Parasitoid Wasps Project: Annotation Workbook Directions

Modified by Marisol Santisteban from Version 2.2 created by Marisol Santisteban & Charles Hauser from Version 2.1 created by Nate Mortimer, Melanie Van Stry, Rebecca Burgess, Michael Rubin, Rivka Glaser, & Chris Jones

Resources & Tools

- All links for this lesson can be found on the Parasitoid Wasps Project page of the GEP website (https://thegep.org/wasps/).

Table of Contents

Introduction

For the wasp venom gene annotation project, your project submission will consist of a completed annotation workbook, a screenshot of the user track on the genome browser, and sequence files for the entire transcript and encoded protein.

The annotation workbook is an Excel file with space to fill out all the necessary information for the gene annotation. If it is done correctly, filling out the workbook will also auto-populate a GFF file that can be used for testing the accuracy of gene models (described in the "An alternative to GMC" section below). Please note that everything should be filled out exactly as described: the vocabulary and formatting need to be very precise to make a functional GFF file.

You can download the annotation workbook, including an example venom gene annotation from the "Project Curriculum" section of the Parasitoid Wasps Project page (https://thegep.org/wasps/) on the GEP website.

Workbook Instructions and Transcript Sheets

The first sheet of the Excel document will have the instructions (Figure 1) and the third sheet has the template for inputting your data. Use a separate sheet within the workbook for each transcript of the gene that you annotate.

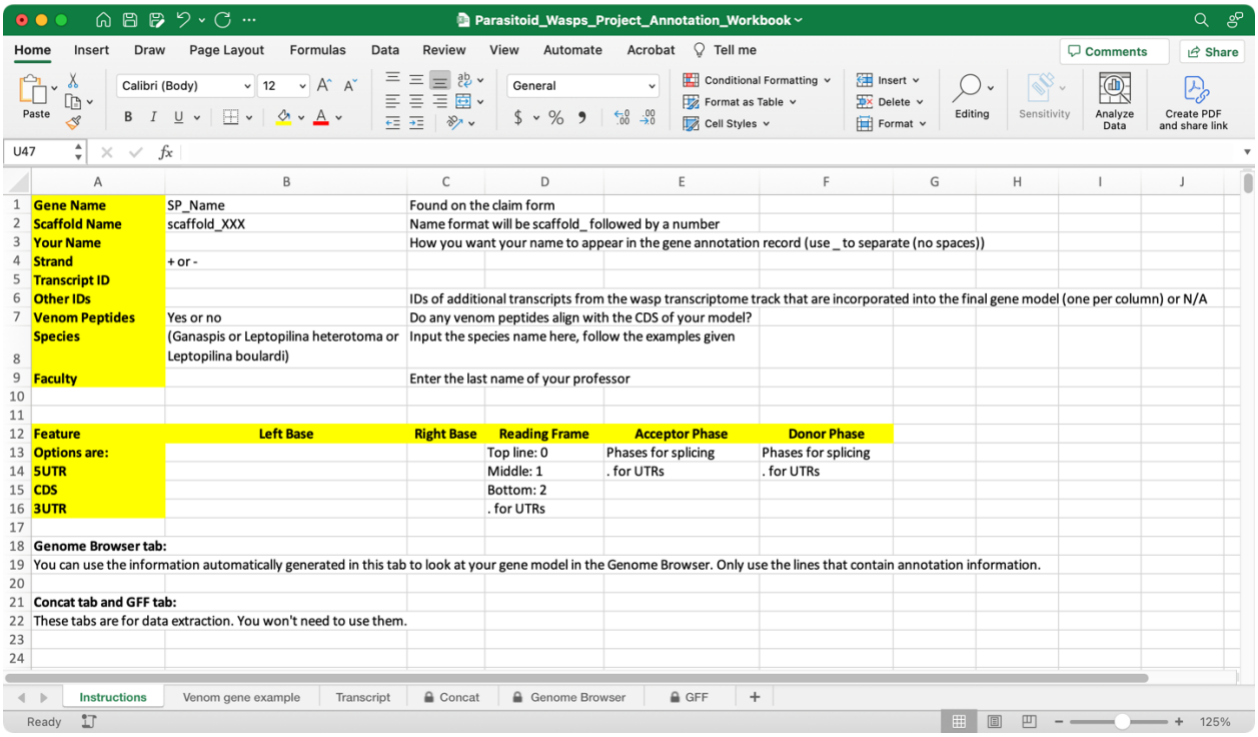

Figure 1. The first sheet contains the instructions. The fields that will need to be filled out are highlighted in the above figure for illustrative purposes only.

Start by filling in the basic information on the third tab (Transcript). Be very careful to follow this guide for formatting.

- **Gene Name**: The name of the gene; this is given to you by your instructor with your assignment.
- **Scaffold Name**: This is found in the top left corner of the Genome Browser (Figure 2). Be careful to fill this out exactly (lower cases in scaffold, underscore rather than a space).
- **Your Name**: Write your name how you want it to appear in the gene annotation record with underscores to separate (no spaces) (e.g., John_Doe or John_C_Doe).
- **Strand**: Use + or – to indicate the DNA strand the gene is on.
    - Note: Do NOT use the transcriptome track to determine + or – strand. Use the other data (e.g., StringTie, gene predictors) instead.
    - Hint: type ‘+ (apostrophe +) or ‘- (apostrophe -) to prevent Excel from thinking you are typing a formula.
- **Transcript ID**: ID corresponding to the Species Transcriptome track. This is what you use to search for the gene and is given to you with your assignment (Figure 2).

