Annotation Files Merger User Guide

Wilson Leung

Table of Contents

Introduction

In order to submit an annotation project to the Genomics Education Partnership (GEP; https://thegep.org), you must complete the annotation report form and include three supplemental files (i.e. a GFF file, a transcript sequence file, and a peptide sequence file). These supplemental files contain different types of information for all the genes and isoforms that you have annotated in your project. While the *Gene Model Checker* will generate these supplemental files for each gene model, these individual files must be combined into a single project file. To prepare for project submission, you will need to create a GFF file which contains the GFF entries for all the genes and isoforms in your project. Similarly, you will need to create a transcript sequence file which contains the transcript sequences for all the gene models in your project, and a peptide sequence file which contains all the protein sequences in your project.

The *Annotation Files Merger* is designed to help you combine the individual files generated by the *Gene Model Checker* into a single project file suitable for project submission. This tool also performs additional checks to verify that all the isoforms have been annotated and it allows you to view all the annotated gene models on the *GEP UCSC Genome Browser*. For projects that contain errors in the consensus sequence, the *Annotation Files Merger* can also combine VCF files generated by the *Sequence Updater*.

This user guide will provide you with a general overview of the *Annotation Files Merger*. It will also illustrate how you can use this tool to identify and resolve common annotation errors.

Acknowledgements

The *Annotation Files Merger* is developed by Wilson Leung at Washington University in St. Louis for the Genomics Education Partnership (GEP).

Questions about the *Annotation Files Merger*

Please contact Wilson (wleung@wustl.edu) if you have any questions or encounter any problems with the *Annotation Files Merger*.

Availability

The *Annotation Files Merger* is available under the "**Resources & Tools**" section of the F Element project page and the Pathways project page on the GEP website.

Overview of the *Annotation Files Merger*

To illustrate the key functionality of the *Annotation Files Merger*, we will prepare the supplemental files for the *Pp2A-29B* gene in *Drosophila suzukii* as part of the *Drosophila* Pathways project. The gene model is located at scaffold KI419149 in the *D. suzukii* Sep. 2013 (BGI/Dsuz_1.0) assembly with the *UCSC Genome Browser* assembly ID **DsuzGB1**. This tutorial assumes you have already used the *Gene Model Checker* to verify all the gene models, and that you have saved the corresponding GFF, transcript, and peptide sequence files. Please see page 31 of the *Gene Model Checker* User Guide for details on how to download the GFF, transcript, and peptide sequence files produced by the *Gene Model Checker*.

Organize the output files from the *Gene Model Checker*

While the *Annotation Files Merger* can handle files with arbitrary names, we recommend that you name these files based on the *UCSC Genome Browser* assembly ID, followed by the name of the gene and the name of the isoform (e.g., DsuzGB1_Pp2A-29B-PA). We also recommend using file extensions to distinguish the different types of file (for example, use ".gff" for GFF files, ".fasta" for transcript sequence files, and ".pep" for peptide sequence files), and store the different types of files in separate folders. Organizing the files in this manner will enable you to easily keep track of your annotation progress and to quickly select all the files that should be combined together.

In this tutorial, all the GFF files are stored in the "GFF" folder, the transcript sequence files are in the "transcripts" folder, and the peptide sequence files are in the "peptides" folder (Figure 1).

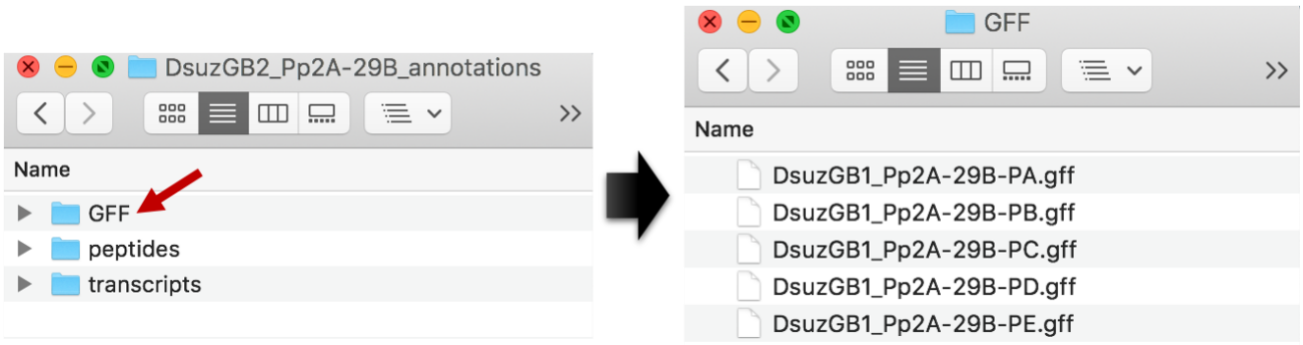


Figure 1. (Left) Organize the supplemental files for the *D. suzukii Pp2A-29B* gene using the "GFF", "peptides", and "transcripts" folders. (Right) The "GFF" folder contains the GFF files for the five isoforms of *Pp2A-29B* in *D. suzukii* produced by the *Gene Model Checker*.

Upload the annotation files to the *Annotation Files Merger*

Open a web browser window and navigate to the Pathways project page on the GEP website. Click on the "Annotation Files Merger" link under the "Resources & Tools" section.

