GEP Project Management System: Annotation Project Submission

Author

Wilson Leung   wleung@wustl.edu

Document History

Initial Draft	06/04/2007
First Revision	01/11/2009
Second Revision	01/08/2010
Third Revision	08/14/2011
Fourth Revision	01/07/2012
Fifth Revision	08/19/2012
Sixth Revision	08/05/2013
Seventh Revision	08/21/2015

Version

GEP Project Management System (Version alpha)

Introduction

In order to facilitate the verification and reconciliation of the gene models submitted to the GEP, all GEP annotation project submissions should contain a completed annotation report and three supplemental files. The project report is part of the annotation package that you have downloaded when you claimed the project. (The annotation report form is also available under the "Specific Issues in GEP Annotation Projects" section of the GEP web site.) The three supplemental files contain data for **all the gene models** in your project in three different formats: General Feature Format (GFF), transcript sequences, and peptide sequences. You can use the Gene Model Checker to validate each gene model and generate these supplemental files for each isoform. For each type of supplemental file, you can use the Annotation Files Merger to combine the individual files for each isoform into a single project file.

Preparing Supplemental Files for Submission

Using the Gene Model Check to create supplemental files for each isoform

In order to submit projects with gene annotations, you must provide three files besides the annotation report: the GFF, transcribed sequences, and peptide sequences for all gene models (Figure 1). The Gene Model Checker will create these supplemental files for each of the gene model you have checked. You can download these files under the "Downloads" tab. Please refer to the

Gene Model Checker User Guide in the documentations section of the GEP web site for more information on how to use the Gene Model Checker to verify your gene models.

Figure 1 Project Management System interface for submitting annotation projects.

Using the Annotation Files Merger to combine all the supplemental files

While the Gene Model Checker will create the supplemental files for each of the isoform you have checked, you must combine these individual files into a single project file prior to submitting a project. After combining all the files, you should have three supplemental files: a project GFF file with all the GFF entries, a project transcript sequence file with all the transcript sequences, and a project peptide sequence file with all the peptide sequences.

Besides creating the merged supplemental files, the Annotation Files Merger also performs additional checks on the merged GFF file to help identify some of the common annotation errors. There are two additional sections in the output when you are merging GFF files: the "Isoform Checklist" and "Show merged GFF File in Genome Browser." The Isoform Checklist section shows the list of genes found in all the GFF files you have submitted. The checklist will issue a warning if the number of isoforms in the merged GFF file differs from the expected number of isoforms found in the corresponding *D. melanogaster* ortholog. The checklist will issue an error if it fails to find the ortholog name in *D. melanogaster* (Figure 2). **Please provide an explanation for any warnings or errors shown in the Isoforms Checklist in your GEP Annotation Report.**

Figure 2 The Annotation Files Merger will issue a warning if the number of isoforms in the merged GFF file differs from the *D. melanogaster* ortholog. It will issue a failure if it cannot find the gene symbol in *D. melanogaster*.