- **Other IDs**: Any other transcripts from the Species Transcriptome track that form part of your final gene model. Many of the Transcriptome track gene models are incomplete so you may have to combine multiple transcripts to make a complete model.
- **Venom Peptides**: Indicate whether there any peptides in the "Species Venom Peptides" track that align with the CDS of your gene.
- **Species**: The species of parasitoid that the transcript comes from; this is given to you by your instructor with your assignment.
- **Faculty**: Input the last name of your instructor.

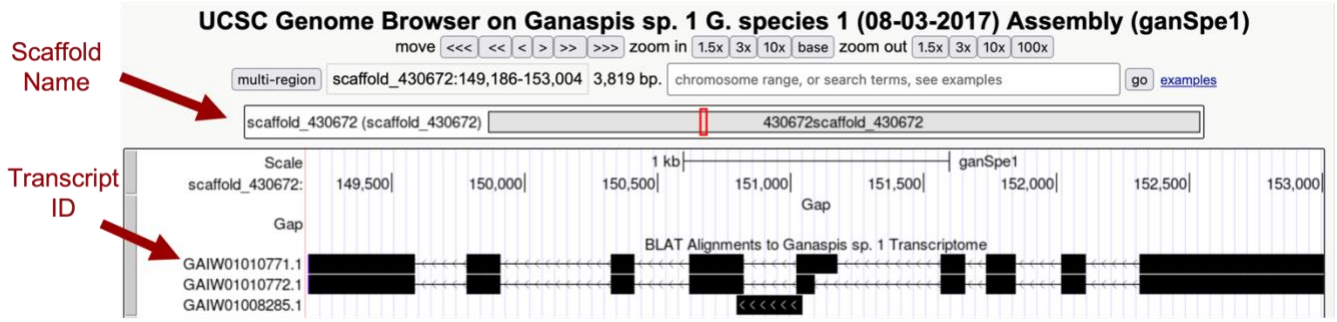

Figure 2. Parasitoid Wasps Genome Browser showing where to find the "Scaffold Name" and "Transcript ID" to input in your workbook. NOTE: You may have to set the "G1 Transcriptome" track to "full" in order to see the Transcripts IDs.

Then start annotating the gene as you learned previously. Add more rows to the annotation region using the insert command in Excel if you need more lines (and sorry that you have so many exons!).

Feature:

You will want to annotate the 5' UTR, CDS exons, and 3' UTR. Be sure to specify each feature by using:

Feature
Options are:
5UTR
CDS
3UTR

The above three features are the only options that will work for later analysis so use these names *exactly*.

Note: The transcriptome track does not attempt to differentiate between UTR and CDS. You will have to use what you've learned about these features to determine the UTR and CDS regions of your gene.

First Base, Last Base: the first and last bases of the feature

For 5UTR, the last base should be the final base before the ATG of the start codon. The first CDS will start with the A of the ATG and the final CDS will end with the final base of the STOP codon. The 3UTR then starts with the first base following the STOP.

Note: Regardless of which strand your gene is on, the first base should be a lower number than the last base. So, if you are annotating a gene on the minus strand the numbers will have to go in "backwards" so that the "first base" is the smaller number. You'll use the strand field as described above to designate + or – strand.

Reading Frame:

Your options are 0, 1, 2. What we have referred to as Frame 1 (either +1 or -1) now becomes 0, Frame 2 becomes 1, and Frame 3 becomes 2. Because the 5UTR and 3UTR don't have a reading frame, use a period (.) in these boxes.

Acceptor Phase and Donor Phase:

These fields are for you to keep track of your splicing phases as determined in your annotation. Because the 5UTR and 3UTR don't have splicing phases, use a period (.) in these boxes. Also, the first CDS exon won't have an acceptor phase and the last CDS exon won't have a donor phase so use a period (.) in these boxes as well.

Once you have filled all the data for the gene in this example, the Transcript tab would look like this (Figure 3):

Gene Name	G1_PHGPx.right				
Scaffold Name	scaffold_430672				
Your Name	John_Doe				
Strand	+				
Transcript ID	GAIW01010771.1				
Other IDs	N/A				
Venom Peptides	Yes				
Species	Ganaspis				
Faculty	Mortimer				
Feature	**First Base**	**Last Base**	**Reading Frame**	**Acceptor Phase**	**Donor Phase**
5UTR	150642	150822	.	.	.
5UTR	151023	151100	.	.	.
CDS	151101	151176	2	.	1
CDS	151565	151658	0	2	2
CDS	151736	151845	2	1	1
CDS	152019	152110	1	2	0
CDS	152311	152526	0	0	.
3UTR	152527	153004	.	.	.

Figure 3. Workbook "Transcript" sheet filled in.

As you are filling out the data in the third tab of the Excel file (Transcript), the other tabs—Concat, Genome Browser, and GFF—will automatically populate (Figures 4, 5, and 6).

	A	B	C	D	E	F
1	gene_id "G1_PHGPx.right"					
2	transcript_id "GAIW01010771.1"					
3						
4	Feature	First Base	Last Base	Reading Frame	Acceptor Phase	Donor Phase
5	5UTR	150642	150822	.	.	.
6	5UTR	151023	151100	.	.	.
7	CDS	151101	151176	2	.	1
8	CDS	151565	151658	0	2	2
9	CDS	151736	151845	2	1	1
10	CDS	152019	152110	1	2	0
11	CDS	152311	152526	0	0	.
12	3UTR	152527	153004	.	.	.
13	0	0	0	0	0	0

Instructions | Venom gene example | Transcript | Concat | Genome Browser | GFF | +

Figure 4. The "Concat" sheet will automatically populate based on the information you enter in Figure 3.

