



# An Introduction to Web Apollo

*Manual Annotation Workshop at Kansas State University*

Monica Munoz-Torres, PhD | @monimunozto

Berkeley Bioinformatics Open-Source Projects (BBOP)  
Genomics Division, Lawrence Berkeley National Laboratory

IX Arthropod Genomics Symposium. Manhattan, KS. 17 June, 2015

# TODAY

---

Recommended Browsers: Google Chrome, Firefox.

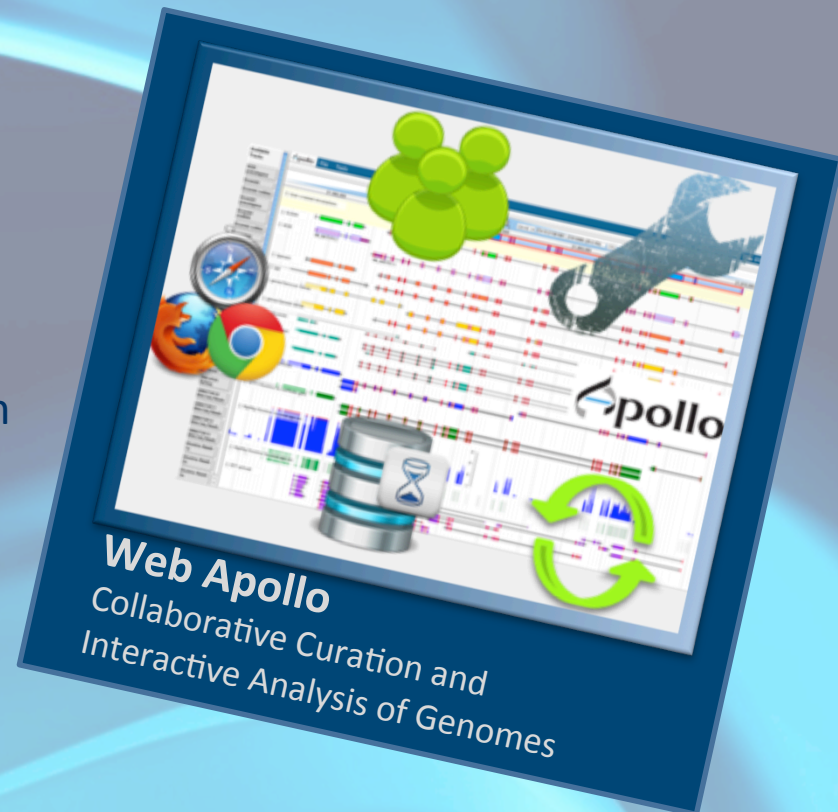
Exercises file available at Basecamp

Workshop slides and answers to exercises will be available on Basecamp next week.

# OUTLINE

---

- **GENOME CURATION**  
steps involved
- **COMMUNITY BASED CURATION**  
our experience
- **APOLLO**  
empowering collaborative curation
- **APOLLO on THE WEB**  
becoming acquainted
- **PRACTICE**  
demonstration and exercises



# DURING THIS WORKSHOP

you will

---

- ❖ Understand the process of genome curation in the context of annotation:  
assembled genome → automated annotation → manual annotation
- ❖ Become familiar with the environment and functionality of the Web Apollo genome annotation editing tool.
- ❖ Learn to identify homologs of known genes of interest in a newly sequenced genome of interest.
- ❖ Learn how to corroborate and modify automatically generated gene models using available biological evidence (in Apollo).

# I INVITE YOU TO:

---

- ❖ Observe details in figures
- ❖ Listen to explanations
- ❖ Ask questions at any time
- ❖ Use Twitter & share your thoughts: I am @monimunocto  
A few tags & users: #WebApollo #annotation #biocuration #GMOD  
#genome @JBrowseGossip
- ❖ Take brakes:  
LBL's ergo team suggests I should not work at the computer for >45 minutes without a break; neither should you! We will be here for ~2.5 hours: please get up and stretch your neck, arms, and legs as often as you need.

## I kindly ask that you refrain from:

---

- ❖ Reading all the text I wrote.

Think of the text on these slides as your “class notes”. You will use them during exercises.

- ❖ Checking email.

You have my undivided attention, I’d like to receive yours in exchange.

**Warning:** If you get \*caught\*, you will read it out loudly for everyone to hear, we may contribute to the response.

Let Us Get Started

# REMEMBER, REMEMBER...

from intro webinar last week

---

- CENTRAL DOGMA  
in molecular biology
- WHAT IS A GENE?  
let's think computationally
- TRANSCRIPTION  
mRNA in detail
- TRANSLATION  
and many definitions
- GENOME CURATION  
steps involved
- WHAT TO LOOK FOR  
training the annotators

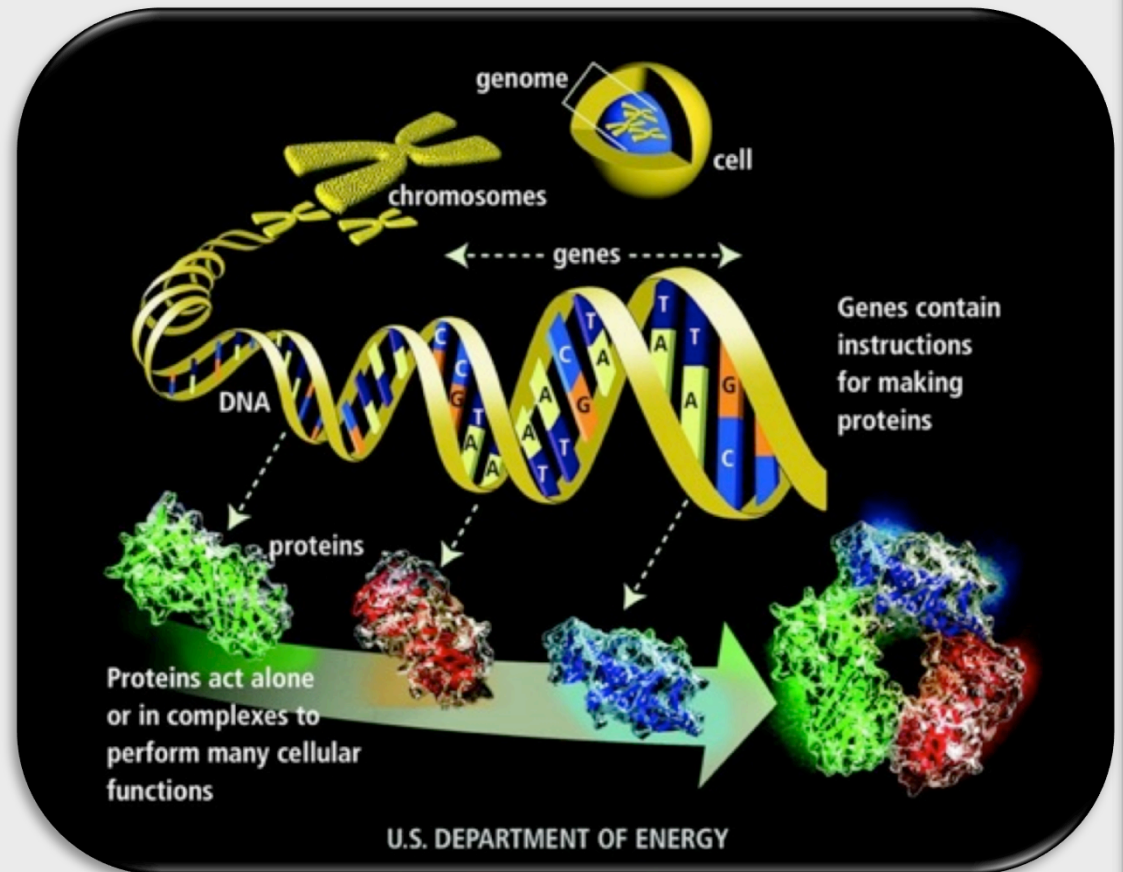




# CURATING GENOMES

steps involved

- 1 Generation of Gene Models**  
calling ORFs, one or more rounds of gene prediction, etc.
- 2 Annotation of gene models**  
Describing function, expression patterns, metabolic network memberships.
- 3 ★ Manual annotation ★**



# GENE PREDICTION

---

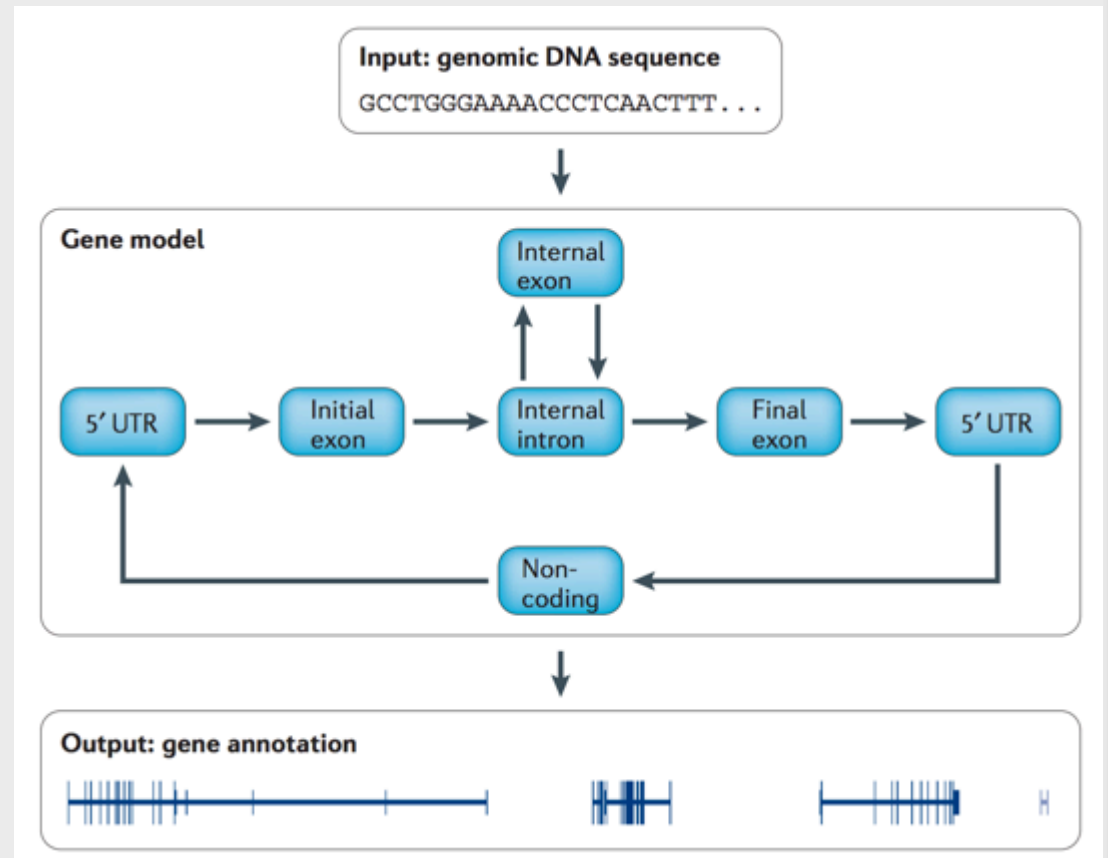
- ❖ The identification of structural features of the genome.
  - Primarily protein-coding genes.
  - Also transfer RNAs (tRNA), ribosomal RNAs (rRNA), regulatory motifs, long and small non-coding RNAs (ncRNA), repetitive elements (masked), etc.

# GENE PREDICTION

❖ Methods for discovery:

**1) *Ab initio*:** based on DNA composition, deals strictly with genomic sequences and makes use of statistical approaches to search for coding regions and typical gene signals.

- E.g. Augustus, GENSCAN, geneid, fgenesh, etc.

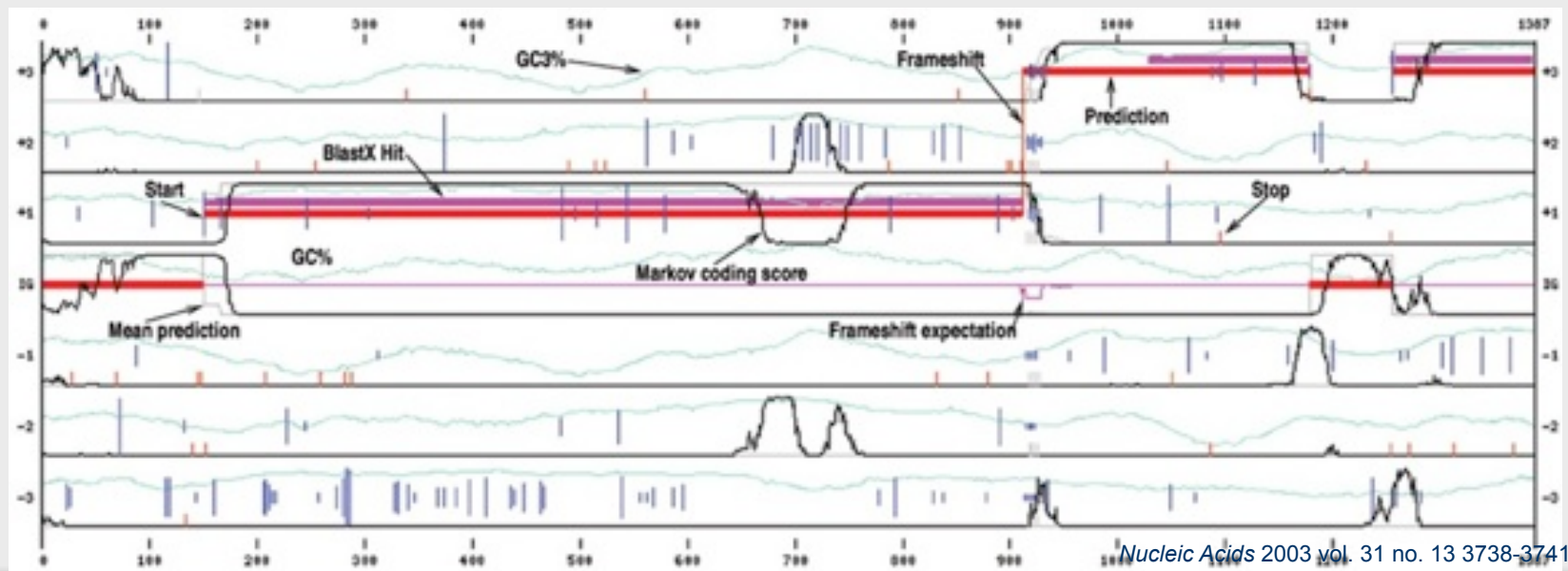


# GENE PREDICTION

❖ Methods for discovery:

**2) Homology-based:** evidence-based; finds genes using either similarity searches in the main databases or experimental data including RNAseq, expressed sequence tags (ESTs), full-length complementary DNAs (cDNAs), etc.

- E.g: SGP2, fgenesh++



# GENE ANNOTATION

---

Integration of data from prediction tools to generate a **reliable set of structural annotations**: involves *ab initio* predictions, assessment of biological evidence to drive the gene prediction process, and the synthesis of these results to produce a set of consensus gene models.

- ❖ Models may be organized using:
  - ❖ automatic integration of predicted sets; e.g: GLEAN
  - ❖ packaged tools from pipeline; e.g: MAKER

In some cases algorithms and metrics used to generate consensus sets may actually reduce the accuracy of the gene's representation; in such cases it is usually better to use an *ab initio* model to create a new annotation.

# NOT PERFECT

automated annotation remains an imperfect art

Unlike the more highly polished genomes of earlier projects, today's genomes have:

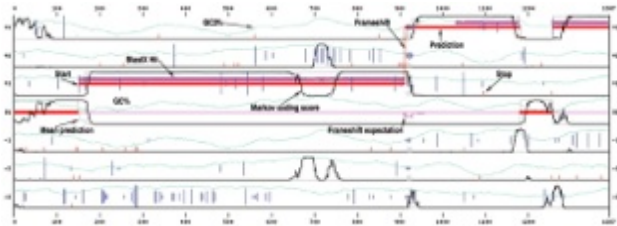
1. lower coverage.
2. more frequent assembly errors and annotation of genes across multiple scaffolds.