In addition to the Isoform Checklist, the Annotation Files Merger also allows you to view all the gene models in the merged GFF file in the GEP UCSC Genome Browser mirror (Figure 3). Under the "Show merged GFF File in the Genome Browser" section, select the project region for your project (e.g. *D. erecta* 2nd 3L control) and then click on the "Show Track" button. **Note that this is one of the required screenshot in the GEP Annotation Report.**
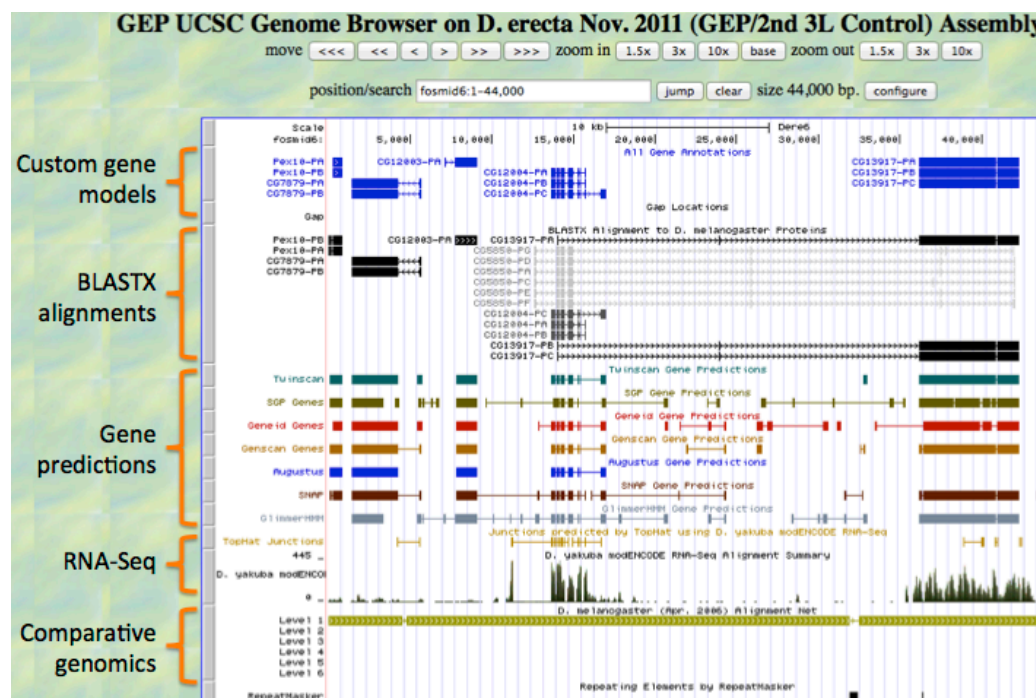

Figure 3 Custom gene models (blue) from the merged GFF file on the GEP UCSC Genome Browser mirror. This is one of the required screenshots for the GEP Annotation Report.
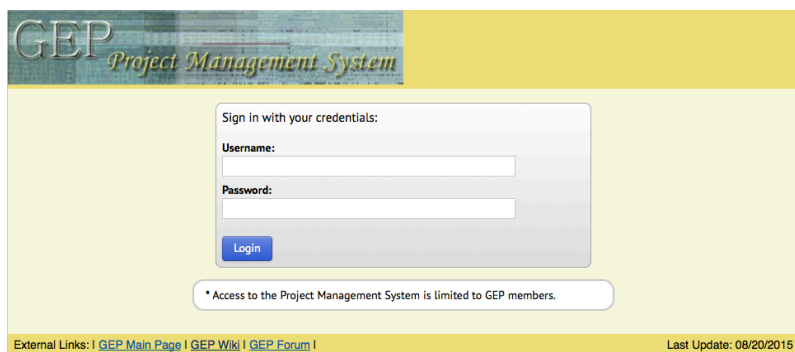
Please refer to the Annotation Files Merger User Guide in the documentations section of the GEP web site for additional instructions on how to use the Annotation Files Merger.

Submitting Annotation Projects with at Least One Gene

Once you have generated the supplemental files and completed the annotation report, you are ready to submit the annotation project to the GEP. In this tutorial, we will submit the fosmid6 project from the *D. erecta* 2nd 3L control region (derecta_2nd3Lcontrol_Nov2011_fosmid6) using the GEP Project Management System.

Procedure:

1. Login to the GEP Project Management System (Figure 4)
   - Navigate to the GEP Project Management System through the GEP home page (http://gep.wustl.edu ➔ "Projects" ➔ "Project Management System")
   - Enter the credentials issued to you by the GEP:
     o Username: [<first letter of first name><last name>]
     o Password: [<issued by the GEP>]
   - Click "Login"

Figure 4 Login screen for the GEP Project Management System.