	A	B	C	D	E	F	G	H	I
1	GFF info:								
2	scaffold_430672	John_Doe	5UTR	150642	150822	.	+	.	G1_PHGPx.right
3	scaffold_430672	John_Doe	5UTR	151023	151100	.	+	.	G1_PHGPx.right
4	scaffold_430672	John_Doe	CDS	151101	151176	2	+	2	G1_PHGPx.right
5	scaffold_430672	John_Doe	CDS	151565	151658	0	+	0	G1_PHGPx.right
6	scaffold_430672	John_Doe	CDS	151736	151845	2	+	2	G1_PHGPx.right
7	scaffold_430672	John_Doe	CDS	152019	152110	1	+	1	G1_PHGPx.right
8	scaffold_430672	John_Doe	CDS	152311	152526	0	+	0	G1_PHGPx.right
9	scaffold_430672	John_Doe	3UTR	152527	153004	.	+	.	G1_PHGPx.right
10	scaffold_430672	John_Doe	0	0	0	0	+	0	G1_PHGPx.right
11	scaffold_430672	John_Doe	0	0	0	0	+	0	G1_PHGPx.right

Instructions | Venom gene example | Transcript | Concat | Genome Browser | GFF | +

Figure 5. The "Genome Browser" sheet will automatically populate based on the information you enter in Figure 3.

	A	B	C	D	E	F	G	H	I
1	scaffold_430672	John_Doe	5UTR	150642	150822	.	+	.	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
2	scaffold_430672	John_Doe	5UTR	151023	151100	.	+	.	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
3	scaffold_430672	John_Doe	CDS	151101	151176	2	+	.	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
4	scaffold_430672	John_Doe	CDS	151565	151658	0	+	2	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
5	scaffold_430672	John_Doe	CDS	151736	151845	2	+	1	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
6	scaffold_430672	John_Doe	CDS	152019	152110	1	+	2	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
7	scaffold_430672	John_Doe	CDS	152311	152526	0	+	0	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
8	scaffold_430672	John_Doe	3UTR	152527	153004	.	+	.	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";
9	scaffold_430672	John_Doe	0	0	0	0	+	0	gene_id "G1_PHGPx.right"; transcript_id "GAIW01010771.1";

Instructions | Venom gene example | Transcript | Concat | Genome Browser | GFF | +

Figure 6. The "GFF" sheet will automatically populate based on the information you enter in Figure 3.

The “Genome Browser” sheet will allow you to add your own annotation into the Genome Browser. This can be used to check the accuracy of your annotation and to get the predicted sequence of your gene as detailed below.

Checking Accuracy of the Gene Model (Alternative to Gene Model Checker)

To see your gene model in the context of the other evidence on the browser, you will need to upload the data from your workbook. Go to the browser window and click on the “add custom tracks” button found under the viewer window (Figure 7).

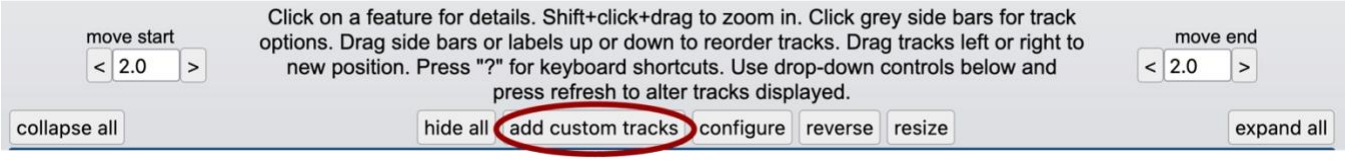

Figure 7. Click on the “add custom tracks” button to add your gene model to the browser window.

In your Annotation Workbook, click on the “Genome Browser” tab and copy the lines that contain the full GFF information (shaded) (Figure 8).

	A	B	C	D	E	F	G	H	I
1	GFF info:								
2	scaffold_430672	John_Doe	5UTR	150642	150822	.	+	.	G1_PHGPx.right
3	scaffold_430672	John_Doe	5UTR	151023	151100	.	+	.	G1_PHGPx.right
4	scaffold_430672	John_Doe	CDS	151101	151176	2	+	2	G1_PHGPx.right
5	scaffold_430672	John_Doe	CDS	151565	151658	0	+	0	G1_PHGPx.right
6	scaffold_430672	John_Doe	CDS	151736	151845	2	+	2	G1_PHGPx.right
7	scaffold_430672	John_Doe	CDS	152019	152110	1	+	1	G1_PHGPx.right
8	scaffold_430672	John_Doe	CDS	152311	152526	0	+	0	G1_PHGPx.right
9	scaffold_430672	John_Doe	3UTR	152527	153004	.	+	.	G1_PHGPx.right
10	scaffold_430672	John_Doe	0	0	0	0	+	0	G1_PHGPx.right
11	scaffold_430672	John_Doe	0	0	0	0	+	0	G1_PHGPx.right
12	scaffold_430672	John_Doe	0	0	0	0	+	0	G1_PHGPx.right
13	scaffold_430672	John_Doe	0	0	0	0	+	0	G1_PHGPx.right
14	scaffold_430672	John_Doe	0	0	0	0	+	0	G1_PHGPx.right

Instructions | Venom gene example | Transcript | Concat | Genome Browser | GFF | +

Figure 8. The “Genome Browser” sheet of the Annotation Workbook.

Paste the data into the “Paste URLs or data” field of the Add Custom Tracks screen and click on “Submit” (Figure 9).

