**Title**

Diaci v3.0: Chromosome-level assembly, *de novo* transcriptome and manual annotation of *Diaphorina citri,* insect vector of Huanglongbing

**Author Information**

Dr. Teresa D. Shippy[1], Dr. Prashant S. Hosmani[2], Dr. Mirella Flores-Gonzalez[3], Ms. Marina Mann[4], Ms. Sherry Miller[1,5], Dr. Matthew T. Weirauch[6], Mr. Chad Vosburg[7,8], Ms. Crissy Massimino[8], Mr. Will Tank[1], Mr. Lucas de Oliveira[8], Mr. Chang Chen[3], Ms. Stephanie Hoyt[3], Ms. Rebekah Adams[9], Mr. Samuel Adkins[8], Mr. Samuel T. Bailey[9], Mr. Xiaoting Chen[6], Ms. Nina Davis[8], Ms. Yesmarie DeLaFlor[8], Ms. Michelle Espino[8], Ms. Kylie Gervais[8], Ms. Rebecca Grace[8], Mr. Douglas Harper[8], Ms. Denisse L. Hasan[8], Ms. Maria Hoang[8], Ms. Rachel Holcomb[8], Ms. Margaryta R. Jernigan[8], Ms. Melissa Kemp[8], Ms. Bailey Kennedy[1], Mr. Kyle Kercher[8], Mr. Stefan Klaessan[8], Dr. Angela Kruse[4], Ms. Sophia Licata[8], Ms. Andrea Lu[1], Mr. Ron Masse[8], Mr. Anuja Mathew[8], Ms. Sarah Michels[8], Ms. Elizabeth Michels[8], Mr. Alan Neiman[8], Mr. Seantel Norman[9], Mr. Jordan Norus[8], Ms. Yasmin Ortiz[8], Mr. Naftali Panitz[2], Mr. Thomson Paris[10,11], Ms. Kitty M. R. Perentesis[9], Mr. Michael Perry[9], Mr. Max Reynolds[8], Ms. Madison M. Sena[9], Ms. Blessy Tamayo[8], Ms. Amanda Thate[1], Ms. Sara Vandervoort[1], Ms. Jessica Ventura[9], Mr. Nicholas Weis[1], Mr. Tanner Wise[8], Dr. Robert G. Jr. Shatters[10], Dr. Michelle Heck[12,4], Dr. Joshua B. Benoit[8], Dr. Wayne B. Hunter[10], Dr. Lukas A. Mueller[2], Dr. Susan J. Brown[1], Dr. Tom D'Elia[8], Dr. Surya Saha[2]

[1]Division of Biology, Kansas State University, Manhattan, KS, 66506, USA, [2]Boyce Thompson Institute, 533 Tower Road, Ithaca, NY, 14853, USA, [3]Boyce Thompson Institute, 534 Tower Road, Ithaca, NY, 14853, USA, [4]Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, 14853, USA, [5]Allen County Community College, Burlingame, KS, 66413, USA, [6]The Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 15012, USA, [7]Department of Plant Pathology and Environmental Microbiology, Pennsylvania State University, PA, 16802, USA, [8]Department of Biological Sciences, Indian River State College, Fort Pierce, FL, 34981, USA, [9]Department of Biological Sciences, University of Cincinnati, Cincinnati, OH, 45221, USA, [10]U.S. Horticultural Research Laboratory, USDA-ARS, Fort Pierce, FL, 34945, USA, [11]Entomology and Nematology Department, North Florida Research and Education Center, University of Florida, Fort Pierce, FL, 32351, USA, [12]Emerging Pests and Pathogens Research Unit, Robert W. Holley Center, USDA-ARS, Ithaca, NY, 14850, USA

Teresa Shippy <tshippy@ksu.edu>; Prashant Hosmani <psh65@cornell.edu>; Mirella Flores-Gonzalez <mrf252@cornell.edu>; Marina Mann <mam734@cornell.edu>; Sherry Miller <smiller@allencc.edu>; Matthew Weirauch <matthew.weirauch@cchmc.org>; Chad Vosberg <cvv5156@psu.edu>; Crissy Massimino <emassimi@irsc.edu>; Will Tank <wtank@ksu.edu>; Lucas de Oliveira <lmdeoliv@ncsu.edu>; Chang Chen <cc2589@cornell.edu>; Stephanie Hoyt <shh89@cornell.edu>; Rebekah Adams <adams2rh@mail.uc.edu>; Samuel Adkins <adkinssc@mail.irsc.edu>; Samuel Bailey <bailest@mail.uc.edu>; Xiaoting Chen <xiaoting.chen@cchmc.org>; Nina Davis <ndavis@irsc.edu>; Yesmarie DeLaFlor <delafloryt@mail.irsc,edu>; Michelle Espino <espinomichelle87@gmail.com>; Kylie Gervais <gervaiskb@mail.irsc.edu>; Rebecca Grace <rlg262@cornell.edu>; Douglas Harper <douglasharper711@gmail.com>; Denisse Hasan <hasandl@mail.irsc.edu>; Maria Hoang <mariatramhoang@gmail.com>; Rachel Holcomb <holcombrm@mail.irsc.edu>; Margaryta Jernigan <margarytax3@gmail.com>; Melissa Kemp <kempma@mail.irsc.edu>; Bailey Kennedy <baileyk7@ksu.edu>; Kyle Kercher <kkercher2017@my.fit.edu>; Stefan Klaessan <klaesensr@mail.irsc.edu>; Angela Kruse <ark253@cornell.edu>; Sophia Licata

<licatase@mail.irsc.edu>; Andrea Lu <andrealu@vet.k-state.edu>; Ron Masse <masserd@mail.irsc.edu>; Anuja Mathew <mathewat@mail.irsc.edu>; Sarah Michels <michelssy@mail.irsc.edu>; Elizabeth Michels <michelsey@mail.irsc.edu>; Alan Neiman <neimanat@mail.irsc.edu>; Seantel Norman <normanst@mail.uc.edu>; Jordan Norus <norusjm@mail.irsc.edu>; Yasmin Ortiz <ortiz.yasmin@ufl.edu>; Naftali Panitz <np298@cornell.edu>; Thomson Paris <thomsonparis@ufl.edu>; Kitty Perentesis <perentem@mail.uc.edu>; Michael Perry <perry2ma@mail.uc.edu>; Max Reynolds <reynoldsm1@mail.irsc.edu>; Madison Sena <senamm@mail.uc.edu>; Blessy Tamayo <tamayoba@mail.irsc.edu>; Amanda Thate <athate22@ksu.edu>; Sara Vandervoort <vandervoortsara@ksu.edu>; Jessica Ventura <venturjr@mail.uc.edu>; Nicholas Weis <weisnicholas@gmail.com>; Tanner Wise <wisetf@maail.irsc.edu>; Robert Shatters <robert.shatters@usda.gov>; Michelle Heck <michelle.heck@usda.gov>; Joshua Benoit <benoitja@ucmail.uc.edu>; Wayne Hunter <wayne.hunter@usda.gov>; Lukas Mueller <lam87@cornell.edu>; Susan Brown <sjbrown@ksu.edu>; Tom D'Elia <tdelia@irsc.edu>; Surya Saha <ss2489@cornell.edu>

**Abstract**

**Background**

*Diaphorina citri* is a vector of "*Candidatus* Liberibacter asiaticus" (*C*Las), the gram-negative bacterial pathogen associated with citrus greening disease. Control measures rely on pesticides with negative impacts on the environment, natural ecosystems and human and animal health. In contrast, gene-targeting methods have the potential to specifically target the vector species and/or reduce pathogen transmission.

