



Dec 16, 2020

## 🌐 Annotating genes in *Diaphorina citri* genome version 3

Teresa D Shippy<sup>1</sup>, S Miller<sup>2</sup>, C Massimino<sup>3</sup>, C Vosburg [Indian River State College<sup>4</sup>, PS Hosmani<sup>5</sup>, M Flores-Gonzalez<sup>5</sup>, LA Mueller<sup>5</sup>, WB Hunter<sup>6</sup>, JB Benoit<sup>7</sup>, SJ Brown<sup>1</sup>, T D'elia<sup>3</sup>, S Saha<sup>5</sup>

<sup>1</sup>Kansas State University; <sup>2</sup>Kansas State University, Allen County Community College; <sup>3</sup>Indian River State College;

<sup>4</sup>Indian River State College, The Pennsylvania State University; <sup>5</sup>Boyce Thompson Institute;

<sup>6</sup>USDA-ARS U.S. Horticultural Research Laboratory; <sup>7</sup>University of Cincinnati

2

Works for me

dx.doi.org/10.17504/protocols.io.bniimcce

D. citri annotation

Teresa Shippy  
Kansas State University

DOI

[dx.doi.org/10.17504/protocols.io.bniimcce](https://dx.doi.org/10.17504/protocols.io.bniimcce)

PROTOCOL CITATION

Teresa D Shippy, S Miller, C Massimino, C Vosburg [Indian River State College, PS Hosmani, M Flores-Gonzalez, LA Mueller, WB Hunter, JB Benoit, SJ Brown, T D'elia, S Saha 2020. Annotating genes in *Diaphorina citri* genome version 3. **protocols.io**  
<https://dx.doi.org/10.17504/protocols.io.bniimcce>

LICENSE

— This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

CREATED

Oct 16, 2020

LAST MODIFIED

Dec 16, 2020

PROTOCOL INTEGER ID

43306

- 1 Gather background information about the gene you have chosen. Use [PubMed](https://pubmed.ncbi.nlm.nih.gov/) to find relevant publications about orthologs in other insects. Reading these papers will help you know what to expect when you begin annotating the *D. citri* gene (i.e. exon number, domains, isoforms, etc.).

You can also use [DiaphorinaCyc](https://www.ncbi.nlm.nih.gov/datasets/genome/all_genomes.shtml), [OrthoDB](https://orthodb.org/) or other databases to gather information about your gene of interest.

- 2 Take thorough notes (preferably in a Word or PowerPoint document) throughout the annotation process. Document the results of each step and any choices or changes you make. Screenshots are often helpful.
- 3 Find orthologous protein sequences for your gene of interest by starting with a well-established gene (perhaps from *Drosophila melanogaster*) and then performing BLASTs and reciprocal BLASTs using the [NCBI BLASTp](https://blast.ncbi.nlm.nih.gov/Blast.cgi) function. Suggested insects for orthologs include:
  - *Drosophila melanogaster* (fruit fly; Order: Diptera)
  - *Anopheles gambiae* (African Malaria mosquito; Order: Diptera)

- *Tribolium castaneum* (red flour beetle; Order: Coleoptera)
- *Apis mellifera* (Western honey bee; Order: Hymenoptera)
- *Nasonia vitripennis* (parasitic jewel wasp; Order: Hymenoptera)
- *Acyrtosiphon pisum* (pea aphid; Order: Hemiptera)
- *Bemisia tabaci* (tobacco or silverleaf whitefly; Order: Hemiptera).

Copy the FASTA-formatted orthologous sequences into a text file for future use.


- Before you begin searching for *D. citri* orthologs, it may be helpful to build a multiple sequence alignment and phylogenetic tree with the orthologs from other insect species using Mega7 or MegaX. These analyses can give insight into which regions of the protein are likely to be conserved and how the hemipteran orthologs compare to those of other insects (although keep in mind that the sequences you have gathered, particularly the hemipteran, may just be computer predictions on a draft assembly and thus may not be completely accurate).