Image: [www.BroadInstitute.org](http://www.BroadInstitute.org)

# MANUAL ANNOTATION

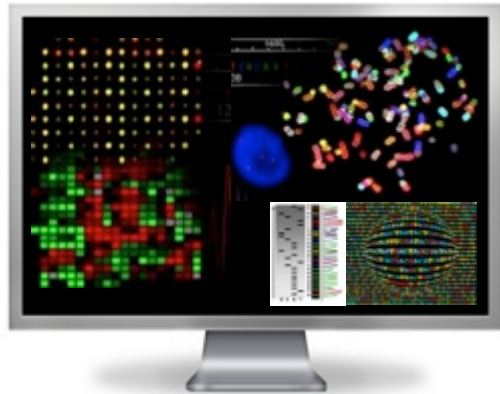
working concept

---



Schiex et al. *Nucleic Acids* 2003 (31) 13: 3738-3741

## Automated Predictions



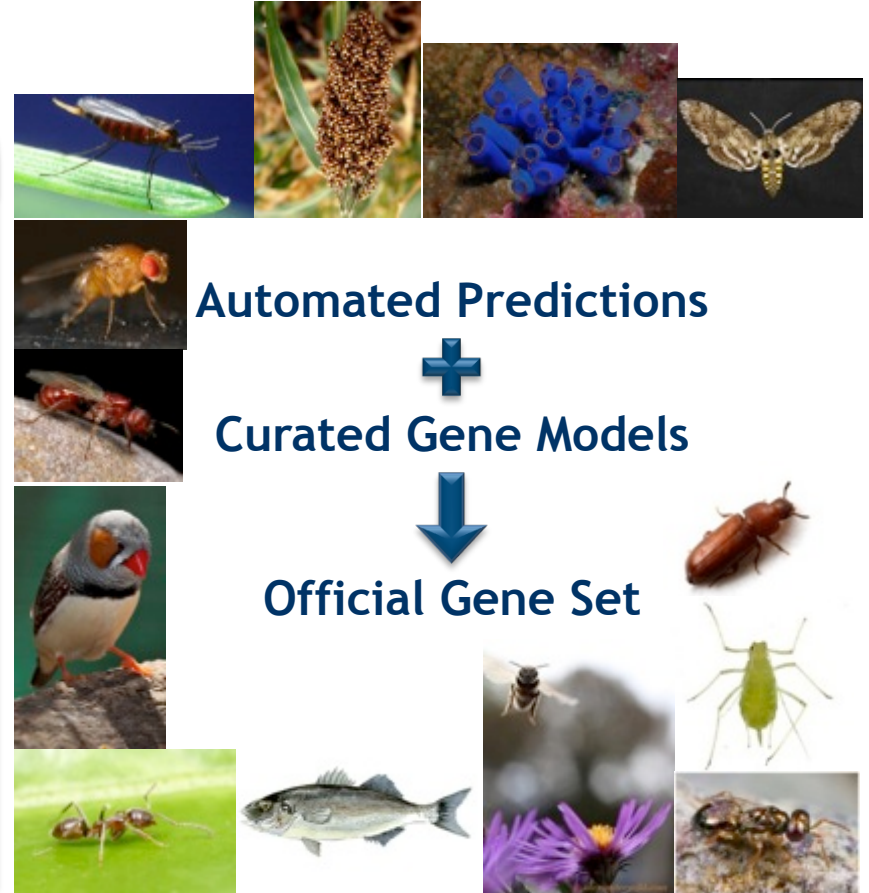
## Experimental Evidence

cDNAs, HMM domain searches, RNAseq,  
genes from other species.

- ❖ Precise elucidation of biological features encoded in the genome requires careful examination and review.

# MANUAL ANNOTATION is necessary

- ❖ Evaluate all available evidence and corroborate or modify genome element predictions.
- ❖ Determine functional roles through comparative analysis using literature, databases, and experimental data.
- ❖ Resolve discrepancies and validate automated gene model hypotheses.
- ❖ **Desktop version of Apollo** was designed to fit the manual annotation needs of genome projects such as fruit fly, mouse, zebrafish, human, etc.

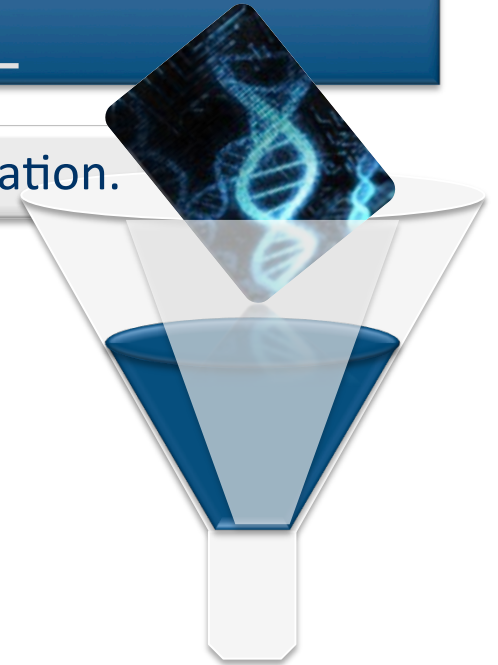


*"Incorrect and incomplete genome annotations will poison every experiment that uses them".*  
- M. Yandell



# BUT, MANUAL CURATION did not always scale well

Too many sequences and not enough hands to approach curation.



1

## Museum Model

A small group of highly trained experts; e.g. GO

Elsik et al. 2006. *Genome Res.* 16(11):1329-33.

2

## Old-time Jamborees

A few very good biologists and a few very good bioinformaticians camp together, during intense but short periods of time.

3

## Cottage Model

Researchers work by themselves, then may or may not publicize results; ... may be a dead-end with very few people ever aware of these results.

# POWER TO THE CURATORS

augment existing tools

---

## Give more people the power to curate!

Fill in the gap for all the things that won't be easy to cover with these approaches; this will allow researchers to better contribute their efforts.

Big data are not a substitute for, but a supplement to traditional data collection and analysis.

The Parable of Google Flu. Lazer et al. 2014. *Science* 343 (6176): 1203-1205.



- ❖ Enable more curators to work
- ❖ Enable better scientific publishing
- ❖ Credit curators for their work

# IMPROVING TOOLS FOR MANUAL ANNOTATION

our plan

---

“More and more sequences”: more genomes, within populations and across species, are now being sequenced.

This begs the need for a universally accessible genome curation tool:

To produce accurate sets of genomic features.

To address the need to correct for more frequent assembly and automated prediction errors due to new sequencing technologies.

# GENOME ANNOTATION

an inherently collaborative task

---

Researchers often turn to colleagues for second opinions and insight from those with expertise in particular areas (e.g., domains, families). To facilitate and encourage this, we continue to improve Apollo.

Apollo is a web-based, collaborative genomic annotation editing platform.

*We need annotation editing tools to modify and refine the precise location and structure of the genome elements that predictive algorithms cannot yet resolve automatically.*

<http://GenomeArchitect.org>



# APOLLO

genome annotation editing tool

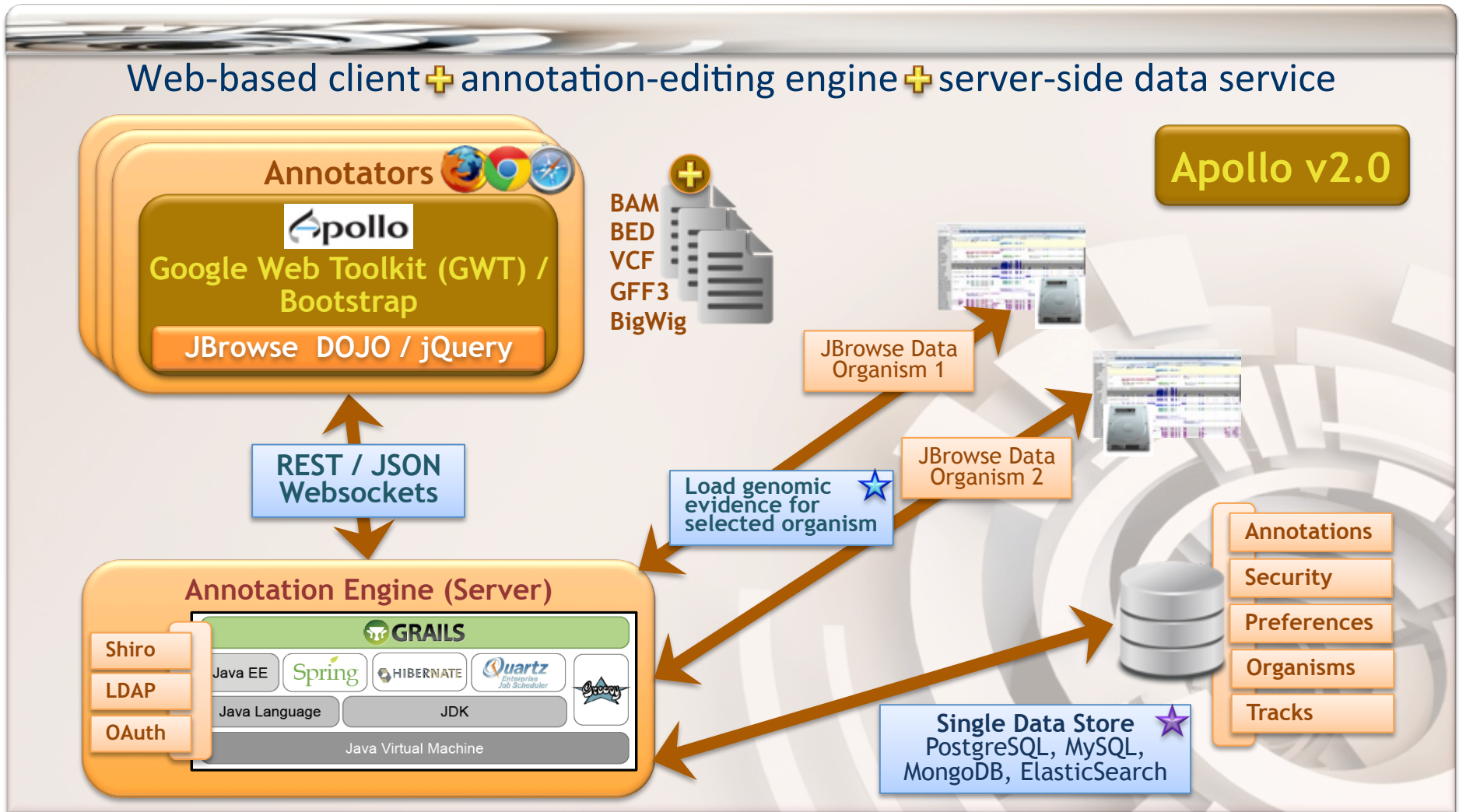


- ❖ Web based, integrated with JBrowse.
- ❖ Supports real time collaboration!
- ❖ Automatic generation of ready-made computable data.
- ❖ Supports annotation of genes, pseudogenes, tRNAs, snRNAs, snoRNAs, ncRNAs, miRNAs, TEs, and repeats.
- ❖ Intuitive annotation, gestures, and pull-down menus to create and edit transcripts and exons structures, insert comments (CV, freeform text), GO terms, etc.

# NEW APOLLO ARCHITECTURE

simpler, more flexible

Web-based client + annotation-editing engine + server-side data service



# DISPERSED COMMUNITIES

## collaborative manual annotation efforts

---

We continuously train and support hundreds of geographically dispersed scientists from many research communities to conduct manual annotations, recovering coding sequences in agreement with all available biological evidence using Web Apollo.

- ❖ Gate keeping and monitoring.
- ❖ Tutorials, training workshops, and “genebores”.
- ❖ Personalized user support.



# CURATION

## how it works

1

Identifies elements that best represent the underlying biology (including missing genes) and eliminates elements that reflect systemic errors of automated analyses.

2

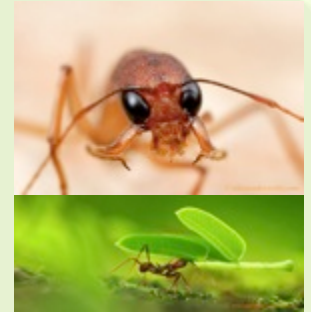
Assigns function through comparative analysis of similar genome elements from closely related species using literature, databases, and researchers' lab data.

APOLLO

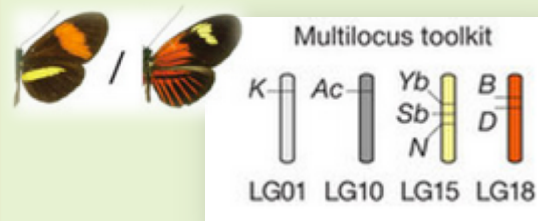
### Examples

Comparing 7 ant genomes contributed to better understanding evolution and organization of insect societies at the molecular level; e.g. division of labor, mutualism, chemical communication, etc.

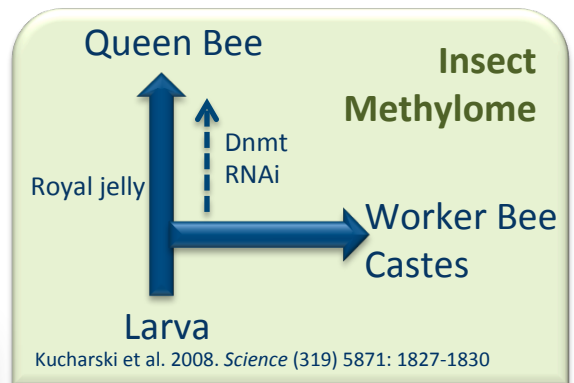
Libbrecht et al. 2012. *Genome Biology* 2013, 14:212



Anchoring molecular markers to reference genome pointed to chromosomal rearrangements & detecting signals of adaptive radiation in *Heliconius* butterflies.



Joron et al. 2011. *Nature*, 477:203-206





# CURRENT COLLABORATIONS

training and contributions

## Partnerships



UNIVERSITY



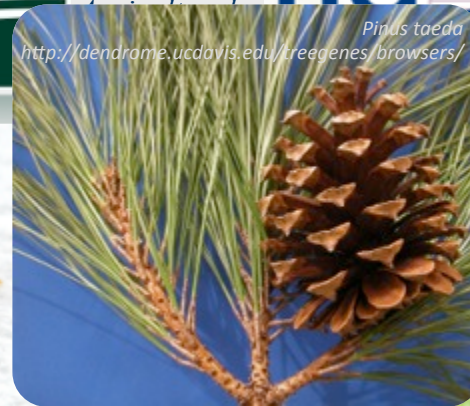
*Phlebotomus papatasi*



National



*Wasmania auropunctata*



*Pinus taeda*

<http://dendrome.ucdavis.edu/treegenes/browsers/>



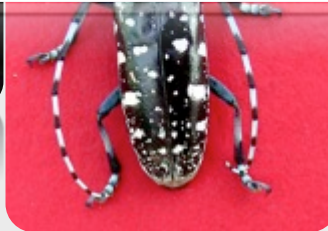
Nature Reviews Genetic



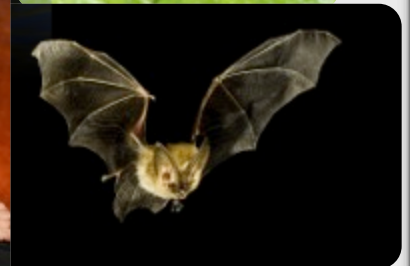
Norwegian Spruce <http://congenie.org/>



*Tallapoosa darter* <http://darter2.westga.edu/>



*Homo sapiens* hg19



# LESSONS LEARNED

## What we have learned:

- Collaborative work distills invaluable knowledge
- We must enforce strict rules and formats
- We must evolve with the data
- A little training goes a long way
- NGS poses additional challenges



# THE COLLABORATIVE CURATION PROCESS AT I5K

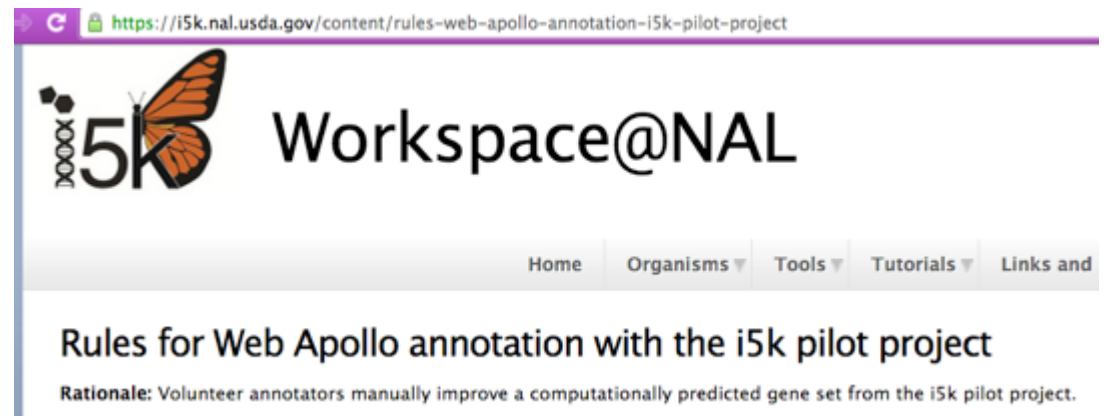
---

- 1) In some cases a computationally predicted consensus gene set is generated using multiple lines of evidence. In other cases, more than one gene set are made available for analysis: e.g. Primary Gene Sets: HAZT\_v0.5.3-Models, Augustus gene set.
- 2) i5K Projects will integrate consensus computational predictions with manual annotations to produce an updated Official Gene Set (OGS):
  - » If it's not on either track, it won't make the OGS!
  - » If it's there and it shouldn't, it will still make the OGS!