**Results**

To improve the genomic resources needed for targeted pest control, we assembled a *D. citri* genome based on PacBio long reads followed by proximity ligation-based scaffolding. The 474 Mb genome has 13 chromosomal-length scaffolds. 1,036 genes were manually curated as part of a community annotation project, composed primarily of undergraduate students. We also computationally identified a total of 1,015 putative transcription factors (TFs) and were able to infer motifs for 337 TFs (33 %). In addition, we produced a genome-independent transcriptome and genomes for *D. citri* endosymbionts.

**Conclusions**

Manual annotation provided more accurate gene models for use by researchers and also provided an excellent training opportunity for students from multiple institutions. All resources are available on CitrusGreening.org and NCBI. The chromosomal-length *D. citri* genome assembly serves as a blueprint for the development of collaborative genomics projects for other medically and agriculturally significant insect vectors.

Figure 1. Graphical abstract of the major findings from a multi-year, multi-institutional *Diaphorina citri* genome assembly project.

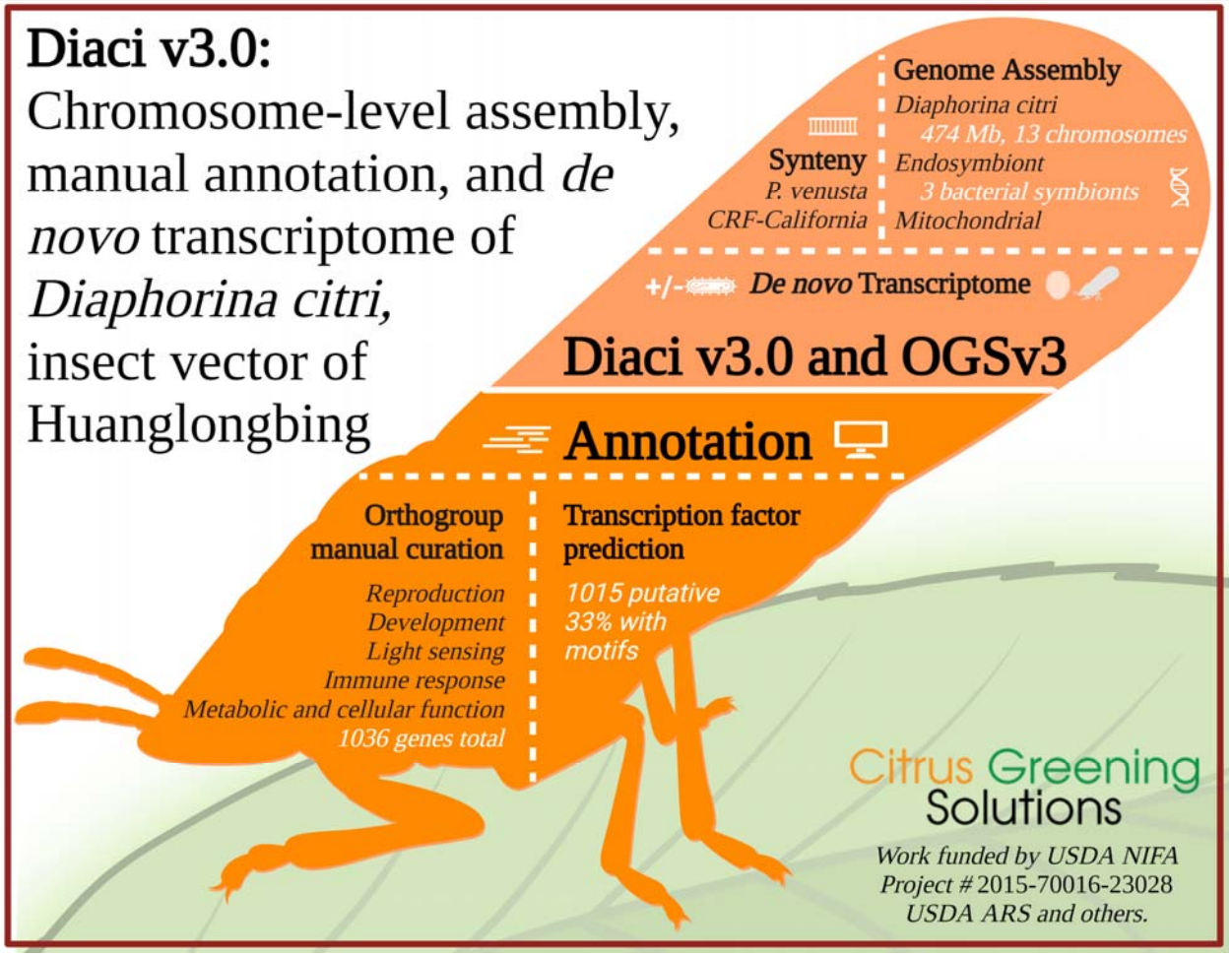



Figure 1 Legend: The most recent version (3.0) of the *Diaphorina citri* genome assembly (Diaci v3.0) is available on CitrusGreening.org and NCBI, along with its Official Gene Set v3, *de novo* transcriptome and extensive manual annotations covering major pathways and hundreds of genes. We have also predicted transcription factors and protein-coding genes, and compared Diaci v3.0 to the *Pachypsylla venusta* psyllid genome, as well as other *D. citri* genome assemblies recently published. Lastly, during our genome assembly, we also created draft genomes for multiple *D. citri* endosymbionts. Our work took place over eight years, multiple institutions and dozens of students, graduate students and faculty.

## 1. Background

High throughput methods in genomics, transcriptomics and proteomics coupled with recent developments in single cell technologies have allowed for better understanding of processes at the cellular, organismal and ecological levels. However, a high quality genome assembly and annotation of both coding and non-coding genes is critical to the success of ‘omics assays, as well as interdiction methods that are based on gene targets. Easily accessible genomic and transcriptomic resources and web-based tools are essential for developing

therapies and finding long-term management strategies, especially for difficult diseases of agriculture. Huanglongbing is an example of how genomic and transcriptomic resources [1] can help provide real management strategies for an otherwise terminal agricultural industry.