MEGA 7 [↗](#)

MEGA X [↗](#)

- To find your gene of interest in *D. citri*, we recommend using the *T. castaneum* or *A. pisum* ortholog to BLAST against *D. citri* sequences at [www.citrusgreening.org](http://www.citrusgreening.org).

⊖ Input parameters

Categories	Psyllid Databases 
Database	<input checked="" type="checkbox"/> Diaphorina citri MCOT proteins <a href="#">db details</a> <input type="checkbox"/> Diaphorina citri NCBI v100 proteins <input type="checkbox"/> Diaphorina citri OGS v1.0 proteins <input type="checkbox"/> Diaphorina citri OGS v1.0 proteins <input type="checkbox"/> Diaphorina citri OGS v2.0 CDS <input type="checkbox"/> Diaphorina citri OGS v2.0 proteins <input type="checkbox"/> Diaphorina citri OGS v2.0 transcripts <input type="checkbox"/> Diaphorina citri de novo transcriptome <input type="checkbox"/> Diaphorina citri genome v1.1 <input type="checkbox"/> Diaphorina citri genome v1.9 <input type="checkbox"/> Diaphorina citri genome v1.91 <input type="checkbox"/> Diaphorina citri genome v2.0 <input type="checkbox"/> Diaphorina citri genome v2.0 ALT <input type="checkbox"/> Diaphorina citri genome v3.0 <input type="checkbox"/> Diaphorina citri genome v3.0 ALT
Program	
Query	

**BLAST**

Recommended databases are

- Diaphorina citri OGS v2.0 proteins (use blastp)

Official gene set proteins from genome v 2. It includes computationally predicted proteins and some manually curated proteins.

- Diaphorina citri Isoseq HQ (use tblastn)

High-quality consensus transcripts from PacBio IsoSeq long reads

- Diaphorina citri MCOT proteins (use blastp)




A set of proteins resulting from the combination of several transcriptomes, followed by selection of the best transcript for each locus. MCOT names ending with O or T come from *de novo*-assembled transcriptomes and thus represent a genome-independent source of probable protein models. Names ending with other letters are based on *D. citri* genome v1.1 and are not as reliable.

- Diaphorina citri de novo transcriptome (use tblastn)

Contains a both IsoSeq long read transcripts (names contain i) and transcripts assembled from Illumina short reads RNASeq data (names contain r).

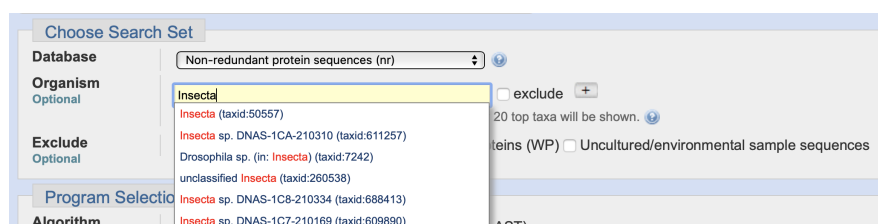
You should also do a tBLASTn against Diaphorina citri genome v3.0 and Diaphorina citri genome v3.0 ALT (probable duplicate sequences removed from the main assembly) to make sure you have located all the matches to your protein of interest. This is especially important if you are not able to find a *D. citri* ortholog in any of the other databases.