# CONSENSUS SET: REFERENCE AND START POINT

---

- Isoforms: drag original and alternatively spliced form to *'User-created Annotations'* area.
- If an annotation needs to be removed from the consensus set, drag it to the *'User-created Annotations'* area and label as *'Delete'* on Information Editor.
- Overlapping interests? **Collaborate** to reach agreement.
- Follow guidelines for i5K Pilot Species Projects as shown at <http://goo.gl/LRu1VY>



Apollo

# WEB APOLLO

## the sequence selection window

Sort

Show  10 entries  
25  
50  
100

Filter: scaffold527

Length

Organism	Name	Length
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold527</a>	1635968
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5270</a>	31767
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5271</a>	15707
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5272</a>	19345
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5273</a>	25468
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5274</a>	16904
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5275</a>	15324
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5276</a>	18237
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5277</a>	12610
<input type="checkbox"/> Leptinotarsa decemlineata	<a href="#">Scaffold5278</a>	13516

Showing 1 to 10 of 11 entries (filtered from 24,393 total entries)

◀ Previous Next ▶



# WEB APOLLO

## graphical user interface (GUI) for editing annotations

'File':  
Upload your own  
evidence: GFF3, BAM,  
BigWig, VCF\*. Add  
combination and  
sequence search  
tracks.

'View': change  
color by CDS,  
toggle strands,  
set highlight.

'Tools':  
Use BLAT to query the  
genome with a protein  
or DNA sequence.

Navigation tools:  
pan and zoom

Search box: go to  
a scaffold or a  
gene model.

Grey bar of coordinates  
indicates location. You can also  
select here in order to zoom to  
a sub-region.

The screenshot shows the Web Apollo interface with several key components:

- Navigation Bar:** Includes the Apollo logo, a menu (File, View, Tools, Help), a search box containing 'Chr10 Chr10:22209226..22213685 (6.46 Kb)', and a 'Go' button.
- Coordinate Bar:** A grey bar at the top showing genomic coordinates from 20,560,000 to 22,760,000.
- Available Tracks:** A sidebar on the left lists track types such as 'NCBI pseudogene', 'Ensembl miRNA', 'Ensembl pseudogene', 'Ensembl snoRNA', 'Ensembl snRNA', 'Fgenesh', 'fgeneshpp', 'GenelD', 'SGP', and 'cDNA spliced'.
- 'User-created Annotations' Track:** A yellow track containing blue annotations for 'MGC139957' and 'APEX1'.
- 'Evidence Tracks Area':** A track below showing purple annotations for 'MGC139957' and 'APEX1'.
- Navigation Tools:** A set of icons (back, forward, search, zoom in, zoom out) for navigating the tracks.
- User Profile:** A 'demo' user profile icon in the top right corner.

Available Tracks

Login

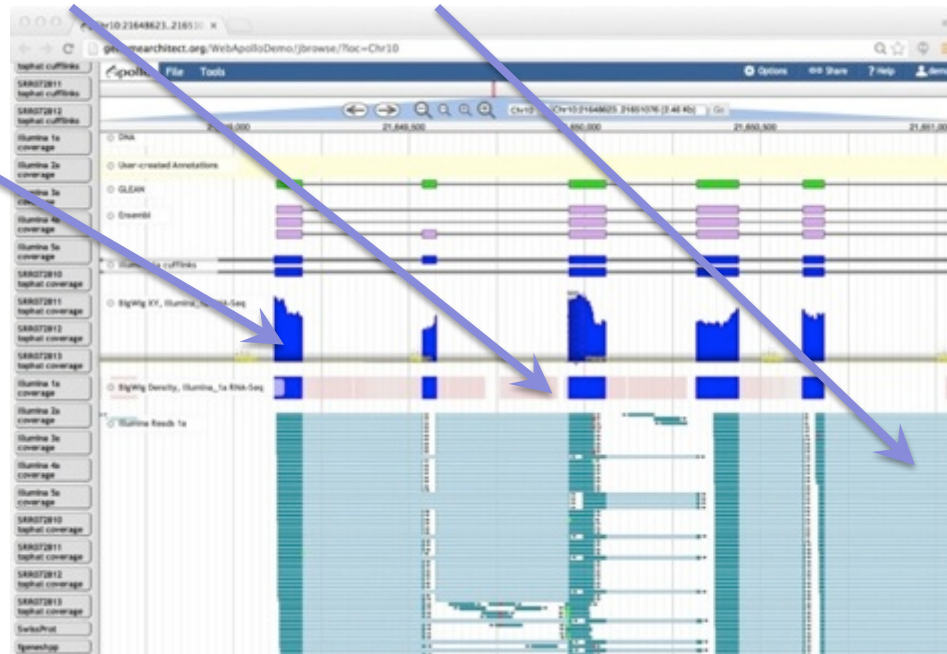


# WEB APOLLO

## additional functionality

In addition to protein-coding gene annotation that you know and love.

- Non-coding genes: ncRNAs, miRNAs, repeat regions, and TEs
- Sequence alterations (less coverage = more fragmentation)
- Visualization of stage and cell-type specific transcription data as coverage plots, heat maps, and alignments





# GENERAL PROCESS OF CURATION

## steps to remember

---

1. Select a chromosomal region of interest, e.g. scaffold.
2. Select appropriate **evidence tracks**.
3. Determine whether a feature in an existing evidence track will provide a reasonable gene model to start working.
  - If yes: select and **drag** the feature to the 'User-created Annotations' area, **creating an initial gene model**. If necessary use editing functions to adjust the gene model.
  - If not: *let's talk*.
4. Check your edited gene model for integrity and accuracy by comparing it with available homologs.

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

# USER NAVIGATION

The screenshot displays the Apollo genome browser interface. On the left, the 'Available Tracks' panel lists various evidence tracks, including 'Abdomen 454 Contigs', 'Acep\_OGSv1.2', 'Aech\_OGSv3.8', 'Aplis cerana reads', 'Augustus Set 12', 'Augustus Set 9', 'Brain and Ovary 454 contigs', 'Cflo\_OGSv3.3', 'Dmel\_r5.42', 'Embryo 454 contigs', 'Fgenesh', 'Fgenesh++ with RNASeq training data', 'Fgenesh++ without RNASeq training data', 'Forager Bee Brain Illumina Contigs', 'Forager RNA-Seq HeatMap', 'Forager RNA-Seq XY Plot', 'Forager RNA-Seq reads', 'GeneID', and 'Heat'. The main view shows a genomic track with coordinates from 0 to 300,000. A 'User-created Annotations' track is highlighted in yellow. Below it, the 'Official Gene Set v3.2' track shows gene models for GB40018-RA, GB40019-RA, GB40020-RA, GB40031-RA, GB40032-RA, and GB40017-RA. A right-click menu is open over an mRNA annotation, showing options: 'View details', 'Highlight this mRNA', and 'Create new annotation'. The 'Create new annotation' menu is expanded to show a list of biological features: gene, pseudogene, tRNA, snRNA, snoRNA, ncRNA, rRNA, miRNA, repeat\_region, and transposable\_element.

Choose (click or drag) appropriate evidence tracks from the list on the left.

Click on an exon to select it. Double click on an exon or single click on an intron to select the entire gene.

Select & drag any elements from an evidence track into the curation area: these are editable and considered the curated version of the gene. Other options for elements in evidence tracks available from right-click menu.

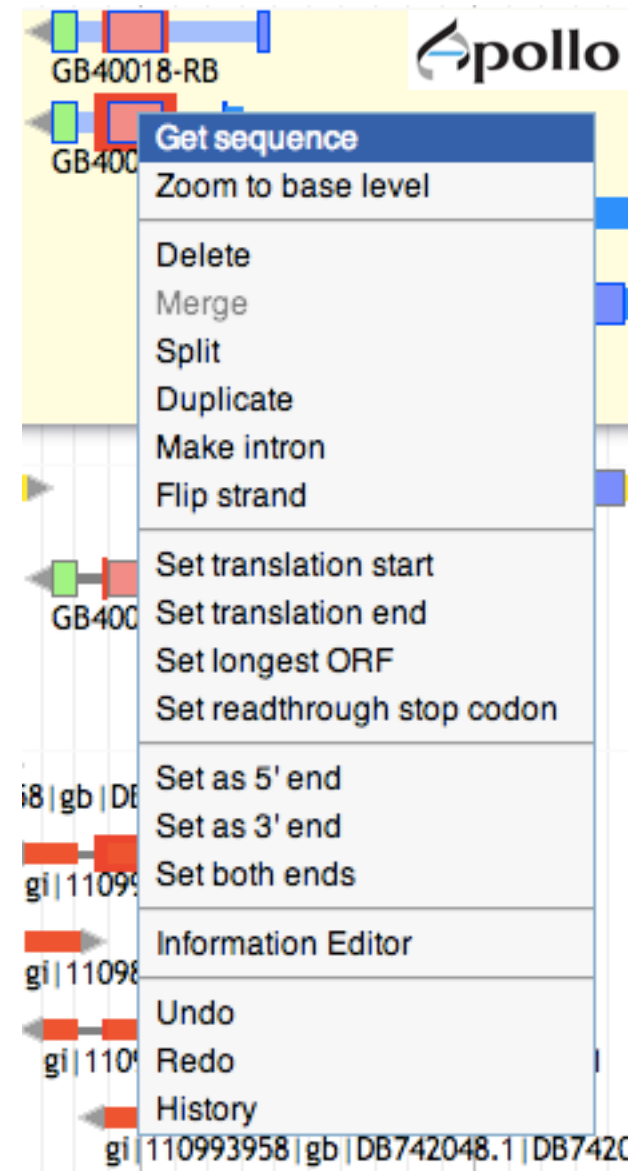
If you select an exon or a gene, then every track is automatically searched for exons with exactly the same co-ordinates as what you selected. Matching edges are highlighted red.

Hovering over an annotation in progress brings up an information pop-up.

# USER NAVIGATION

## Right-click menu:

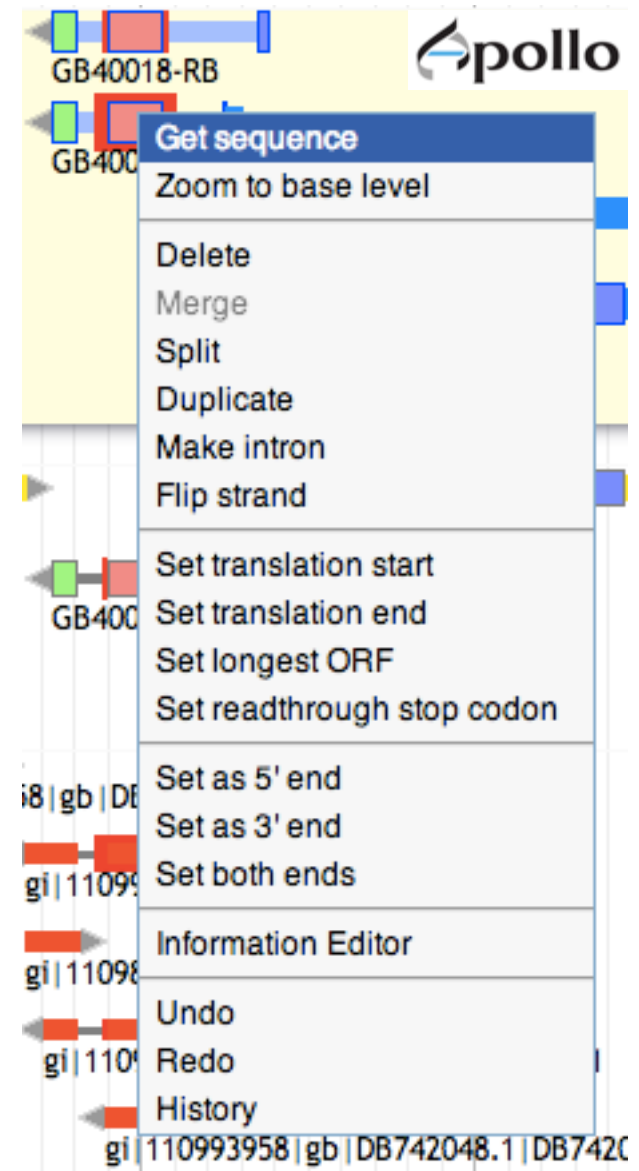
- With the exception of deleting a model, all edits can be reversed with 'Undo' option. 'Redo' also available. All changes are immediately saved and available to all users in real time.
- 'Get sequence' retrieves peptide, cDNA, CDS, and genomic sequences.
- You can select an exon and select 'Delete'. You can create an intron, flip the direction, change the start or split the gene.



# USER NAVIGATION

Right-click menu:

- If you select two gene models, you can join them using 'Merge', and you may also 'Split' a model.
- You can select 'Duplicate', for example to annotate isoforms.
- Set translation start, annotate selenocysteine-containing proteins, match edges of annotation to those of evidence tracks.



# USER NAVIGATION

Annotations, annotation edits, and **History**: 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_READTHROUGH_STOP_CODON	mmtorres	5/13/14 10:51 AM
UNSET_READTHROUGH_STOP_CODON	mmtorres	5/13/14 10:52 AM
SET_READTHROUGH_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_READTHROUGH_STOP_CODON	mmtorres	5/13/14 10:51 AM
UNSET_READTHROUGH_STOP_CODON	mmtorres	5/13/14 10:52 AM
SET_READTHROUGH_STOP_CODON	mmtorres	5/13/14 10:55 AM



# USER NAVIGATION

## The Annotation Information Editor

The screenshot shows a web application window titled "Annotation Info Editor" with a close button in the top right corner. The window is split into two panels: "Gene" on the left and "Transcript" on the right, separated by a vertical dashed line. Both panels have identical fields for Name, Symbol, and Description, with the same values: "Apurinic-Apyrimidinic Endonucleas", "Apex-1", and "Multifunctional DNA Repair Enzym". Below these is a "Status" section with two radio buttons: "Approved" (selected) and "Needs review". At the bottom of each panel is a "DBXRefs" table with two columns: "DB" and "Accession". The "Gene" panel's table is empty, while the "Transcript" panel's table contains two entries: "WormBase" with "WB\_000123" and "FlyBase" with "FB0000456". The "FlyBase" row is highlighted with a dashed blue border. Below the table in each panel are "Add" and "Delete" buttons.

Gene	
Name	Apurinic-Apyrimidinic Endonucleas
Symbol	Apex-1
Description	Multifunctional DNA Repair Enzym
Status	
<input checked="" type="radio"/>	Approved
<input type="radio"/>	Needs review
DBXRefs	
DB	Accession

Transcript	
Name	Apurinic-Apyrimidinic Endonucleas
Symbol	Apex-1
Description	Multifunctional DNA Repair Enzym
Status	
<input checked="" type="radio"/>	Approved
<input type="radio"/>	Needs review
DBXRefs	
DB	Accession
WormBase	WB_000123
FlyBase	FB0000456

DBXRefs are database crossed references: if you have reason to believe that this gene is linked to a gene in a public database (including your own), then add it here.



# USER NAVIGATION

## The Annotation Information Editor

The screenshot displays two main sections of the Annotation Information Editor. The top section is titled "Gene Ontology IDs" and contains a list of two entries: "GO:0004984" and "GO:0005549". Below this list are two buttons: "Add" and "Delete". The bottom section is titled "Comments" and contains two entries: "Annotation type: Modify an existing gene model" and "Set readthrough to skip STOP signal in third exon.". Below this list are also two buttons: "Add" and "Delete". A vertical dashed line is positioned to the left of the form.

- Add PubMed IDs
- Include GO terms as appropriate from any of the three ontologies
- Write comments stating how you have validated each model.



# USER NAVIGATION

- ‘Zoom to base level’ option reveals the DNA Track.
- Change color of exons by CDS from the ‘View’ menu.
- The reference DNA sequence is visible in both directions as are the protein translations in all six frames. You can toggle either direction to display only 3 frames.

Zoom in/out with keyboard:  
shift + arrow keys up/down





# Web Apollo User Guide

(Fragment)

[http://genomearchitect.org/web\\_apollo\\_user\\_guide](http://genomearchitect.org/web_apollo_user_guide)

# ANNOTATING SIMPLE CASES

In a “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.

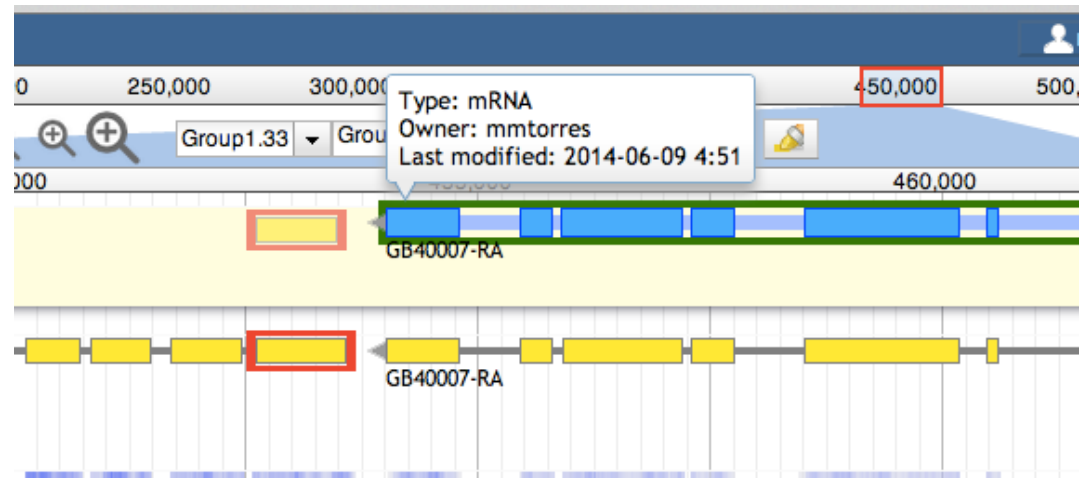
The following sections describe simple modifications.

