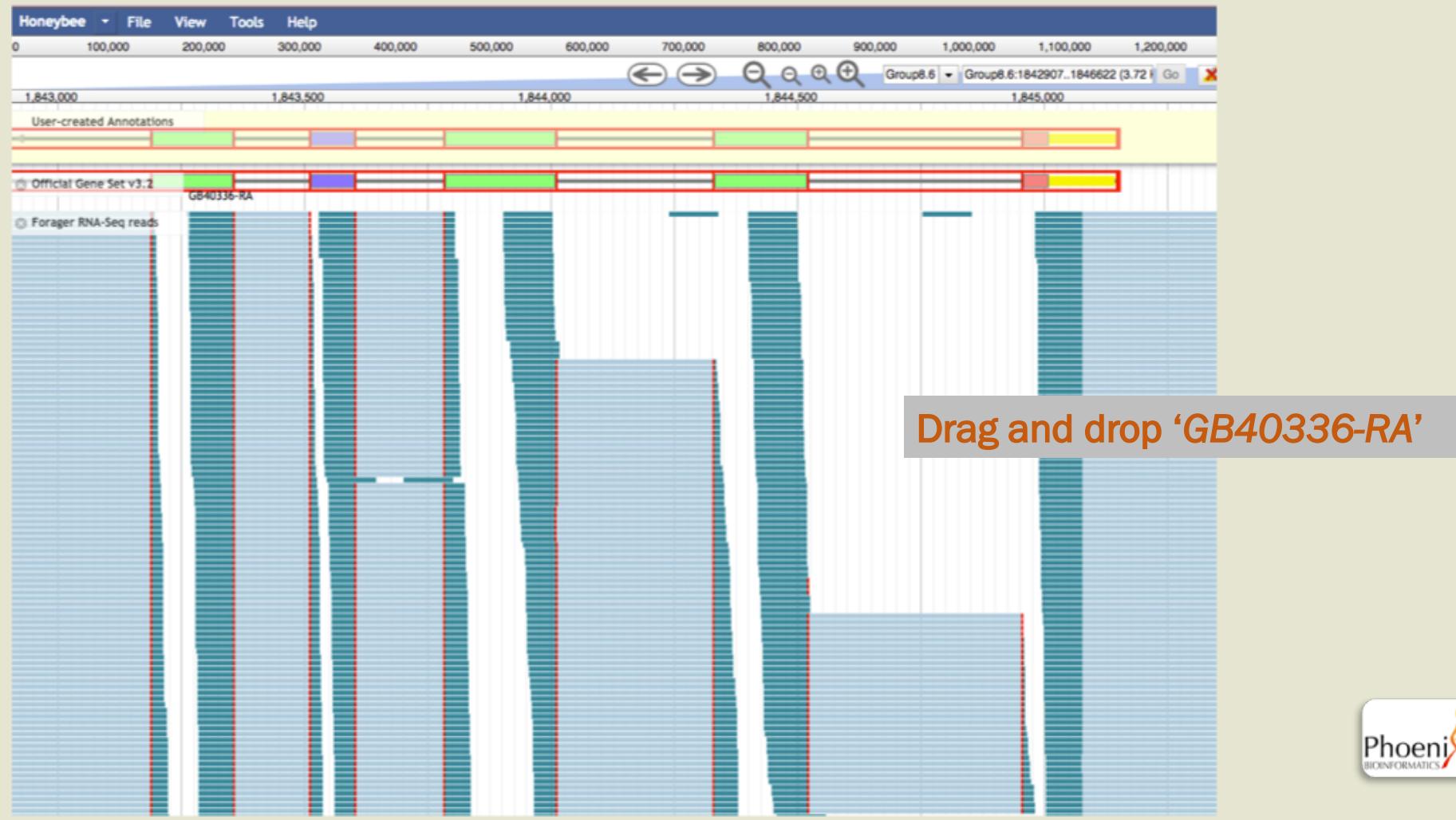
Drag and drop 'GB40335-RA'