Huanglongbing (HLB), also known as citrus greening disease, is the most devastating of all citrus diseases and there is currently no effective control strategy. In the state of Florida alone, HLB has caused over $7 billion in lost revenue since 2005 and thousands of lost jobs [2]. HLB affects all genotypes of *Citrus* and some other Rutaceae species, causing early fruit drop, uneven ripening, unpleasant fruit taste, leaf mottling, and tree death in 3-5 years. In Asia, North and South America, Oceania and the Arabian Peninsula, HLB is caused by a phloem-limited, gram-negative bacterium, "*Candidatus* Liberibacter asiaticus" (*C*Las) and its insect vector, the Asian citrus psyllid, *Diaphorina citri* Kuwayama (Hemiptera: Liviidae) moves the pathogen from tree to tree. Asymptomatic but infectious trees and minimal vector control have contributed to the rapid spread of *C*Las, making disease monitoring and control difficult, emphasizing the need to find vector-control strategies to stop or slow the spread to new, healthy trees [3]. While the vector-pathogen interactions of HLB have been reviewed and compared in multiple articles [4,5], this study aims to update the current genomic resources available for the vector, *D. citri.*

A short read-based genome assembly (Diaci v1.0) from a Florida *D. citri* population has been an important resource for the HLB research community since 2011, with multiple updates and improvements adding to it over the following 10 years (Diaci v1.1, Diaci v1.9, and Diaci v2.0) [6]. The early Diaci v1.1 assembly proved to be fragmented, with a contig N50 of 34.4 Kb and scaffold N50 of 109.8 Kb. Moreover, a benchmarking universal single-copy orthologs (BUSCO) [7] analysis indicated that a significant number of conserved single-copy markers were missing, deeming improvements necessary. In 2018, a higher quality genome assembly (Diaci v2.0) was released on citrusgreening.org [1,8] using PacBio and Dovetail technologies, reducing the contig count and improving the contig N50 to 749.5 Kb. Ongoing improvements to the genome including community-driven manual annotations [9] and transcriptomic and proteomic analyses of psyllid mRNAs and proteins differentially expressed during *C*Las transmission have also identified a number of mis-assemblies such as tandem duplications and a high degree of gene fragmentation [10]. The Diaci v1.1 assembly was also lacking in key areas that are needed for a thorough understanding of *C*Las-*D. citri* interactions, as well as interactions between *D. citri* and its primary, obligate, and facultative intracellular endosymbionts.

*D. citri* harbors two obligate bacterial endosymbionts with highly reduced genomes ("*Candidatus* Profftella armatura" and "*Candidatus* Carsonella ruddii") within a specialized organ called the bacteriome, which evolved specifically to host these symbionts. As shown by their "*Candidatus"* status, "*Ca.* P. armatura" and "*Ca.* C. ruddii" are unculturable and are limited to one specific organ within *D. citri*. The endosymbionts of *D. citri* are critical to psyllid survival and reproduction - "*Ca.* C. ruddii" produces essential vitamins allowing *D. citri* to feed from the nutrient poor phloem [11–14]. Suppression of "*Ca.* P. armatura" in *D. citri* resulted in increased psyllid mortality [15,16] and the enosymbiont's defensive properties [17] are also of interest for vector control strategies. *D. citri* is also host to a specific strain of the facultative endosymbiont, *Wolbachia* (*w*Di), though its role has yet to be clearly determined [18,19]. Separating these

endosymbionts from the insect host can only be done computationally, and typically the endosymbiont genomes must be removed from sequencing data prior to *de novo* assembly or other downstream analyses of the *D. citri* genome. Because psyllid biology is intimately connected to, and regulated by, the bacteriome symbionts, strategies to target and suppress psyllid endosymbionts are emerging as treatments that can reduce psyllid vector populations while limiting the damage to non-target species. The endosymbionts also interact with pathogens and can affect the acquisition and transmission of the bacterial pathogen *C*Las [15,16,20].

Here, we report an improved assembly of a Florida *D. citri* genome which we call Diaci v3.0. Diaci v3.0 has chromosomal-length scaffolds (produced by *de novo* assembly of PacBio long reads and Dovetail Chicago and Hi-C scaffolding) and a new Official Gene Set v3 (OGSv3, generated by automated gene prediction using Illumina RNAseq and PacBio Iso-Seq data as evidence). This genome assembly is most notable for the high-quality gene annotation effort conducted in parallel. Through a community annotation effort, primarily involving undergraduate students mentored by senior scientists, gene sets (including chitin metabolism, cuticle formation, segmentation and segmental identity, signal transduction, protein degradation, chromatin remodeling, phototransduction, circadian rhythm, carbohydrate metabolism, melanization and spermatogenesis) were manually curated using the Apollo annotation editor [21].

In order to better understand genome architecture, we compared Diaci v3.0 to that of the hackberry petiole gall psyllid, *Pachypsylla venusta,* and to three *D. citri* genomes that were reported after our genome was publicly released [22]. We also describe a comprehensive and genome-independent *de novo* transcriptome based on a variety of available transcript evidence from different tissues and conditions. Additionally, high quality and complete genome assemblies for the endosymbionts "*Ca.* P. armatura" and "*Ca.* C. ruddii", and high quality draft genomes for two *w*Di strains, provide genomics resources to further understand complex organ-specific, intermicrobial interactions.

Data Description

**Background and Purpose of data collection**

High quality genome assemblies and accurate annotation of both coding and non-coding genes are critical to the success of 'omics assays. While the vector-pathogen interactions of Huanglongbing (Citrus Greening disease) have been reviewed and compared in multiple articles [4,5], this study aims to update the current genomic and transcriptomic resources available for the vector, *Diaphorina citri.*

**General methods**

DNA was obtained from a *D. citri* colony originating from psyllids collected in Indian River County, Florida and maintained in the U.S. Horticultural Research Laboratory, USDA, Fort Pierce, Florida. DNA for genome assembly was extracted from pooled *D. citri* adults and

sequenced using PacBio long read technology followed by multiple rounds of assembly and scaffolding, followed by duplication reduction, error correction and exclusion of non-target organism reads produced the Diaci v3.0 genome assembly. Genome assemblies of three *D. citri* endosymbionts were generated using reads excluded during the Diaci v3.0 assembly process, then cleaned and verified. We visually compared synteny of Diaci v3.0 with another psyllid genome, and with another recently published *D. citri* genome assembly. Our Iso-Sequencing used RNA sequencing data and an established computational pipeline to assemble gene isoforms independent of the genome. Additionally, a *de novo D. citri* transcriptome was generated using publicly available short read RNA sequencing data, and was combined with the Iso-seq isoforms. Protein-coding genes were predicted and annotated following the MAKER annotation pipeline and results were informed using Mikado and RNAseq datasets. We also predicted Transcription factors and their motifs using previously published pipelines and performed preliminary analysis of the effect of *C*Las infection on transcription factors and their targets. Our manual curation was a huge part of this genome project and involved teams of student annotators, graduate students and faculty across multiple institutions. We have previously published the annotation pipeline used [23]. Finally, we compared the predicted proteins of 12 different insect species to create orthogroups and assign GO terms to *D. citri* genes. See Figure 5 for an overview of the methods.

### Data Availability

The data sets supporting the results of this article are available in NCBI via BioProject accessions No. PRJNA29447 and PRJNA609978.

