Apollo[1]

Contents

1. *Access points: Apollo, Genome Browser and BLAST*
2. *How to identify genes that need to be annotated?*
3. *Gene manual annotations*
4. *Metadata*

1. Access points

Apollo front page

From the tools menu or directly typing its URL.

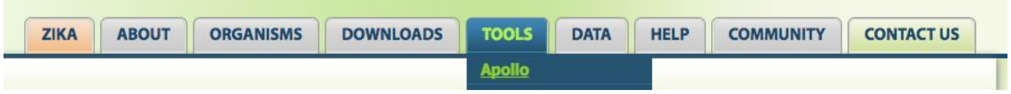
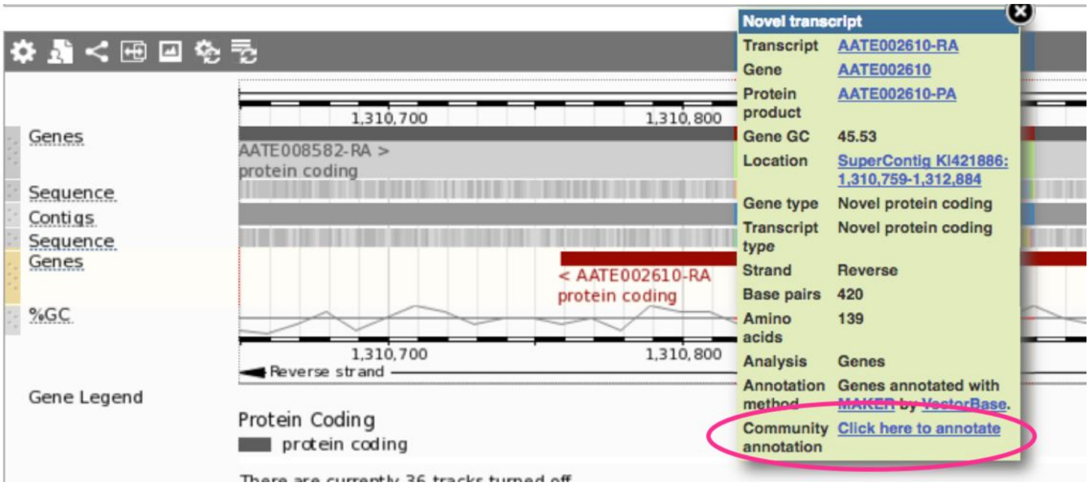


Figure 1. VectorBase navigation menu. Access Apollo from the Tools menu or typing its its https://www.vectorbase.org/apollo

Genome Browser

In the gene, transcript or location display, click on the track of your gene of interest. In the pop out window click on ‘Click here to annotate’, this link will take you to Apollo in the region of the gene of interest. This link is also available from the gene tree as shown in Figure 13.



Novel transcript

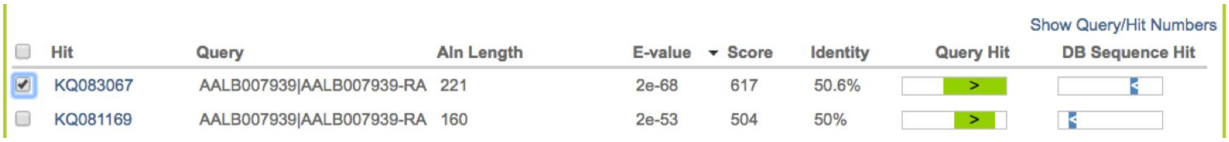
Transcript	AATE002610-RA
Gene	AATE002610
Protein product	AATE002610-PA
Gene GC	45.53
Location	SuperContig KI421886: 1,310,759-1,312,884
Gene type	Novel protein coding
Transcript type	Novel protein coding
Strand	Reverse
Base pairs	420
Amino acids	139
Analysis	Genes
Annotation method	Genes annotated with MAKER by VectorBase.
Community annotation	Click here to annotate

Figure 2. Track of interest (in color) to display the pop-out window with the link to Apollo.

[1] For a more complete Apollo user guide created by the developers of this tool visit: http://genomearchitect.github.io/users-guide/

BLAST

1. Perform a BLAST job (*e.g.*, tBLASTn) and select a hit of interest from the output

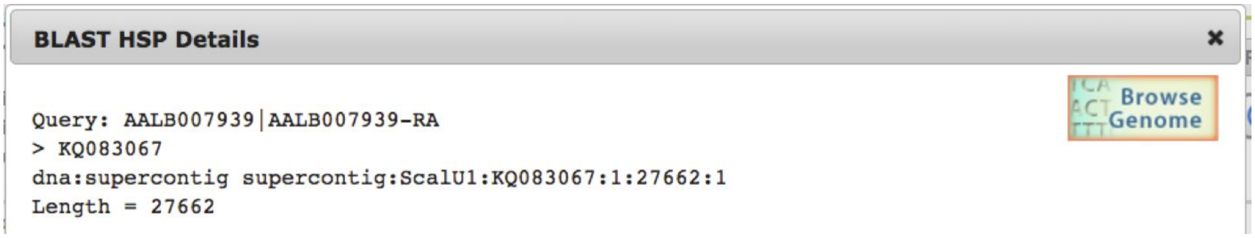


Show Query/Hit Numbers

☐ Hit	Query	Aln Length	E-value	▾ Score	Identity	Query Hit	DB Sequence Hit
☑ KQ083067	AALB007939|AALB007939-RA	221	2e-68	617	50.6%	>	<
☐ KQ081169	AALB007939|AALB007939-RA	160	2e-53	504	50%	>	<

Figure 3.