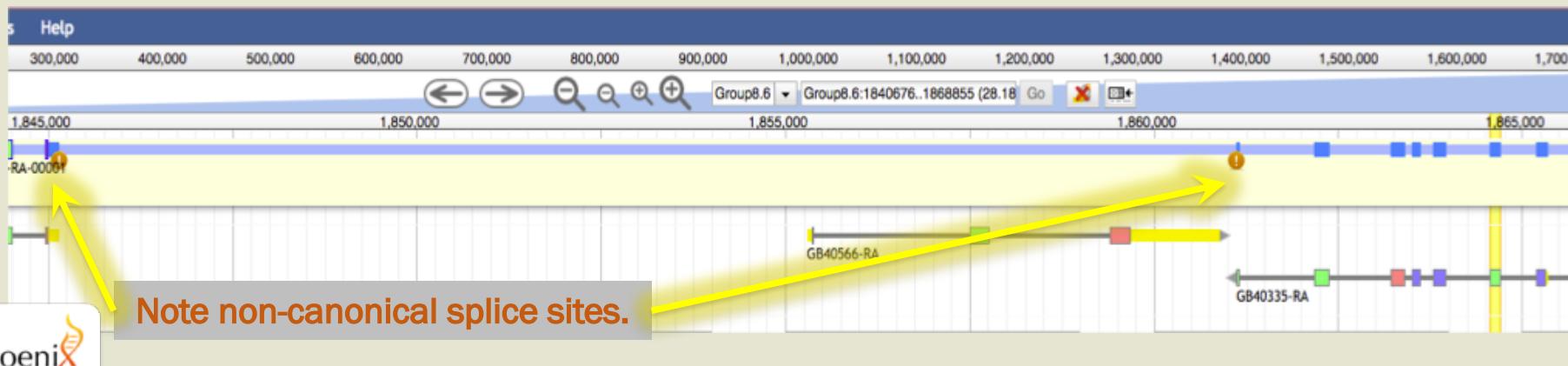
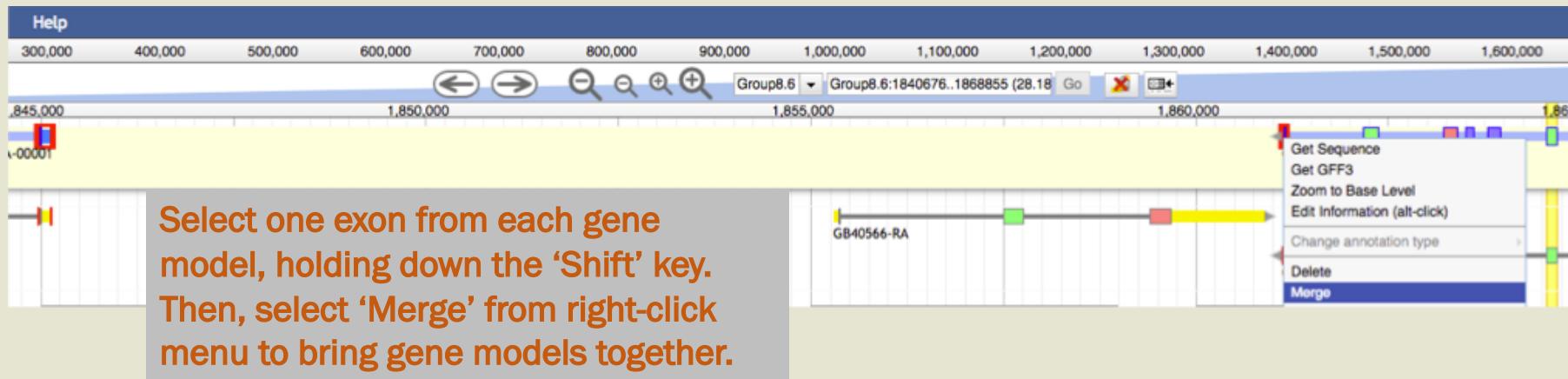
Transcriptomic data support a longer gene