# ADDING EXONS

Select and drag the putative new exon from a track, and add it directly to an annotated transcript in the 'User-created Annotations' area.

- Click the exon, hold your finger on the mouse button, and drag the cursor until it touches the receiving transcript. A dark green highlight indicates it is okay to release the mouse button.
- When released, the additional exon becomes attached to the receiving transcript.

- A confirmation box will warn you if the receiving transcript is not on the same strand as the feature where the new exon originated.



# ADDING EXONS

---

Each time you add an exon region, whether by extension or adding an exon, **Web Apollo recalculates the longest ORF**, identifying 'Start' and 'Stop' signals and allowing you to determine whether a 'Stop' codon has been incorporated after each editing step.

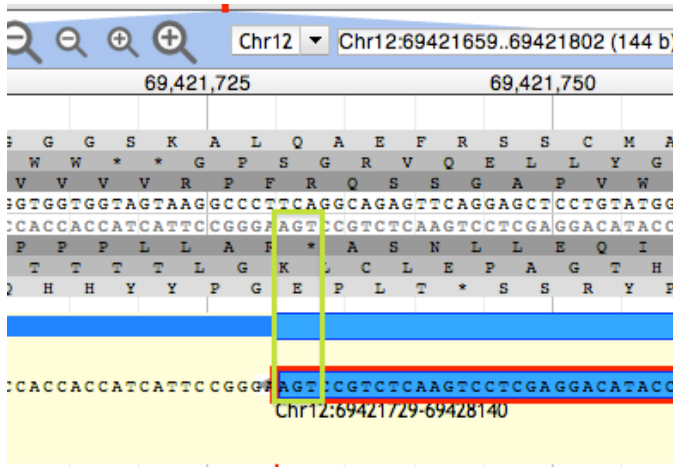
Web Apollo demands that an exon already exists as an evidence in one of the tracks. You could provide a text file in GFF format and select File → Open. GFF is a simple text file delimited by TABs, one line for each genomic 'feature': column 1 is the name of the scaffold; then some text (irrelevant), then 'exon', then start, stop, strand as + or -, a dot, another dot, and Name=some name

Example:

```
scaffold_88 Qratore exon21 2111+ . . Name=bob
scaffold_88 Qratore exon22015111+ . . Name=rad
```



# ADDING UTRs



View zoomed to base level. The DNA track and annotation track are visible. The DNA track includes the sense strand (top) and anti-sense strand (bottom). The six reading frames flank the DNA track, with the three forward frames above and the three reverse frames below. The User-created Annotation track shows the terminal end of an annotation. The green rectangle highlights the location of the nucleotide residues in the 'Stop' signal.

Gene predictions may or may not include UTRs. If transcript alignment data are available and extend beyond your original annotation, you may extend or add UTRs.

1. Position the cursor at the beginning of the exon that needs to be extended 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.

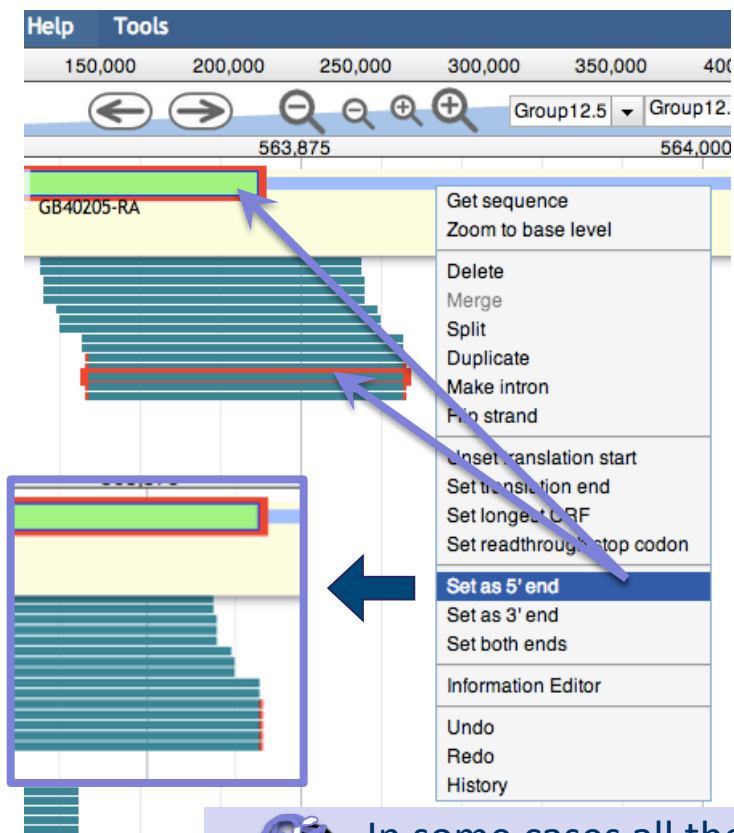


To add a new spliced UTR to an existing annotation follow the procedure for adding an exon.


# EXON STRUCTURE INTEGRITY

1. Zoom in sufficiently to clearly resolve each exon as a distinct rectangle.
2. Two exons from different tracks sharing the same start and/or end coordinates will display a red bar to indicate the matching edges.
3. 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.
4. Note if there are cDNA / RNAseq reads that lack one or more of the annotated exons or include additional exons.

# EXON STRUCTURE INTEGRITY



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

 In some cases all the data may disagree with the annotation, in other cases some data support the annotation and some of the data support one or more alternative transcripts. Try to annotate as many alternative transcripts as are well supported by the data.

# EDITING LOGIC

The editing logic in the server:

- selects longest ORF as CDS
- flags non-canonical splice sites

The screenshot displays the Apollo genome browser interface. The top navigation bar includes the Apollo logo, menu items (File, View, Help, Tools), and a user profile labeled 'demo'. The main view shows a genomic region on Chromosome 10 (Chr10) from 22,210,000 to 22,215,000. A track labeled 'User-created Annotations' shows a gene model with exons in blue and introns in light blue. A specific feature is highlighted with a red box and labeled 'Chr10:22212449-22214982'. Below this, the 'Evidence Tracks Area' shows various evidence tracks, including 'NCBI' and 'SGP', with a track for 'APEX1' visible. Annotations are color-coded: blue for exons, light blue for introns, purple for cDNA not spliced, and pink for cDNA pseudogene. A yellow box highlights a region of the gene model. Three blue arrows point from this region to a magnified inset. The inset shows a close-up of the gene model with a red box around a splice site. A yellow circle with an exclamation mark is placed on the splice site, indicating a non-canonical splice site. The inset also shows the 'Edge-matching' process between the gene model and the evidence tracks.

Available Tracks

- NCBI pseudogene
- Ensembl miRNA
- Ensembl pseudogene
- Ensembl snoRNA
- Ensembl snRNA
- Fgenesh
- GeneID
- SGP
- cDNA not spliced
- cDNA pseudogene

filter by text

File View Help Tools

Chr10 Chr10:22208951..22215080 (6.13 Kb) Go

22,210,000 22,211,250 22,212,500 22,213,750 22,215,000

User-created Annotations MGC139957

'User-created Annotations' Track

Evidence Tracks Area

NCBI

SGP

cDNA not spliced

cDNA pseudogene

MGC139957

APEX1

Chr10:22212449-22214982

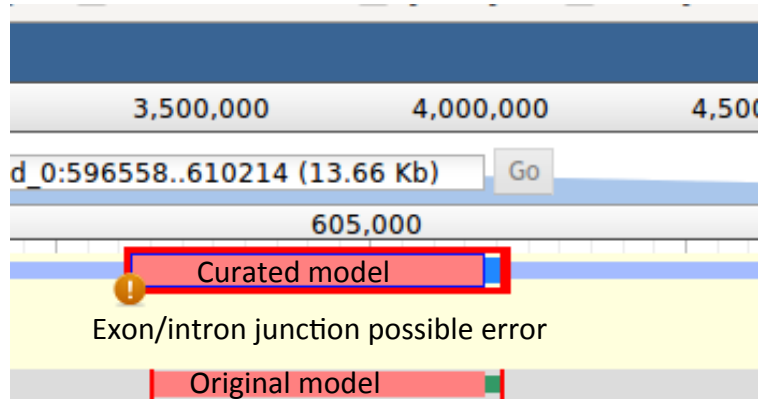
Flags non-canonical splice sites.

Selection of features and sub-features

Edge-matching



# SPLICE SITES



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

Most insects, have a valid non-canonical site GC-AG. Other non-canonical splice sites are unverified. Web Apollo flags GC splice donors as non-canonical.

Zoom to base level to review non-canonical splice site warnings. These do not necessarily need to be corrected, but should be flagged with the appropriate comment.

## Canonical splice sites:

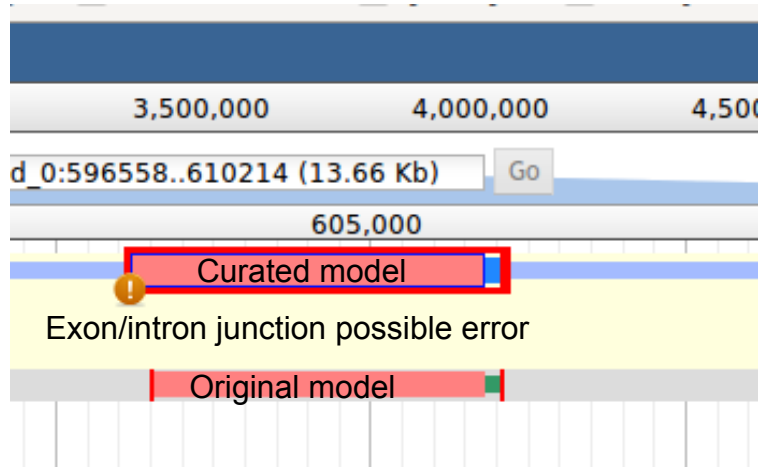
forward strand  
5'-...exon]GT / AG[exon...-3'

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



# SPLICE SITES

keep this in mind



Some gene prediction algorithms do not recognize GC splice sites, thus the intron/exon junction may be incorrect. For example, one such gene prediction algorithm may ignore a true GC donor and select another non-canonical splice site that is less frequently observed in nature.

Therefore, if upon inspection you find a non-canonical splice site that is rarely observed in nature, you may wish to search the region for a more frequent in-frame non-canonical splice site, such as a GC donor. If there is an in-frame site close that is more likely to be the correct splice donor, you may make this adjustment while zoomed at base level.

## Canonical splice sites:

forward strand

5'-...exon]GT / AG[exon...-3'

reverse strand, not reverse-complemented:

3'-...exon]GA / TG[exon...-5'



Use RNA-Seq data to make a decision.

# 'START' AND 'STOP' SITES

Web Apollo calculates the longest possible open reading frame (ORF) that includes canonical 'Start' and 'Stop' signals within the predicted exons.

If it appears to have calculated an incorrect 'Start' signal, you may modify it selecting an in-frame 'Start' codon further up or downstream, depending on evidence (protein database, additional evidence tracks). An upstream 'Start' codon may be present outside the predicted gene model, within a region supported by another evidence track.

The screenshot displays a sequence alignment interface. At the top, a DNA sequence is shown with amino acid translations above it. A black box highlights a specific start codon (ATG) in the sequence. A context menu is open over this codon, listing several actions: 'Get sequence', 'Zoom back out', 'Delete', 'Merge', 'Split', 'Duplicate', 'Make intron', 'Flip strand', 'Set translation start', 'Set translation end', 'Set longest ORF', and 'Set readthrough stop codon'. The 'Set translation start' option is currently selected and highlighted in blue.

# 'START' AND 'STOP' SITES

keep this in mind

---

Note that the 'Start' codon may also be located in a non-predicted exon further upstream. If you cannot identify that exon, add the appropriate note in the transcript's 'Comments' section.

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

In some cases, a 'Stop' codon may not be automatically identified. Check to see if there are data supporting a 3' extension of the terminal exon or additional 3' exons with valid splice sites.



**Aw, Snap!**

Something went wrong while displaying this webpage. To continue, press Reload or go to another page.

[Learn more](#)

# 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. Drag and drop each gene model to '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. Record the IDs of both starting gene models in 'DBXref' and add comments to record that this annotation is the result of a merge.

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

GB40205-RA

GB40204-RA

GB40204-RA

GB40204-RA

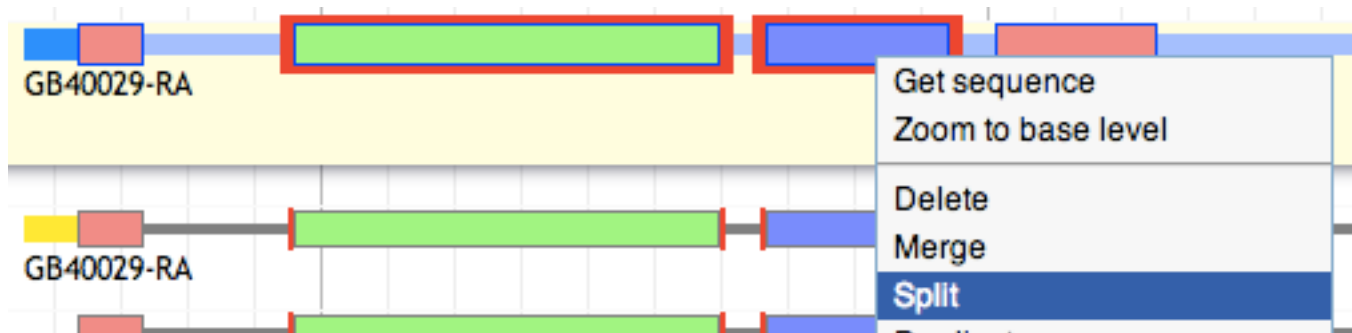
- Get sequence
- Zoom to base level
- Delete
- Merge**
- Split
- Duplicate

# COMPLEX CASES

## split a gene prediction

---

One or more splits may be recommended when different segments of the predicted protein align to two or more different families of protein homologs, and the predicted protein does not align to any known protein over its entire length. Transcript data may support a split (if so, verify that it is not a case of alternative transcripts).



# COMPLEX CASES

frameshifts, single-base errors, and selenocysteines

The screenshot displays the Apollo genome browser interface. At the top, navigation icons (back, forward, zoom in, zoom out) and a coordinate field showing 'Chr10' and 'Chr10:22213112..22213112' are visible. Below this is a scale bar with coordinates 22,213,175, 22,213,200, and 22,213,250. The main area contains a DNA track with a highlighted sequence: 'GGAGCCGGACGTAATCCGTGTATCCGACAGTGGTTCGTGTTTAC'. Below the DNA track is a protein translation track with amino acid sequences: 'P R P A L G T \* A V T S T N A S Y S A T I T R P S \* S C S S S P', 'S A C I R H I G C H Q H K C V I F S H N H S A F L I M F F L T', 'L G L H \* A H R L S P A Q M R H I Q P Q S L G L P D H V L P H L', 'CCTCGGCCTGCATTAGG CACATAGGCTGTCACCAGCACAAAT GCGTCATATTCAGCCACAATCACTC GGCCTTCCTGATCATGTTCTTCCTCACC', 'GGAGCCGGACGTAATCC GTGTATCCGACAGTGGTTCGTGTTTAC CCGGA', 'G R G A N P V Y A T V L V F', 'R P R C \* A C L S D G A C I', 'V E A Q M L C M P Q \* W C L H'. A yellow bar below the protein track is labeled 'User-created Annotations' Track. A context menu is open over the DNA track, listing options: 'Toggle Reverse Strand', 'Toggle Protein Translation', 'Create Genomic Insertion' (highlighted), 'Create Genomic Deletion', and 'Create Genomic Substitution'. Three dialog boxes are also present: 'Add Substitution' with fields for '+ strand' and '- strand' and an 'Add' button; 'Add Deletion' with a 'Length' field and an 'Add' button; and 'Add Insertion' with fields for '+ strand' and '- strand' and an 'Add' button. A zoomed-in view of the DNA sequence is shown at the bottom, with a green box highlighting the 'G' in 'GGAGCCGGACGTAATCCGTGTATCCGACAGTGGTTCGTGTTTAC' and a vertical green line extending through the protein translation track.





# COMPLEX CASES

## frameshifts, single-base errors, and selenocysteines

1. Web Apollo allows annotators to make single base modifications or frameshifts that are reflected in the sequence and structure of any transcripts overlapping the modification. Note that these manipulations do NOT change the underlying genomic sequence.
2. 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.
3. 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.
4. Once you have entered the modifications, Web 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.
5. 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.

# COMPLETING THE ANNOTATION

---

Follow our checklist until you are happy with the annotation!

Then:

- Comment to validate your annotation, even if you made no changes to an existing model. Your comments mean you looked at the curated model and are happy with it; think of it as a vote of confidence.
- Or add a comment to inform the community of unresolved issues you think this model may have.