Under the "File Type" field, select the type of files you would like to merge (i.e. GFF, transcript sequence, peptide sequence, or VCF files). In this tutorial, we will merge the GFF files for the different isoforms of *Pp2A-29B* in *D. suzukii*. Select the "GFF Files (.gff)" option under "File Type".

Open the “GFF” folder with all the GFF files for *Pp2A-29B*. Select all the GFF files in the folder, and then drag the selected files onto the grey box labeled “Drag and drop the files you want to merge here”. Release the mouse when the grey box turns green and the label changed to “Drop the selected files here” (Figure 2).

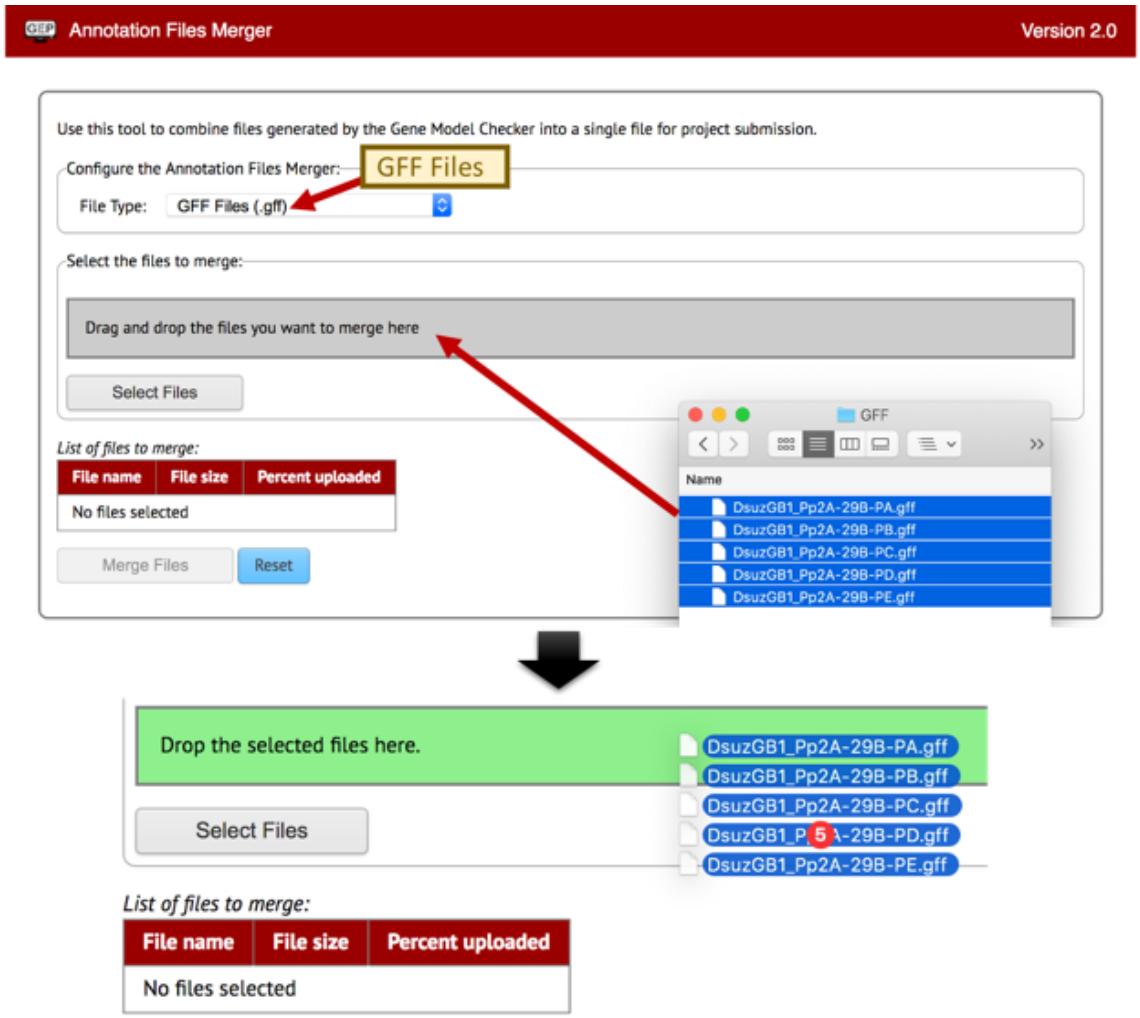


Figure 2. Drag and drop the files you want to merge onto the *Annotation Files Merger* web page.

Note: You can also use the “Select Files” button under the “Select the files to merge” section to select the files that you would like to merge.

The selected files will be added to the table with the title “List of files to merge” (Figure 3). You can select additional files (using either drag-and-drop or the “Select Files” button), and they will be appended to the table. After you have selected all the files you would like to merge, click on the “Merge Files” button to upload and merge the files.

List of files to merge:

File name	File size	Percent uploaded
DsuzGB1_Pp2A-29B-PA.gff	1694	Queued
DsuzGB1_Pp2A-29B-PB.gff	1694	Queued
DsuzGB1_Pp2A-29B-PC.gff	1694	Queued
DsuzGB1_Pp2A-29B-PD.gff	1903	Queued
DsuzGB1_Pp2A-29B-PE.gff	1694	Queued

Merge Files Reset

Figure 3. The GFF files for the A, B, C, D, and E isoforms of *Pp2A-29B* in *D. suzukii* will be uploaded to the *Annotation Files Merger*.

