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

Gregg Helt[1], Ed Lee[1], Justin T Reese[2], Monica C Munoz-Torres[1], Christopher P Childers[2], Robert Buels[3], Ian Holmes[3], Christine G Elsik[2,4], Suzanna E Lewis[1].
1 Genomics Division, Lawrence Berkeley National Laboratory. 2 Division of Animal Sciences, University of Missouri. 3 Department of Bioengineering, University of California, Berkeley. 4. Division of Plant Sciences, University of Missouri.

This guide allows users to:

- Become familiar with the environment of the Web Apollo annotation tool.
- Understand Web Apollo'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 Web Apollo.

CONTENTS

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 annotating, 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 the NCBI Non Redundant (NR) database. When annotating gene models using Web Apollo, 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 Web Apollo at
http://genomearchitect.org/WebApolloDemo

You may choose to go to the reference sequence selection screen (*'Select Track'* page), which is the main entry point for Web Apollo and allows users to view the available reference sequences as well as perform bulk operations on those sequences (e.g. exporting data). Alternatively you may jump directly into edit mode.

If prompted, use the credentials provided below
Username | Password: demo | demo

3.1.1 Select a Scaffold / Chromosome to Begin your Work

If you know the identifier of the scaffold (or chromosome) you wish to work on, find it under the *'Name'* list. Scaffolds will be listed in groups of 10, 25, 50 or 100 entries. You may scroll through the options on the upper left drop-down menu to select your preference and facilitate your search. The *'Filter'* box on the upper right allows users to find their scaffold of interest by entering letters and or numbers from its identifier. The filter search is not case-sensitive. Once you find the desired scaffold, click on its highlighted name; the Web Apollo 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 and have the sequence of a transcript or protein homolog related to your gene of interest, you might use the *'Search Sequence'* feature to run a BLAT (BLAST-Like Alignment Tool) search. Querying the assembled genome using BLAT will determine the existence of a gene model prediction that is putatively homologous to your gene of interest. Click the *'Tools'* item on the Web Apollo menu bar, 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 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 Web Apollo Main Window.

3.2 The Web Apollo Main Window:

Figure 1 highlights the main elements of the Web Apollo annotation window. The blue bar at the top includes the following functions:

- The *'File'* tab allows users add any combination of data files (including GFF3, BAM, BigWig) and URLs. Web Apollo automatically suggests tracks to display their contents.
- '*Tools'* allows users to perform BLAT searches (see section 3.1.2).

- *'Options'* allows users to color all exons in the window by CDS frame.
- By clicking on the *'Share'* option, users can obtain a permanent URL to share with collaborators. Web Apollo will display the visible region at the time the URL link was generated.
- The *'?Help'* tab brings up a list of helpful commands.
- Finally, the *'Login'* option allows users to enter and exit the Web Apollo editing tool. When logged in, the user name will be displayed instead of the word *'Login'*.

The *'Navigation Panel'* at the top of the window holds the controls for localization within a chromosome or scaffold (group), as well controls to move to a different scaffold/chromosome. A list of *'Available Tracks'* is available on the left. Dragging and dropping a label onto the *'Evidence'* panel allows visualization of all gene predictions and evidence tracks. The light yellow stripe on top is the *'User-created Annotations'* area, where users will drag the gene models/exons to be modified. All transactions performed on the *'User-created Annotations'* area can be reversed with the *'Undo'* option. Select the annotation in progress, and use right/Apple-click over it to display a menu that includes this and the *'Redo'* options.

Figure 1. Web Apollo Annotation Editor overview. This view shows 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 panel. To turn a track on or off, click and drag the track title from the column on the left, in or out of the main panel. **A.** The *'Navigation Panel'* runs along the top of the main panel; it includes buttons to move left and right, and two levels of zooming. The dropdown box is used to select the group (scaffold or chromosome) for annotation, and the textbox is used to manually enter its coordinates. **B.** The *'User-created Annotations'* panel contains the manual annotations. **C.** The *'Evidence'* 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 3.1.2), you may do so at this point using the *'Sequence search'* feature from the *'Tools'* tab on the menu bar. You may also navigate along the scaffold using the navigation arrows. Your gene of interest may appear on the forward (sense) or reverse (anti-sense) strand. Gene predictions are labeled with identifiers, and users may retrieve additional information by selecting the entire model and using the right/Apple menu to select the *'Information'* item.