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

# HOW TO BEGIN

---

To find the gene region you wish to annotate, you may use:

- a) a protein sequence of a homolog from another species
- b) a sequence from a similar gene in species of interest (e.g. another gene family member)
- c) on your own, you aligned your gene models or transcriptomic data to the genome.
- d) you used high quality proteins and/or gene family alignments (multi or single species) and are able to identify conserved domains.

**Option 1** – You have a sequence but don't know where it is in this genome:

- Use BLAT in the Apollo window, or BLAST at NAL's i5k BLAST server, available at: <http://i5k.nal.usda.gov/blastn>
- You may also use other tools for annotation and contribute your data from those efforts.

**Option 2** – The genome has already been annotated with your sequences and you have a gene identifier that has been indexed in Apollo.

- That is, you know where to look, so type the ID in the Search box of Apollo.
  - Apollo autocompletes using a case-insensitive search anchored on the left-hand side of the word. For example "HaGR" will show all "hagr" objects (up to 30).
- Choose one of the genes and click "Go".
- You can do that with Domains, Alignments or Gene names provided to you (if they have been indexed).

**Option 3** – Find genes based on functional ontology terms or network membership identifiers.

# GENERAL PROCESS OF CURATION

---

1. Select the chromosomal region of interest, e.g. scaffold.
2. Select appropriate **evidence tracks**.
3. Determine whether a feature in an existing evidence track will provide a reasonable gene model to start working.
  - If yes: select and **drag** the feature to the 'User-created Annotations' area, **creating an initial gene model**. If necessary use editing functions to adjust the gene model.
  - Nothing available to you? *Let's have a talk*.
4. Check your edited gene model for integrity and accuracy by comparing it with available homologs.

*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.

# WHAT ANNOTATORS SHOULD LOOK FOR

pay attention to these details

---

- ❖ **Annotating a simple case:** WHEN “The official prediction is correct, or nearly correct, assuming that no aligned data extends beyond the gene model and if so, it is not likely to be coding sequence, and/or the gene prediction matches what you know about the gene”:
  - a. Can you add UTRs?
  - b. Check exon structures.
  - c. Check splice sites: ...]5'-GT/AG-3'[...
  - d. Check ‘start’ and ‘stop’ sites.
  - e. Check the predicted protein product(s).
  - f. If the protein product still does not look correct, go on to “Annotating more complex cases”.

# WHAT ANNOTATORS SHOULD LOOK FOR

## continued

---

- ❖ **Additional functionality.** You may also need to learn how to:
  - a. Get genomic sequence
  - b. Merge exons
  - c. Add/Delete an exon
  - d. Create an exon de novo (within an intron or outside existing annotations).
  - e. Right/apple-click on a feature to get feature ID and additional information
  - f. Looking up homolog descriptions going to the accession web page at UniProt/Swissprot

# WHAT ANNOTATORS SHOULD LOOK FOR

## continued

---

### ❖ **Annotating more complex cases:**

- a. Incomplete annotation: protein integrity checks, indicate gaps, missing 5' sequences or missing 3' sequences.
- b. Merge of 2 gene predictions on same scaffold
- c. Merge of 2 gene predictions on different scaffolds (uh-oh!).
- d. Split of a gene prediction
- e. Frameshifts, Selenocysteine, single-base errors, and other inconvenient phenomena

# WHAT ANNOTATORS SHOULD LOOK FOR

## continued

---

### ❖ Adding important project information in the form of Canned and/or Customized Comments:

- a. NCBI ID, RefSeq ID, gene symbol(s), common name(s), synonyms, top BLAST hits (GenBank IDs), orthologs with species names, and anything else you can think of, because you are the expert.
- b. Type of annotation (e.g.: whether or not the gene model was changed)
- c. Data source (for example if the Fgeneshpp predicted gene was the starting point for your annotation)
- d. The kinds of changes you made to the gene model, e.g.: split, merge
- e. Functional description
- f. Whether you would like for your MOD curator to check the annotation
- g. Whether part of your gene is on a different scaffold.



# THE CHECK LIST

## for accuracy and integrity

---

1. Can you add UTRs (e.g.: via RNA-Seq)?
2. Check exon structures
3. Check splice sites: most splice sites display these residues ...]5'-GT/AG-3'[...
4. Check 'Start' and 'Stop' sites
5. Check the predicted protein product(s)
  - Align it against relevant genes/gene family.
  - blastp against NCBI's RefSeq or nr
6. If the protein product still does not look correct then check:
  - Are there gaps in the genome?
  - Merge of 2 gene predictions on the same scaffold
  - Merge of 2 gene predictions from different scaffolds
  - Split a gene prediction
  - Frameshifts
    - error in the genome assembly?
  - Selenocysteine, single-base errors, and other inconvenient phenomena
7. Finalize annotation by adding:
  - Important project information in the form of canned and/or customized comments
  - IDs from GenBank (via DBXRef), gene symbol(s), common name(s), synonyms, top BLAST hits (with GenBank IDs), orthologs with species names, and everything else you can think of, because you are the expert.
  - Whether your model replaces one or more models from the official gene set (so it can be deleted).
  - The kinds of changes you made to the gene model of interest, if any. E.g.: splits, merges, whether the 5' or 3' ends had to be modified to include 'Start' or 'Stop' codons, additional exons had to be added, or non-canonical splice sites were accepted.
  - Any functional assignments that you think are of interest to the community (e.g. via BLAST, RNA-Seq data, literature searches, etc.)

Example

# Apollo Example

- Introductory demonstration using the *Hyalella azteca* genome (amphipod crustacean).

The screenshot displays the Apollo genome browser interface. The top navigation bar includes the Apollo logo, menu items (File, View, Tools, Help), and a user profile icon labeled 'mmt'. Below the navigation bar is a genomic scale from 0 to 1,500,000. A search bar shows 'Scaffold527' and a zoomed-in region 'Scaffold527:917451..978950 (61.5 Kb)'. The main view area contains several tracks: 'User-created Annotations' (yellow), 'BlastResults' (with HSP: antp\_Tcas), 'LDEC\_v0.5.3-Models' (with LdecTmpB004337-RA), and 'augustus\_masked' (with augustus\_masked-Scaffold527-abinit-gene-9.0-mRNA-1 and augustus\_masked-Scaffold527-abinit-gene-9.1-mRNA-1). The left sidebar lists available tracks for various taxonomic groups.

A public Apollo Demo using the Honey Bee genome is available at <http://genomearchitect.org/WebApolloDemo>

# What do we know about this genome?

- Currently publicly available data at NCBI:
  - >37,000 nucleotide seqs → scaffolds, mitochondrial genes
  - 300 amino acid seqs → mitochondrion
  - 53 ESTs
  - 0 conserved domains identified
  - 0 “gene” entries submitted
- Data at i5K Workspace@NAL
  - 10,832 scaffolds, 23,288 transcripts, 12,906 proteins

# PubMed Search: what's new?

PubMed  RSS Save search Advanced

Summary ▾ 20 per page ▾ Sort by Most Recent ▾ Send to: ▾

**Results: 1 to 20 of 130** << First < Prev Page 1 of 7 Next > Last >>

- [A comparison of the sublethal and lethal toxicity of four pesticides in \*Hyalella azteca\* and \*Chironomus dilutus\*.](#)  
1. Hasenbein S, Connon RE, Lawler SP, Geist J.  
Environ Sci Pollut Res Int. 2015 Mar 26. [Epub ahead of print]  
PMID: 25804662  
[Related citations](#)
- [Responses of \*Lyngbya wollei\* to algaecide exposures and a risk characterization associated with their use.](#)  
2. Calomeni AJ, Iwinski KJ, Kinley CM, McQueen A, Rodgers JH Jr.  
Ecotoxicol Environ Saf. 2015 Mar 12;116:90-98. doi: 10.1016/j.ecoenv.2015.03.004. [Epub ahead of print]  
PMID: 25770656
- [Toxicity of fluoride to aquatic species and evaluation of toxicity modifying factors.](#)  
3. Percy K, Elphick J, Burnett-Seidel C.  
Environ Toxicol Chem. 2015 Mar 2. doi: 10.1002/etc.2963. [Epub ahead of print]  
PMID: 25732700  
[Related citations](#)
- [10-Day survival of \*Hyalella azteca\* as a function of water quality parameters.](#)  
4. Javidmehr A, Kass PH, Deanovic LA, Connon RE, Werner I.  
Ecotoxicol Environ Saf. 2015 May;115:250-6. doi: 10.1016/j.ecoenv.2015.02.008. Epub 2015 Feb 26.  
PMID: 25725458  
[Related citations](#)
- [An evaluation of the residual toxicity and chemistry of a sodium hydroxide-based ballast water treatment system for freshwater ships.](#)  
5. Elskus AA, Ingersoll CG, Kemble NE, Echols KR, Brumbaugh WG, Henquinet JW, Watten BJ.  
Environ Toxicol Chem. 2015 Feb 18. doi: 10.1002/etc.2943. [Epub ahead of print]  
PMID: 25693486  
[Related citations](#)
- [Toxicity of Cúspide 480SL® spray mixture formulation of glyphosate to aquatic organisms.](#)  
6. Currie Z, Prosser RS, Rodriguez-Gil JL, Mahon K, Poirier D, Solomon KR.  
Environ Toxicol Chem. 2015 Feb 5. doi: 10.1002/etc.2913. [Epub ahead of print]  
PMID: 25655706  
[Related citations](#)

# PubMed Search: what's new?

PubMed : ((["2010/01/01"[Date - Publication] : "3000"[Date - Publi  
RSS Save search Advanced

Summary ▾ 20 per page ▾ Sort by Most Recent ▾

Results: 1 to 20 of 130 << First <

[A comparison of the sublethal and lethal toxicity of four pesticides](#)  
1. [Chironomus dilutus](#).  
Hasenbein S, Connon RE, Lawler SP, Geist J.  
Environ Sci Pollut Res Int. 2015 Mar 26. [Epub ahead of print]  
PMID: 25804662  
[Related citations](#)

[Responses of \*Lyngbya wollei\* to algaecide exposures and a risk characterization associated with their use](#).  
2. [Calomeni AJ, Iwinski KJ, Kinley CM, McQueen A, Rodgers JH Jr.](#)  
Ecotoxicol Environ Saf. 2015 Mar 12;116:90-98. doi: 10.1016/j.ecoenv.2015.03.004. [Epub ahead of print]  
PMID: 25770656

[Toxicity of fluoride to aquatic species and evaluation of toxicity modifying factors](#).  
3. [Pearcy K, Elphick J, Burnett-Seidel C.](#)  
Environ Toxicol Chem. 2015 Mar 2. doi: 10.1002/etc.2963. [Epub ahead of print]  
PMID: 25732700  
[Related citations](#)

[10-Day survival of \*Hyalella azteca\* as a function of water quality parameters](#).  
4. [Javidmehr A, Kass PH, Deanovic LA, Connon RE, Werner I.](#)  
Ecotoxicol Environ Saf. 2015 May;115:250-6. doi: 10.1016/j.ecoenv.2015.02.008. Epub 2015 Feb 26.  
PMID: 25725458  
[Related citations](#)

[An evaluation of the residual toxicity and chemistry of a sodium hydroxide-based ballast water treatment system for freshwater ships](#).  
5. [Elskus AA, Ingersoll CG, Kemble NE, Echols KR, Brumbaugh WG, Henquinet JW, Watten BJ.](#)  
Environ Toxicol Chem. 2015 Feb 18. doi: 10.1002/etc.2943. [Epub ahead of print]  
PMID: 25693486  
[Related citations](#)

[Toxicity of Cúspide 480SL® spray mixture formulation of glyphosate to aquatic organisms](#).  
6. [Currie Z, Prosser RS, Rodríguez-Gil JL, Mahon K, Poirier D, Solomon KR.](#)  
Environ Toxicol Chem. 2015 Feb 5. doi: 10.1002/etc.2913. [Epub ahead of print]  
PMID: 25655706  
[Related citations](#)

## Multiple origins of pyrethroid insecticide resistance across the species complex of a nontarget aquatic crustacean, *Hyalella azteca*

Donald P. Weston<sup>a,1</sup>, Helen C. Poynton<sup>b</sup>, Gary A. Wellborn<sup>c</sup>, Michael J. Lydy<sup>d</sup>, Bonnie J. Blalock<sup>b</sup>, Maria S. Sepulveda<sup>e</sup>, and John K. Colbourne<sup>f</sup>

<sup>a</sup>Department of Integrative Biology, University of California, Berkeley, CA 94720; <sup>b</sup>School for the Environment, University of Massachusetts, Boston, MA 02125; <sup>c</sup>Department of Biology, University of Oklahoma, Norman, OK 73019; <sup>d</sup>Center for Fisheries, Aquaculture and Aquatic Sciences, Southern Illinois University, Carbondale, IL 62901; <sup>e</sup>Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907; and <sup>f</sup>The Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405

“Ten populations (3 laboratory cultures, 7 California water bodies) differed by at least 550-fold in sensitivity to pyrethroids.”

“By sequencing the primary pyrethroid target site, the voltage-gated sodium channel (*vpsc*), we show that point mutations and their spread in natural populations were responsible for differences in pyrethroid sensitivity.”

“The finding that a non-target aquatic species has acquired resistance to pesticides used only on terrestrial pests is troubling evidence of the impact of chronic pesticide transport from land-based applications into aquatic systems.”

# How many sequences for our gene of interest?

## And what do we know about them?

- Para, (voltage-gated sodium channel alpha subunit; *Nasonia vitripennis*).
- NaCP60E (Sodium channel protein 60 E; *D. melanogaster*).
- MF: voltage-gated cation channel activity (IDA, GO:0022843).
- BP: olfactory behavior (IMP, GO:0042048), sodium ion transmembrane transport (ISS, GO:0035725).
- CC: voltage-gated sodium channel complex (IEA, GO:0001518).

Gene (voltage gated sodium channel) AND "arthropods"[porgn: \_\_txid6656]  
Save search Advanced

Display Settings: [x] Tabular, 20 per page, Sorted by Relevance

Results: 1 to 20 of 112 << First < Prev Page 1 of 6 Ne

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> <a href="#">Para</a> ID: 100101930	voltage-gated sodium channel alpha subunit [ <i>Nasonia vitripennis</i> (jewel wasp)]	Chromosome 2, NC_015868.2 (23670859..23709551)	
<input type="checkbox"/> <a href="#">Nav</a> ID: 100101199	voltage-gated sodium channel alpha subunit [ <i>Bombyx mori</i> (domestic silkworm)]	NW_004582014.1 (5225762..5257092)	PARA
<input type="checkbox"/> <a href="#">CpipJ_CPIJ017894</a> ID: 6052303	voltage-gated sodium channel [ <i>Culex quinquefasciatus</i> (southern house mosquito)]		CpipJ_CPIJ017894
<input type="checkbox"/> <a href="#">NaCP60E</a> ID: 37981	Na channel protein 60E [ <i>Drosophila melanogaster</i> (fruit fly)]	Chromosome 2R, NT_033778.4 (24891614..24921677)	Dmel_CG34405, CG34405, CG9071, C DIC60, DSC, DSC1, Dmel_CG34405, Dme Dmel_CG9071, Na-C dsc1, smi60E
<input type="checkbox"/> <a href="#">CpipJ_CPIJ017896</a> ID: 6052300	voltage-gated sodium channel [ <i>Culex quinquefasciatus</i> (southern house mosquito)]		CpipJ_CPIJ017896
<input type="checkbox"/> <a href="#">CpipJ_CPIJ007596</a> ID: 6040379	voltage-gated sodium channel [ <i>Culex</i> ]		CpipJ_CPIJ007596

# BLAST at i5K

<https://i5k.nal.usda.gov/blast>

**PNAS**

**Multiple origins of pyrethroid insecticide resistance across the species complex of a nontarget aquatic crustacean, *Hyalella azteca***

Donald P. Weston<sup>1,2</sup>, Helen C. Poynton<sup>3</sup>, Gary A. Wellborn<sup>4</sup>, Michael J. Lydy<sup>5</sup>, Bonnie J. Blalock<sup>6</sup>, Maria S. Sepulveda<sup>7</sup>, and John K. Colbourne<sup>1</sup>

<sup>1</sup>Department of Integrative Biology, University of California, Berkeley, CA 94720; <sup>2</sup>School for the Environment, University of Massachusetts, Boston, MA 02125; <sup>3</sup>Department of Biology, University of Oklahoma, Norman, OK 73019; <sup>4</sup>Center for Fisheries, Aquaculture and Aquatic Sciences, Southern Illinois University, Carbondale, IL 62901; <sup>5</sup>Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907; and <sup>6</sup>The Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405

## D. Segment 3: Domain II S4-S6

```

UCB      RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
BK       RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
PGC-D    RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
PGC-B-a.1 RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
PGC-B-a.2 RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
PGC-B-b  RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
MS       RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
MO       RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
CH-b     RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
CH-a     RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
GC       RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQ
    
```

14

Figure S4:  
A.