2. Click on the 'Browse Genome' icon in the top of the pop out window

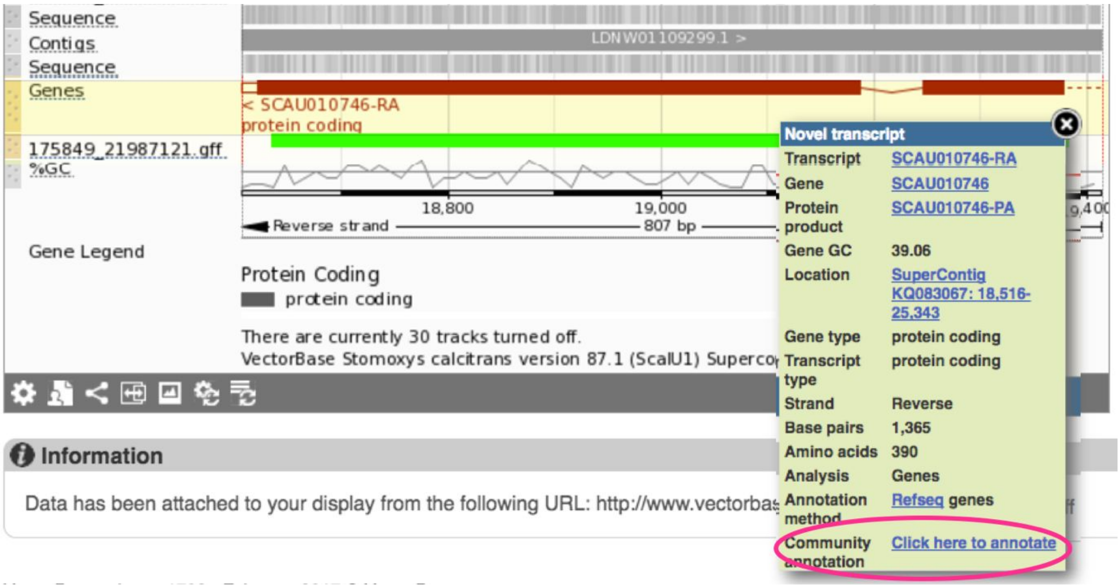


BLAST HSP Details ✖

```
Query: AALB007939|AALB007939-RA
> KQ083067
dna:supercontig supercontig:ScalU1:KQ083067:1:27662:1
Length = 27662
```

Figure 4.

3. In the Genome Browser scroll to the bottom panel. The green track is the BLAST hit. The red track is the gene, click on it. In the pop out window select 'Click here to annotate'. Apollo will open in the region of the gene of interest.



There are currently 30 tracks turned off.
VectorBase Stomoxys calcitrans version 87.1 (ScalU1) Superco

Information

Data has been attached to your display from the following URL: http://www.vectorbas

Novel transcript

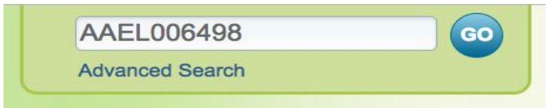
Transcript	SCAU010746-RA
Gene	SCAU010746
Protein product	SCAU010746-PA
Gene GC	39.06
Location	SuperContig KQ083067: 18,516-25,343
Gene type	protein coding
Transcript type	protein coding
Strand	Reverse
Base pairs	1,365
Amino acids	390
Analysis	Genes
Annotation method	Refseq genes
Community annotation	Click here to annotate

Figure 5.

2. How to identify genes that need to be annotated?

To illustrate the use of Apollo we have chosen the opsin genes in mosquitoes as a sample case. VectorBase host 24 mosquito genomes. It is known from the literature that *Anopheles gambiae, Aedes aegypti* and *Culex quinquefasciatus*, have 11, 10 and 13 opsin genes respectively, that had been already manually annotated.

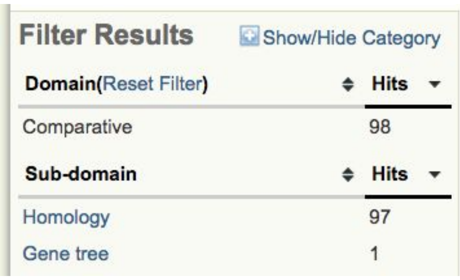
1. Type the opsins gene AAEL006498 (*Ae. aegypti*, GPRop1) in VectorBase Search and click GO



AAEL006498

Figure 6.

2. In the 'Filter Results' box select Comparative (Domain) > Gene tree (Sub-domain). Because there is a single hit.



Filter Results

Domain(Reset Filter)	Hits
Comparative	98

Sub-domain	Hits
Homology	97
Gene tree	1

Figure 7.

3. The landing page will be one with a fully expanded gene tree. The tips of the tree are labelled with the genes with a name (or symbol) or its gene ID, and the species. Notice how different genus or taxonomic groups are grouped by different background colors:
    - *Anopheles*
    - *Aedes* and *Culex*
    - *Lutzomyia* and *Phlebotomus*
    - *Glossina, Musca and Stomoxys*
    - *Drosophila*
    - *Rhodnius, Cimex, Pediculus*, *Ixodes*
    - *Biomphalaria*.

The white blocks with green bars, located at the right of the tree, represent aligned amino acids. The legend at the bottom explains all the image conventions.