Interpret the *Annotation Files Merger* output

Once the files have been merged successfully, the "Download merged file" section should appear with a link to the merged file. Right click (control-click on macOS) on the link labeled "Merged File link" and select "Save Link As…" or "Download Linked File As…" to save the merged file. We recommend using the following naming convention for the merged files: <species prefix>_<gene_name>.<file_type>.

For example, we will call the merged GFF file "dsuz_Pp2A-29B.gff" because the four letter species prefix for *D. suzukii* is "dsuz", the name of the annotated gene is "*Pp2A-29B*", and the file type is GFF.

In addition to the "Download merged file**"** section, there are two additional sections that will appear when you merge GFF files: the "Isoform checklist**"** section and the "Show merged GFF File in the Genome Browser" section (Figure 4).

GEP Annotation Files Merger Version 2.0

File Merge Results

Download merged file

Right click on the Merged File link and select "Save Link As..." to save the merged file.

Isoforms checklist

Gene	Status	Message
Pp2A-29B	Pass	

Show merged file in the Genome Browser

Click on the "Show Track" button to view the merged file in the Genome Browser

Show Track

Merge another set of files

Figure 4. After the GFF files have been combined into a single file, the File Merge Results page consists of three sections: "Download merged file", "Isoform checklist", and "Show merged file in the Genome Browser".

The “Isoform checklist” section checks each gene name in the merged GFF file to ensure that they are valid FlyBase gene symbols. For each gene in the merged GFF file, the *Annotation Files Merger* will determine the total number of isoforms based on the FlyBase gene annotation in *D. melanogaster*, and verify that the merged GFF file contains annotations for all the isoforms.

The “Show merged GFF File in the Genome Browser” section allows you to view gene models in the merged GFF file as a custom track on the *GEP UCSC Genome Browser*. The *Annotation Files Merger* will automatically infer the genome assembly that should be used based on the metadata in the original GFF files.

Note: If the *Annotation Files Merger* cannot determine the genome assembly based on the provided GFF files, a drop-down box will appear in the “Show merged GFF File in the Genome Browser” section where you can manually select the project region.

Click on the “Show Track” button to view the all the gene models in the merged GFF file on the *GEP UCSC Genome Browser* (Figure 5). Zoom out 1.5x. The gene models for the five isoforms of *Pp2A-29B* are generally consistent with the RNA-Seq read coverage from the adult females and adult males samples, as well as the *Genscan* and *Geneid* gene predictions.

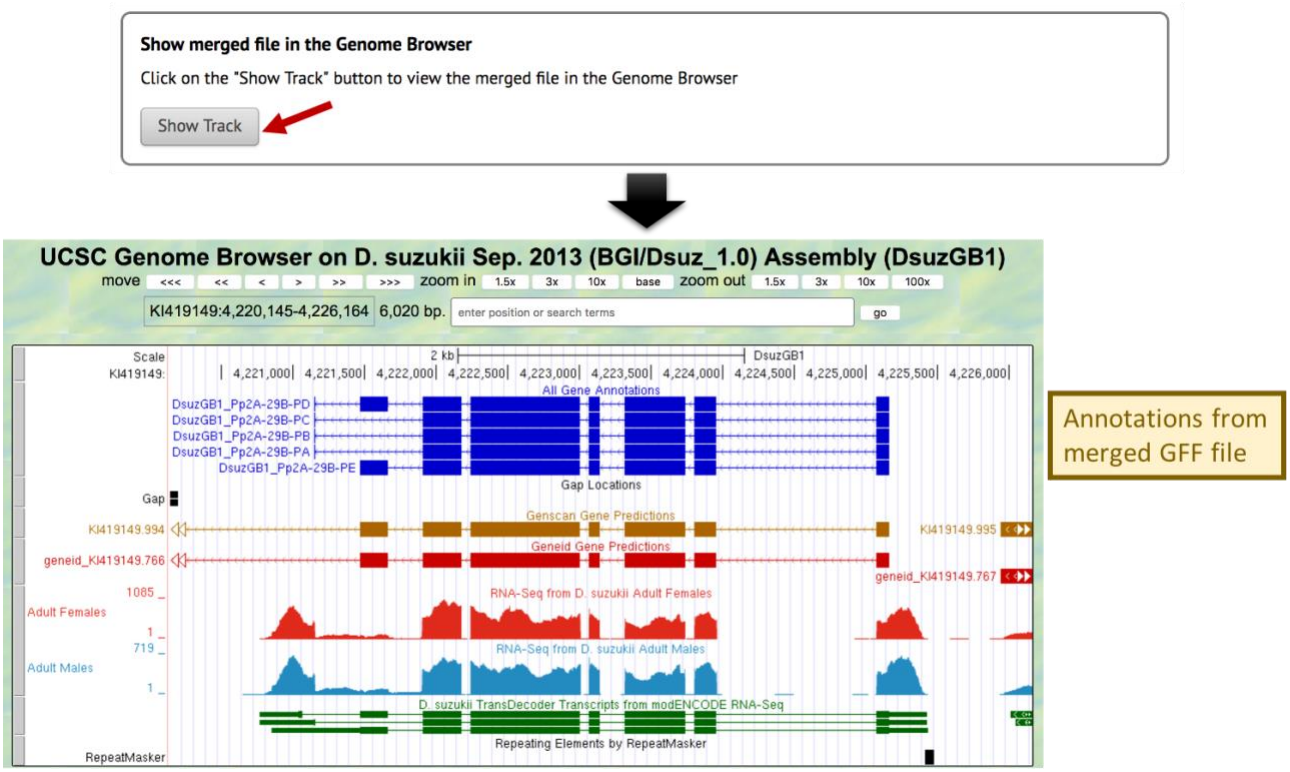


Figure 5. Use the “Show Track” button on the *Annotation Files Merger* results page to view the gene models in the merged GFF file on the *GEP UCSC Genome Browser*.