## 2. Analyses and Discussion

### 2.1. Genome assembly

The Diaci v.1.1 genome was generated and assembled with Illumina short reads, and thus had limitations to its usefulness for genomics-based research. Assembly of large eukaryotic genomes using short read technology results in fragmentation and misassembly of the genome because the short reads do not provide the contiguity required to assemble long and high-quality scaffolds [24]. The development of long read sequencing techniques, such as PacBio [25], and chromosome structure capture technologies such as Dovetail Chicago and Hi-C [26,27], afforded an opportunity to generate a chromosomal-length *D. citri* genome to be used by researchers in the citrus greening community.

We generated 36.2 Gb of PacBio long reads. The Canu assembler was used to create a PacBio-only *de novo* assembly with 38,263 high confidence contigs, referred to as unitigs. The N50 of the unitigs (the length of the shortest contig for which longer and equal length contigs cover at least 50 % of the assembly) was 29 kb. Intermediate range scaffolding was performed with the Dovetail Chicago [24] method. Scaffolding with the HiRise assembler [24] based on paired-end reads from Chicago libraries with 22.53 X coverage of the *D. citri* genome increased the N50 from 29 Kb to 383 Kb. This step resulted in 12,369 joins and corrected 48

misassemblies in the unitig set. These Chicago scaffolds were subjected to gap-filling with PBjelly [28], available on GitHub [29] to create an intermediate Diaci v2.0 assembly with 1906 scaffolds and a scaffold N50 of 749.5 Kb (Table 1). A second round of long range scaffolding was performed on the non-gap-filled Chicago scaffolds using paired-end reads from a Hi-C library [30], which had higher coverage and larger inserts than the Chicago library. The HiRise assembler made 1003 joins in the Chicago assembly and increased the scaffold N50 from 383 Kb to 26.7 Mb.

After Hi-C scaffolding, the assembly contained 13 chromosomal-length scaffolds (likely representing the 12 putative chromosomes [31] and the sex chromosome in the *D. citri* genome) and 24,930 unplaced scaffolds. We expected a high rate of duplication in the unplaced scaffolds in this assembly due to the pooling of multiple individuals to obtain the long read data, so we used Redundans [32] to classify the the unplaced scaffolds as unique or redundant compared to the rest of the genome assembly. We joined unique scaffolds (with 1000 Ns separating adjacent scaffolds) to create chromosome 00 in the Diaci v3.0 assembly. The 23,672 redundant scaffolds may represent alternative loci in the *D. citri* population and are reported as alternate (ALT) contigs. We subsequently removed 12 unplaced scaffolds from chromosome 00 and 2.3 Mb from chromosomal-length scaffold 9 that were identified as microbial contamination. The final Diaci v3.0 assembly has a total size of 474 Mb and an N50 of 40.6 Mb (Table 1). The complete history of *D. citri* genome sequencing can be found at citrusgreening.org [8].

Table 1: *Diaphorina citri* genome assembly statistics compared across three version releases.

| ***D. citri* genome assembly version*** | **Diaci v1.1** | **Diaci v2.0** | **Diaci v3.0** |
|---|---|---|---|
| **Genome total size** | 485.7 Mb^ | 498.8 Mb | 474 Mb |
| **Scaffold N50** | 109.8 Kb | 749.5 Kb | 40.6 Mb |
| **Maximum scaffold length** | 1.0 Mb | 4.2 Mb | 50.3 Mb |
| **Contig N50** | 34.4 Kb | 29.0 Kb | 29.0 Kb |
| **Maximum contig length** | 431.2 Kb | 2.1 Mb | 409.7 Kb |
| **Total scaffolds** | 161,988 | 1,906 | 1,257** |
| **Number of N's** | 19.3 Mb | 4.5 Mb | 13.4 Mb |
| **Complete BUSCO (arthropoda)** | 82.6 % | 92.1 % | 93.5 % |
| **Repeat %** | 26.4 % | 31.9 % | 30.2 % |

Table 1 Legend: Comparison of statistics for three versions of the Florida *Diaphorina citri* (Diaci) genome: Diaci v1.1 [9], Diaci v2.0 (see Section 2.1) and the current version Diaci v3.0. BUSCO (Benchmarking Universal Single-Copy Orthologs) percentages were obtained by comparing

each assembly to the Arthropoda (1013) gene set. Repeat percentages were calculated using RepeatMasker. *Each genome assembly is named as such: Diaci = *Diaphorina citri;* v = version. **13 chromosomal-length scaffolds with 1244 shorter non-redundant unplaced scaffolds, where shorter scaffolds are combined into a 14th "chromosome" with 1000 N's between each scaffold. Additional, redundant and unplaced scaffolds which did not map to any of the current chromosomal-length scaffolds are not included in this number. ^Mb = Megabytes, Kb=Kilobytes.

## Endosymbiont and mitochondrial assembly

The genomic DNA isolated from *D. citri* included bacterial sequences from which we obtained near complete assemblies for the obligate *D. citri* endosymbionts "*Candidatus* Profftella armatura" and "*Candidatus* Carsonella ruddii", and for two strains of the facultative endosymbiont *Wolbachia*. We were able to identify specific regions of the *Wolbachia* genome that are shared between the two strains, as well as regions that were unique to each (Supplemental Figure 1). We compared our assemblies to those previously published [18,33] through orthology analysis based on sequence identity and found the assemblies to contain the majority of orthogroups present in all published genomes (Table 2). We utilized the mitochondrial genome of *D. citri* sourced from China [34] to determine the mitochondrial sequences for our Florida *D. citri* samples.

Table 2: Genome statistics for *Diaphorina citri* endosymbionts including two obligate bacteria and two *Wolbachia* strains.

| ***D. citri* Endosymbionts** | ***Wolbachia* Strain 1** | ***Wolbachia* Strain 2** | **"*Candidatus* Profftella armatura"** | **"*Candidatus* Carsonella ruddii"** |
|---|---|---|---|---|
| **Reference Genomes** | 8 | 8 | 2 | 9 |
| **Conserved Orthogroups** | 559 | 559 | 307 | 116 |
| **Present in Genome** | 534 | 503 | 307 | 106 |
| **Unique Orthogroups** | 96 | 93 | ND | 10 |
| **Total Orthogroups** | 1038 | 1038 | 307 | 217 |
| **Number Missing Proteins** | 122 | 150 | 1 | 42 |
| **Orthogroups with Duplicated Proteins** | 181 | 159 | 1 | 23 |

Table 2 Legend: An orthology analysis of endosymbiont genomes from *Diaphorina citri* compares the number of conserved orthogroups (the set of genes predicted to be derived from the last common ancestor) across published genomes to those present in our assemblies, as well as the number of orthogroups unique to our assemblies. We also present the number of

proteins that may be duplicated or missing in each of our assemblies. Our assembly of the *D. citri Wolbachia* symbiont suggests two unique strains are present. The two other obligate symbionts are unculturable (thus, "*Candidatus")* and play roles that are well documented in the literature. ND = no data.