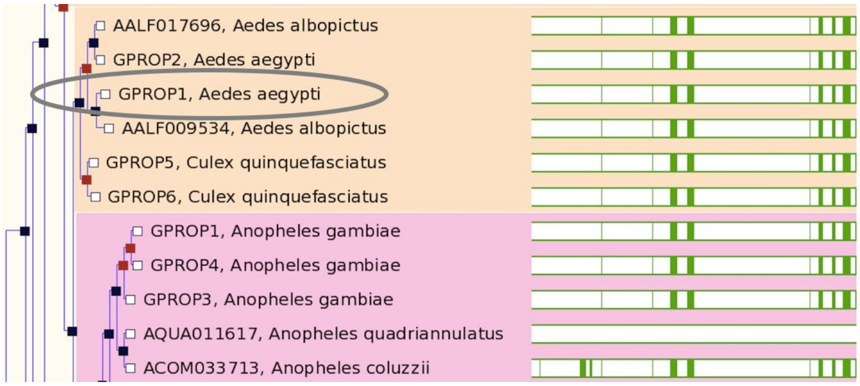

AALF017696, Aedes albopictus
GPROP2, Aedes aegypti
GPROP1, Aedes aegypti
AALF009534, Aedes albopictus
GPROP5, Culex quinquefasciatus
GPROP6, Culex quinquefasciatus
GPROP1, Anopheles gambiae
GPROP4, Anopheles gambiae
GPROP3, Anopheles gambiae
AQUA011617, Anopheles quadriannulatus
ACOM033713, Anopheles coluzzii

Figure 8. Fragment of the gene tree with the gene used as query marked with a gray oval.

4. Which gene models seem wrong or incomplete?

Using well annotated genes, investigate if your genes of interest need manual annotation, with the help of the amino acid alignments next to the gene tree. After an initial visual inspection of the genes that may be wrong or incomplete, confirm your observations downloading the genes sequence and creating a multiple sequence alignment. Look Figures 9 and 10, including the legended, for examples.

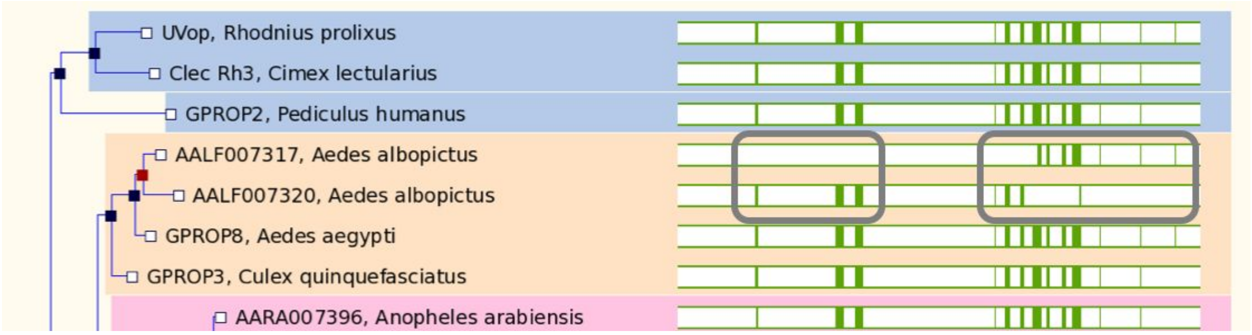

UVop, Rhodnius prolixus
Clec Rh3, Cimex lectularius
GPROP2, Pediculus humanus
AALF007317, Aedes albopictus
AALF007320, Aedes albopictus
GPROP8, Aedes aegypti
GPROP3, Culex quinquefasciatus
AARA007396, Anopheles arabiensis

Figure 9. The amino acid alignment of AALF007317 and AALF007320, when compared with A. aegypti GPRop8 and C. quinquefasciatus GPRop3, seems to indicate that those are two pieces of the same gene, the missing pieces in one are found in the other. From the literature, we know that both AaGPRop8 and CqGPRop3 are ultraviolet opsin genes, and we predict that this new merged gene model is an ultraviolet opsin too.