Merging other supplemental files

Click on the "Merge another set of files" button to reset the web page. Repeat the same procedure (as described on pages 3–5) to merge the transcript sequence files and the peptide sequence files. The merged transcript sequence file will be called dsuz_Pp2A-29B.fasta and the merged peptide sequence file will be called dsuz_Pp2A-29B.pep. At this point, you have created the three supplemental files required for project submission.

Merging VCF files

For projects that contain errors in the project sequence, you can use the *Sequence Updater* to document the changes to the project sequence using the Variant Call Format (VCF). If the project sequence contains multiple errors, you might have generated multiple VCF files that would need to be combined into one project VCF file prior to project submission.

For example, a previous analysis has identified two consensus errors within the *Tor* gene located on scaffold_6500 in the *D. mojavensis* May 2011 (Agencourt dmoj_caf1/DmojCAF1) assembly. The consensus errors were located at 26,044,753 (A ➔ AT) and 26,045,153 (T ➔ TG).

We have previously used the *Sequence Updater* to generate two VCF files to describe the changes to the consensus sequence. To combine these VCF files using the *Annotation Files Merger*, select the "Variant Call Format (.vcf.txt)" option under the "File Type" field, and then drag the two VCF files onto the "Drag and drop the files you want to merge here" grey box. Click on the "Merge Files" button (Figure 6).

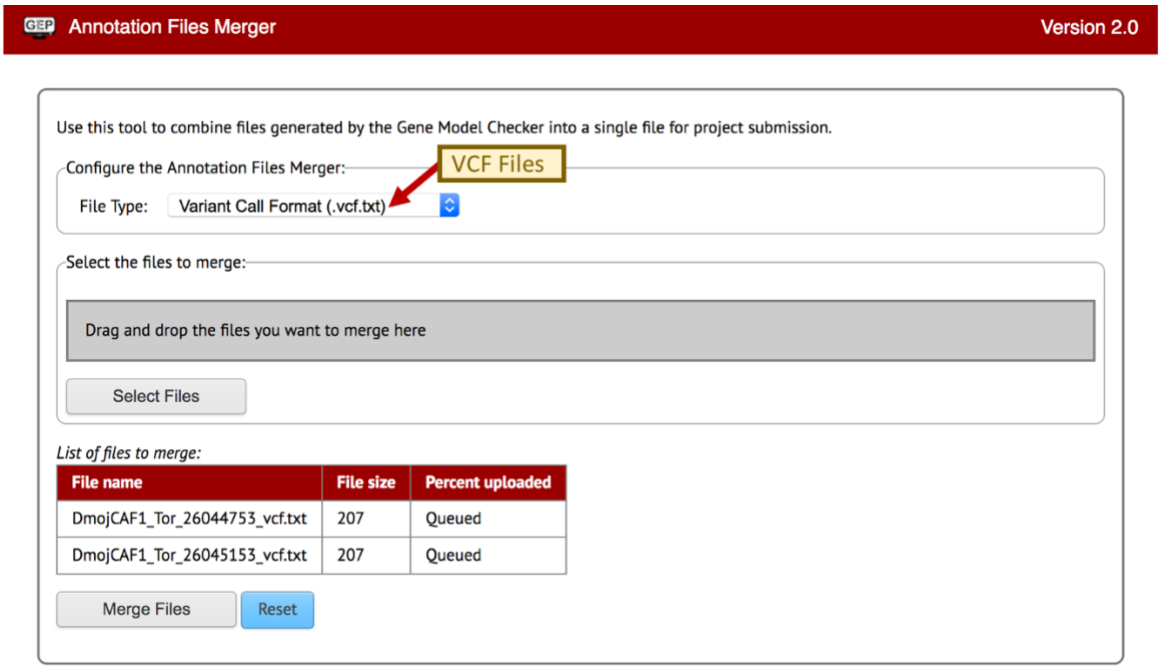


Figure 6. Configuring the *Annotation Files Merger* to combine multiple VCF files.

The “File Merge Results” panel consists of two sections. The “Download merged file” section contains a link to download the merged VCF file. The “Show merged file in the Genome Browser” section allow us to view the sequence changes in the *GEP UCSC Genome Browser* in conjunction with the other evidence tracks.

Because the *Tor* gene annotation was from the *D. mojavensis* DmojCAF1 assembly, we will change the “Select a Project Region…” field to “D. mojavensis May 2011 (Agencourt dmoj_caf1/DmojCAF1)”, and then click on the “Show Track” button (Figure 7).