## Comparative genomic analysis (synteny)

We compared the Diaci v3.0 assembly to the genome of the hackberry petiole gall psyllid, *Pachypsylla venusta,* [35] to identify regions of synteny (Figure 2). Three other members of the genus *Diaphorina* have a male 2N count of 25 (12 autosome pairs, plus the X chromosome) as is common among psyllids [36,37]. The 13 chromosomal-length scaffolds in the *D. citri* genome assembly are consistent with these observations and match the number of chromosomal-length scaffolds in three recently published *D. citri* genomes [22]. In contrast, *P. venusta* has only 11 autosome pairs in addition to the X chromosome [38]. Most of the chromosomal-length scaffolds showed a one-to-one correspondence between the two species, suggesting their overall genome structure is highly conserved. The one major exception is that approximately half of *P. venusta* chromosome 3 (pv3, Figure 2) aligns with *D. citri* chromosomal-length scaffold 12, while the other half shares synteny with *D. citri* chromosomal-length scaffold 13 (dc12 and dc13, Figure 2). Thus, it seems likely that the reduction in autosome number in *P. venusta* is due to a fusion event between the chromosomes homologous to *D. citri* scaffolds 12 and 13.

Figure 2. Parallel plot showing synteny between *Diaphorina citri* Diaci v3.0 and *Pachypsylla venusta* genome assemblies.

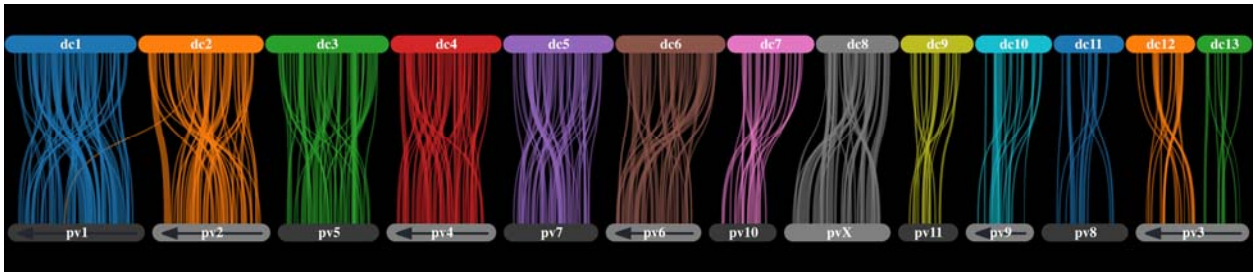



Figure 2 Legend: *Diaphorina citri* chromosomal-length scaffolds from Diaci v3.0 are numbered in order of size (dc1-dc13) and make up the top row in the parallel plot. *Pachypsylla venusta* (the Hackberry petiole gall psyllid) chromosomal-length scaffolds are numbered as previously described (pv1-pv11) [35]. The *P. venusta* scaffolds marked with arrows (pv2, pv3, pv4, pv6, and pv9) have been reversed to optimize synteny to Diaci v3.0. We see that dc8 is likely the *D. citri* X chromosome, while pv3 appears to be a fusion of sequences homologous to dc12 and dc 13.

We also used conserved synteny with *P. venusta* to identify the *D. citri* sex chromosome. The identity of the X chromosome scaffold of *P. venusta* has been established by comparing read mapping coverage of genomic sequence data from single males and females [35]. *D. citri*

chromosomal-length scaffold 8 (dc8) is highly syntenic with the *P. venusta* X chromosome (pvX, Figure 2), allowing us to identify it as the probable *D. citri* X chromosome. Carlson et al [22] also compared one of their *D. citri* genomes with *P. venusta* and obtained the same results, identifying scaffold 7 of the CRF-California genome (which corresponds to scaffold 8 in the Diaci v3.0 genome, see Figure 3) as the *D. citri* X chromosome.

Despite both efforts yielding chromosomal-length scaffolds, the 474 Mb Diaci v3.0 Florida genome is about 200 Mb larger than each of the three *D. citri* genomes described recently by Carlson et al [22]: CRF-Uruguay (266.76 Mb), CRF-California (282.67 Mb), and CRF-Taiwan (282.89 Mb). Moreover, flow cytometry analysis in the same paper gave genome size estimates of 274.4 Mb for males and 287.0 Mb for females. To investigate the reasons for the apparently inflated size of the Diaci v3.0 genome, we used MCScanX to align syntenic blocks between the Diaci v3.0 assembly and the CRF-California genome. The two genomes show one-to-one correspondence of chromosome-length scaffolds, with a few differences in numbering due to differences in relative scaffold size. All Diaci v3.0 scaffolds are larger than the corresponding scaffolds of the CRF-California genome, suggesting that the additional sequence in the Diaci v3.0 assembly is distributed throughout the genome (Figure 3). These results along with close examination of the v3.0 genome during manual annotation suggest that the Diaci v3.0 genome contains many heterologous loci that have been assembled into the main scaffolds as false tandem duplications. This issue likely stems from the use of multiple psyllids for DNA isolation during the Diaci v3.0 genome assembly process, a standard procedure at the time.

Figure 3. Parallel plot of two chromosomal-length *Diaphorina citri* genomes: Diaci v3.0 (dc) and CRF-California (dcca).

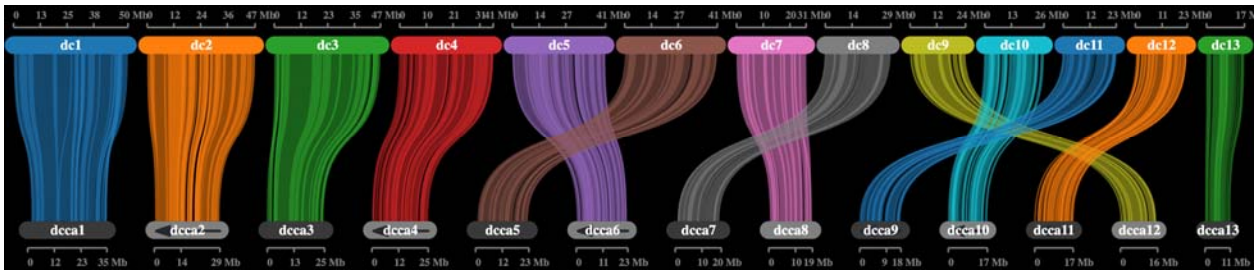



Figure 3 Legend: The 474 Mb *Diaphorina citri* Diaci v3.0 Florida genome is about 200 Mb larger than the CRF-California (282.67 Mb) genome from Carlson et al [22]. All Diaci v3.0 chromosomal-length scaffolds (dc) are larger than the corresponding scaffolds of the CRF-California genome (dcca), suggesting that the additional sequence in the Diaci v3.0 assembly is distributed throughout the genome. Lines connecting scaffolds indicate syntenic blocks of genes. Scaffold sizes are shown in megabytes (Mb). Gray ovals with arrows indicate CRF-California scaffolds that have been reversed to match the orientation of the corresponding Diaci v3.0 scaffold.