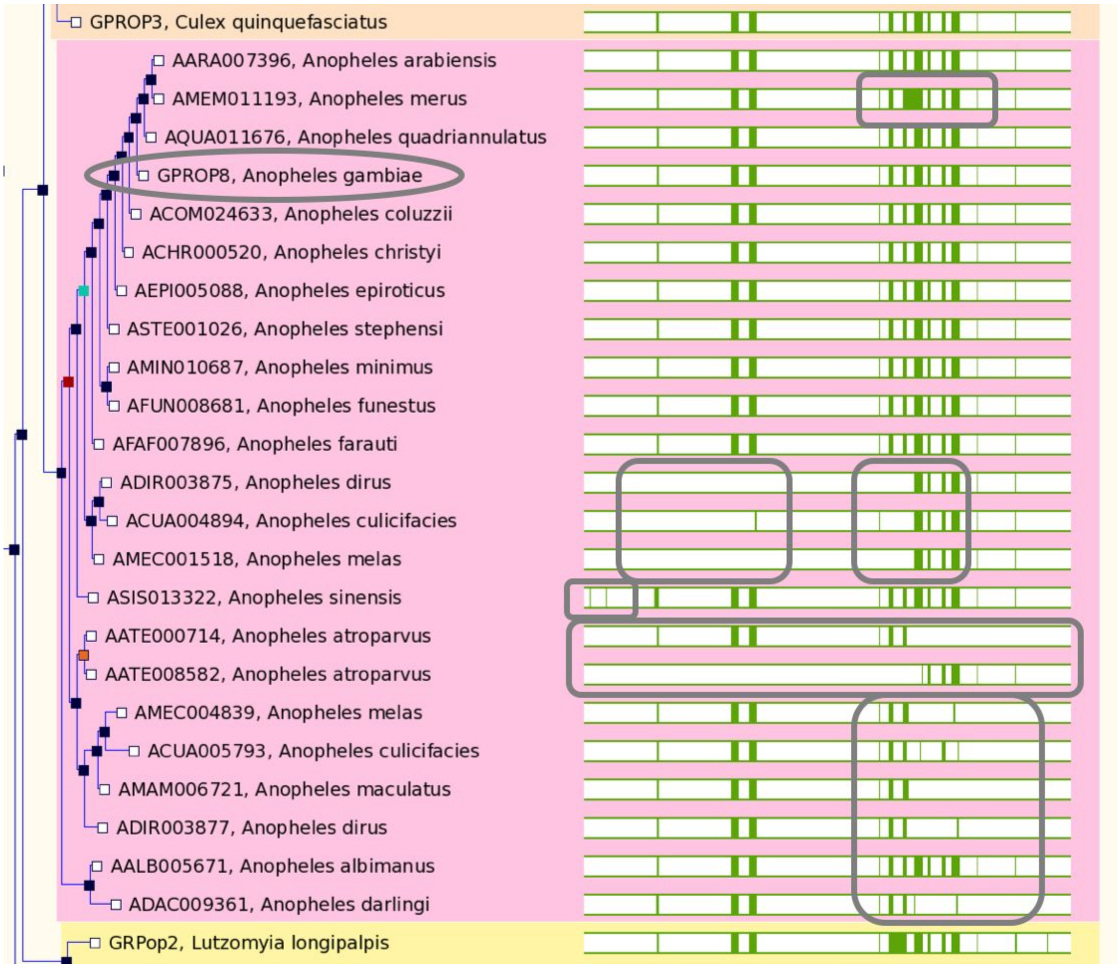

GPROP3, Culex quinquefasciatus
AARA007396, Anopheles arabiensis
AMEM011193, Anopheles merus
AQUA011676, Anopheles quadriannulatus
GPROP8, Anopheles gambiae
ACOM024633, Anopheles coluzzii
ACHR000520, Anopheles christyi
AEPI005088, Anopheles epiroticus
ASTE001026, Anopheles stephensi
AMIN010687, Anopheles minimus
AFUN008681, Anopheles funestus
AFAF007896, Anopheles farauti
ADIR003875, Anopheles dirus
ACUA004894, Anopheles culicifacies
AMEC001518, Anopheles melas
ASIS013322, Anopheles sinensis
AATE000714, Anopheles atroparvus
AATE008582, Anopheles atroparvus
AMEC004839, Anopheles melas
ACUA005793, Anopheles culicifacies
AMAM006721, Anopheles maculatus
ADIR003877, Anopheles dirus
AALB005671, Anopheles albimanus
ADAC009361, Anopheles darlingi
GRPop2, Lutzomyia longipalpis

Figure 10. This clade of the tree has one gene per Anopheles species, the two copies from A. atroparvus seem two fragments of the same gene like in the A. albopictus case above. Because AgGPRop8 is an ultraviolet opsin, we predict that all other Anopheles genes in this clade are ultraviolet too. In addition to the A. atroparvus case, the other rectangles show cases of longer, missing and extra conserved areas (each block could represent more than one exon).

5. Continue doing the same visual inspection of the tree until you identify all the candidate gene models that need to be improved manually

3. Gene manual annotations

Once the list of genes to be manually annotated has been identified, How to submit to VectorBase Apollo the annotations and the metadata (gene symbol and putative function)? This worked example provides the step-by-step process.

1. Use VectorBase Search and type this query for *Anopheles atroparvus*: AATE000714[2]

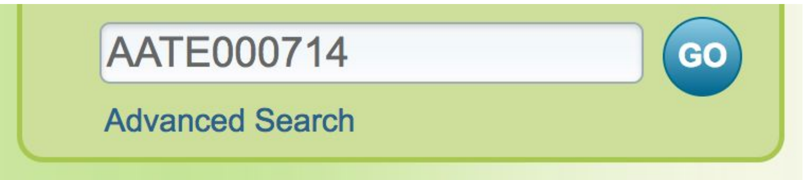


Figure 11.

2. This gene is the top hit, click on it. In the genome browser page click on the 'Gene tree' link.

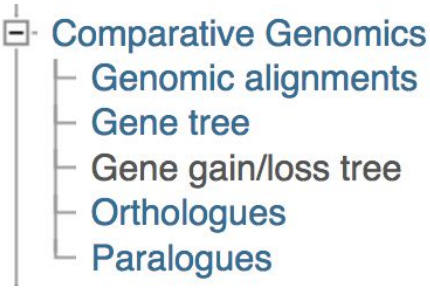
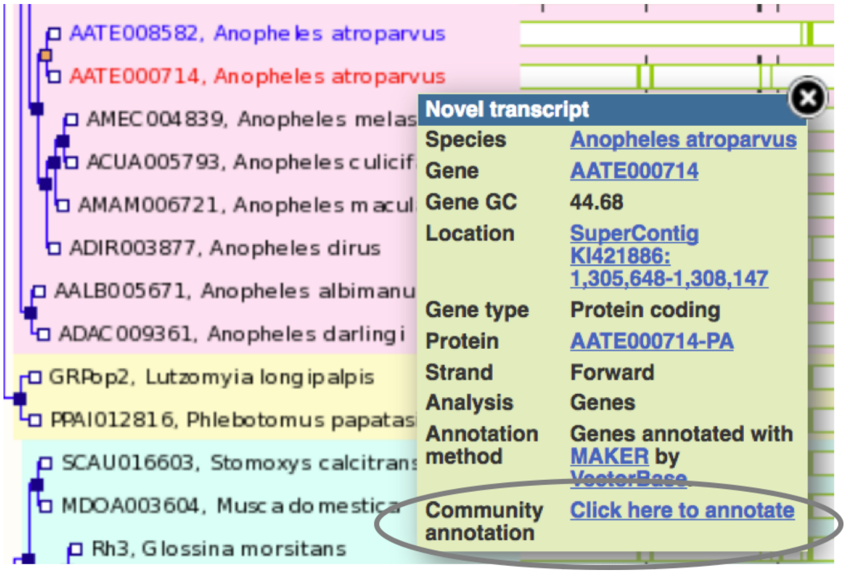


Figure 12.

3. Scroll to the bottom of the page to see the tree. Locate the genes AATE000714 and AATE008582. In the options below the tree click on 'View fully expanded tree'. As mentioned above (Figure 10), these two genes seem to be fragments of a single gene. Click on each, to call their pop-up windows. Both genes are located in supercontig KI421886. In any of the two pop-up windows select 'Click here to annotate'.


