WebApollo: A WEB-BASED SEQUENCE ANNOTATION EDITOR FOR COMMUNITY ANNOTATION
User Guide

Gregg Helt[1], Ed Lee[1], Justin T Reese[2], Christopher P Childers[2], Monica C Munoz-Torres[1], Robert Buels[3], Ian Holmes[3], Christine G Elsik[2,4], Suzanna E Lewis[1].

1 Bioinformatics Open-source Projects, Lawrence Berkeley National Laboratories. 2 Division of Animal Sciences, University of Missouri. 3 Department of Bioengineering, University of California at Berkeley. 4. Division of Plant Sciences, University of Missouri

This guide allows users to:

- Become familiar with the environment of the WebApollo annotation tool.
- Understand WebApollo's functionality for the process of manual annotation.
- Learn to corroborate and modify computationally predicted gene models using all available gene predictions and biological evidence using WebApollo.

CONTENTS

Page

I. GENERAL INFORMATION

II. GETTING STARTED: LOGGING IN AND NAVIGATION

III. ANNOTATION

I. GENERAL INFORMATION

1. General Process of Manual Annotation

The major steps of manual annotation are 1) locate a chromosomal region of interest, 2) determine whether a feature in an existing evidence track will provide a reasonable gene model to start working with, 3) drag the selected feature to the User Annotation area, creating an initial gene model, 4) use editing functions to edit the gene model if necessary, 5) check your edited gene model for consistency with existing homologs by exporting the fasta formatted sequence and searching a protein sequence database, such as UniProt or NCBI NR. When annotating gene models using WebApollo, remember that you are looking at a 'frozen' version of the genome assembly and you will not be able to modify the assembly itself.

2. Evidence Provided in this Demo

2.1 Evidence that supports a protein coding gene model

2.1.1 Consensus Gene Set:

GLEAN

2.1.2 Protein Coding Gene Prediction Supported by Biological Evidence:

NCBI
Ensembl
fgeneshpp

2.1.3 Ab initio protein coding gene prediction:

fgenesh
geneid
sgp

2.1.4 Transcript Sequence Alignment:

EST spliced
EST not spliced
cDNA spliced
cDNA not spliced
RNA-Seq SRR072810
RNA-Seq SRR072811
RNA-Seq SRR072812
RNA-Seq SRR072813
Illumina Reads 1a
Illumina Reads 2a
Illumina Reads 3a
Illumina Reads 4a
Illumina Reads 5a

2.1.5 Protein homolog alignment:

Swissprot exonerate
VEGA exonerate
human Ensembl exonerate
human RefSeq exonerate
mouse Ensembl exonerate
mouse RefSeq exonerate
dog Ensembl exonerate
rat RefSeq exonerate
zebrafish Ensembl exonerate
zebrafish RefSeq exonerate

2.2 Evidence suggesting a region is not a protein coding gene

2.2.1 Non-protein coding gene prediction:

Ensembl miRNA
Ensembl snoRNA
Ensembl snRNA

2.2.2 Pseudogene prediction:

NCBI pseudogene
Ensembl pseudogene
EST pseudogene
cDNA pseudogene

2.2.3 Repetitive DNA:

RepeatMasker

II. NAVIGATION

3. Initial reconnaissance and adjustments.

3.1 Log in

To begin annotating a gene, log in to WebApollo at,
http://icebox.lbl.gov:8080/WebApolloDemo

Username | Password: demo | demo

You will be directed to the *'Select Track'* page.

3.1.1 Select a Scaffold to Work On

If you know the identifier of the scaffold you wish to work on, find it in the scaffold id list. You may use your browser's *'find'* function to quickly find a scaffold id in a long list if you enter the exact scaffold id. Once you find the scaffold, click the *'Edit'* button next to it; the WebApollo Main Window will open in a new browser tab.

3.1.2 Search for a specific sequence:

If you do not know the scaffold id, but have a transcript or protein homolog sequence related to your gene of interest, you may use the *'Search Sequence'* feature to run a BLAT (BLAST-Like Alignment Tool) search of the assembled genome and determine the existence of a gene model prediction that is putatively homologous to your gene of interest. Click the Tools item on the WebApollo menubar, and select *'Sequence Search'* from the dropdown choices. Choose to run a Protein or Nucleotide BLAT search from the drop down menu, and paste the string of residues to be used as query. Check the box labeled '*Search all genomic sequences*' to search against the entire genome. The existence of paralogs may cause your query to match more than one scaffold or genomic range. Select the desired genomic range to be displayed in the WebApollo Main Window.