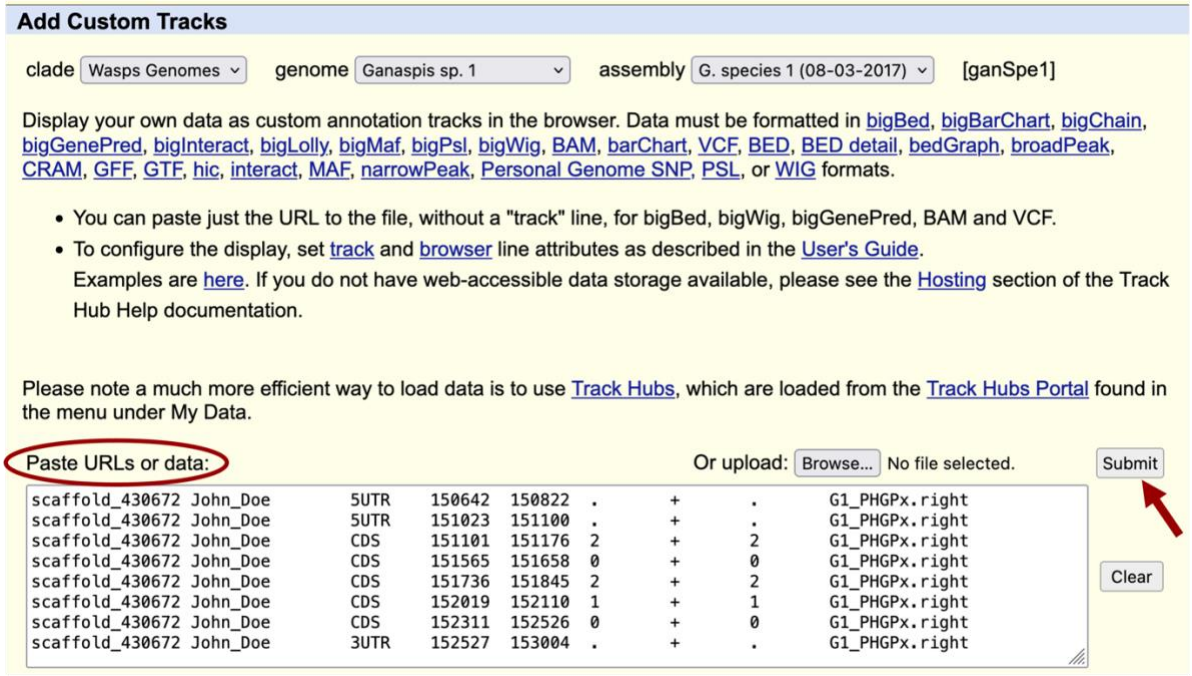

Figure 9. Pasting the coordinates of the annotated gene in the “Add Custom Tracks” window.

This will bring you to the Manage Custom Tracks screen. You can rename your gene model if you wish by clicking on the “User Track” link (Figure 10, arrow) then replacing the highlighted text following “track name=” in the “Edit configuration” field (Figure 10, shaded).

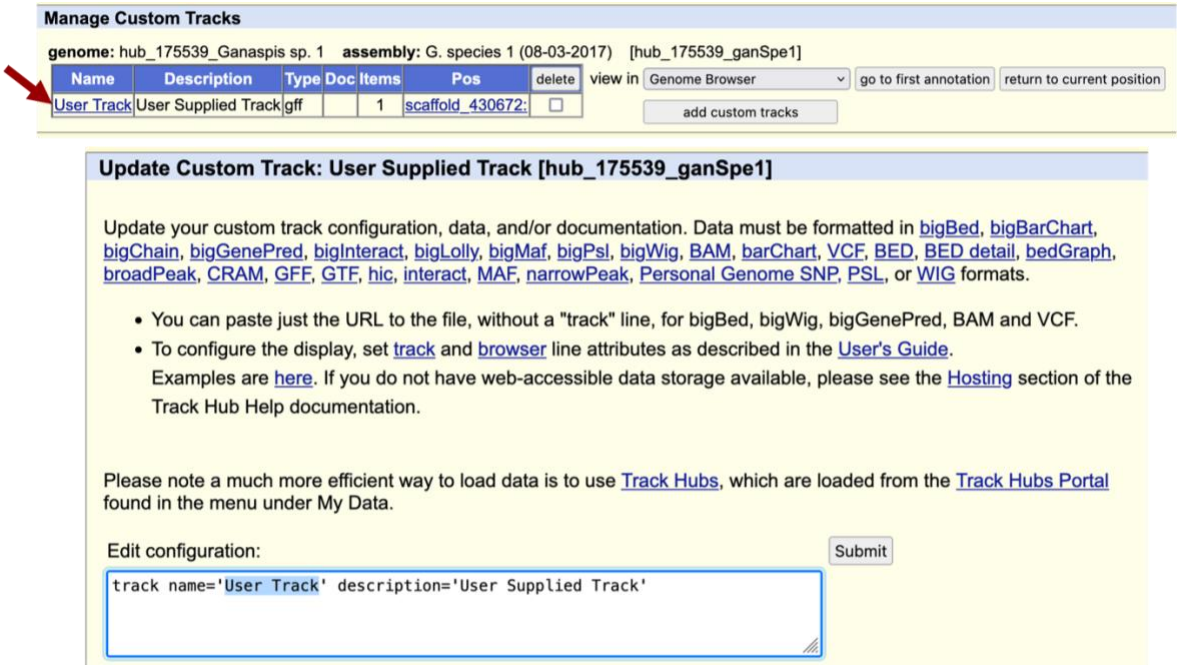

Figure 10. Renaming your gene model.