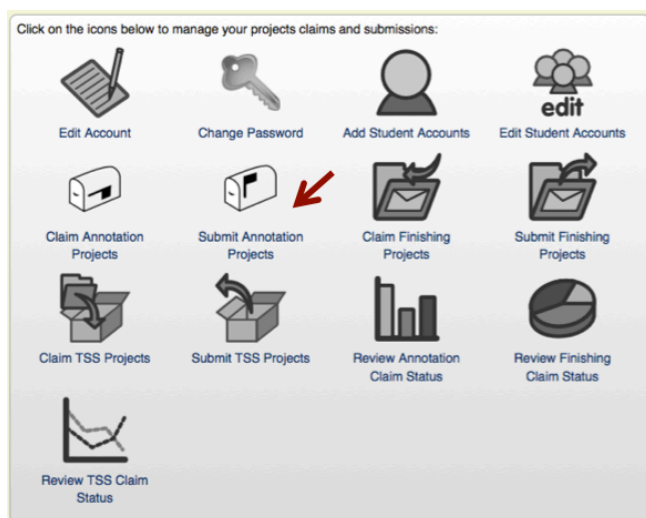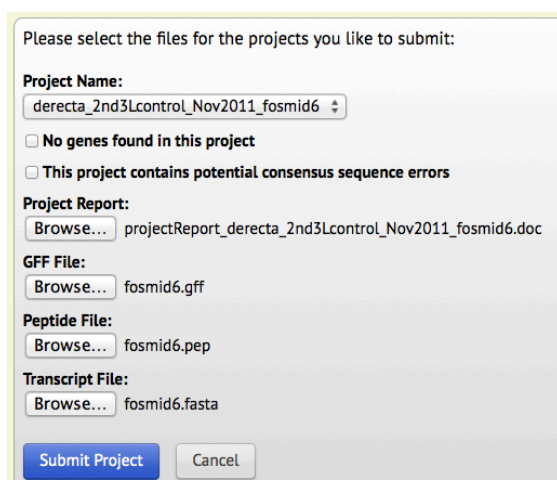2. Click on the link "Submit Annotation Projects" link on the main page (Figure 5).

Figure 5 Navigate to the "Submit Annotation Project" section to submit annotation projects.

3. Select the appropriate project and the four required files (i.e. the GEP Annotation Report, GFF, transcript and peptide sequence files). Then click on the "Submit Project" button (Figure 6).

Figure 6 Submitting the fosmid6 project using the Project Management System by selecting the project and uploading the four required project files (i.e. annotation report, GFF, peptide and transcript sequence files).

4. If all goes well, you should see a message that confirms the project has been successfully submitted to the GEP (Figure 7).
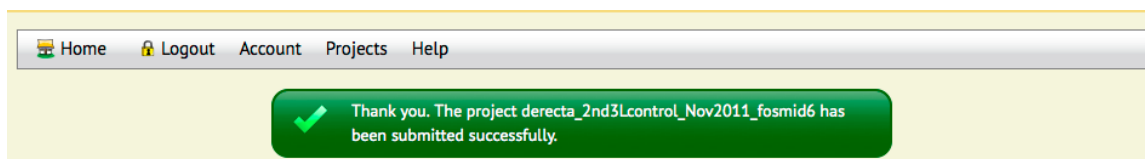

Figure 7 Confirmation message from the Project Management System indicating that the project has been submitted successfully.

> **Note:** Please send an email to Wilson (wleung@wustl.edu) to confirm that the GEP has received your project submission and that it contains all the information required for project reconciliation.

Submitting Annotation Projects with No Genes

If your project does not have any genes, the only file required for project submission is the completed GEP Annotation Report. However, before submitting a project with no gene annotations, please verify that you have completed all the sequence searches described at the bottom of page 1 of the GEP Annotation Report and that you have included all the required evidence (*e.g.*, screenshots of BLAST alignment results) to support this hypothesis.