## 2.2. Annotation

We identified more repeats in the Diaci v3.0 genome assembly compared to the previously published Diaci v1.1 assembly (Table 1). This can largely be attributed to the use of

long-read PacBio sequences in the Diaci v3.0 assembly. Computational annotation of genes was carried out with the MAKER annotation pipeline after masking repetitive elements. In total, we identified 18,947 protein-coding genes with 21,231 alternatively spliced isoforms (Table 3) in the Official Gene Set version 3 (OGSv3) gene annotation. OGSv3 includes automatically predicted as well as manually curated gene models. Use of PacBio Iso-Seq enabled accurate predictions of complete gene models and their isoforms. Predicted protein-coding genes were longer compared to OGSv1 gene models, with more exons predicted per gene (Table 3). Curated genes represent models that were manually refined using the Apollo annotation editor [21], based on evidence from orthology and expression data. Curation was performed by distributed groups based on the previously described workflow [23,39]. Detailed reports of curated gene families in pathways of interest have been published and are summarized in section 2.5.2 below. Overall, curated genes are longer with more exons compared to predicted genes (Table 3). We also observe fewer non-canonical 5' and 3' splice sites in the curated genes and OGSv3 gene models, indicating higher quality of gene models compared to previous annotations. During the curation process, we removed 258 predicted genes that were not supported by expression data or represented false duplications. In addition to manual curation, we also generated functional annotation for OGSv3 gene models using the AgBase functional annotation pipeline [40] to assign experimentally validated GO terms, InterProScan domains and pathways based on sequence similarity.

Table 3: *Diaphorina citri* annotated gene counts for Official Gene Sets v1, v2, v3 and manually curated.

| **Annotation Set** | **OGSv1** | **OGSv2** | **OGSv3** | **Curated** |
|---|---|---|---|---|
| **Total Genes** | 19,311 | 20,793 | 18,947 | 1,036 |
| **Total Transcripts** | 20,966 | 25,292 | 21,231 | 1,143 |
| **Average Transcript Length (bp)** | 1,317 | 1,944 | 2,034 | 2,503 |
| **Exons Per Transcript (avg)** | 5.42 | 7.06 | 7.29 | 7.87 |
| **Average Exon Length (bp)** | 243 | 275 | 279 | 318 |
| **Non-Canonical Splice Sites** | 6.05 % | 3.13 % | 2.47 % | 1.91 % |

Table 3 Legend: The *Diaphorina citri* Official Gene Set versions 1, v2 and v3 include automatically predicted gene models (columns "OGSv1", "OGSv2" and "OGSv3"), of which, over 1000 have been manually curated (column "Curated"). bp = base pairs. avg = average.

2.3. Transcription factor prediction

Transcription factors (TFs) are an important class of proteins that regulate gene expression levels by interacting with short DNA binding "motifs" in the genome. Knowledge of the TF repertoire and, when possible, the DNA binding motifs recognized by these TFs, thus

represents a critical component of any genomics toolkit. We identified a total of 1,015 putative TFs in Diaci v3.0 (Figure 4, panel A). This value is similar to other insects [41–44]. e were able to infer motifs for 337 (33 %) *D. citri* TFs based on orthology (Supplemental File 1), mostly based on available DNA binding specificity data from *D. melanogaster* (228 TFs), but also from species as distant as human (77 TFs) and mouse (12 TFs). Many of the largest TF families have inferred motifs for a substantial proportion of their TFs, including Homeodomain (93 of 104, 89 %), bHLH (57 of 63, 90 %) and Forkhead box (24 of 31, 77 %). As expected from other insect systems [41–44], the largest gap in binding specificity knowledge is for C2H2 zinc fingers (only 39 of 377, ~10 %), which evolve quickly by shuffling the regions encoding the many zinc finger DNA binding domains contained in their protein sequences..

Along with prediction of the TFs and their motifs, we used a combination of weighted gene co-expression network analysis (WGCNA) and TF binding enrichment analysis to identify specific gene targets of interest. The WGCNA identified specific modules of genes associated with infection with *C*Las (Figure 4, panel B). Based upon these modules, we used HOMER analysis [45] to identify the specific TFs whose binding sites were enriched upstream of the *C*Las-infection associated genes when compared to the whole gene set. From these analyses, there were 24 TFs with increased enrichment in binding sites for genes with differential expression in the whole body or the guts (Figure 4, panel C) within 2000 bp (Supplemental File 2) or 500 bp (Supplemental File 3) from the start of transcription. Transcriptional levels were found to be relatively high for four of these TF genes, with one increased and three decreased during *C*Las infection. Comparison with other insect systems revealed that the three genes with decreased expression are orthologs of *Myb*, *Mothers against dpp* (*Mad*), and *Suppressor of Hairless (Su(H))*, and the one that shows increased expression is the ortholog of *ventral nervous system defective (vnd)*. These TFs are involved in multiple aspects of development and could have critical roles in the response of *D. citri* to *C*Las and potential biological trade-offs that may occur during infection. As an example, *Drosophila melanogaster Mad* loss of function alleles result in defects in midgut and fat body morphogenesis [46], two organs known to be involved in *C*Las-*D. citri* interactions [14,47,48]. The other TFs identified have been associated with many processes that range from neuronal development to cell differentiation [49], but functional studies will be necessary to confirm the roles in relation to *C*Las-*D*. *citri.*

Figure 4. Comparison of transcription factors across insects and response to infection in *Diaphorina citri*.