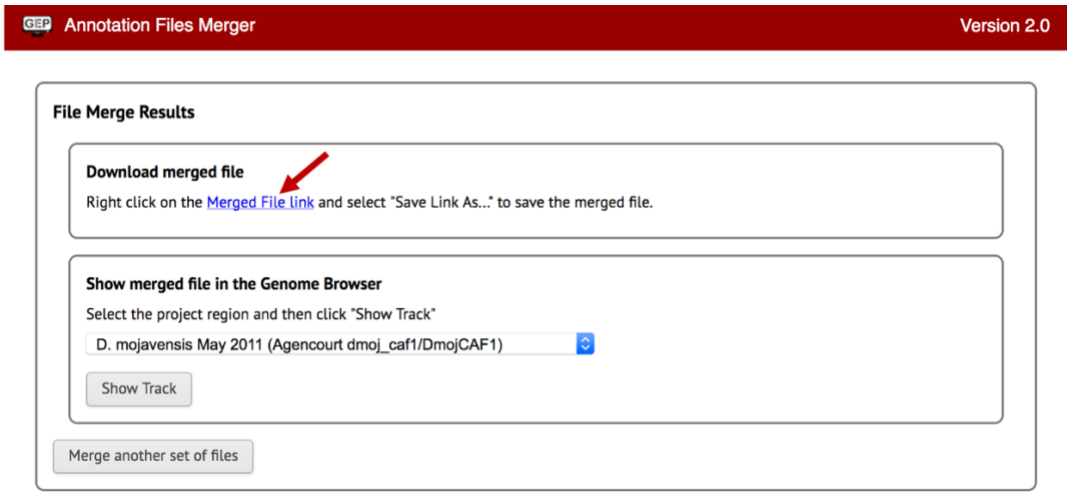


Figure 7. Right click (control-click on macOS) on the "Merged File link" and select "Save File As..." or “Download Linked File As…” to save the merged VCF File. Select the project region and then click on the "Show Track" button to view the merged VCF file on the *GEP UCSC Genome Browser*.

The sequence changes described by the merged VCF files will appear under the “Potential Consensus Errors” custom track on the *GEP UCSC Genome Browser*. To navigate to the region surrounding the *Tor* ortholog, enter “scaffold_6500:26,038,100-26,047,000” into the “enter position or search term” field, and then click on the “go” button. Click on the “hide all” button and then change the display settings for the following evidence tracks:

Under “Custom Tracks”:

- Potential Errors: **pack**

Under “Genes and Gene Prediction Tracks”:

- RefSeq Genes: **pack**
- D. mel Proteins: **pack**
- *GeMoMa* Genes: **pack**
- *Genscan* Genes: **pack**
- *Geneid* Genes: **pack**
- *Augustus*: **pack**

Under “RNA-Seq Tracks”:

- RNA-Seq Coverage: **full**

Click on one of the “refresh” buttons to update the Genome Browser view (Figure 8).

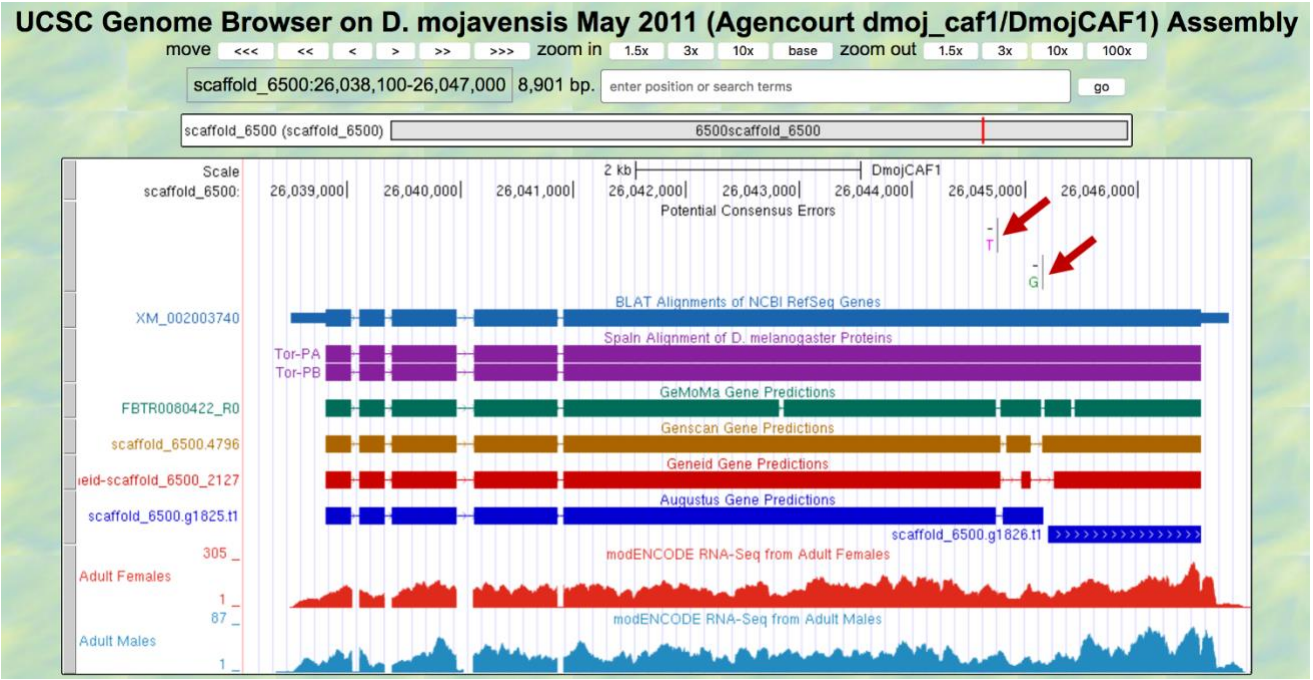


Figure 8. The changes described by the merged VCF file will appear as a custom track labeled "Potential Consensus Errors" on the *GEP UCSC Genome Browser*.