After you have located the gene of interest, add as many gene prediction and evidence tracks as you consider necessary to inform your annotation by dragging and dropping them from the list of 'Available Tracks' on the left. 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 BLAST alignment or other aligned data (e.g.: alignments of protein homologs, cDNAs and, RNAseq reads). To highlight your preferred gene model, double click on any exon or click on one of the introns, select the entire gene model. Drag the highlighted model/evidence into to the *'User-created Annotations'* area.

At this point you may wish to download the protein sequence (see section 4.3.1) to query a protein database and help you determine if the selected gene model is a biologically accurate choice. For example, you may perform a protein sequence search of UniProt or NCBI NR. If you have knowledge of protein domains in your gene of interest, you may perform a protein domain search of the InterPro databases to verify that your selected gene model contains the expected domains. If further investigation suggests that you have not selected the best gene model to start annotating, delete it by highlighting it (as above) and using the *'Delete'* function from the right/Apple-click menu.

Once a gene model is selected as the best starting point for annotation, the annotator must decide whether it needs further modification. Protein or domain database searches may have already informed this decision. Scroll down the evidence tracks to see if splice sites in transcript alignments agree with the selected gene model, or if evidence suggests addition or modification of an exon is necessary. Transcript (cDNA/EST) alignments that are significantly longer than the gene model may indicate the presence additional coding sequence or untranslated regions (UTRs). 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 well, resulting in a short protein alignment or one with gaps. Protein and transcript alignments in regions with tandem closely related genes might also be problematic, with partial alignments to one gene, then skipping over to align the rest to a second gene.

4.2 Simple Cases:

In this guide, a *'simple case'* occurs when 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.

4.2.1 Add UTRs:

Gene predictions may or may not include UTRs. If transcript alignment data are available and extend beyond your original annotation, you may add or extend UTRs. First, position the cursor at the beginning of the exon that needs to be extended, use the right/Apple-click to display the menu and choose the *'Zoom to base level'* option. 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.

4.2.2 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, a red bar appears 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. Use the square bracket keys [and] to jump to the next exon splice junction or coding sequence (CDS). The curly bracket keys { and } allow users to jump to the next transcript. Check whether there are any ESTs that lack one or more of the annotated exons or include additional exons.

To correct an exon boundary to match data in the evidence tracks use the '*Zoom to base level*' option, 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 the evidence data support.

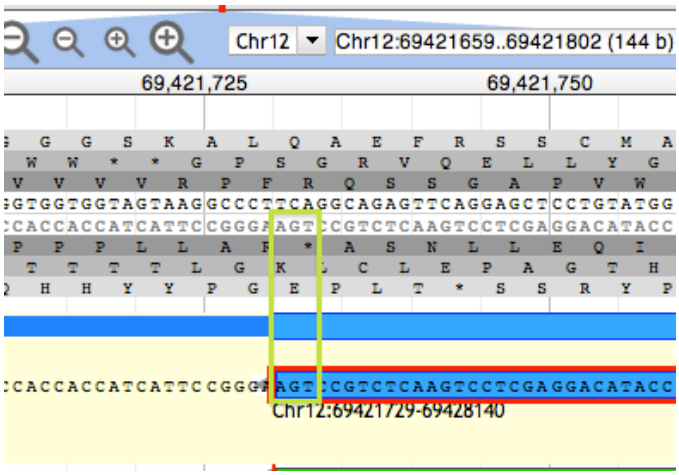

Figure 2. View zoomed to base level. The DNA track and annotation track are visible. The DNA track includes the sense strand (top) and anti-sense strand (bottom). The six reading frames flank the DNA track, with the three forward frames above and the three reverse frames below. The User-created Annotation track shows the terminal end of an annotation. The green rectangle highlights the location of the nucleotide residues in the *'Stop'* signal.

4.2.3 Splice sites:

In most Eukaryotes the majority of splice sites at the exon/intron boundaries appear as 5'-…exon]GT/AG[exon…-3'. All other splice sites are called '*non-canonical'* and are indicated in Web Apollo with an orange circle with a white exclamation point inside, placed over the edge of the offending exon. When alternative transcripts are added, be sure to inspect each transcript to verify that all changes are saved correctly.

If a non-canonical splice site is present, zoom to base level to review it. No all non-canonical splice sites must be corrected, and in such cases they should be flagged with the appropriate comment. (Adding a '*Comment'* is addressed in Section 5). Prior knowledge about the organism of interest may help the user decide whether a predicted non-canonical splice site is likely to be real. For instance, GC splice donors have been observed in many organisms, but less frequently than the GT splice donors described above. As mentioned above Web Apollo 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. In such cases 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 a non-canonical splice site that is rarely observed in nature is present, 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 a close in-frame site that is more likely to be the correct splice donor, make this adjustment while zoomed at base level. To assist in the decision to modify a splice site, download the translated sequences and use them to search well-curated 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.