#### Input parameters

Categories	Psyllid Databases	    
Database	Diaphorina citri genome v3.0	<a href="#">db details</a>
Program	tblastn (protein to translated nucleotide db)	
Query	autodetect	<a href="#">Show example</a>

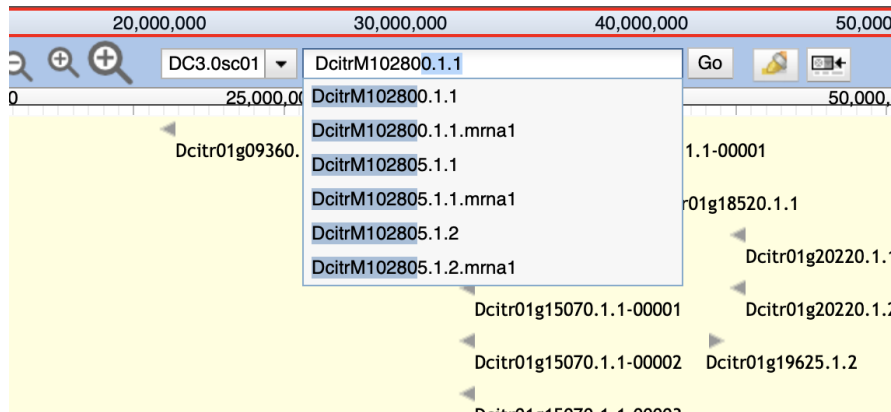
While you can sometimes find your *D. citri* gene of interest using the BLAT function in Apollo with an insect ortholog, BLAT is more stringent than BLAST and therefore it will not necessarily allow you to detect *D. citri* sequence using a sequence from another organism. Therefore, we don't recommend going immediately to Apollo if the only sequence you have is from other insects.

- 6 Any *D. citri* proteins or transcripts you identify as candidates should be reciprocal blasted back to insects using [NCBI BLAST](#) to ensure you have identified true orthologs. You may want to limit the search to Insecta or a particular insect using the Organism box.

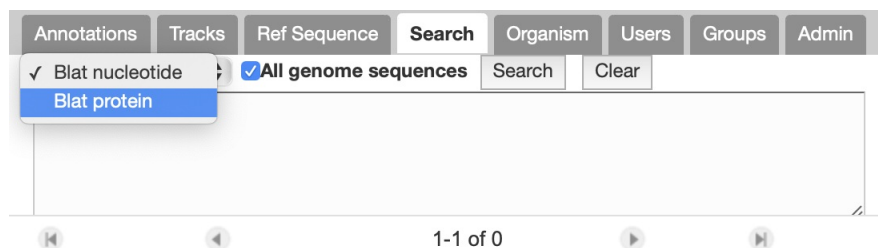


Be sure to check the coordinates for the query and subject in your BLAST results. Significant differences in the coordinate values could suggest that one of the sequences is incomplete.

- 7 Use [MUSCLE](#) (multiple alignment program) to compare your *D. citri* protein sequence(s) to its orthologs in other insects. This will help determine if your protein is complete.
- 8 Login to [Apollo](#). Enter the ID of the *D. citri* sequence you identified into the Search bar. This will take you directly to the that gene.

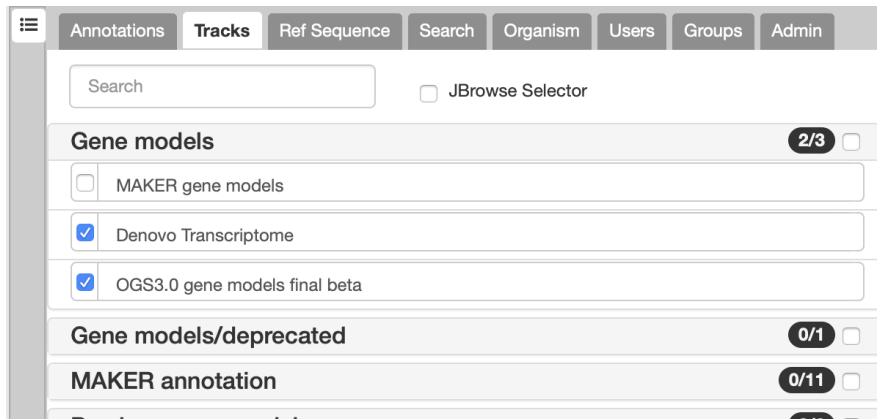


An alternative way to find your sequence of interest is to use your best *D. citri* protein to BLAT the *D. citri*v3.0 genome. Click on the Search tab in the right-hand panel. Be sure to choose “BLAT protein” and check the “search all genomic sequences” box to search all scaffolds.