Use the *Annotation Files Merger* to identify annotation errors

In addition to preparing the supplemental files for submission, the *Annotation Files Merger* also performs additional checks to help you identify common annotation errors. This part of the user guide will illustrate how you can use the different sections of the *Annotation Files Merger* output to identify and resolve common annotation issues.

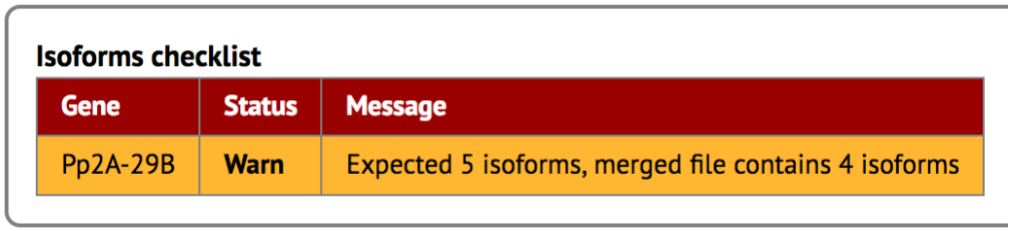
The Isoform Checklist

When merging GFF files, one of the sections in the “File Merge Results” output panel is the “Isoform checklist”. The *Annotation Files Merger* will validate the gene and isoform names in the GFF file against the FlyBase release available at the *Gene Record Finder*.

The merged file should contain gene models for all isoforms

For each gene in the merged GFF file, the *Annotation Files Merger* will check against the corresponding *D. melanogaster* record in the *Gene Record Finder* to determine the expected number of isoforms. The “Isoform checklist” will show a “Warn” status (highlighted in yellow) for any genes where the number of isoforms in *D. melanogaster* differs from the number of isoforms found in the merged GFF file. While some isoforms might not exist and novel isoforms might appear in the species you are annotating, these are unusual cases that require detailed explanations in the annotation report for the GEP project.

For example, if the annotator did not include the GFF file for one of the isoforms of *Pp2A-29B* when they upload the GFF files to the *Annotation Files Merger*, the "Isoforms checklist" will include a warning indicating that the number of submitted isoforms (4) differs from the expected number of isoforms (5) found in the *D. melanogaster* ortholog (Figure 9).



Isoforms checklist

Gene	Status	Message
Pp2A-29B	**Warn**	Expected 5 isoforms, merged file contains 4 isoforms

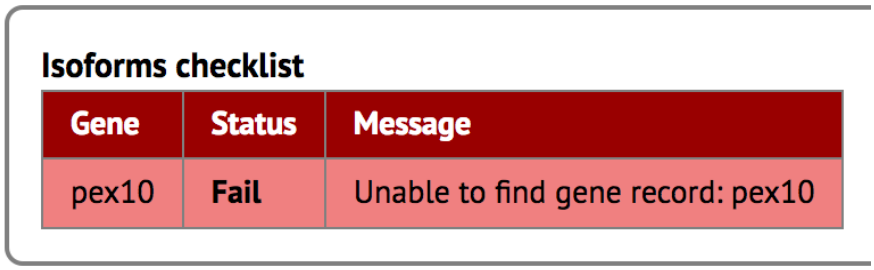
Figure 9. The "Isoforms checklist" section will display a warning when the number of isoforms found in the merged GFF file for the target species differs from the number of isoforms found in the *D. melanogaster* ortholog.

Note: The supplemental files should contain the annotations for **all isoforms**, irrespective of whether these isoforms have identical coding regions.

Gene names in *Drosophila* are case-sensitive

The "Isoform checklist" will show a "Fail" status (highlighted in red) for gene symbols that do not exist in FlyBase. While novel genes and new paralogs can occasionally be found in the different *Drosophila* species, it is unusual and requires detailed explanations in the annotation report for the GEP project.

Note that gene names in *Drosophila* are case-sensitive. Gene names that begin with a lowercase letter indicate that the gene is first named for a phenotype observed in a mutant recessive allele. Gene names that begin with an uppercase letter indicate either the phenotype is first named for a phenotype observed in a mutant dominant allele or it describes the molecular function of the gene. For example, the "Isoform checklist" in Figure 10 shows that the *Annotation Files Merger* cannot find the gene record for *pex10* because the correct gene symbol for the Peroxin 10 gene is *Pex10*.



Isoforms checklist

Gene	Status	Message
pex10	**Fail**	Unable to find gene record: pex10

Figure 10. The *Annotation Files Merger* displays an error message if the gene symbol (e.g., *pex10*) does not exist in the FlyBase gene records from the *Gene Record Finder*.