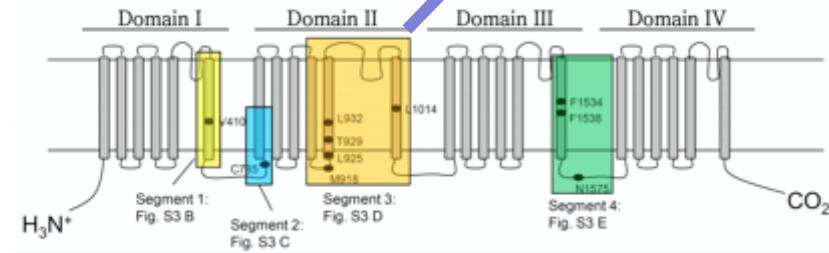


Figure S4 continued:

```

UCB      DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
BK       DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
PGC-D    DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
PGC-B-a  DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
PGC-B-b  DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
MS       DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
MO       DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
CH-b     DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
CH-a     DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
GC       DQMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR
    
```

>vgsc-Segment3-DomainII

RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFIFAVMGMQLFGKNYTEKVTKFKWSQDG  
QMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGDVDFSCVPPFLATVVIIGNLVVFMHR



# BLAST at i5K

<https://i5k.nal.usda.gov/blast>

## BLAST Databases

### Organisms

- Eurytemora affinis*
- Fopius arisanus*
- Frankliniella occidentalis*
- Gerris buenoi*
- Halyomorpha halys*
- Homalodisca vitripennis*
- Hyalella azteca*
- Ladona fulva*
- Latrodectus hesperus*
- Leptinotarsa decemlineata*
- Limnephilus lunatus*
- Loxosceles reclusa*
- Manduca sexta*

### Hyalella azteca

#### Nucleotide

- Genome Assembly - *Hyalella\_azteca\_scaffolds*
- Transcript - *Hyalella\_azteca\_BCM\_v0.5.3\_transcripts*

#### Peptide

- Protein - *Hyalella\_azteca\_BCM\_v0.5.3\_proteins*

## Query Sequence

Enter sequence below in FASTA format:

```
>vgsc-Segment3-DomainII
RVFKLAKSWPTLNLLISIMGKTVGALGNLTFV
LCIIIFAVMGMQLFGKNYTEKVTKFKWSQD
GQMPRWNFVDFHFSMIVFRVLCGEWIESM
WDCMYVGDVFCVFFLATVWIGNLVVFSMHR
```

Or load it from disk

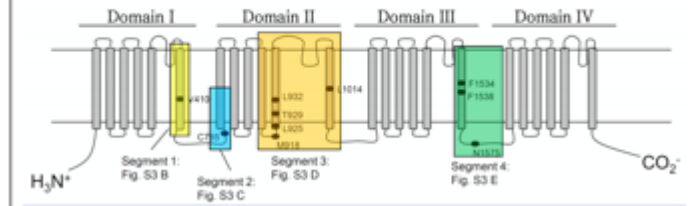
Choose File No file chosen

## Program

blastn  tblastn  tblastx  blastp  blastx

**PNAS**  
Multiple origins of pyrethroid insecticide resistance across the species complex of a nontarget aquatic crustacean, *Hyalella azteca*  
Donald P. Wilson<sup>1,2</sup>, Helen E. Poynton<sup>3</sup>, Gary A. Wellborn<sup>4</sup>, Michael J. Ludy<sup>5</sup>, Bonnie J. Mahan<sup>6</sup>, Maria S. Sepúlveda<sup>7</sup>, and John K. Colbourne<sup>1</sup>  
<sup>1</sup>Department of Integrative Biology, University of California, Berkeley, CA 94720; <sup>2</sup>Center for the Environment, University of Massachusetts, Boston, MA 02125; <sup>3</sup>Department of Biology, University of California, San Diego, CA 92092; <sup>4</sup>Center for Invasive Species and Ecosystem Health, Louisiana State University, Baton Rouge, LA 70803; <sup>5</sup>Department of Biology and Natural Resources, North Carolina State University, Raleigh, NC 27695; <sup>6</sup>The Center for Biomass and Biofuels, Indiana University, Bloomington, IN 47405

Figure S4:  
A.



>vgsc-Segment3-DomainII

```
RVFKLAKSWPTLNLLISIMGKTVGALGNLTFVLCIIIFAVMGMQLFGKNYTEKVTKFKWSQD
QMPRWNFVDFHFSMIVFRVLCGEWIESMWDCMYVGDVFCVFFLATVWIGNLVVFSMHR
```

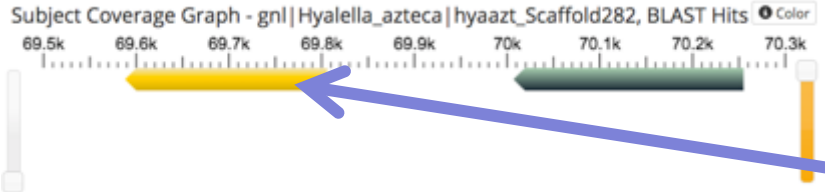
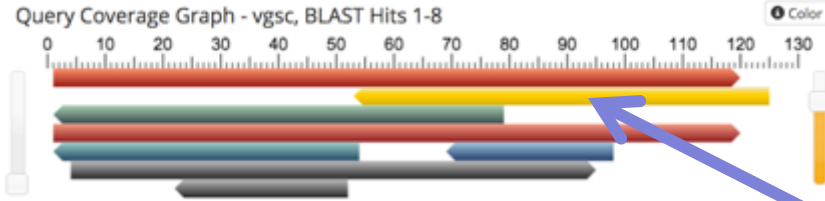
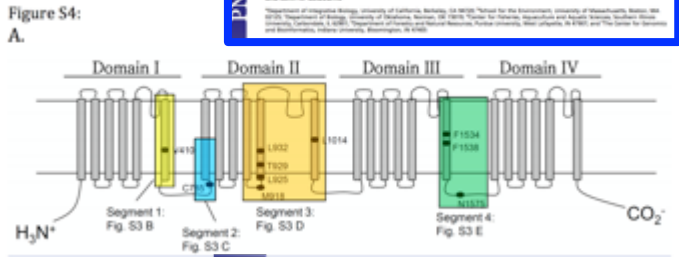
# BLAST at i5K

<https://i5k.nal.usda.gov/blast>

**Multiple origins of pyrethroid insecticide resistance across the species complex of a nontarget aquatic crustacean, *Hyalella azteca***

Donald P. Wilson<sup>1,2</sup>, James E. Poppe<sup>3</sup>, Gary A. Wellborn<sup>4</sup>, Michael J. Ludy<sup>5</sup>, Dennis J. Mahan<sup>6</sup>, Maria S. Sepúlveda<sup>7</sup>, and John A. Colbourne<sup>1</sup>

<sup>1</sup>Department of Integrative Biology, University of California, Berkeley, CA 94720; <sup>2</sup>Center for the Environment, University of Massachusetts, Boston, MA 02125; <sup>3</sup>Department of Biology, University of Louisiana, Monroe, LA 70002; <sup>4</sup>Center for Invasive Species and Ecosystem Health, Louisiana State University, Baton Rouge, LA 70803; <sup>5</sup>Department of Biology and Natural Resources, Weber State University, Ogden, UT 84403; <sup>6</sup>The Center for Genetic and Bioinformatics, Indiana University, Bloomington, IN 47405



BLAST Report FASTA

gnl|Hyalella\_azteca|hyaazt\_Scaffold282  
 Length=1631645

Score = 156 bits (395), Expect = 2e-42, Method: Compositional matrix adjust.  
 Identities = 73/73 (100%), Positives = 73/73 (100%), Gaps = 0/73 (0%)  
 Frame = -1

Query 53 TEKVTKFKWSQDGMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGFSCVPFFLAT 112  
 54 TEKVTKFKWSQDGMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGFSCVPFFLAT  
 55 Subject 69806 TEKVTKFKWSQDGMPRWNFVDFHFSFMIVFRVLCGEWIESMWDCHYVGFSCVPFFLAT 69627

Query 113 VVIGNLVVSPFMHR 105  
 58 VVIGNLVVSPFMHR  
 59 Subject 69596 VVIGNLVVSPFMHR 69589

Showing 2 to 8 of 8 entries

blastdb	qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	eval
hyazt	vgsc	Scaffold282	100	73	0	0	53	125	69806	69588	2e-42
hyazt	vgsc	Scaffold282	73.49	83	18	2	1	79	70255	70007	5e-25
hyazt	vgsc	HAZT004012-RA	57.72	123	44	3	1	120	502	855	1e-41
hyazt	vgsc	Scaffold85	70.37	54	16	0	1	54	400102	399941	9e-14
hyazt	vgsc	Scaffold85	76.67	30	7	0	69	98	399712	399623	1e-9
hyazt	vgsc	HAZT008744-RA	21.9	105	69	3	4	95	7	321	0.22

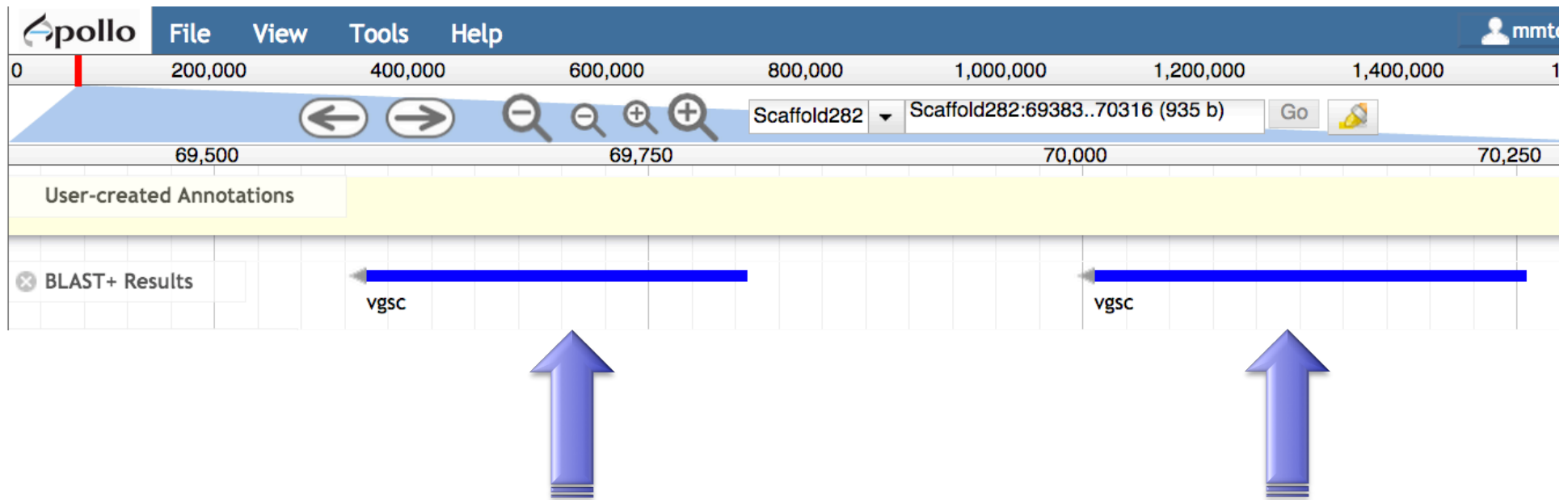
Download

To Web Apollo



# BLAST at i5K:

high-scoring segment pairs (hsp) in “BLAST+ Results” track



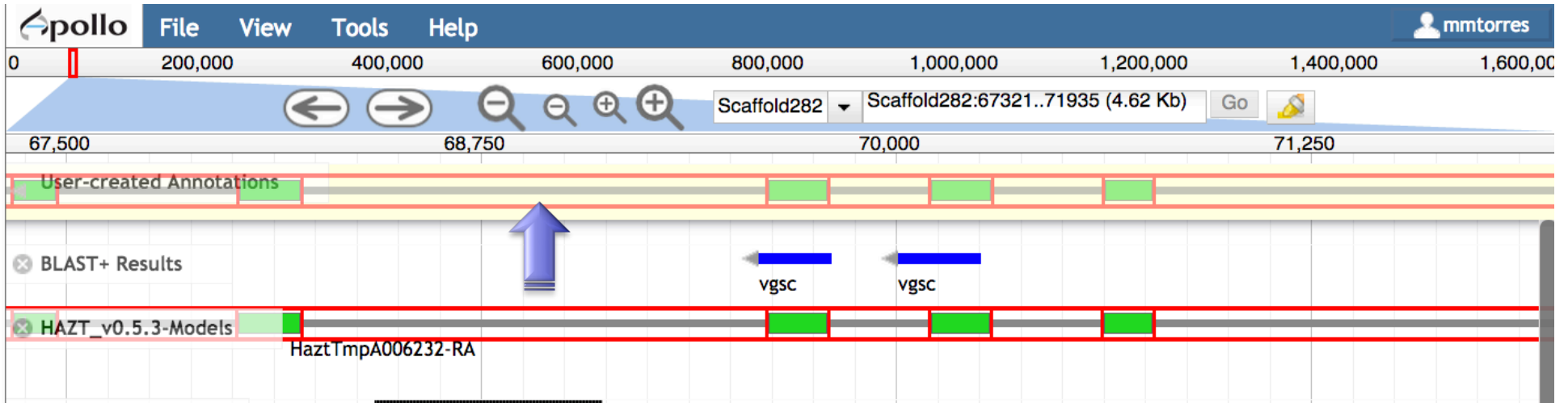
# Available Tracks

## Available Tracks

- ▼ 0. Reference Assembly 3
  - GC Content
  - Gaps in assembly
  - BLAST+ Results
- ▼ BCM\_v0.5.3 45
  - ▼ 1. Gene Sets 2
    - ▼ Primary Gene Sets: Protein Coding 1
      - HAZT\_v0.5.3-Models
    - ▼ Supplementary Gene Predictions 1
      - augustus\_masked
  - ▼ 2. Evidence 2
    - ▼ Repeats 2
      - repeatmasker
      - repeatrunner

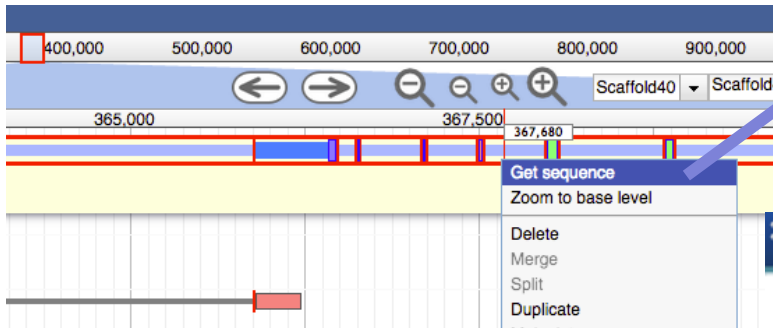
- ▼ 3. Mapped Proteins 41
  - ▼ Other 1
    - cegma
  - ▼ Protein2genome 20
    - protein2genome\_Annelida
    - protein2genome\_Arthropoda
    - protein2genome\_Atelocerata
    - protein2genome\_Cephalochordata
    - protein2genome\_Chelicerata
    - protein2genome\_Cnidaria
    - protein2genome\_Craniata
    - protein2genome\_Crustacea
    - protein2genome\_Echinodermata
    - protein2genome\_Mollusca
    - protein2genome\_Nemata
    - protein2genome\_Nematomorpha
    - protein2genome\_Onychophora
    - protein2genome\_Parazoa
    - protein2genome\_Placozoa
    - protein2genome\_Platyhelminthes
    - protein2genome\_Priapulida
    - protein2genome\_Tardigrada
    - protein2genome\_Tunicat
    - protein2genome\_UNCATEGORISED
  - ▼ Supplementary: BlastX 20
    - blastx\_Annelida
    - blastx\_Arthropoda
    - blastx\_Atelocerata
    - blastx\_Cephalochordata

# Creating a new gene model: drag and drop



- Web Apollo automatically calculates the longest open reading frame (ORF). In this case, the ORF includes the high-scoring segment pairs (hsp).