When the results appear, you will likely see multiple hits representing individual exons. Click on one of the hits (should be close to 100% identity) to find the location of your gene in the *D. citri* genome assembly. Be aware that this version of the genome (v3.0) may still have some duplicate regions. Thus, it is possible that you will have multiple hits for the same exon or even more than one partial or complete copy of a gene. If this happens, annotate the most complete version possible and mark any extra OGS3.0 models for removal (see Step 10.4).

- 9 Select evidence tracks that you want to use in the curation process.

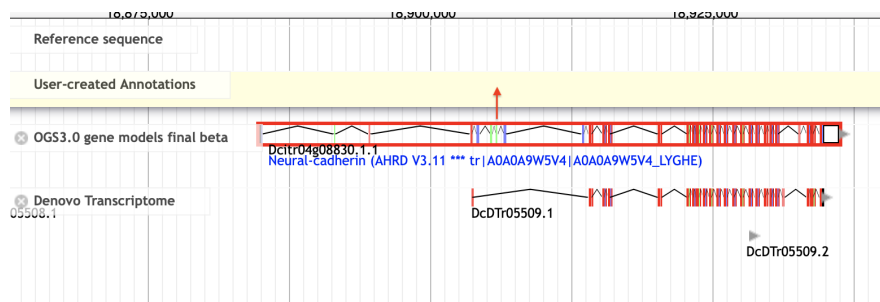


The attached document discusses the pros and cons of many of the evidence sources available for *D. citri*.

 [evidence hierarchy.docx](#)

- 10 Curate your gene in Apollo by following the steps below. More information on using Apollo can be found [here](#). The Apollo version currently in use at citrusgreening.org is 2.6.1

- 10.1 Choose a beginning gene model from one of the evidence tracks (one that seems to best fit the available data) and drag it to the upper section of Apollo. This is only a copy of the gene model, so any changes you make will not affect the original model.



- 10.2 Make changes to the gene model as needed based on the evidence available. Document all changes you make from the beginning gene model, as well as the reason for the changes. Perform multiple alignments as needed, to compare the model to data from various evidence tracks and to determine if changes to the model have improved the alignment.

A note on the accuracy of the v3 genome sequence:

Because this genome was produced with PacBio reads (which are longer but less accurate than Illumina reads) there may be errors in the genome which affect the reading frame of your gene. Comparing the genome sequence with genome-independent sequences (e.g. *denovo* transcriptome data or the Illumina-Single psyllid 10x track (under Reference sequence)) is helpful for evaluating sequence discrepancies. There will probably be many SNPs in these sequences compared to the genome, but changes to the genome are only necessary when an obvious error in the genome is significantly affecting the gene structure or reading frame. To make a change to

the genome, zoom to base level and right click on the genome sequence at the position you want to change.

10.3 If you find evidence for multiple models for your gene, be aware that there are several reasons this could occur. First, the differences could indicate multiple isoforms of the gene. Second, the differences could result from inaccurate computer predictions or incorrect manual annotation. Third, the differences could be caused by misassemblies (usually duplications) in the genome that cause similar models to map to the genome differently. To distinguish between these possibilities, it will be helpful to BLAT with various parts of the model to check for duplications. Iso-Seq transcripts are particularly useful in these cases, because they contain the full-length transcript in a single sequencing read and can provide definitive evidence of alternative transcripts.

10.4 If you find an OGS3.0 model that needs to be completely removed from the official gene set (e.g. a duplicate model resulting from a genome misassembly), you should enter information for that model in the “To remove OGSv3” tab of the ACP Gene curation targets [spreadsheet in Google Docs](#).