AATE008582, Anopheles atroparvus
AATE000714, Anopheles atroparvus
AMEC004839, Anopheles melas
ACUA005793, Anopheles culicif
AMAM006721, Anopheles macul
ADIR003877, Anopheles dirus
AALB005671, Anopheles albimanu
ADAC009361, Anopheles darlingi
GRPop2, Lutzomyia longipalpis
PPAI012816, Phlebotomus papatas
SCAU016603, Stomoxys calcitrans
MDOA003604, Musca domestica
Rh3, Glossina morsitans

Novel transcript

Species	Anopheles atroparvus
Gene	AATE000714
Gene GC	44.68
Location	SuperContig KI421886: 1,305,648-1,308,147
Gene type	Protein coding
Protein	AATE000714-PA
Strand	Forward
Analysis	Genes
Annotation method	Genes annotated with MAKER by VectorBase
Community annotation	Click here to annotate

[2] This example is no longer available to follow along in VectorBase. Starting with June 2017 release, the genes AATE000714 and AATE008582 have been merged in the gene AATE021780.

Figure 13.

4. This is Apollo.

***Note:* Learn how to use Apollo with our tutorial and test (or sandbox) organisms. Once you are proficient with the tool and are ready to annotate please request one account to VectorBase help desk sending an email to info@vectorbase.org.**

5. First, log in. Click on 'Tracks' and type any letter in the Search box to visualize all available tracks below.

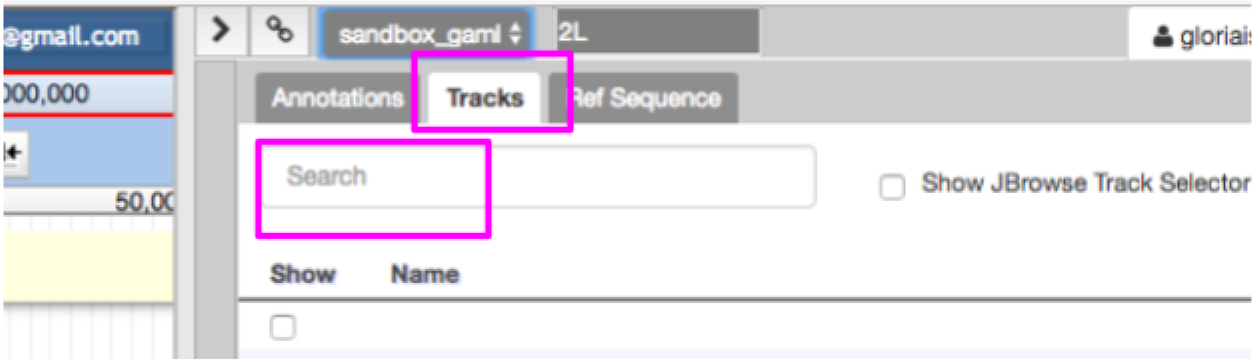


Figure 14.

What are the available data sets for your species of interest? Or for one of the two sandboxes available? *e.g.*, protein alignment, sequence gaps, BAM files.

Select VectorBase automatic gene models, *e.g.*, for *Anopheles gambiae* is AgamP4.4.

6. <u>The instructors will give you a live demo</u>.

In *Anopheles atroparvus* Apollo zoom out until you see the complete gene models for both AATE000714 and AATE008582. Drag them to the 'User-created Annotations' area (select them by any exons to move the whole model).

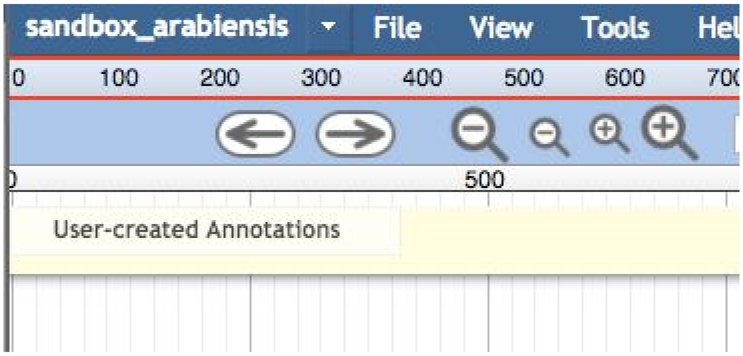


Figure 15.

Remember that based on our previous observations, these two genes seem to be a single gene model. We know from the literature that *A. gambiae* GPRop8 (AGAP006126), in located in the same tree clade and is well annotated. Identify this gene in Apollo *A. gambiae* protein alignment track.

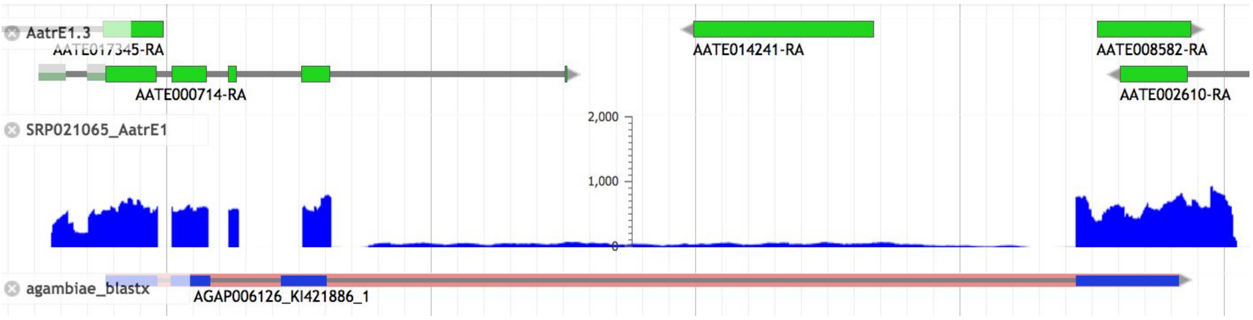


Figure 16.