There is no need to change the track name every time you make a track, but this does let you view two or more of your own gene models at the same time which is often useful. Clicking on “Submit” will take you back to the Manage Custom Tracks screen.

Click on the link with the project name, scaffold_430672 in our example, to view your annotation in the browser (Figure 11).

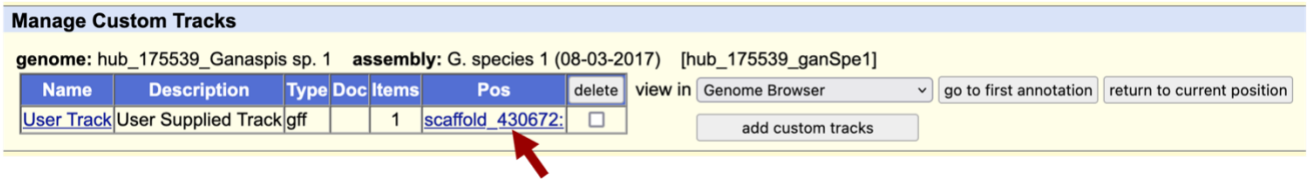

Figure 11. Add a custom track containing your gene model to the Genome Browser window.

The next screen will have a new track with your gene model centered around the coordinates you annotated. In this example, because we did not change the Name, the track appears as “User Supplied Track” (Figure 12, red box). You can customize the number of tracks shown on the viewing window.

At a minimum, you would want to have the following tracks turned on:

- SPALN Alignments
- At least one Gene Predictor track (in this example Augustus, N-SCAN, Geneid)
- At least one RNA-Seq track (in this example Unpaired and Paired-end RNA-Seq Coverage, StringTie)
- Ganaspis Venom Peptides

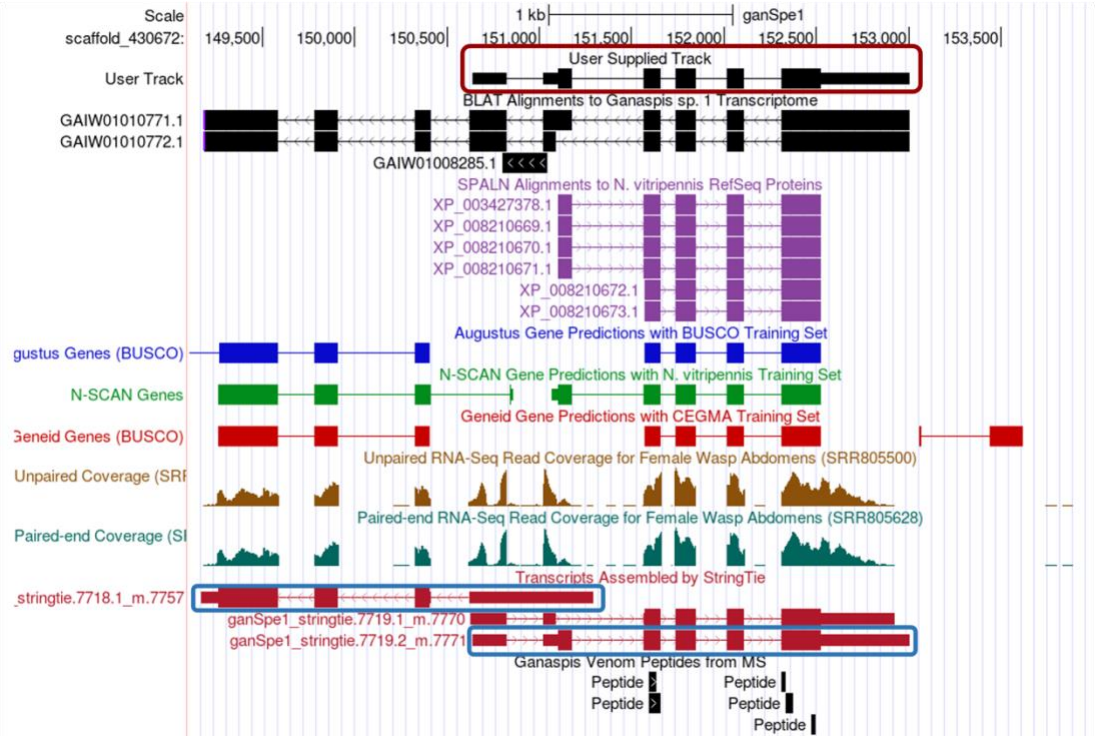

Figure 12. Your custom gene model visualized in the Genome Browser window.

Under this zoomed-out view, it can be appreciated that the *de novo* transcriptome GAIW01010771.1 prediction is probably bringing together two separate genes. This commonly happens when the genes have overlapping UTRs (one of the big issues with unstranded RNA-Seq).

Both Augustus and N-SCAN predict a gene on the left part of the region and StringTie predictions fall into two separate sections (Figure 12, blue boxes).

For this example, we have annotated the gene on the right side of the transcript ID—corresponds to the StringTie prediction on the right. However, the gene to the left side would also need to be annotated using a separate workbook.

For the submission of your project, you will need to take a screenshot of the browser with your gene model in it.

Getting the Sequence Data for the pep and fasta files

In the main navigation menu at the top of the Genome Browser, click on “Tools” and then “Table Browser” (Figure 13).