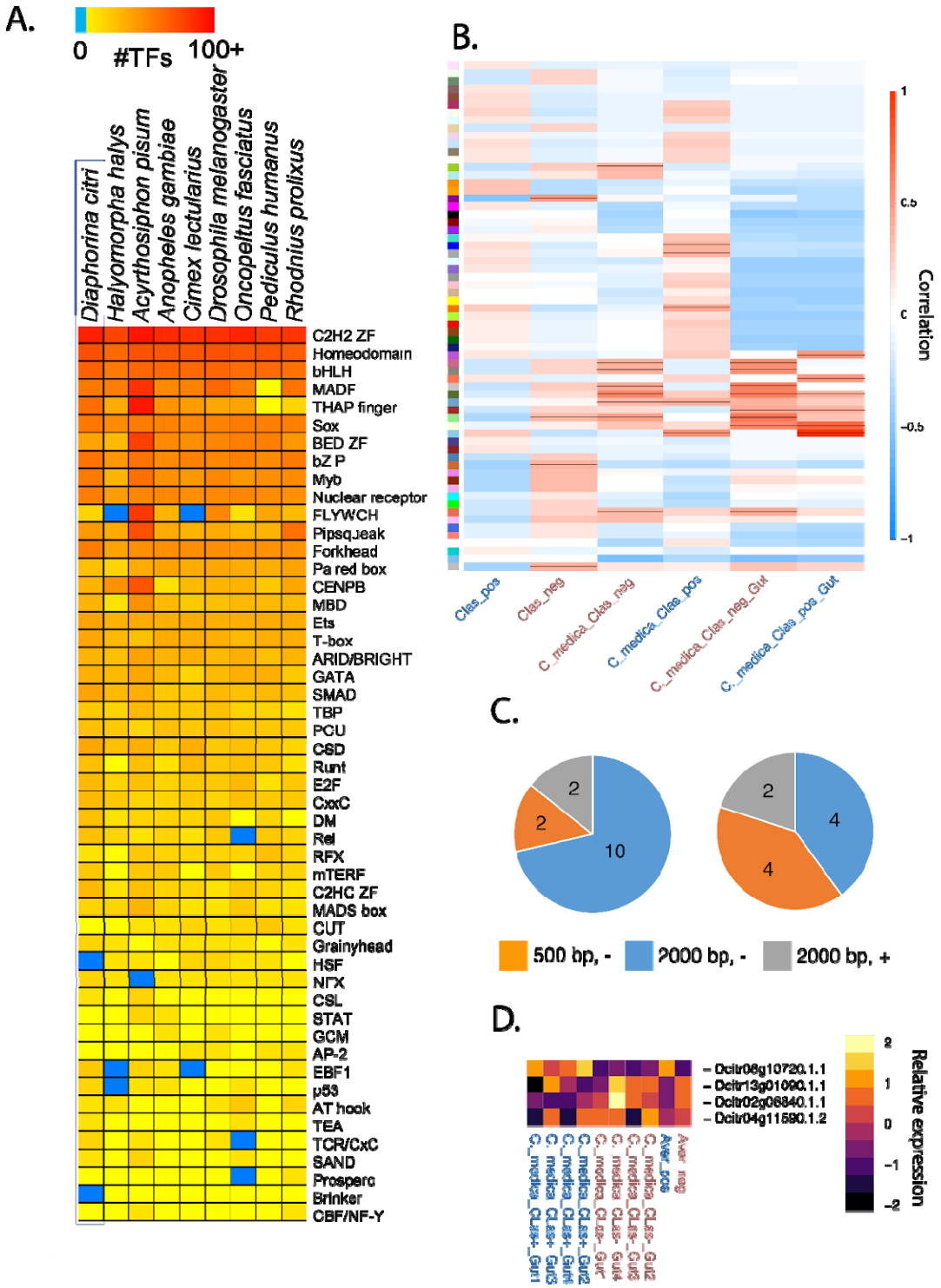



Figure 4 Legend: A) The top 50 families of transcription factors (TF, one shown per row) identified in *D. citri* (column 1) and in other insects (columns 2-9), where red is higher numbers of inferred TF motifs, and blue is no data. B) Weighted gene co-expression network analysis (based on FPKM values) and TF binding enrichment analysis identified specific modules of gene targets associated with *"Candidatus* Liberibacter asiaticus" (*C*Las) infection. Association between modules are shown as colors on the y-axis of co-expressed genes. CLas_pos are all

*C*Las-positive (infected) and Clas_neg are all *C*Las-negative (healthy) samples. Samples denoted as C_medica_Clas_pos are CLas-infected *D. citri* raised on *Citrus medica* plants and those as C_medica_Clas_neg are healthy, non-infected raised on healthy *Citrus medica* plants. Lastly, C_medica_Clas_pos_Gut and C_medica_Clas_neg_Gut are the same as above but data comes only from guts of *D. citri* in both treatments. Red indicates higher correlation between module and *C*Las treatment with a black line indicating significant correlation at $P < 0.05$. C) Expression patterns of four TFs of interest show increased enrichment in binding sites for genes with differential expression during *C*Las infection. Increased expression shown as lighter yellow and decreased expression in dark blue. Specific genes listed include: Dcitr02g06840.1.1 (*protein mothers against dpp*, *Mad*); Dcitr08g10720.1.1 (ortholog of *ventral nervous system defective, vnd*); Dcitr13g01090.1.1 (transcription factor *Myb*); Dcitr04g11590.1.2 (*suppressor of hairless, Su(H)*). Aver_pos and Aver_neg are averages of all four *C*Las-positive gut samples and all four *C*Las-negative gut samples. The averages show altered TF binding prediction and differential expression within the gut. Transcript levels in A), B), and C) are reported from the Psyllid Expression Network [1] using Diaci v3.0 and the *de novo* transcriptome.

## 2.4 *De novo* Transcriptome

We identified a number of missing and incomplete genes during our detailed manual curation and orthology analysis of the *D. citri* genome, so we created a genome-independent *de novo* transcriptome to generate a more comprehensive set of transcripts. The transcriptome set was generated using 1.4 billion short and long individual reads from a range of experimental conditions and tissue types (Supplemental Table 3). This resulted in 40,637 genes and 60,261 transcripts with an average length of 1,736.1 bp and contig N50 of 3,657 bp. To differentiate transcripts in downstream analysis we assigned transcript identifiers according to the transcript evidence source. The DcDTr identifier prefix corresponds to 41,457 transcripts assembled from short read RNA sequencing data. And, the DcDTi identifier prefix was assigned to 18,804 transcripts supported by long read Iso-Seq data. Functional descriptions were assigned to 27,813 transcripts. We later added nine genome-independent transcripts (from seven genes) that were identified during manual curation. Most of these came from the MCOT transcriptome we described previously [6]. The transcriptome was validated with the hemipteran BUSCO data set and has 94.2% complete single-copy orthologs, and only 3.2 % missing single-copy orthologs (Supplemental Table 2).

## 2.5. Pathway-based manual curation

Community curation of *D. citri* genes was initiated with the Diaci v1.1 genome, resulting in more than 500 curated genes in Official Gene Set v1.0 [6], and has continued with subsequent genome versions. This initiative is particularly noteworthy because it is primarily driven by undergraduate students with input and supervision from scientists at multiple research institutions. The students were mentored by senior peer student annotators as well as expert annotators. Students had regular interactions with scientists from the insect genomics

community working on genes and pathways under curation. Implementation of rigorous and consistent annotation practices across a virtual team of highly diverse annotators required project management tools and regular video conferences in addition to extensive documentation that was continuously updated in response to user feedback.

This group annotation strategy has been successful on multiple fronts (Table 4) and is likely to serve as a framework for integrating students into genomics to provide a high-quality genome resource for species of importance while educating the next generation of scientists in practical applications of bioinformatics methods. One of the most important lessons learned during this project is that our genome annotation and the student experience both benefit when students are recruited earlier in their academic career (first or second year students) and supported through a longer period than traditional undergraduate research experiences where students typically start research as juniors or seniors. These students can be involved from the start of analysis until publication, with a few generating their own first-author publications [50–55]. Students can join and contribute to genome annotation after first year introductory biology courses that cover basic molecular biology, which provides the necessary background to understand gene structure and function. The students can learn more complex biology through their annotation research experience with help from faculty and peer annotators. Students that are well trained and available to contribute for longer time frames reduce turnover and create a sense of a research community at local institutions that extends to having colleagues at other research institutions. Simply put, they learn the value of collaboration early in their careers and function with no silos around their mission and communication. Students that persist for two years or longer are extremely valuable for their ability to facilitate the training of new students through peer training and peer mentoring [6,39,56]. Several studies have shown that an extended duration for an undergraduate research experience increases student gains, both as part of the research project and individually in relation to general academic success [57]. This extended time frame also allows for higher level processing of data and the ability to interact with senior scientists, which helps students apply their knowledge of concepts and develop as independent research scientists [58]. Underrepresented students have been shown to specifically benefit from longer research experiences, with students reporting greater gains in scientific thinking, confidence to contribute to science, confidence working in teams and working independently [59].