Based on what we have seen in the tree for all the other *Anopheles* species, in the GPRop8 gene seen here and the RNAseq evidence, we decide to merge the two pieces in a gene.

7. To merge the gene fragments, select both (with the Shift key) and right click. In the pop up window select 'merge'.

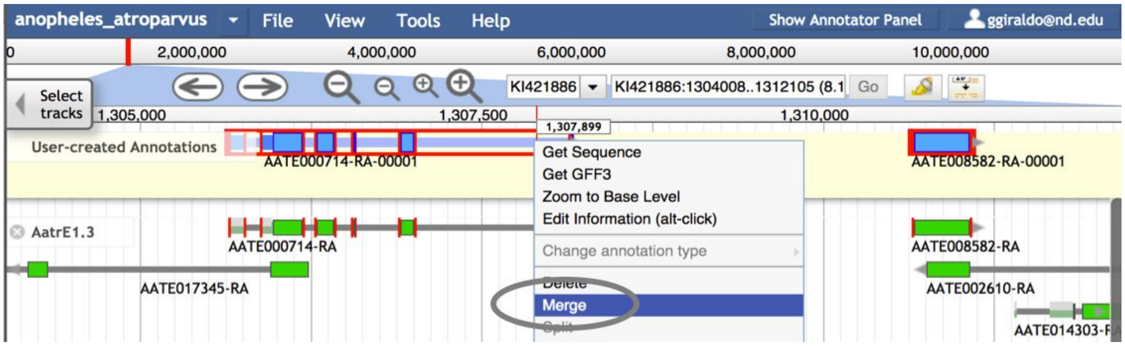


Figure 17.

8. Merging the models recalculates the longest open reading frame, or ORF. In this case, the change in coding sequence obtained needs new splice sites, as flagged here by the exclamation marks.

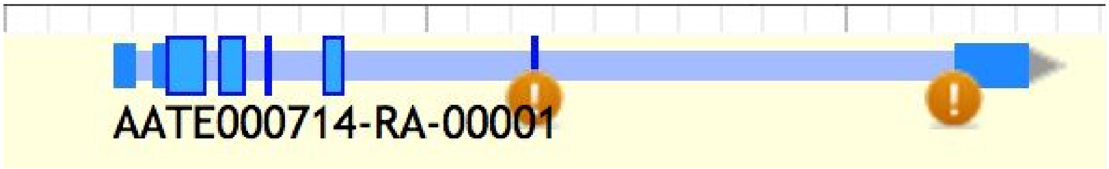


Figure 18.

9. Use the available expression evidence, RNAseq and BAM tracks, to confirm if the gene has all the expected exons. In this case, it seems the small exon labeled with the exclamation mark can be deleted. Select only that exon, open the pop-up window with the right click and select 'Delete'.

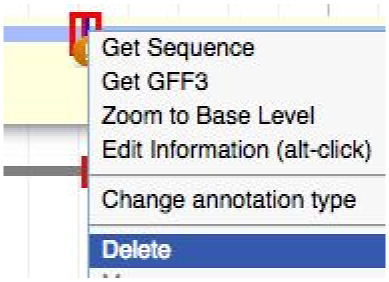
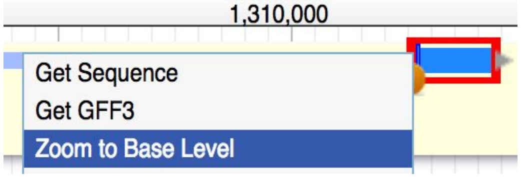


Figure 19.

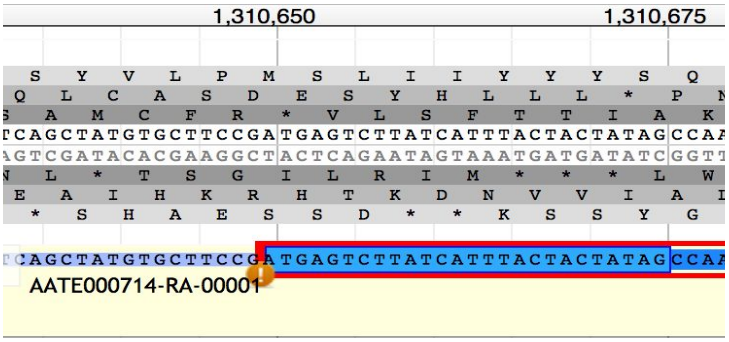
10. How to correctly annotate the intron/exon boundaries?



Right click and select the last exon (labeled with the exclamation mark). Zoom to the base level.

Figure 20.

Notice the nucleotides at the splice site



```
1,310,650                                   1,310,675
S  Y  V  L  P  M  S  L  I  I  Y  Y  Y  S  Q
Q  L  C  A  S  D  E  S  Y  H  L  L  L  *  P  N
S  A  M  C  F  R  *  V  L  S  F  T  T  I  A  K
TCAGCTATGTGCTTCCGATGAGTCTTATCATTTACTACTATAGCCAA
AGTCGATACACGAAGGCTACTCAGAATAGTAAATGATGATATCGGTT
N  L  *  T  S  G  I  L  R  I  M  *  *  *  L  W
E  A  I  H  K  R  H  T  K  D  N  V  V  I  A  I
*  S  H  A  E  S  S  D  *  *  K  S  S  Y  G
TCAGCTATGTGCTTCCGATGAGTCTTATCATTTACTACTATAGCCAA
AATE000714-RA-00001
```

Figure 21.

For your convenience, this drawing shows the nomenclature and the most common nucleotides at the splice sites.