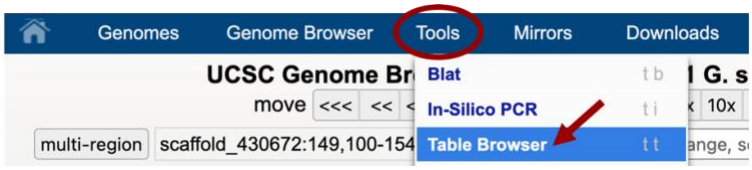

Figure 13. Navigate to the “Table Browser.”

Within Table Browser, set the "group" field to “All Tracks,” the "track” field to “User Track,” the “region” field to “position,” and “output format” to “sequence.” Then click on the “get output” button (Figure 14).

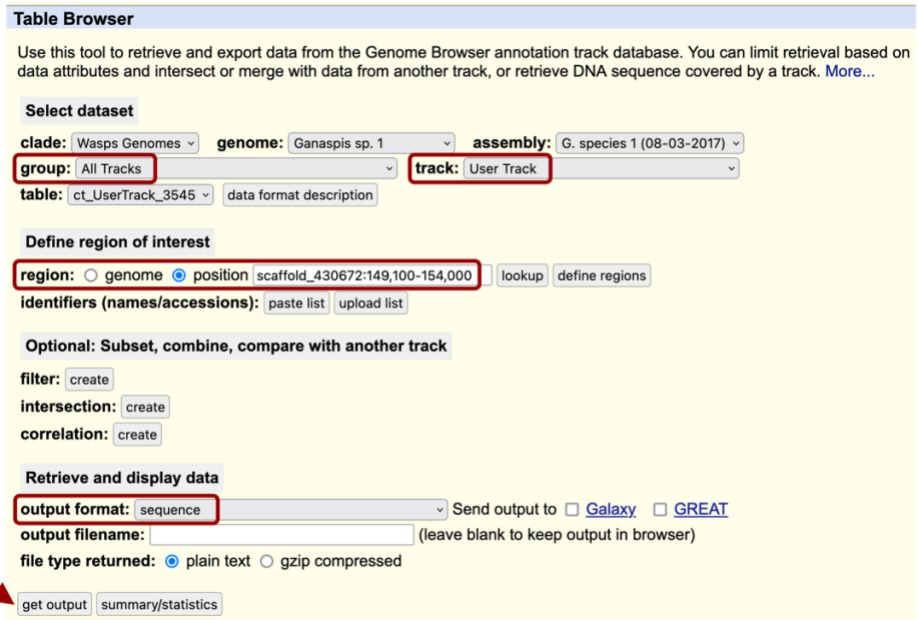

Figure 14. Configure the Table Browser.

To retrieve the protein sequence, select “CDS Exons”, “One FASTA record per gene” and “All upper case”, then click on “get sequence” (Figure 15).

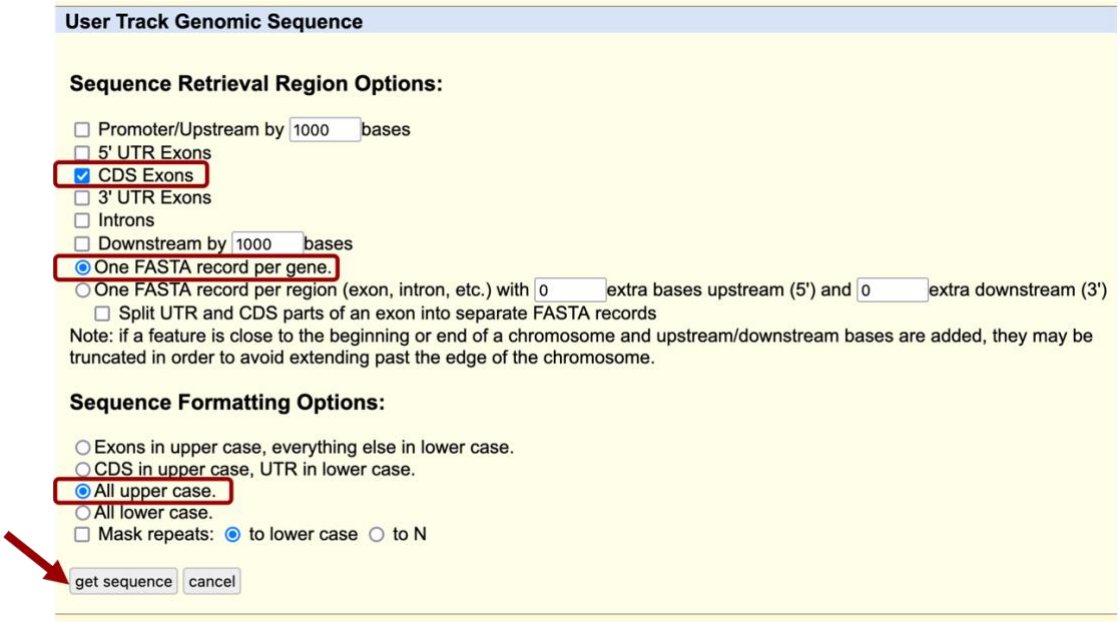

Figure 15. Retrieve the protein sequence.

Copy and paste the nucleotide sequence from the next screen into [NCBI’s ORFfinder](). Then set “Minimal ORF length (nt)” to “300” and check the box next to “Ignore nested ORFs.” Click on the “Submit” button (Figure 16).