Detect overlapping features

One of the common annotation errors for the F element project is an over-reliance on the *blastx* alignment track on the *GEP UCSC Genome Browser*. In some cases where *blastx* hits from multiple genes are aligned to the same part of the genomic sequence (e.g. because of conserved domains), some students might create gene models for all the genes listed on the *blastx* track.

Because nested genes (and overlapping exons in particular) are rare in *Drosophila*, the *Annotation Files Merger* will compare the exons in the merged GFF file and identify exons from different genes that overlap with each other. If any overlapping exons were found, a section labeled "Overlapping features found in merged GFF file" will appear in the "File Merge Results" panel. These overlapping features will require detailed explanations of the supporting evidence in the annotation report for the GEP project.

For example, the *D. melanogaster* gene *PlexB* contains five conserved domains that are also found in the gene *PlexA* (i.e. Semap_dom, Plexin_repeat, TIG1_plexin, TIG2_plexin, IPT_dom, Plexin_cytoplasmic_RasGAP_dom). As a result, the *blastx* alignment track shows multiple isoforms from both genes aligning to the same part of the *D. ananassae* Muller F element project contig37 (Figure 11). Based on the E-value and percent identity of the two sets of *blastx* hits, the gene predictions, and RNA-seq data, the annotator should have concluded that this region of contig37 contains the putative ortholog of *PlexB* and not *PlexA*.

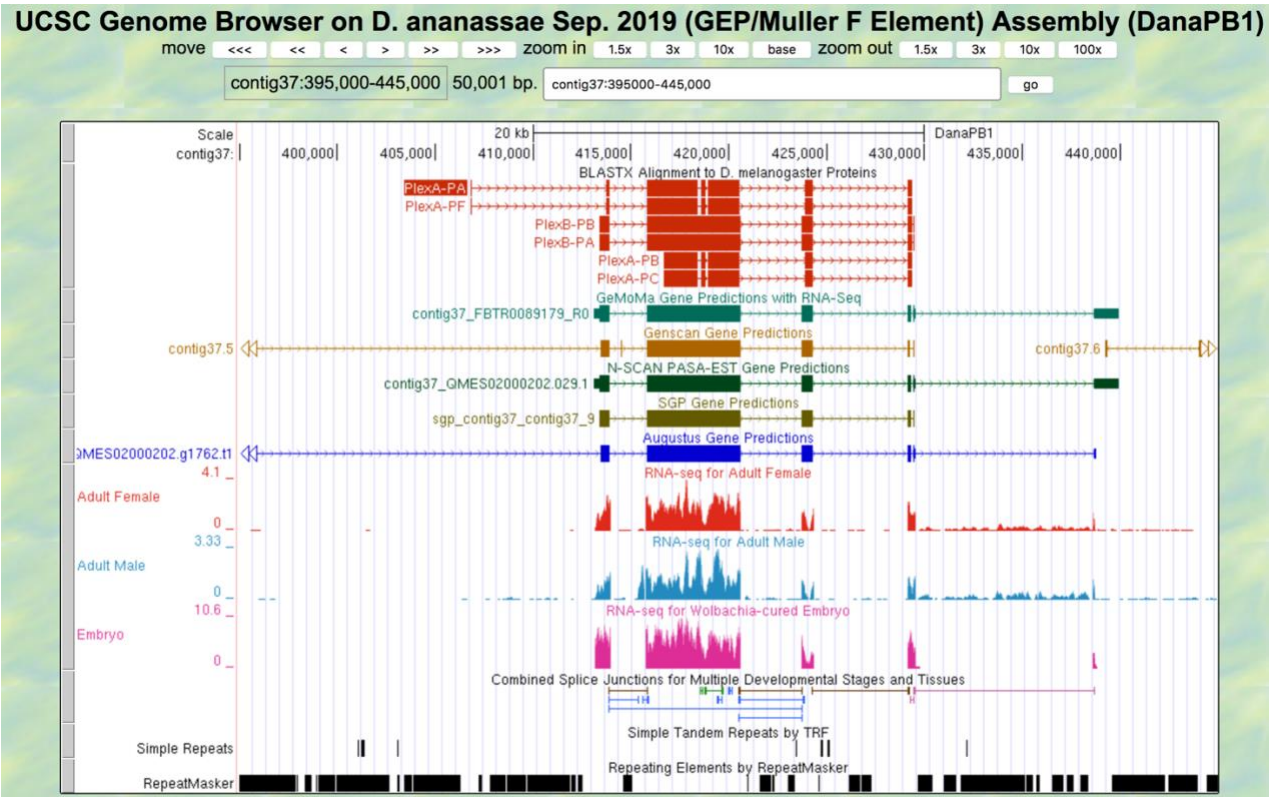


Figure 11. The *blastx* alignment track in the *GEP UCSC Genome Browser* shows different sets of *blastx* hits to two different genes (*PlexB* and *PlexA*) because they shared multiple conserved domains.

However, if the annotator annotated both *PlexB* and *PlexA* and try to merge the GFF annotation files using the *Annotation Files Merger*, the **Overlapping features found in merged GFF file** section will highlight the two genes with overlapping exons (Figure 12).