# Get Sequence



**Alignments**

Download ▾ GenPept Graphics Sort by: E value

voltage-gated sodium channel [Cancer borealis]  
Sequence ID: [gb|ABL10360.2](#) Length: 1989 Number of Matches: 4

Range 1: 634 to 1989 GenPept Graphics

Score	Expect	Method	Identities	Positives
2051.00	(5314)	0.0	Compositional matrix adjust.	1058/1390(76%) 1167/1390(83)

Query 14 EEEALALKLKNPETNPFIDPIQKQTVVDMRDVVMVNDIIEQAAQHGRLSRASDHG---  
Sbjct 634 +++ LALKLKNP+ NPPIDPIQKQTVVDMRDV VL DIIEQAA GR+SRASDHG  
D+DDMLALKLKNPDANPFIQKQTVVDMRDVRLQDIIEQAAVNAGRMSRASDHGVS

Query 71 -----EDEEEPLAAKIKKKVIEYGRLLIILCVWDCWLWIKIQHILGLIVDFPFVLF  
Sbjct 694 YYFSAEEEEESFKEKLLKLLLEYLLKAIDIFCVWDCSYWNKFAKLVELLVDFPFVLF  
E+EEE K+KKK+EY + I+I CVWDC W K ++ L+VDFPFVLF

Query 126 TLCIVVNTLFMAMDHGMRQSFDFLKMGNYSIACISPTFTIEAFTKLMAMSPKIFYO  
Sbjct 754 TLCIVVNTLFMAMDH+GM SF FLKMGN Y A TP+IE F K++AMSPK+Y Q  
TLCIVVNTLFMAMDHYGMNCSFDHFLKMGNYYFTA-----TFSIECLKI IAMSPKYYLQ

Query 186 EGWNIFFDFIVALSLIELGLEGVQLSVLRSFRLGIQRTIRVDVQLRVFKLAKSWPTLN  
Sbjct 809 EGWNIFFDF I VLSLLELGLANVGLSVLRSFRL-----LAKSWPTLN

Query 246 LLISIMGKTVGALGNLTFVLCIIIFIFAVMGMLFQGNYTEKVTFKFWSQDGMQRWNV  
Sbjct 857 LLISIMGKTVGALGNLTFVLCIIIFIFAVMGMLFQGNYTEKVMKFPNPDGQLPRWNT

Query 306 DFFHSFMIVFRVLCGEWIESMWDVMYVDFSCVPPFLATVVGIVLNLFLALLSSFG  
Sbjct 917 DFFHSFMIVFRVLCGEWIESMWDVMYVDFSCVPPFLATVVGIVLNLFLALLSSFG

**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI BLAST/ blastp suite/ Formatting Results - J8537GVE016

http://blast.ncbi.nlm.nih.gov/Blast.cgi

Hazt\_6232 1380 residues [peptide]

RID J8537GVE016 (Expires on 04-09 06:55 am)

Query ID |c|Query\_65336  
Description Hazt\_6232 1380 residues [peptide]  
Molecule type amino acid  
Query Length 1380

Database Name J8537GVE016  
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects  
Program BLASTP 2.2.31+ Citation

Other reports: Search Summary Taxonomy reports Distance tree of results Multiple alignment

DELTA-BLAST, a more sensitive protein-protein search

**Graphic Summary**

Show Conserved Domains

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

Query seq. [Diagram showing conserved domains: Ion\_trans, Na\_trans\_assoc, Ion\_trans, Ion\_trans]

Specific hits: Na\_trans\_assoc  
Superfamilies: Na\_trans\_assoc superfamily  
Multi-domains: Ion\_trans, Ion\_trans, Ion\_trans

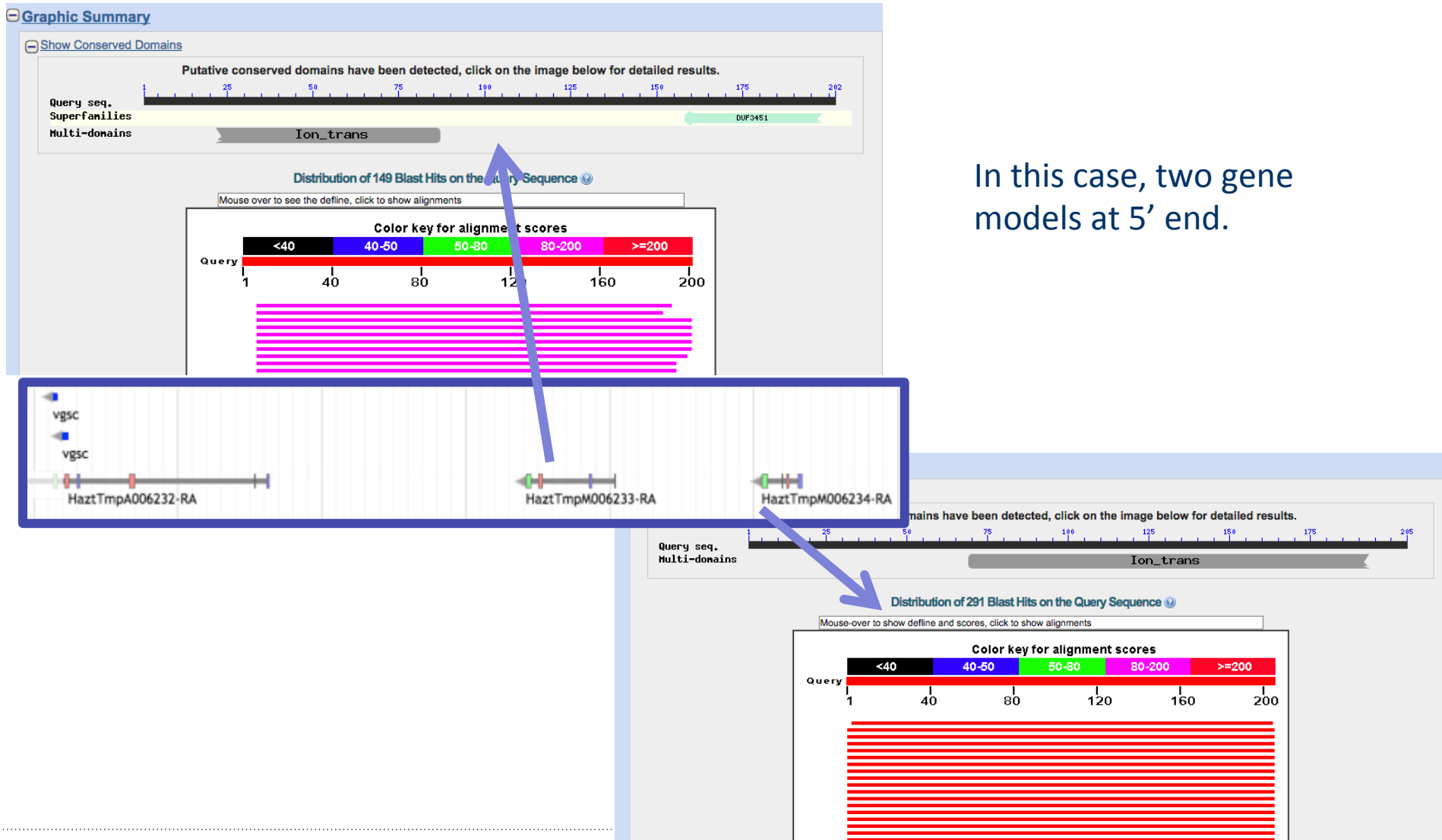
Distribution of 200 Blast Hits on the Query Sequence

Color key for alignment scores

Score Range	Color
<40	Black
40-60	Green
60-80	Yellow
80-200	Pink
>=200	Red

Query [Diagram showing alignment scores across the sequence]

# Flanking sequences (other gene models) vs. NCBI nr



# Review alignments



Download v GenPept Graphics Sort by: E value

**voltage-gated sodium channel [Cancer borealis]**  
Sequence ID: [gb|ABL10300.2](#) Length: 1909 Number of Matches: 4

Range 1: 634 to 1989 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
2051 bits(5314)	0.0	Compositional matrix adjust.	1058/1390(76%)	1167/1390(83%)	57/1
Query 14	EPEEALALKLNKPNPFPIDPIQKQTVVDMRDVMDVNDIIEQAAQHGRLSRASDHG---	70			
Sbjct 634	++ LALKLKNP+ NPFIDPIQKQTVVDMRDV VL DIIEQAA GR+SRASDHG	693			

Download v GenPept Graphics Sort by: E value

**voltage-gated sodium channel [Cancer borealis]**  
Sequence ID: [gb|ABL10300.2](#) Length: 1909 Number of Matches: 2

**HaztTmpM006234**

Range 1: 35 to 244 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
300 bits(767)	1e-89	Compositional matrix adjust.	154/210(73%)	179/210(85%)	9/210(4%)
Query 4	VISKGKIDFRFSATDALWLLSPFSLRRTAIYILINPLFNFFIICTILSNCLMMRPSTE	63			
Sbjct 35	VISKG+DIFRFSATDA+W+LSPF+P+RRTA++ILI+PLFNFFIICTIL+NC+LM+ P+ E	94			

Download v GenPept Graphics Sort by: E value

**para sodium channel [Blattella germanica]**  
Sequence ID: [gb|AAC47484.1](#) Length: 2031 Number of Matches: 4

Range 1: 681 to 2031 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
1823 bits(4721)	0.0	Compositional matrix adjust.	964/1398(69%)	1094/1398(78%)	84/1398(6%)
Query 20	ALKLNKPNPFPIDPIQKQTVVDMRDVMDVNDIIEQAAQHGRLSRASDHG-----	70			
Sbjct 681	A+K K+ + NPF+ +Q+ T+VDM DVMVLNDIIEQAA G+ SRAS+HG	736			

Download v GenPept Graphics Sort by: E value

**AAEL008297-PA, partial [Aedes aegypti]**  
Sequence ID: [tel|AF\\_001033100.1](#) Length: 910 Number of Matches: 2

Range 1: 65 to 276 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
269 bits(687)	2e-80	Compositional matrix adjust.	144/212(68%)	170/212(80%)	8/212(3%)
Query 2	FTVISKGDIFRFSATDALWLLSPFSLRRTAIYILINPLFNFFIICTILSNCLMMRPS	61			
Sbjct 65	F V+SKGKIDFRFSAT+AL+L PF+P+RR AIYIL++PLF+FFII TIL+NC+LM+ PS	124			

Download v GenPept Graphics Sort by: E value

**AAEL008297-PA, partial [Aedes aegypti]**  
Sequence ID: [tel|AF\\_001033100.1](#) Length: 910 Number of Matches: 1

Range 1: 312 to 484 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
188 bits(478)	1e-51	Compositional matrix adjust.	112/188(60%)	133/188(70%)	17/1
Query 8	QCPRGFLCQYGNPDYGYTSFDSFVSWALLSAFRMLTQDYWENLYQQVLRAGPWEHIV				
Sbjct 312	QC G++CLQYGNP+YGYTSFD+F WA LSAFRMLTQDYWENLYQ VLR+AGPWEH++F				

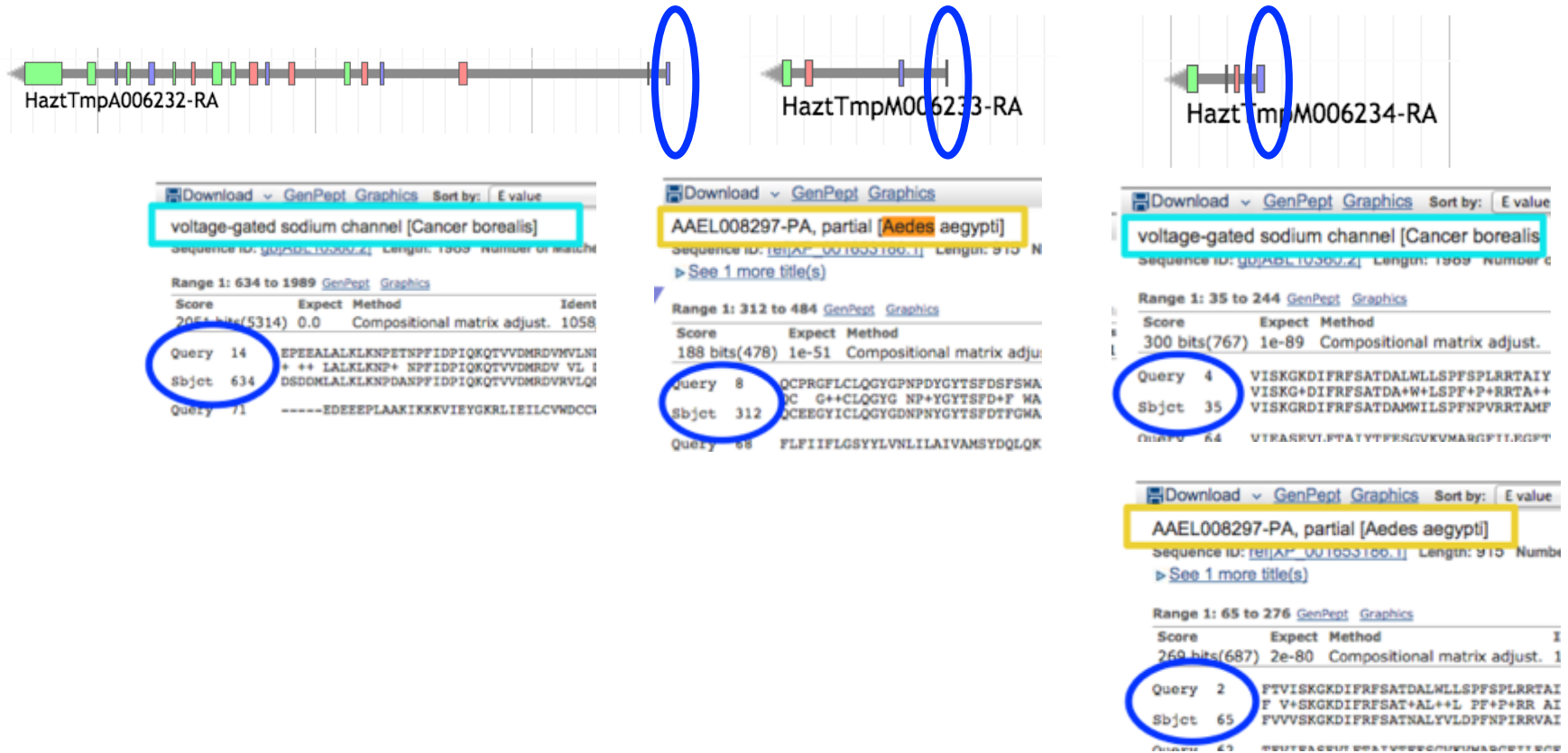
HaztTmpM006232

HaztTmpM006233

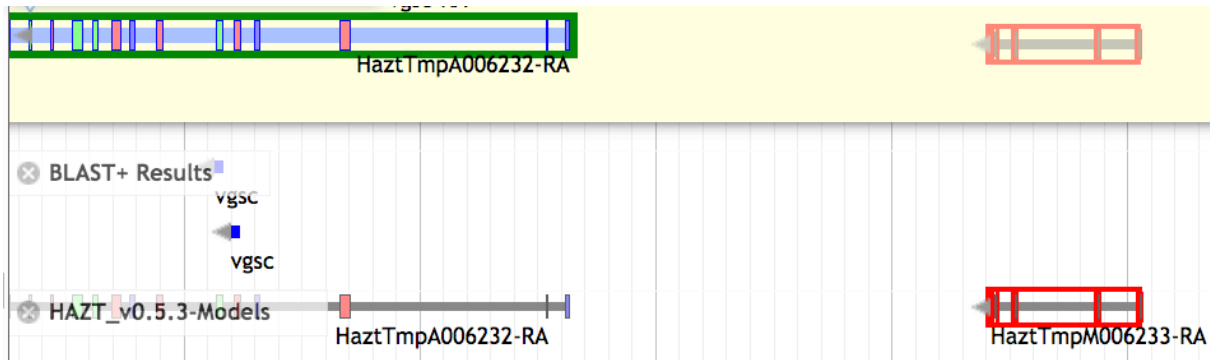




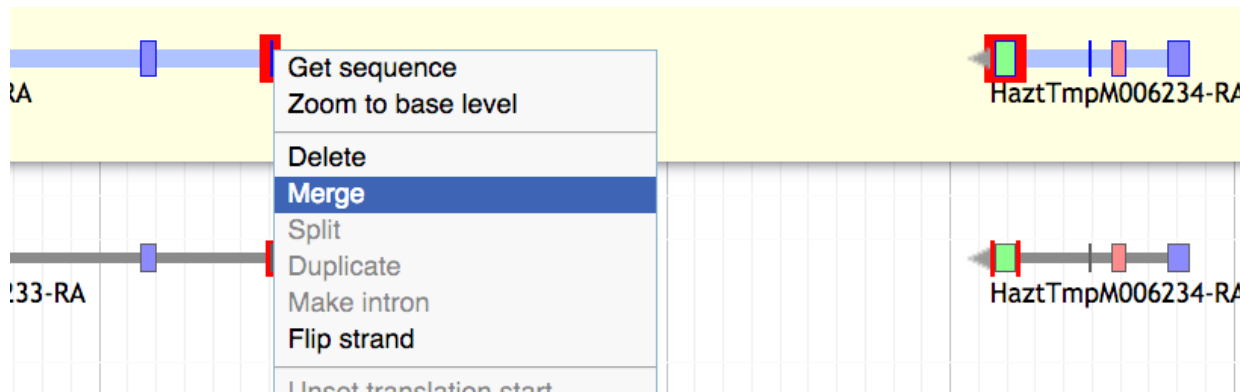
# Hypothesis for *vgsc* gene model



# Editing: merge

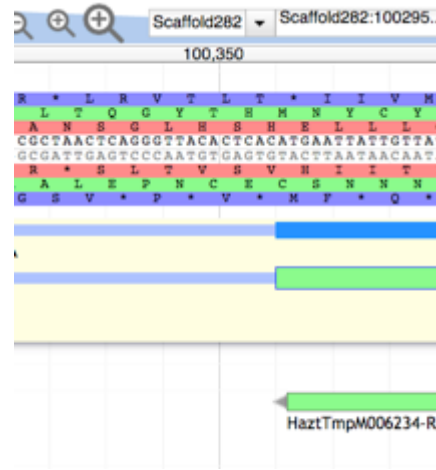
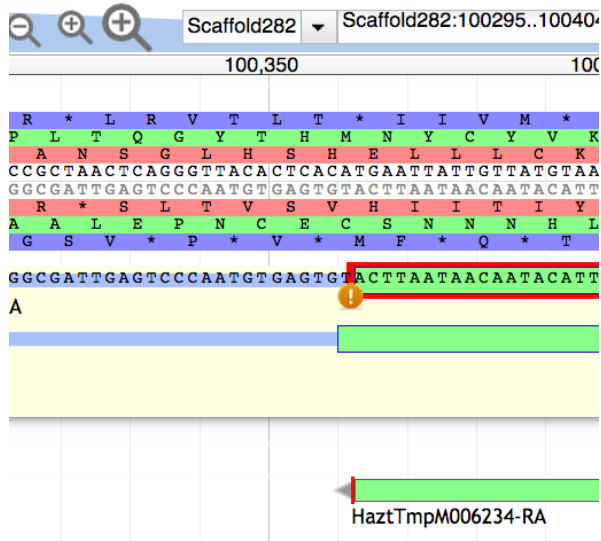


Merge by dropping an exon or gene model onto another.