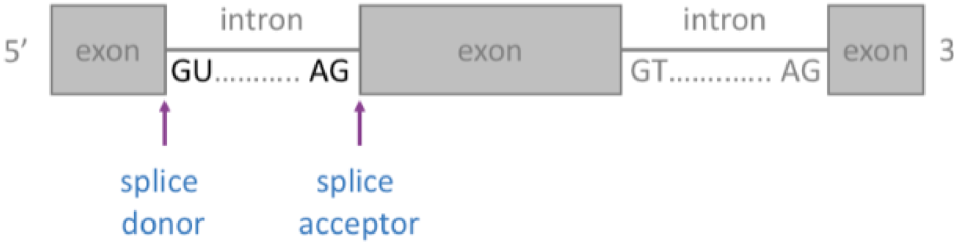
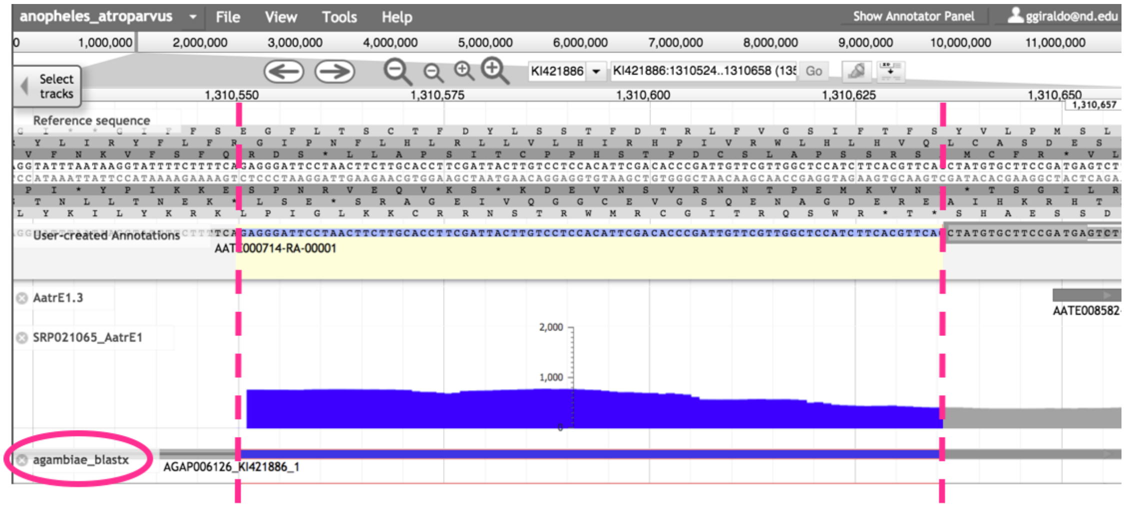


Figure 22.

Based on the RNAseq evidence and the homology with the *A. gambiae*, *A. atroparvus* gene is missing the piece marked with the dotted lines. Turn off *A. gambiae* proteins with a click in the 'x' of the track (we will bring this track back soon).



KI421886:1310524..1310658 (135

```
GGTATTTAATAAGGTATTTTCTTTTCAGAGGGATTCCTAACTTCTTGCACCTTCGATTACTTGTCCTCCACATTCGACACCCGATTGTTCGTTGGCTCCATCTTCACGTTCACCTATGTGCTTCCGATGAGTCT
CCATAAATTATTCCATAAAAGAAAAGTCTCCCTAAGGATTGAAGAACGTGGAAGCTAATGAACAGGAGGTGTAAGCTGTGGGCTAACAAGCAACCGAGGTAGAAGTGCAAGTCGATACACGAAGGCTACTCAGA
TCAGAGGGATTCCTAACTTCTTGCACCTTCGATTACTTGTCCTCCACATTCGACACCCGATTGTTCGTTGGCTCCATCTTCACGTTCACCTATGTGCTTCCGATGAGTCT
AATE000714-RA-00001
```

Figure 23.

Select one of the BAM tracks (green) that more closely resembles the RNAseq tracks (blue). With a right click open the menu and select 'Set as 5' end'.

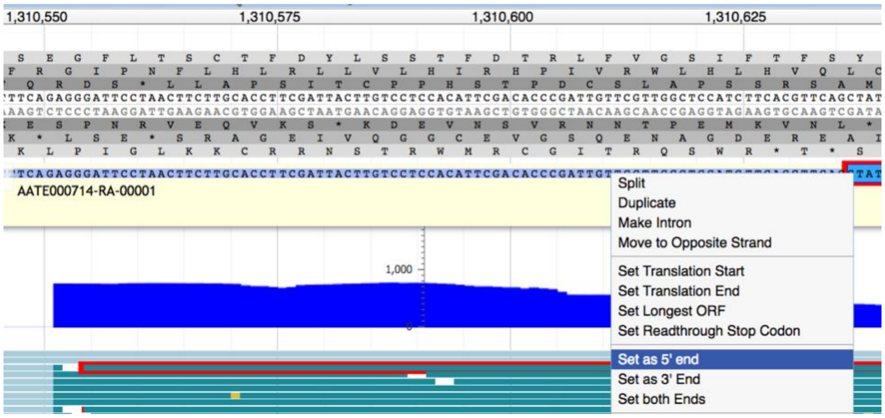


Figure 24.