Transcriptomic data support a longer gene



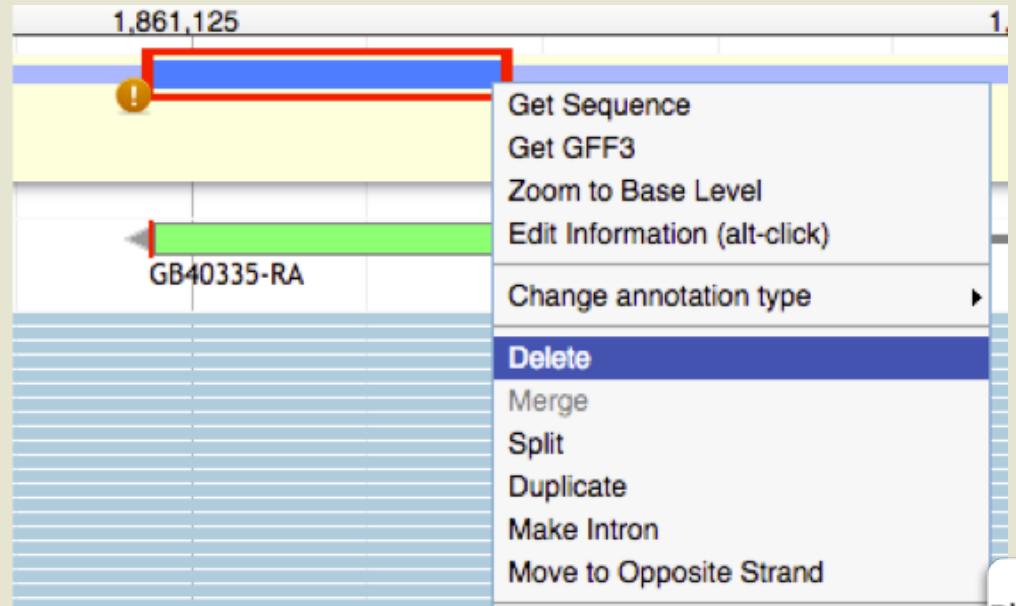
Merge transcripts



Exon not supported by RNA-Seq data

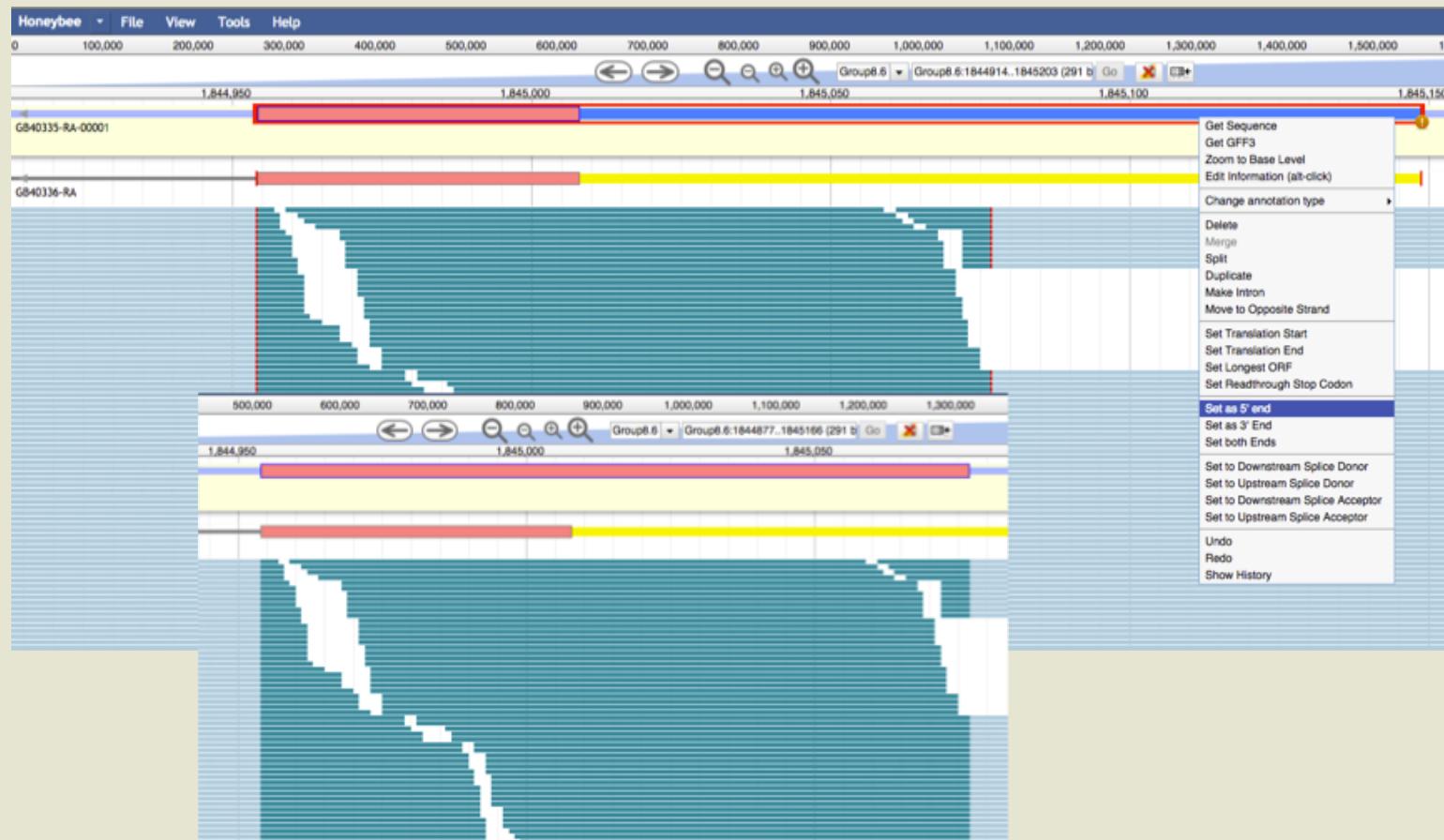


At the end of GB40335-RA, select last exon and right-click to choose the 'Delete' option.