- 11 When you think you have completed the gene annotation, do final alignments in [MUSCLE](#) with insect orthologs and use the protein sequence to do domain analysis (you can use [NCBI Protein BLAST](#) or [InterPro](#)) to ensure all expected parts of the protein are present in your *D. citri* model.
- 12 Open the Information Editor (right-click on the model and choose Open Annotation). The form for entering information should appear in the lower portion of the right-hand panel.

Name the model (see Step 13 for more information) in the Details tab and describe the changes made to the gene model (use Comment tab).

Details	Coding	GO	Gene Product	Provenance	DbXref	Comment	Attributes
<div> <div>Go</div> <div>ID</div> <div>Delete</div> </div>							
Type	mRNA						
Name	Dcitr04g08830.1.1-00001						
Aliases (' ' separated)							
Description							
Location	18884895 - 18936219 strand(+)						
Ref Sequence	DC3.0sc04						
Owner(s)							
Created	Dec 02, 2020 04:48 PM						
Updated	Dec 02, 2020 04:48 PM						

In the current version of Apollo (2.6.1) you have to fill in the name for both the gene and mRNA separately. When you choose an annotation in the right-hand panel, it should expand to show associated mRNAs. You can then click on either the gene or mRNA to access the appropriate information form. The mRNA name is shown on models in the User-created Annotations window, but the gene name must be used when searching the Annotations list.

- 13 Naming convention established by the *D. citri* community is as follows.
  1. Use the insect ortholog name (if there is one) for the gene name. Follow the name with an -RA for first isoform, -RB for second isoform, etc.
  2. The final mRNA name should be the OGS3.0 ID from the computationally predicted model (OGS3.0 gene models final beta track) that most closely matches your annotated model. The gene name should be entered in the description field for the mRNA.
- 14 At this point it may be helpful to build a final phylogenetic tree in MEGA, this time including the protein sequence from your annotation along with several orthologs. Phylogenetic trees do not need to be included in the final pathway report unless they provide support for a particular point you want to make.

**MEGA 7** [↗](#)

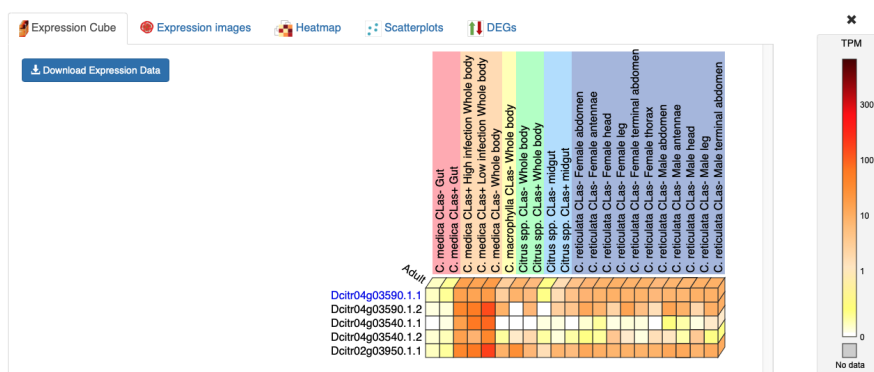
**MEGA X** [↗](#)

- 15 Write a pathway report with an Introduction containing background information obtained from literature searches, followed by a combined Results and Discussion section.

A table showing gene names, IDs, and types of evidence used should be included in the report (example below).

Gene/Isoform	OGSv3 ID	Gene model		Evidence supporting annotation			
		Complete	Partial	MCOT	IsoSeq	RNASeq	Ortholog

You may also want to use the [Citrusgreening Expression Network](#) to examine expression of your annotated genes in a variety of *D. citri* RNA-Seq datasets.



Expression Cube generated by Citrusgreening Expression Network

Please also include a short Methods section detailing any tweaks you used to the basic annotation protocol. You can reference the [Hosmani et al 2019 "Quick Guide"](#) and also include a reference to this protocol.

- Share your pathway report with the group as instructed (usually by uploading it to a shared folder) and prepare a PowerPoint to report your findings at an annotation meeting.