3.2 The WebApollo Main Window:

Fig. 1 highlights the main elements of the WebApollo annotation window. The *'Navigation panel'* at the top of the window holds the controls for localization within a chromosome or scaffold (group), or to move to a different one. The 'Available Tracks' panel runs along the length of the left hand side of the window. All gene predictions and evidence can be visualized in the *'Evidence pane'*. The light yellow stripe on top is the *'User Annotations'* area, where users will drag the gene models/exons to be modified. All transactions performed on the *'User Annotations'* area can be reversed with the *'Undo'* option from the right/apple-click menu, displayed over any part of the annotation in progress. A *'Redo'* option is also available.

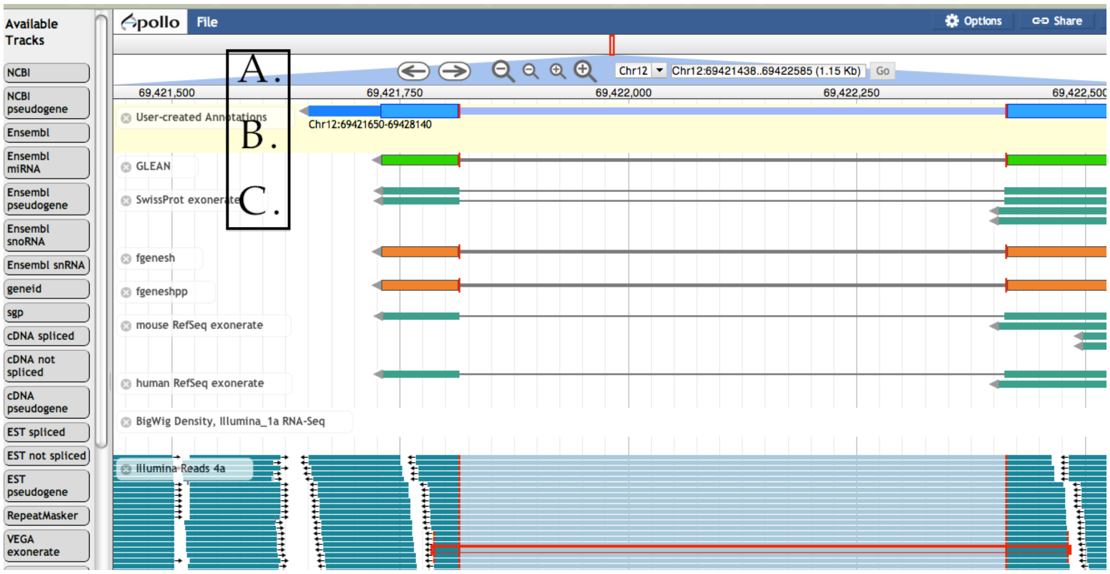

Figure 1. WebApollo Annotation Editor overview. A view showing an annotation in progress. The main interface is similar to JBrowse, with available tracks displayed as a set of tiles along the left side of the main pane. To turn a track on or off, click and drag the track title into or out of the main viewing pane. **A.** The navigation panel runs along the top of the main pane, and includes buttons to pan left and right and two levels of zooming. The dropdown box is used to select the sequence for annotation, and the textbox is used to manually enter the coordinates for viewing. **B.** The User-created Annotations panel contains the manual annotations. **C.** The Evidence panel. As in JBrowse, this panel contains the evidence tracks. Annotators create annotations by first selecting and dragging a model from the Evidence panel to the User-created Annotations panel.

III. ANNOTATION

4. Annotating a gene.

4.1 Initiating an annotation

If you have not already performed a BLAT search to identify your gene of interest (see section 4.3.1), you may do so at this point using the *'Sequence search'* feature from the right/apple-click menu over the empty *'User Annotations'* area, or you may simply navigate along the scaffold using the navigation arrows. Your gene of interest may appear on the forward (sense) or reverse (anti-sense) strand. Predictions are labeled with clear identifiers, and users may retrieve additional information by selecting the entire model and using the right/apple menu to select the *'Information'* feature.