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 a peptide sequence and comparing it against a public database.

Top Panel: A genome browser interface shows 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" context menu (which includes options like "Get GFF3", "Zoom to Base Level", "Edit Information (alt-click)", "Change annotation type", and "Delete") to the sequence alignment area below.

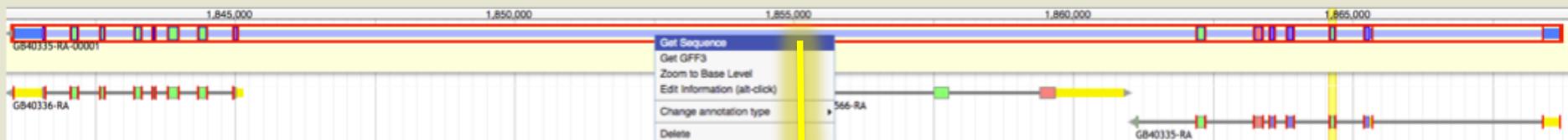
Middle Panel: A "Sequence" window displays the retrieved peptide sequence: >fccb9943-c0dc-4bbc-b43b-74f6dfe33dfe (sequence:mRNA) 715 / residues [Group8.6:1841036-1868721 - strand] [peptide]. The sequence itself is too long to be fully shown here.

Bottom Panel: A BLAST search interface is shown. The query sequence is identical to the one in the middle panel. The search parameters include:

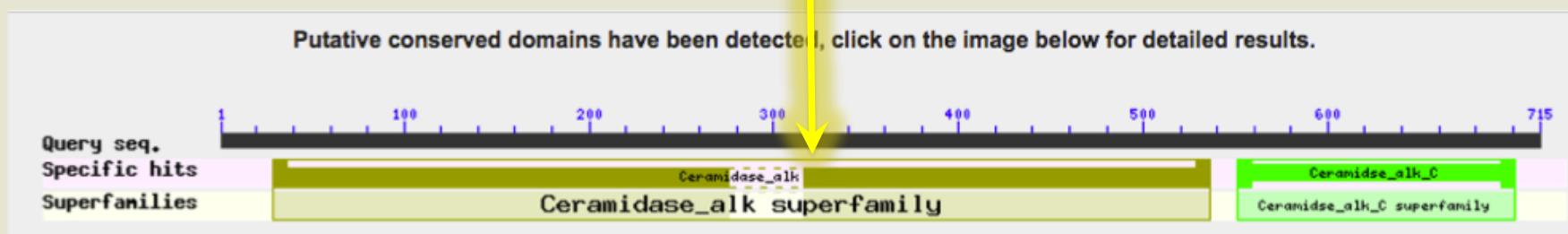
- Job Title:** ceramidase-Amel
- Database:** Non-redundant protein sequences (nr)
- Algorithm:** blast (protein-protein BLAST)
- Program Selection:** Standard Protein BLAST

Logos: The Berkeley Lab logo is in the bottom left corner, and the Phoenix Bioinformatics logo is in the bottom right corner.

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 overfamily

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 i
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

BERKELEY LAB

Phoenix BIOINFORMATICS

Add metadata in ‘Information Editor’

The screenshot shows the 'Information Editor' interface for a gene entry. The main window displays a 'gene' section with details like Name (Neutral ceramidase), Symbol (C0ase), and Description (Enzyme, cleaves fatty acids from...). Below it is an 'mRNA' section with similar fields. A yellow arrow points from the 'Comments' field in the main window to a detailed view of the 'Comments' section at the bottom.

Comments

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

PubMed Identifiers

icebox.lbl.gov says:
Publication title: 'Insights into social insects from the genome of the honeybee *Apis mellifera*'

NCBI Gene LOC409628
BeeBase GB40336

Gene Ontology terms

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

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

Gene Ontology IDs

Comments

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



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

JBrowse. Ian Holmes' Lab
University of California, Berkeley

Christine Elsik's Lab,
University of Missouri



Deepak Unni



Thank You.

Collaborators

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

[Chris Mungall](#)

[Suzanna Lewis](#)

Seth Carbon (Noctua / AmiGO)

Nathan Dunn (Apollo)

- 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
- i5k Community

berkeleybop.org



Berkeley
UNIVERSITY OF CALIFORNIA

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



Useful resources

<https://github.com/GMOD/Apollo>

<http://genomearchitect.github.io/users-guide/>