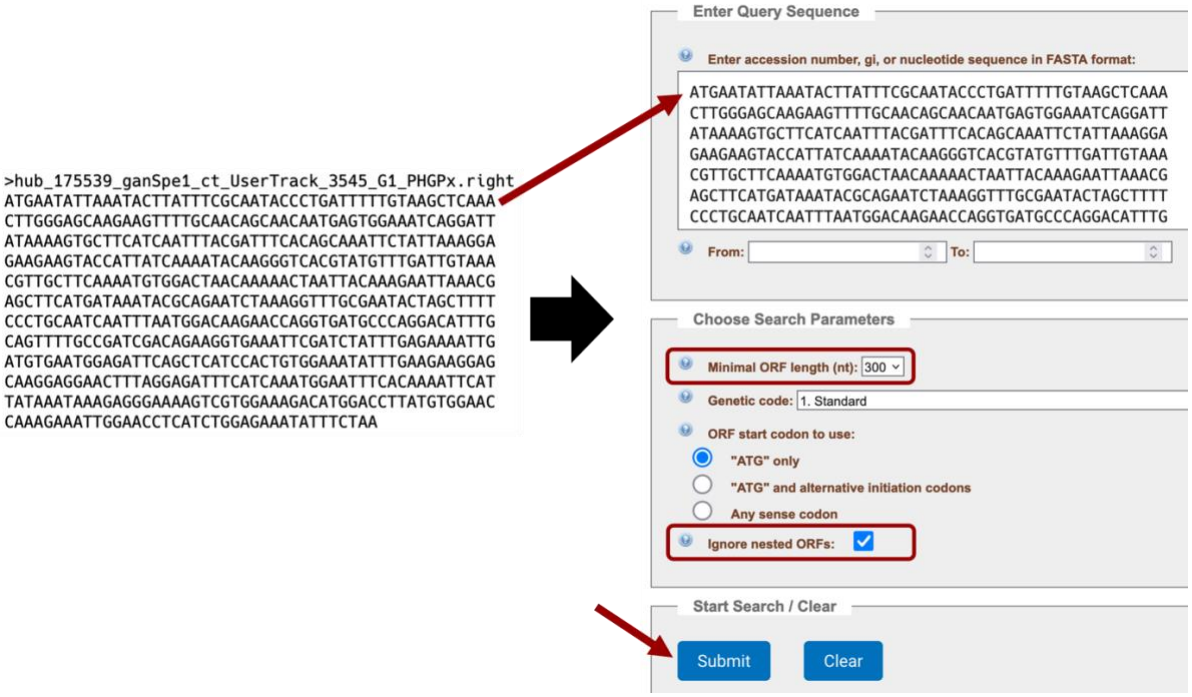

Figure 16. To search for Open Reading Frames in your DNA sequence, paste the nucleotide sequence in the ORFfinder.

If your annotation is correct, your sequence will be translated in the correct frames and you will get a single ORF that spans the entire sequence (Figure 17). Take a screenshot of the ORF map and copy/paste the amino acid sequence into a new text file. Name this file "xxxx_protein." If you get an ORF that is smaller or incomplete, something is annotated wrong. Use the user track on the Genome Browser to verify your exon coordinates.

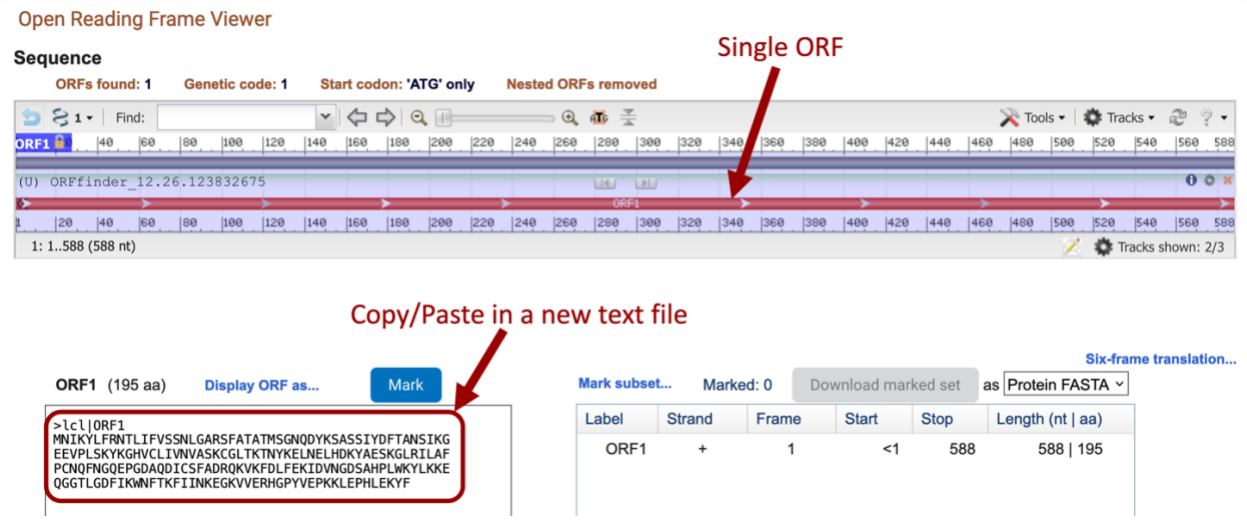

Figure 17. Open Reading Frame Viewer and protein sequence (box).

To get the sequence of the whole transcript, go back to the tab with the Genome Browser, currently showing the nucleotide sequence of the translated portion of the gene (Figure 16, left), and then click on the back button to navigate back to the Figure 15 screen. Now select "5' UTR Exons," "CDS Exons," "3' UTR Exons," "One FASTA record per gene," and "All upper case." Then click on "get sequence" (Figure 18).