4.2.4 'Start' and 'Stop' sites:

By default, Web Apollo 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 Web Apollo did not calculate the correct *Start* signal, the user can 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'* option 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 4.3.3 on how to *'Add 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'* table of the *'Annotation Info Editor'* as described in section 5).

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 manually setting it 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'* table of the *'Annotation Info Editor'*).

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. See section 4.3.3 on how to *'Add an exon'*. Each time you add an exon region, whether by extending an existent exon or adding a new one, Web Apollo recalculates the longest ORF to identity *Start* and *Stop* signals, allowing you to determine whether a *Stop* codon has been incorporated after each editing step.

4.2.5 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, Web Apollo 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, and 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'* option.

4.3.3 Add an exon:

You may select and drag the putative new exon from a track in the *'Evidence'* panel, and add it directly to an annotated transcript in the *'User-created 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 the mouse button is released, the additional exon becomes attached to the receiving transcript. If the receiving transcript is on the opposite strand from the one where you selected the new exon, a warning dialog box will ask you to confirm the change.

As described before, Web Apollo dynamically recalculates the longest ORF for each 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'* option from the right/Apple-click menu. Web Apollo will identify the nearest canonical splice sites (5'-…exon]GT/AG[exon…-3') to modify the model, and will recalculate the longest ORF. If Web Apollo cannot find a set of canonical splice sites within the selected exon, a dialog box will appear with a warning.

If everything you know about the model indicates that this exon should not be preserved in its current form, you may manually disrupt the exon using the *'Split'* option from the right/Apple-click menu, which creates a 1-nucleotide intron without taking into whether or not account the surrounding splice sites are canonical.

4.3.5 Delete an exon:

Select the exon using a single click (double click selects the entire model), and select the '*Delete'* option from the right/Apple-click menu. Check whether deleting one or more exons disrupts the reading frame, inserts premature *Stop* signals, etc.

4.3.6 Flip the strand of annotation:

At times transcript alignments may appear on the strand opposite to the model's coding strand, particularly when the transcript alignment does not include a splice junction, which makes it difficult to determine the coding direction. If such aligned evidence is used to initiate an annotation, and it is later determined that the annotation is on the incorrect strand, the user may choose 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. To begin the annotation select each of the gene models that you would like to merge, then drag them separately from the *'Evidence'* panel onto the '*User-created Annotations*' area. A protein alignment may not be a useful starting point because it may have incorrect splice sites and may lack non-conserved regions.

Select the supporting evidence tracks and drag their 'ghost' over the candidate models to corroborate the overlap. Additionally, zoom in and carefully review edge matching (Figure 3) and coverage across models. When you are certain that these models should be merged, while holding the *'Shift'* key click on an intron from each gene model to highlight both. Then select the *'Merge'* option from the right/Apple-click menu. Get the resulting translation sequence and inspect it by querying a protein database, such as UniProt, using this sequence. Be sure to record the IDs of both starting gene models in the *'Comments'* table, and use the appropriate canned comment to indicate that this annotation is the result of a merge.

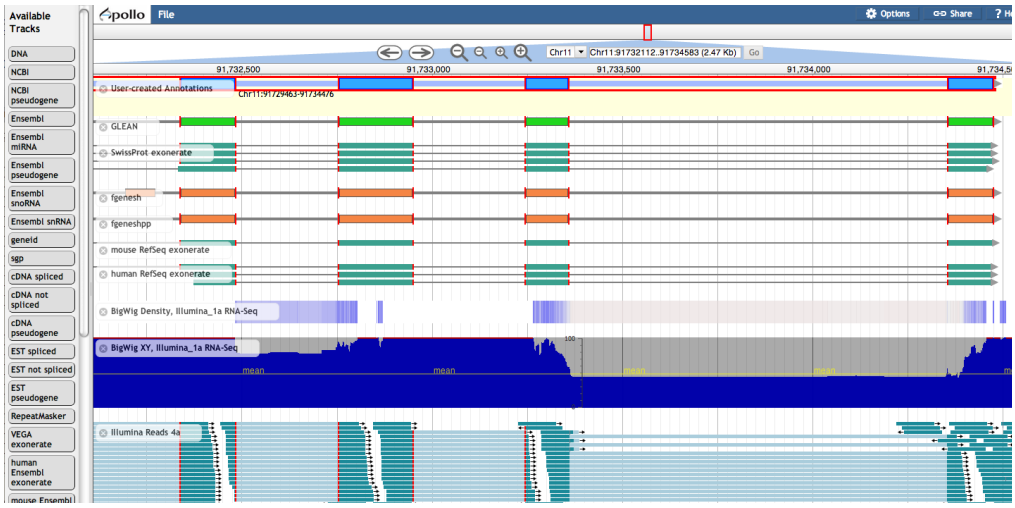

Figure 3. Edge-matching in Web Apollo. When a feature is selected, the exon edges are marked with a red box. All other features that share the same exon boundaries are marked with a red line on the matching edge. This feature allows annotators to confirm that evidence is in agreement without examining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