After you have located the gene of interest, scroll through the different tracks of gene predictions and choose one that you think most closely reflects the structure of the actual gene. You may base your decision on prior knowledge of the reliability of each gene prediction track (e.g., select an evidence based gene model instead of an *ab initio* gene prediction). Alternatively, you may compare the gene prediction tracks to a BLAT alignment or other aligned data (e.g.: alignments of protein homologs, cDNAs and, RNAseq reads). To highlight your preferred gene model, either click on an intron or double click on an exon, so that the whole gene model. Once the gene model is highlighted, drag it to the *'User Annotations'* area.

At this point you may wish to download the protein sequence (see section 4.3.1) so that you may search a protein database to help you determine if you made a wise choice in your gene model selection. For example, you may perform a protein sequence search of UniProt or the NCBI Non Redundant (NR) database. If you have prior knowledge of protein domain content, may perform a protein domain search of the InterPro databases to verify that your selected gene model codes for the expected domains. If further investigation suggests that you have not selected the best gene model to start working with, you may delete it by highlighting it (as above); then right click the gene model to access a menu that provides a *'delete'* function.

Once you are satisfied with the gene model you selected as the best starting point for annotation, you must decide whether it requires modification. You may already know the answer based on a previous protein or domain database search. Scroll down the evidence tracks to see if splice sites in transcript alignments agree with your selected gene model, or if evidence suggests addition or modification of an exon. Transcript (cDNA/EST) alignments that are significantly than the gene model may indicate additional coding sequence or untranslated regions. Keep in mind that transcript alignments may be shorter than the gene model due to the fragmented nature of transcript sequencing. Similarly, protein alignments may not reflect the entire length of the coding region, because divergent regions may not align, resulting in a short protein alignment or one with gaps. Protein and transcript alignments in regions with tandem closely related genes may also be problematic, with alignment in part to one gene and then skipping over to align the rest to a second gene.

4.2 Simple Cases:

In our definition of "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.

4.2.2 Adding UTRs:

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. First, position the cursor at the beginning of the exon that needs to be extended and right click to show the options on the menu and choose to 'Zoom to base level'. Place the cursor over the edge of the exon (5' or 3' end exon as needed) until it becomes a black arrow (Fig. 2) 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, as detailed in section 4.3.3.

4.2.3 Exon structure integrity:

Zoom in sufficiently to clearly resolve each exon as a distinct rectangle. When two exons from different tracks share the same start and/or end coordinates, users will see a red bar appear at the edge of the exon. Use this '*edge-matching'* function by either selecting the whole annotation or one exon at a time. Scrolling along the length of the annotation, exon boundaries may be verified against available EST data. Also note if there are ESTs that lack one or more of the annotated exons or include additional exons.

To change an exon boundary that needs to be corrected to match data in the evidence tracks *zoom to the base pair level*, click on the exon to select it and place the cursor over the edge of the exon. When the cursor changes to an arrow, drag the edge of the exon to the desired new coordinates.

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 alternatives transcripts as are well supported by the data.

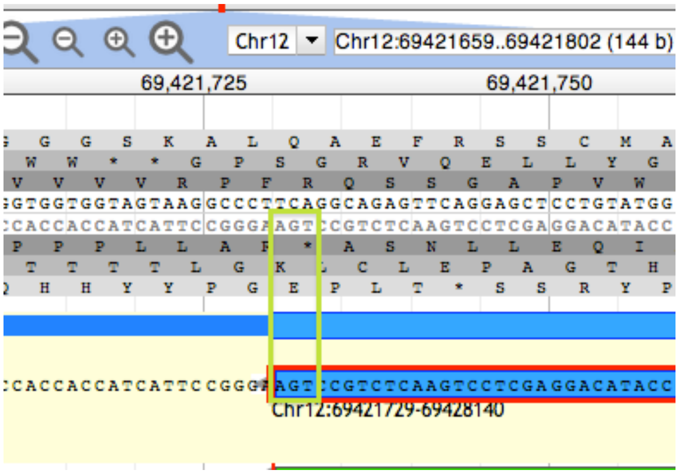
Chr12
Chr12:69421659..69421802 (144 b)
69,421,725
69,421,750
GGTGGTGGTAGTAAGGCCCTTCAGGCAGAGTTCAGGAGCTCCTGTATGG
CACCACCATCATTCCGGGAAGTCCGTCTCAAGTCCTCGAGGACATACC
Chr12:69421729-69428140

Figure 2. 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.