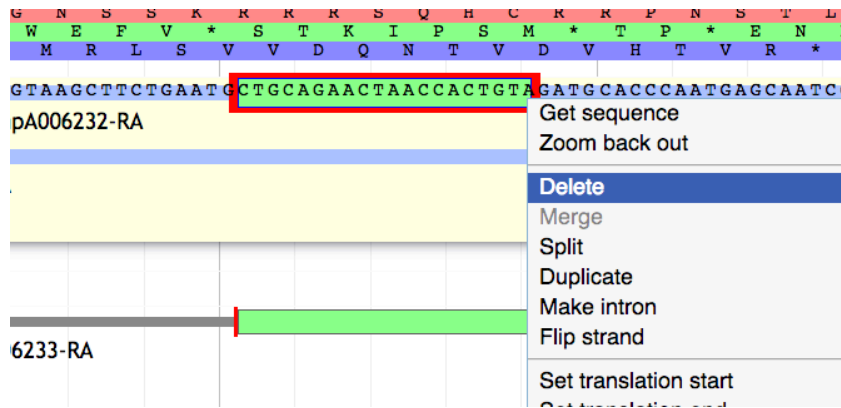


Merge by selecting two exons (holding down “Shift”) and using the right click menu.

# Editing: correct boundaries, delete exons

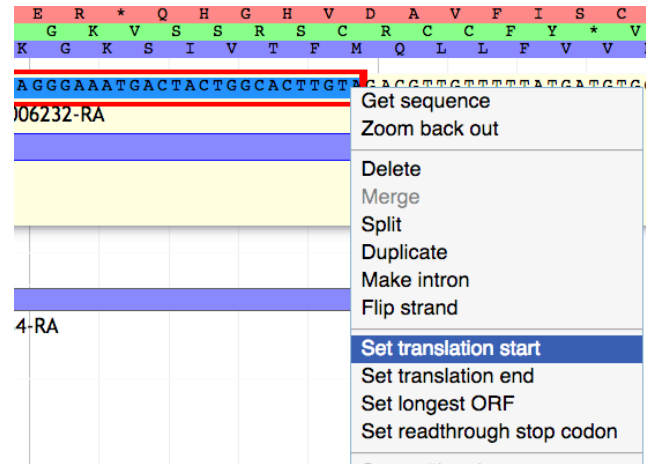


Modify exon / intron boundary by dragging the end of the exon to the nearest canonical splice site.

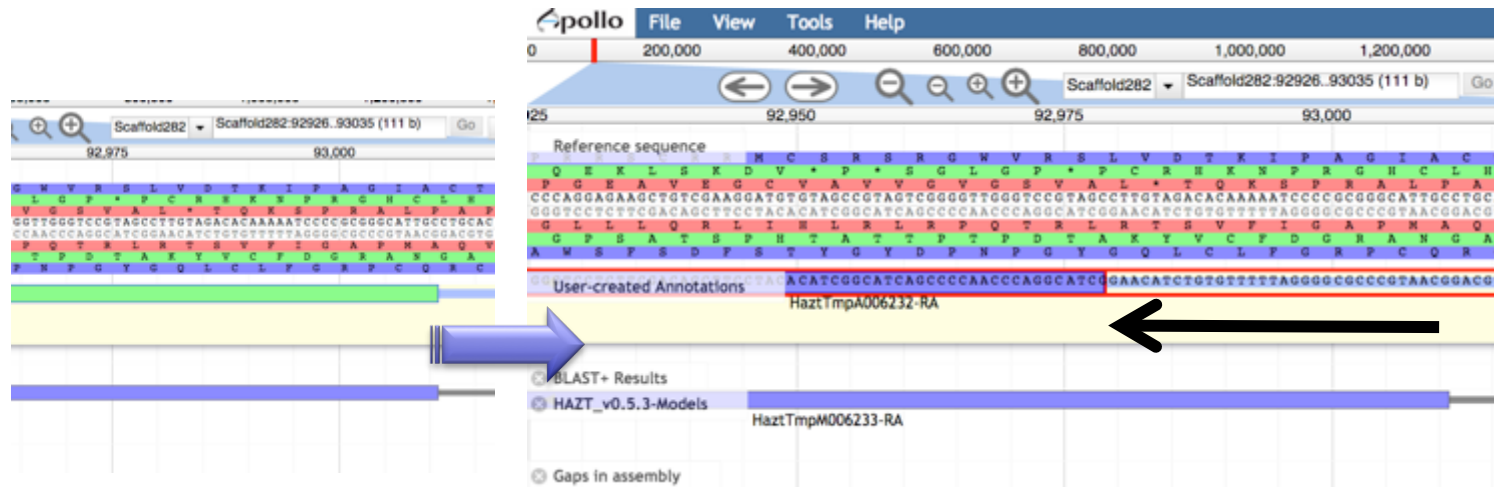


Delete first exon from M006233

# Editing: set translation start, modify boundary

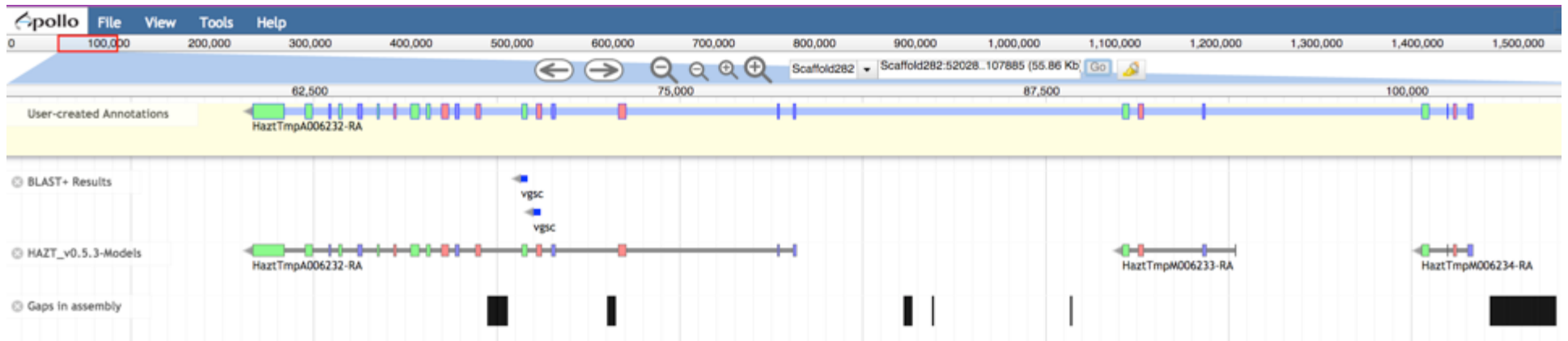


Set translation start



Modify intron /  
exon boundary  
(here and at  
coord. 78,999)

# Finished model



Corroborate integrity and accuracy of the model:

- Start and Stop
- Exon structure and splice sites ...]5'-GT/AG-3'[...
- Check the predicted protein product on NCBI nr

# Information Editor

- DBXRefs: e.g. NP\_001128389.1, *N. vitripennis*, RefSeq
- PubMed identifier: PMID: 24065824
- Gene Ontology IDs: GO:0022843, GO:0042048, GO:0035725, GO:0001518.
- Comments.
- Name, Symbol.
- Approve / Delete radio button.

Annotation type: Modify an existing gene model  
Annotation type: Approve an existing gene model  
Annotation type: Add a new gene model not existing in gene set  
Annotation type: Delete an existing gene model  
Annotation type: Existing gene model is a pseudogene  
Annotation type: Add a gene symbol  
Annotation type: Add a gene description  
Annotation type: Add a comment  
Annotation type: Flag incorrect gene model  
Annotation type: Change translation start site  
Result of: Merging of two or more gene models across scaffolds. Gene involved in merge:  
Result of: Merging of two or more gene models. Gene involved in merge:  
Result of: Splitting a gene model. Original gene model ID:  
Result of: Adding an exon to the gene model  
Result of: Removing an exon from the gene model  
Result of:  
Data source for annotation:  
Note to curator:  
Selenocysteine insertion sequence at coordinates:  
CDS edit: Stop codon readthrough due to selenocysteine insertion  
CDS edit: sequencing error  
Added 5'UTR  
Added 3'UTR

**Comments**  
(if applicable)

Demo

# APOLLO demonstration

---

See Apollo Demonstration Video at:  
[https://youtu.be/VgPtAP\\_fvxY](https://youtu.be/VgPtAP_fvxY)



# Exercises

## Live Demonstration using the *Apis mellifera* genome.

### 1. Evidence in support of protein coding gene models.

#### 1.1 Consensus Gene Sets:

Official Gene Set v3.2  
Official Gene Set v1.0

#### 1.2 Consensus Gene Sets comparison:

OGSv3.2 genes that merge OGSv1.0 and RefSeq genes  
OGSv3.2 genes that split OGSv1.0 and RefSeq genes

#### 1.3 Protein Coding Gene Predictions Supported by Biological Evidence:

NCBI Gnomon  
Fgenesh++ with RNASeq training data  
Fgenesh++ without RNASeq training data  
NCBI RefSeq Protein Coding Genes and Low Quality Protein Coding Genes

#### 1.4 *Ab initio* protein coding gene predictions:

Augustus Set 12, Augustus Set 9, Fgenesh, GeneID, N-SCAN, SGP2

#### 1.5 Transcript Sequence Alignment:

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

# Exercises (continued)

## Live Demonstration using the *Apis mellifera* genome.

### 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

# Web Apollo Workshop Instances

Demo 1: [http://genomes.missouri.edu:8080/Amel\\_4.5\\_demo\\_1](http://genomes.missouri.edu:8080/Amel_4.5_demo_1)

Demo 2: [http://genomes.missouri.edu:8080/Amel\\_4.5\\_demo\\_2](http://genomes.missouri.edu:8080/Amel_4.5_demo_2)

Workshop Documentation can be found at:  
Basecamp

Web Apollo instance for *Diaphorina citri*  
<https://apollo.nal.usda.gov/diacit/selectTrack.jsp>

Register for i5K Workspace@NAL at:  
<https://i5k.nal.usda.gov/web-apollo-registration>

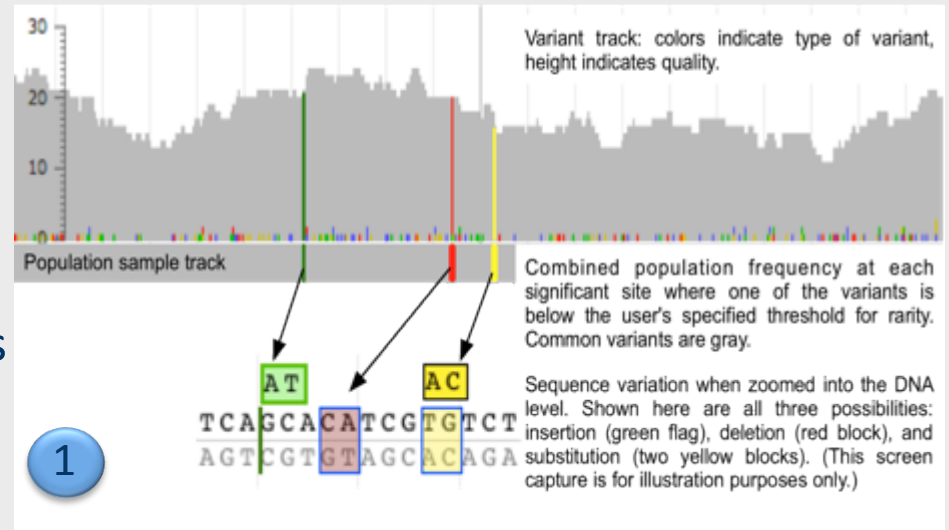
# FUTURE PLANS

## interactive analysis and curation of variants

- ❖ Interactive exploration of VCF files (e.g. from GATK, VAAST) in addition to BAM and GVF.

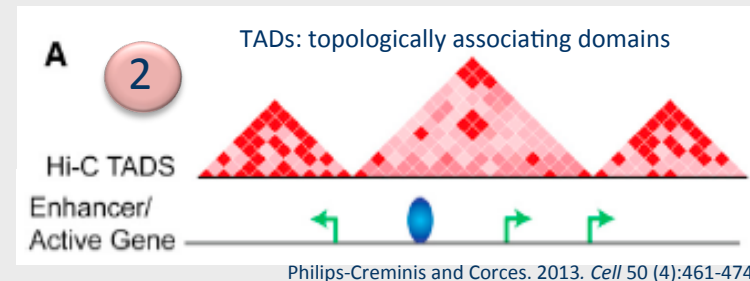
1

Multiple tracks in one: visualization of genetic alterations and population frequency of variants.



- ❖ Clinical applications: analysis of Copy Number Variations for regulatory effects; overlaying display of the regulatory domains.

2



# FUTURE PLANS

## educational tools

---

We are working with educators to make Web Apollo part of their curricula.



In the classroom.

At the lab.



Lecture Series.

Classroom exercises: from genome sequence to hypothesis.

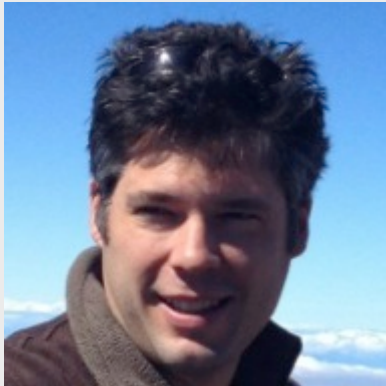


Curation group dedicated to producing education materials for non-model organism communities.

Our team provides online documentation, hands-on training, and rapid response to users.

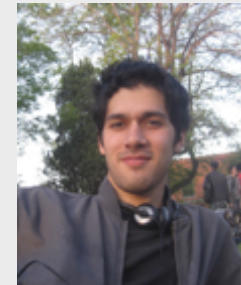
# JOIN US

## Berkeley BOP



Nathan Dunn  
Apollo Technical Lead

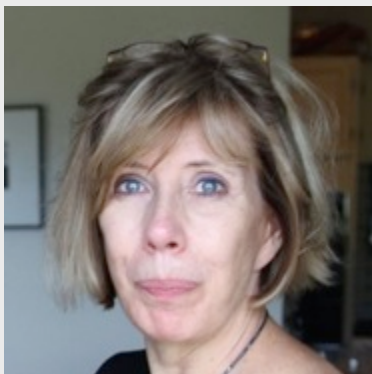
Deepak Unni  
Colin Diesh  
Apollo Developers  
Elsik Lab, University of Missouri



Eric Yao  
JBrowse, UC Berkeley



Suzi Lewis  
Principal Investigator



Please bring your suggestions,  
requests, and contributions to:  
<http://GenomeArchitect.org/>

# Thanks!

## BBOP

### Web Apollo

Nathan Dunn

Colin Diesh §

Deepak Unni §

### Gene Ontology

Chris Mungall

Seth Carbon

Heiko Dietze

---

### JBrowse

Eric Yao \*

---

### NAL at USDA

Monica Poelchau

Christopher Childers

Gary Moore

### HGSC at BCM

fringy Richards

Dan Hughes

Kim Worley

---

Web Apollo: <http://GenomeArchitect.org>

i5K: <http://arthropodgenomes.org/wiki/i5K>

GO: <http://GeneOntology.org>

- **Berkeley Bioinformatics Open-source Projects (BBOP)**, Berkeley Lab: Web Apollo and Gene Ontology teams. *Suzanna E. Lewis (PI)*.
- § *Christine G. Elsik (PI)*. University of Missouri.
- \* *Ian Holmes (PI)*. University of California Berkeley.
- **Arthropod genomics community:** i5K Steering Committee (esp. *Sue Brown* (Kansas State)), *Alexie Papanicolaou* (UWS), BGI, *Oliver Niehuis* (1KITE <http://www.1kite.org/>), and the Honey Bee Genome Sequencing Consortium.
- Apollo is supported by NIH grants 5R01GM080203 from NIGMS, and 5R01HG004483 from NHGRI; by Contract No. 60-8260-4-005 from the National Agricultural Library (NAL) at the United States Department of Agriculture (USDA); and by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.
- Insect images used with permission: <http://AlexanderWild.com>



- **For your attention, thank you!**