4.4.2 Merge two gene predictions from 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'* table for both components. For standardization purposes, please use the following two prepared (canned) comments, adding the name of both models in every case:
"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:

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. Changes are made on the DNA track with the right/Apple-click menu.

If you determine that you need to make one of these changes, zoom in to the nucleotide level, and right/Apple-click over the genomic sequence to access the menu with options

for introducing sequence changes such as insertions, deletions or substitutions. The selected nucleotide must be the starting point for each modification.

The *'Create Genomic Insertion'* option requires a string of nucleotide residues that will be inserted to the right of the cursor's current coordinate. The *'Create Genomic Deletion'* option requires the length of the deletion, starting with the nucleotide where the cursor is positioned. When using the *'Create Genomic Substitution'* option, enter the string of nucleotide residues that will replace the ones on the DNA track.

Once you have entered the modifications, Web Apollo will recalculate the corrected transcript and protein sequences, which can be obtained selecting the *'Get Sequence'* option from the right/Apple-click menu. Since the underlying genomic sequence is reflected in all annotations that include the modified region you should alert the curators of your organism's database using the '*Comments'* section to report these CDS edits.

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

5. Adding more information to your annotations

5.1. Add Comments

When you are satisfied with your annotation, you may provide additional information in the form of *'Comments'*. For example, the ID of the gene prediction that you used to initiate the annotation presents useful information for your database curators. 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 first click to select it, then use the right/Apple-click option to select *'Annotation Info Editor'* from the menu. The *'Annotation Info Editor'* window displays information for both the gene and the transcript; users should determine whether the comment is more appropriate for the gene (e.g. a change in the gene symbol) or an individual transcript (e.g. type of alterations made). In the *'Annotation Info Editor'* window click on the respective *'Add'* button to start a new comment, and a new row, labeled as '*Enter new comment'*, will appear. One click on this row reveals a drop-down menu option on the right, which displays canned comments to choose if they are available for your organism of interest. Alternatively, it is also possible to type custom comments. To edit an existing comment, click over the comment and begin typing, or replace it with a different canned comment. Comments that are no longer relevant or useful may be removed using the *'Delete'* button at the bottom of the box.

5.2. Add database crossed-references

When available, users should also include information to cross-referenced databases by adding the name of the database and the corresponding accession number for each gene or transcript to the *'DBXRefs'* tables, respectively. The process to add information to these tables is the same as described for the *'Comments'* tables.

6. Saving your Annotations

Web Apollo continually and immediately saves your work, automatically recording it on the database. Because of this, your work will not be lost in the event of network disruptions, and no further actions are required in order to save your work.

7. Exporting Data

The user-created annotations may be exported as GFF3 formatted files, either for a single scaffold (or chromosome), or including user annotations from the entire genome assembly. To export the user-created annotations from a single scaffold, click on the track labeled *'User-created Annotations'*, click on the *'Save track data'* option, and then on the *'GFF3'* option. To export GFF3 files from more than one scaffold, or from the entire assembly, go to the scaffold selection page and select as many scaffolds or chromosomes as you need. Then, click on the *'File'* tab at the top right, and choose the options *'Export'* and *'GFF3'* as they appear in your screen.

Data from each of the evidence and prediction tracks can also be exported. GFF3 formatted files of the visible region on the Web Apollo screen, as well as files of the entire scaffold/chromosome can be exported. The data will be formatted according to the original data used to display each track. For instance, RNA-Seq reads could be exported either as GFF3 or BED file formats.

8. Additional information about Web Apollo:

Web Apollo is an open-source project and is under active development. If you have any questions, please contact the Web Apollo development team at **apollo-dev [at] lists [dot] lbl [dot] gov**. Web Apollo is a member of the GMOD project. More details about Web Apollo, can be found at http://gmod.org/wiki/WebApollo, details on the server set-up can be found at http://www.gmod.org/wiki/WebApollo_Installation, the Web Apollo demo is located at http://genomearchitect.org/WebApolloDemo, and this guide can be found at http://genomearchitect.org/webapollo/docs/webapollo_user_guide.pdf.