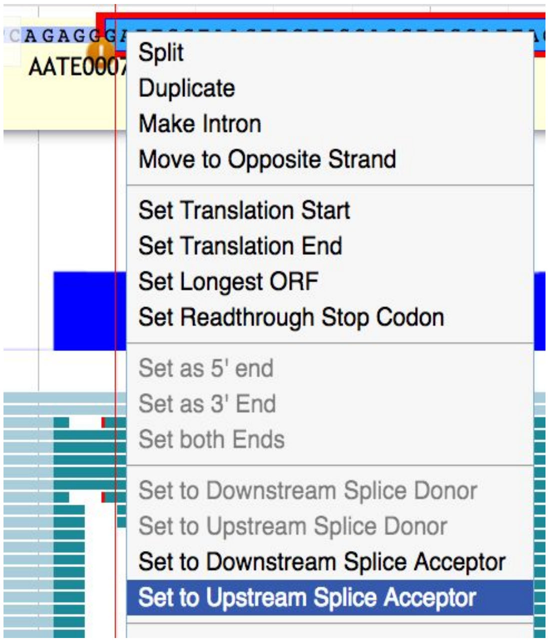


Figure 25.

Again, with a right click open the menu and go to the Splice options. In this specific case, by default, only the 'Splice Acceptor' is available.

Now, will you set it to upstream or downstream? Based on the homology evidence (with *A. gambiae* gene) and the RNAseq transcript evidence it should be upstream. Select that option.

Click one more time in, 'Set to Upstream Splice Acceptor', that sets the beginning of the exon to match with the expression evidence.

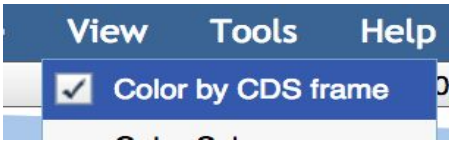


Figure 26.

In the View menu select 'Color by frame'. Go back to 'Select tracks' (Figure 16) and select again *A. gambiae* proteins. Where is the start codon of the gene?

Notice this gene model has 5'UTR based on RNAseq evidence, we are not going to edit that. The frame for the first coding exon is the forward one, indicated by the nucleotides in black (the reverse ones are turn gray). The ORF is the top one in purple, matching the color of the nucleotides in the 'User-created Annotations'.

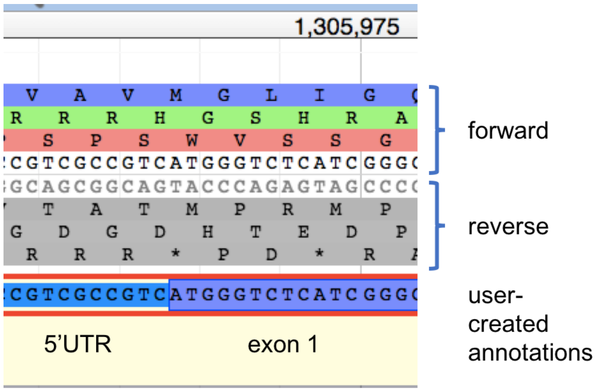


Figure 27.

Finally, with a right click open the menu and select 'Edit information'.

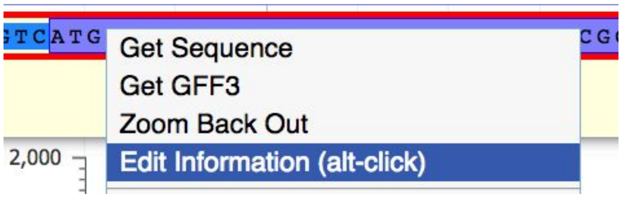
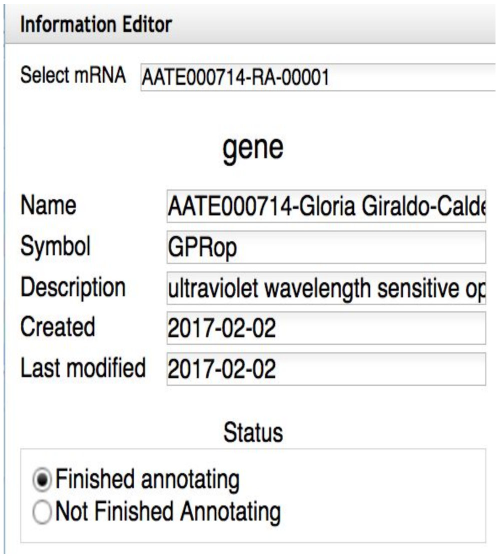


Figure 29.

Filling the gene metadata in the 'Information Editor' is optional and for your own records only, please look at the next section for the recommended metadata submission formats.



Information Editor

Select mRNA AATE000714-RA-00001

gene

Name	AATE000714-Gloria Giraldo-Calde
Symbol	GPRop
Description	ultraviolet wavelength sensitive op
Created	2017-02-02
Last modified	2017-02-02

Status

● Finished annotating
○ Not Finished Annotating

Figure 30.

It you cannot finish a gene in a single working session, just come to the Information Editor and label your gene as 'not finishing annotating'. Finished and in progress annotations are visible to all Web Apollo users.

Once done, your name should be displayed next to the VectorBase gene ID

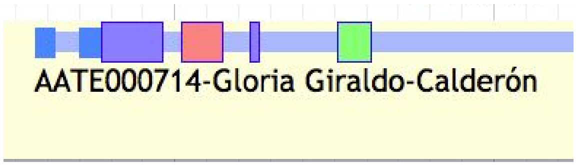


Figure 31.

These new gene models will become part of VectorBase new gene set once a sufficient number of changes has been accumulated and VectorBase curator determines to generate a new gene set for the species[3].

4. Metadata

Metadata are the gene symbol (or name), *e.g*., srp7 or rpl2 and gene description (or function), *e.g*., serpin 7 or ribosomal protein L2. There are two options to submit this data to VectorBase:

- For one or a few gene follow this link and utilize the "Gene Information Capture" form: https://www.vectorbase.org/content/gene-metadata-form
- If you are submitting in batch, e.g., when you are annotating a gene family, please follow our FAQ page here: https://www.vectorbase.org/faqs/how-do-i-batch-submit-annotation-metadata-vectorbase

[3] The gene set pages will provide the information about 'community annotation patch builds' as shown here, https://www.vectorbase.org/organisms/anopheles-atroparvus/ebro/aatre15