4.2.4 Splice sites:

Non-canonical splices (sites other than …]5'-GT/AG-3'[…),will be indicated by an orange circle with a white exclamation point inside, placed over the edge of the offending exon. If you have added alternative transcripts, you should make a second pass to verify that all changes were saved correctly.

If a non-canonical splice site is present, zoom to base level to review it. These do not necessarily need to be corrected, but should be flagged with the appropriate comment. ('*Adding a Comment'* is addressed in Section 5). You may have prior knowledge about the organism that will help you decide whether a predicted non-canonical splice site is likely to be real. For example, GC splice donors have been observed in many organisms, but less frequently than GT splice donors. WebApollo flags GC splice donors as non-canonical. To further complicate the problem, splice sites that are non-canonical, but found in nature, such as GC donors, may not be recognized by some gene prediction algorithms. For example, a gene prediction algorithm that does not recognize GC splice donors may have ignored a true GC donor and selected 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. To aid your decision as to whether you should modify a splice site, you may download translated sequences and use them to search known protein databases, such as UniProt, to see if you can resolve the question using protein alignments. Incorrect splice sites would likely cause gaps in the alignments. If there does not appear to be any way to resolve the non-canonical splice, leave it and add a comment.

4.2.5 'Start' and 'Stop' sites:

By default, WebApollo will calculate the longest possible open reading frame (ORF) that includes canonical *Start* and *Stop* signals within the predicted exons. To check for accuracy of start and stop signals, you may align the translated sequence to a known protein database, such as UniProt, to determine whether the ends of the protein sequence corresponds with those of known proteins.

If it appears that WebApollo did not calculate the correct *Start* signal, you may modify it. To set the *Start* codon manually, position the cursor over the first nucleotide of the candidate *Start* codon and select the *'Set translation start'* feature from the right/apple-click, menu. Depending on evidence from a protein database search or additional evidence tracks, you may wish to select an in-frame start codon further up or downstream. An upstream start codon may be present outside the predicted gene model, within a region supported by another evidence track. See section 4.3.4 on adding an exon.

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 comment (using the transcript comment section in the '*Comments'* option).

In rare cases, the actual *Start* codon may be non-canonical (non-ATG). Check whether a non-canonical *Start* codon is usually present in homologs of this gene, and/or check whether this is a likely occurrence in this organism. If appropriate, you may override the predicted *Start* by setting the it manually to a non-canonical *Start* codon, choosing the one that most closely reflects what you know about the protein, and has the best support from the biological evidence tracks. Add the appropriate comment (using the transcript comment section in the '*Comments'* option).

In some cases, a stop codon may not be automatically be identified. Check to see if there are data supporting 3' extension of the terminal exon or additional 3' exons with valid splice sites. See section 4.3.3 on adding exons. Each time you add an exon region, whether by extension or adding an exon, WebApollo recalculates the ORF to identity start and stop signals, allowing you to determine whether a stop codon has been incorporated after each editing step.

4.2.6 Predicted protein product(s):

If any of your manipulations have thrown an exon out of frame, or caused other drastic changes to the translated sequence, WebApollo will warn you by changing the display of from a light-blue protein-coding stretch to a truncated model shown as a darker blue, narrower rectangle.

If the annotation looks good, obtain the protein sequence (see section 4.3.1) and use it to search a protein database, such as UniProt or NCBI NR. Keep in mind that the best Blast hit may be the exact prediction from which you initiated your annotation (e.g. the RefSeq predicted protein from your organism). You should not consider the identical protein from your organism as external evidence supporting the annotation. Instead, look at alignments to proteins from other organisms.