To submit a project with no genes, select the project under the "Project Name" field, select the checkbox "No genes found in this project", and then select the completed GEP Annotation Report under the "Project Report" field. Click on the "Submit Project" button to submit the project to the GEP (Figure 8).
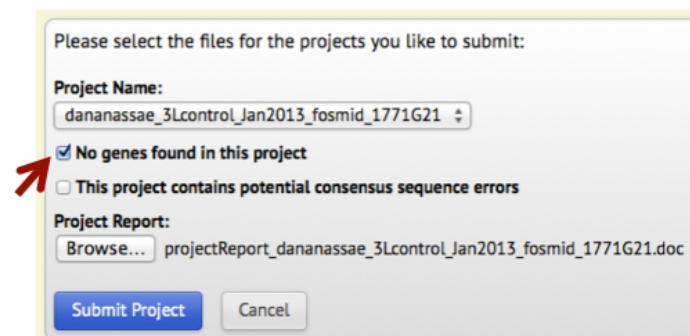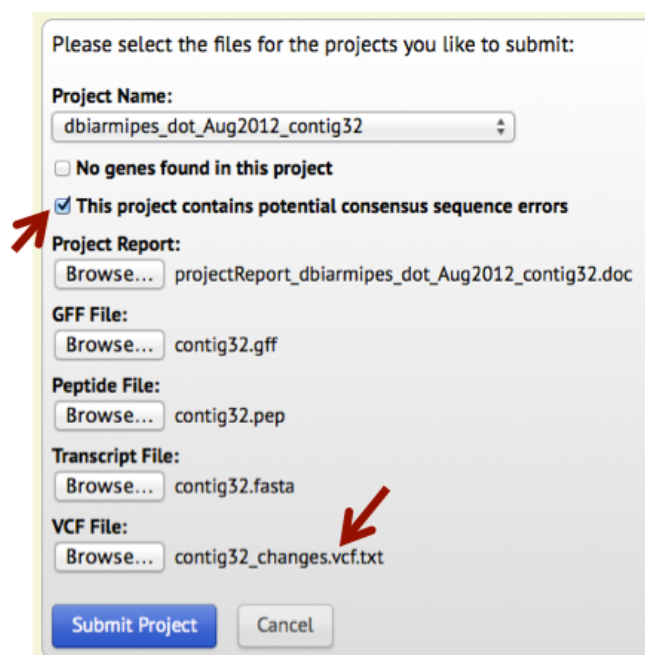

Figure 8 Configure the annotation project submission form to submit a project with no genes.

Submitting Annotation Projects with Consensus Sequence Errors

For projects with errors in the consensus sequence that affect the coding region annotations, students should use the Sequence Updater (available through the GEP home page under "Projects" ➜ "Annotation Resources") to document the changes to the original project sequence. Similar to the GFF, transcript and peptide sequence files, students should use the Annotation Files Merger to combine all the VCF files generated by the Sequence Updater into a single VCF file that describes all the sequence changes in the entire project.

To submit an annotation project with errors in the project sequence, select the project under the "Project Name" field and then select the checkbox next to the "This project contains potential consensus sequence errors" field. A new "VCF File" field will appear on the submission form that will allow you to upload the combined VCF file.

For projects with annotated genes, you will also need to upload the GEP Annotation Report, the combined GFF, transcript and peptide sequence files as described above (Figure 9). Note that all the exon and gene coordinates should always be **relative to the coordinates of the original project sequence** irrespective of whether there are errors in the consensus project sequence.

Figure 9 For projects with errors in the consensus sequence, you should select the "This project contains potential consensus sequence errors" checkbox and upload the combined VCF file.

To submit a project with errors in the consensus sequence but without any genes, you can select the checkboxes for the "No genes found in this project" field and the "This project contains potential consensus sequence errors" field. Then you can select the project under the "Project Name" field and upload both the completed GEP Annotation Report and the combined VCF file.

Additional Resources
Please contact Wilson (wleung@wustl.edu) if you encounter any problems or have any questions regarding the protocol for submitting annotation projects.