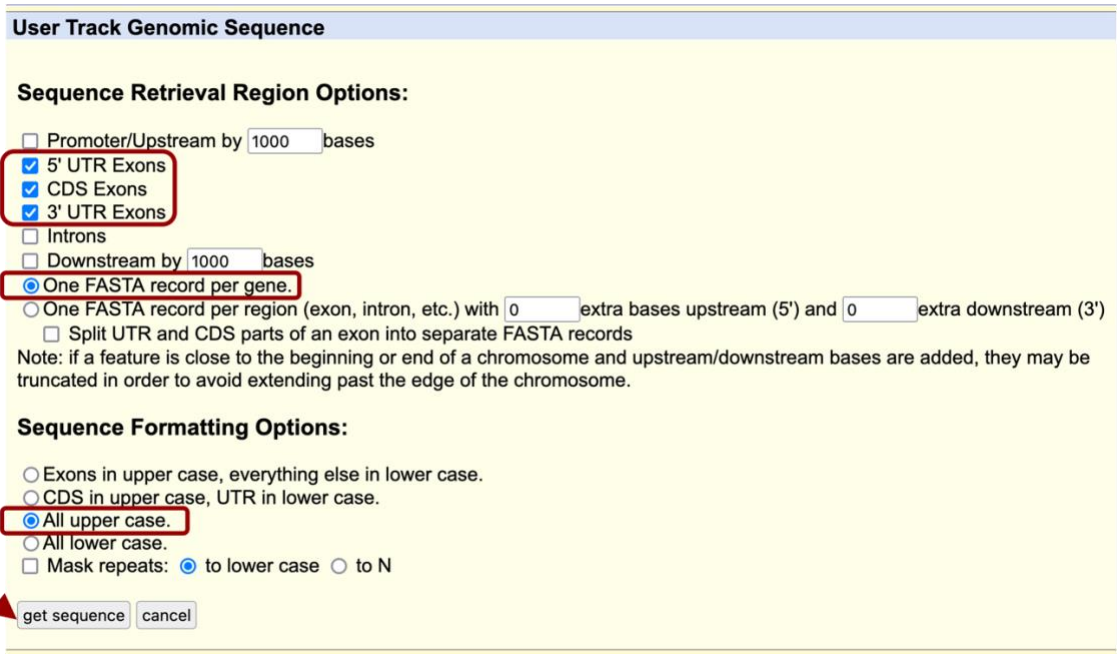

Figure 18. Retrieve the transcript sequence.

Copy/paste the sequence (Figure 19) into a new text file. Name the file “xxxxx_mRNA.”

```
>hub_175539_ganSpe1_ct_UserTrack_3545_G1_PHGPx.right range=scaffold_430672:150642-153004 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ACTTTTTTTCTAGATTCAAACGCTATATATAAATGAAAGAATAAACTTTC
GTCCAATTTTCCAAATTTTCAGGTTATGGTGCAATGTTTACTGCGGTCAT
GCGCAAGCGACGTAATAAAATATATATATAATTGTCACAATGAGTTTCGT
TTTTTAACTGACTACAGTCTCGTTATTTTCGAACTTTAACTATATAGACA
CTCCAAAACATTCTACGGATTCAAAATACTCAAAGGATCAAAAAAAAGTC
GGTGAGAATATGAATATTAAATACTTATTTCGCAATACCCTGATTTTTGT
AAGCTCAAACTTGGGAGCAAGAAGTTTTGCAACAGCAACAATGAGTGGAA
ATCAGGATTATAAAAGTGCTTCATCAATTTACGATTTCACAGCAAATTCT
ATTAAAGGAGAAGAAGTACCATTATCAAAATACAAGGGTCACGTATGTTT
GATTGTAAACGTTGCTTCAAAATGTGGACTAACAAAAACTAATTACAAAG
AATTAAACGAGCTTCATGATAAATACGCAGAATCTAAAGGTTTGCGAATA
CTAGCTTTTCCCTGCAATCAATTTAATGGACAAGAACCAGGTGATGCCCA
GGACATTTGCAGTTTTGCCGATCGACAGAAGGTGAAATTCGATCTATTTG
AGAAAATTGATGTGAATGGAGATTCAGCTCATCCACTGTGGAAATATTTG
AAGAAGGAGCAAGGAGGAACTTTAGGAGATTTCATCAAATGGAATTTCAC
AAAATTCATTATAAATAAAGAGGGAAAAGTCGTGGAAAGACATGGACCTT
ATGTGGAACCAAAGAAATTGGAACCTCATCTGGAGAAATATTTCTAAATT
ACAAATGATCAAGATGCTATATTTTGTATTGCACAAAAGTCACCAGTGAA
TTATGAAATTTGAGTGTGATAAAATTATTAATTCACATATAATCATTTTT
CAAATTAATTACTTTTTATACATACATGGAATATGCAAATTCCTTGTCAA
TAGAGTTAAGAATCGAAGTAGCGAAGGTTCTTTATTTTTATTGCAACGTT
AATCGATAATCACTTTACTTTTTTGTTTATTCTTAAATATTTACATTATA
CAACTCTGTAAACTTGTTGAATAAAACTTCATGAATTATTAATTGAAAAC
GAACACCCACCTTTTACAATTTCAGATATATACGTGTATTTAAATTCTGA
TAAATCAATTTTTATATTCATTTTTAATTTGGTGTATTTTTACTTGATAT
ATTTAAATGAAACTTTCCTAAATATTTTCAAAAATTGTATTTAAATAAAA
GAAAATAAATACGTTTATCAAAAAT
```

Figure 19. Save the transcript sequence as a txt file.