4.3 Additional functionality:

4.3.1 Get sequences:

Select one or more exons, or the entire gene model of interest as needed, retrieve the right/apple-click menu to select the *'Get sequence'* feature. Chose from the options to obtain the protein, cDNA, CDS or genomic sequences.

4.3.2 Merge exons / transcripts:

Select each of the joining exons while holding down the shift key, open the right/apple-click menu and select the '*Merge'* feature.

4.3.3 Add an exon:

You may select and drag the putative new exon from a track in the *'Evidence Pane'*, and add it directly to an annotated transcript in the *'User Annotations'* area. Click the exon and, holding your finger on the mouse button, drag the cursor until it touches the receiving transcript. The receiving transcript will be highlighted in dark green when it is okay to release the mouse button. Once you release the mouse button, the additional exon becomes attached to the receiving transcript. If the receiving transcript is on the opposite strand of the strand on from which you selected the new exon, a dialog box will ask you to confirm.

As described before, WebApollo dynamically recalculates the longest ORF for the model, so you must check whether adding one or more exons disrupts the reading frame, inserts premature *Stop* signals, etc.

4.3.4 Make an intron / split an exon:

Click once on the exon of interest, and select the *'Make intron'* feature from the right/apple-click menu. WebApollo will attempt to identify splice sites such the reading frame is maintained.

4.3.5 Delete an exon:

Select the exon using a single click (a double click will select the whole transcript). Bring up the right/apple-click menu and select '*Delete'*. You must check whether deleting one or more exons disrupts the reading frame, inserts premature *Stop* signals, etc.

4.3.6 Flip the strand of annotation:

Sometimes transcript alignments appear on the strand opposite of the actual coding strand, particularly when the transcript alignment does not encompass a splice junction, making it difficult to determine the coding direction. If you would have used one of these alignments to initiate an annotation and then determined that the annotation is on the incorrect strand, You may use the *'Flip strand'* option from the right/apple-click menu to reverse the orientation of the annotation. As mentioned before, annotators should always reassess the integrity of the translation after modifying an annotation.

4.4 Complex Cases:

4.4.1 Merge two gene predictions on the same scaffold:

Evidence may support the merge of two different gene models. Identify two gene models from the '*Evidence Pane*' that you would like to merge. A protein alignment may not be a useful starting point because it may have incorrect splice sites and may lack non-conserved regions.

Drag and drop each selected gene model to the '*User annotations*' area. You may select the supporting evidence tracks and drag them over the candidate models to corroborate the overlap, and can also zoom in to carefully review edge matching (Figure 3) and coverage across models. Once you are sure you would like to continue with the merge, shift click on an intron from each gene model to highlight both. Then right click and select merge from the drop down menu. You should obtain the resulting translation, and check it by searching a protein database, such as UniProt. Be sure to record the IDs of both starting gene models in the *'Comments'* boxes, and use the appropriate canned comments to record that this annotation is the result of a merge.