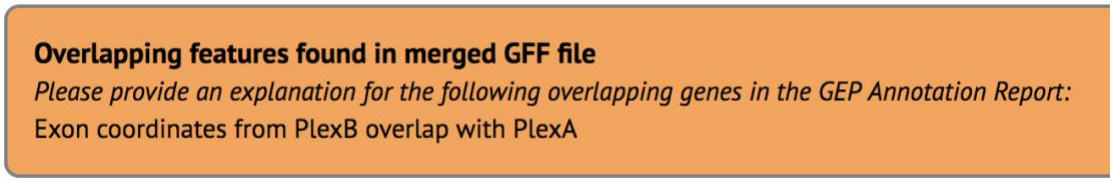


Figure 12. The *Annotation Files Merger* detected overlapping exons between the annotations of *PlexB* and *PlexA*.

Identify additional features that require annotation

The ability to view the merged GFF file as a custom track on the *GEP UCSC Genome Browser* enables you to examine all the annotated gene models in the context of other evidence tracks. This could help you identify regions within your project that warrants further investigation. For example, the region between the genes *Eph* and *CaMKI* on the *D. ananassae* Muller F element project contig50 (at 170,055-171,971) shows significant similarity to msk-PA, which requires further investigations (Figure 13). (In this case, this feature in contig50 is a putative pseudogene derived from *msk*.)

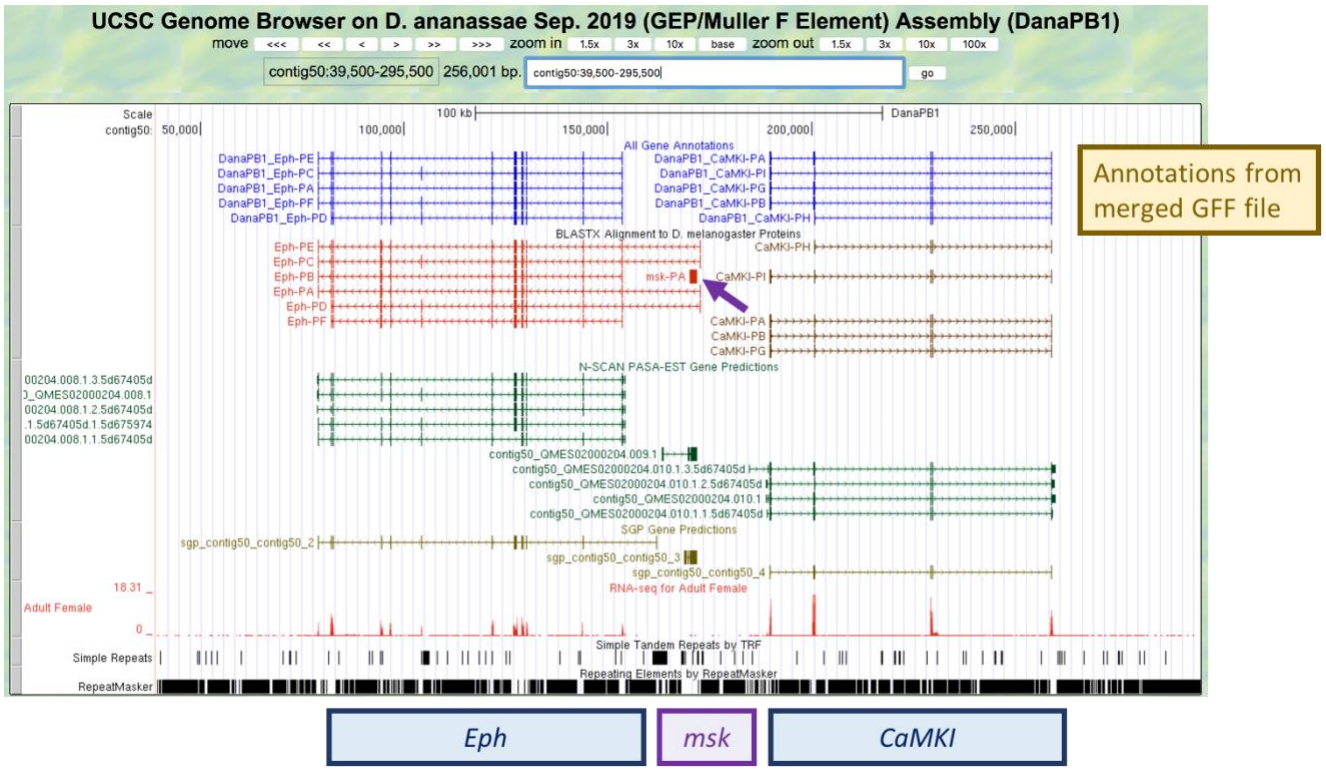


Figure 13. The region between the annotated genes *Eph* and *CaMKI* in *D. ananassae* contig50 shows significant similarity to *msk*. The feature is also supported by gene predictions from *N-SCAN PASA-EST* and *SGP*. Hence this region of contig50 will require further investigations to ascertain if it corresponds to a protein-coding gene.

Conclusion

This user guide demonstrates how you can use the *Annotation Files Merger* to prepare the files generated by the *Gene Model Checker* for project submission. The *Annotation Files Merger* also performs additional checks on the merged GFF file to verify that all the gene symbols are correct and the merged files contains the annotations for all isoforms. Using the *Annotation Files Merger* in conjunction with the *Gene Model Checker* should help detect many common annotation errors and improve the overall accuracy of student annotations.