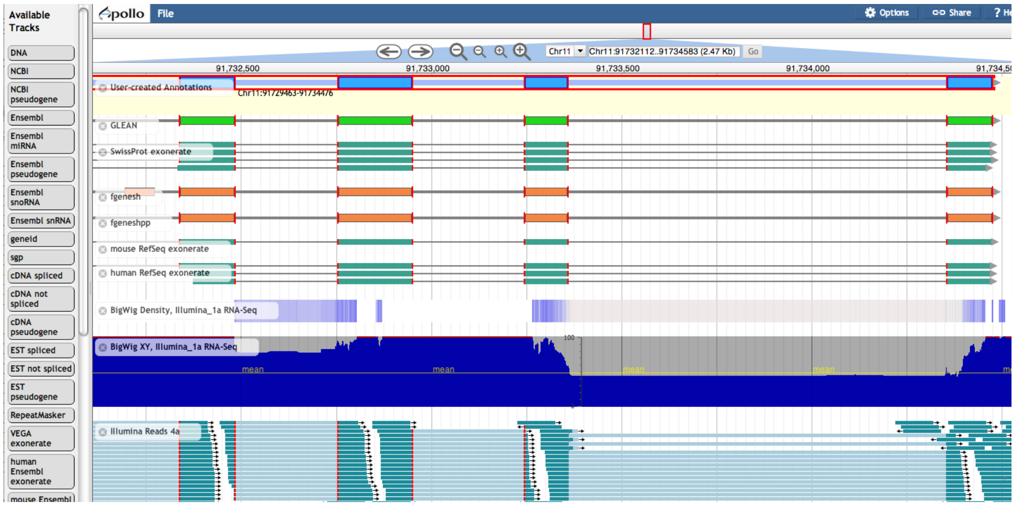

Figure 3. Edge matching in WebApollo. When a feature is selected, the exon edges are marked with a red box. All other features that share the same exon boundaries are marked with a red line on the matching edge. This feature allows annotators to confirm that evidence is in agreement without examining each exon at the base level.

4.4.2 Merge two gene predictions on different scaffolds:

It is not possible to merge two annotations across scaffolds, however annotators should document the fact that the data support a merge in the *'Comments'* section for both components. For standardization purposes, please use the following two canned comments:
"RESULT OF: merging two or more gene models across scaffolds"
"RESULT OF: merging two or more gene models. Gene models involved in merge:"

4.4.3 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 (in this case, verify that it is not a case of alternative transcripts). A split can be created in one of two ways: 1) select the flanking exons using the right/apple-click menu option *'Split'*, or 2) annotate each resulting fragment independently. You should obtain the resulting translation, and check it by searching a protein database, such as UniProt. Be sure to record the original ID for both annotations in the *'Comments'* section.

4.4.4 Frameshifts, single-base errors, and selenocysteines:

WebApollo 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. Changes are made on the DNA track with the right/apple-click menu. Click over a single nucleotide to bring up a menu with options for introducing sequence changes.

If you determine that you need to make one of these changes, zoom in to the nucleotide level and right click on the genomic sequence to access a menu that provides options for creating insertions, deletions or substitutions.

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.

Once you have entered the modifications, WebApollo will recalculate the corrected transcript and protein sequences, which will appear when you use the right/apple-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.

In special cases such as selenocysteine read-throughs, drag the edge of the exon over the position of the prematurely predicted *Stop* signal and add a comment to the transcript's *'Comments'* section. Note that WebApollo does not automatically add the remaining amino acids to the resulting sequence.

5. Adding additional information to your annotations

After you are satisfied with your annotation, you may provide additional information in the form *Comments*. For example, it may be useful to the database curators if you save the ID of the gene prediction that you used to initiate the annotation. Functional information obtained from homologs may also be useful, e.g. homolog ID, description, gene name, gene symbol. You should also indicate the type of changes made to the annotation and whether the gene is split across scaffolds, as described in previous sections.

For each annotated transcript, select the annotation then right/apple click and select *'Comments'* from the menu to begin adding information. Determine if the comment is more appropriate for the gene (e.g. gene symbol) or an individual transcript (e.g. type of changes made). Within either the "*Comments for gene*" menu or "*Comments for transcript*" menu, click on the *'Add Comment'* option. You may choose one of the canned comments from the *'Choose a comment'* drop-down menu. You may also use the *'Add Comment'* feature to add custom comments. To edit an existing comment, click on the *'Edit'* button. Comments that are no longer useful may be removed with the *'Delete'* button.

6. Saving your Annotations

WebApollo saves your work as you go. If your work is interrupted or you are disconnected from the server, your work is automatically recorded in the database. There is no need to perform any additional step to save your work.

7. Exporting GFF3

You may export your annotations as GFF3, either for a single scaffold or for the entire assembly.

On the scaffold selection page, you may export GFF3 for the entire assembly by clicking the "*All to GFF3*" button at the top, or you may export GFF3 for a single scaffold by clicking "*GFF3*" next to the scaffold id.

To export gff3 for a single scaffold within the WebApollo main window, click on the dropdown triangle on the Annotations track's label, and select "*Save track data*" -> "*GFF3*".

8. Additional information about WebApollo:

WebApollo is an open-source project and under active development. If you have any questions, please contact the WebApollo development team at apollo-dev [at] lists [dot] lbl [dot] gov, and we will do our best to help you solve it. WebApollo is a member of the GMOD project. More details about WebApollo can be found at http://gmod.org/wiki/WebApollo and details on the server set-up can be found at http://www.gmod.org/wiki/WebApollo_Installation