Table 4: Overview of student genome annotation recruitment, training, and project outcomes.

| **Institution** | **IRSC** | **UC** | **KSU** | **Cornell/BTI** |
|---|---|---|---|---|
| Number of students | 53 | 32 | 11 | 4 |
| Year of students^ | Second semester freshman to senior | Second semester freshman to senior | Freshman to senior | Freshman to senior |
| Number of Faculty/Senior scientists | 1 | 2 | 3 | 3 |

| Project delivery mode | Course, Capstone, Independent | Course, Capstone, Independent | Small projects | Course, Small projects |
|---|---|---|---|---|
| Annotation process | Pathways and Gene Families | Pathways and Gene Families | Pathways and Gene Families | Pathways and Gene Families |
| # Presentations (National level) | 1 | | | |
| Presentations (local/state level) | 5 | | | |
| Posters (national level) | 10 | | | |
| Posters (local/state level) | 44 | | | |
| Publications* | 12 | | | |
| Careers (STEM professions) | 19 | | | |
| Careers (Graduate/professional school) | 18 | | | |
| Total trainees (undergraduate) | 85 | | | |
| Total trainees (graduate) | 3 | | | |

Table 4 Legend: Across five institutions in multiple US states, dozens of undergraduate students, multiple graduate students, and a few faculty were involved in the manual curation of the predicted genes in the Official Gene Set version 3 (OGSv3). IRSC = Indian River State College (Florida), UC = University of Cincinnati (Ohio), KSU = Kansas State University (Kansas), Cornell = Cornell University (New York), BTI = Boyce Thompson Institute for Plant Research (New York), National level = US multi-state conference, Local/State = only one US state involved or a county-level local organization, STEM = Science, Technology, Engineering, Mathematics. ^Year of students: Typically undergraduate (Bachelors) students are enrolled two semesters a year, for four years (freshman, sophomore, junior, senior), while graduate students are post-Baccalaureate. *Publications with at least one undergraduate author include [6,23,50–55,60–62].

### 2.5.1. Orthogroup-based curation

In order to guide the manual curation, orthology analysis was performed with genes from 12 species (*Bemisia tabaci, Acyrthosiphon pisum, Rhopalosiphum maidis, Diaphorina citri, Pachypsylla venusta, Drosophila melanogaster, Anopheles gambiae, Homalodisca vitripennis, Gerris buenoi, Cimex lectularius, Halyomorpha halys, Oncopeltus fasciatus*) with 82.7 % of those genes being assigned to a total of 14,346 orthogroups. We utilized the orthogroups to confirm if a gene was missing from the Diaci v3.0 genome and verify if all the members of a gene family had been curated. The orthogroups also helped us identify duplicate (partial or complete) models that needed to be removed from the Official Gene Set v3 (OGSv3). We would like to note that using orthogroups is an effective strategy to initiate gene curation projects where an orthogroup containing a gene from *Drosophila* or another well characterized genome is selected followed by curation of the *D. citri* models in that orthogroup. Recently acquired genes from horizontal gene transfer or genes under positive selection that do not share sufficient sequence homology with orthologs will remain unclustered, but also offer valuable targets for manual curation.

### 2.5.2. Specific pathways targets for curation

The community curation team, comprising undergraduate and graduate students and their mentors (Table 4 above), manually annotated several new sets of genes in the Diaci v3.0 official gene set, OGSv3, focusing our efforts on orthogroups, pathways of interest to the *D. citri* research community, and genes that have been targeted for RNAi-based pest control in other insects [63–65]. Detailed reports for these gene sets are included in 11 accompanying articles published as data releases by Gigabyte [50–55,61,62,66–68]. Sequences of the manually curated transcripts and predicted proteins are available for download at citrusgreening.org [8].

#### 2.5.2.1. Development

We annotated *D. citri* orthologs of many genes important for insect development, including genes involved in segmentation, signal transduction and segmental identity.

##### Segmentation

During embryonic development, the insect embryo is divided into segments. The process of segmentation differs somewhat between insects, but the genes involved are well conserved [69]. Out of 33 genes involved in *Drosophila* segmentation, we identified and annotated 25 in *D. citri* [67]. Most of the differences in gene content between *Drosophila* and *D. citri* were expected based on observations in other hemipterans [70,71].

The segmentation genes are traditionally grouped into categories based on their function in *Drosophila*. In *D. citri*, we identified all of the maternal effect genes typically found in hemipterans, including *caudal*, *nanos*, *dorsal*, and one *TGFalpha*. As expected, we did not find *bicoid* or *oskar* [72]. We were unable to find several gap genes in *D. citri*, including *giant*, *buttonhead, orthodenticle 1* and *eagle*. All of these genes, except *eagle*, are also missing in the pea aphid genome. We did, however, find an ortholog of *huckebein*, which was reported absent in the pea aphid [70]. We identified one-to-one orthologs of each of the pair-rule genes in *D. citri*. The segment polarity genes include members of the Wnt and Hedgehog signal transduction pathways, which are described in the next section, as well as the transcription factor genes *engrailed* (*en*) and *gooseberry* (*gsb*). We identified a single copy of the *en* gene in

*D. citri*, but we were unable to find its paralog *invected* (*inv*) despite the presence of both *en* and *inv* in other hemipterans [41,70,73].

### Wnt and other signaling pathways

The Wnt signaling pathway is critical for segment polarity determination during segmentation and plays many other roles during development and in other biological processes [74,75]. Particular ligand genes have been lost in various insects [74] and *D. citri* is no exception. We were unable to find orthologs of *Wnt8/D* or *Wnt9 [54]*, which are also absent in pea aphids. We also did not find *Wnt16*, although it is present in several other hemipterans [70,76]. Conversely, *Wnt6* (contrary to a previous report [77]) and *Wnt10* are present in the *D. citri* genome but apparently absent in pea aphid [70]. The key downstream components of Wnt signaling are all present in the *D. citri* genome. We also annotated 13 genes from the Hedgehog pathway, 16 from the Notch pathway, and five members of the insulin signaling pathway. All of these pathways seem to be highly conserved in *D. citri*.

### Hox genes and their cofactors

Hox genes specify the identity of regions along the body axis. Notably, they are usually arranged in a chromosomal cluster in an order paralleling that of their functional domains along the anterior-posterior body axis, although breaks in the cluster have occurred in several lineages [78]. We identified the full complement of ten Hox genes in *D. citri*, although *labial*, which was present in previous versions of the genome, is missing in Diaci v3.0 due to a local misassembly [66]. The *D. citri* Hox cluster is split into two parts with the breakpoint between *Dfd* and *Scr*, the first time a break between these genes has been reported in insects. Both clusters of Hox genes are on the same chromosome about 6 Mb apart.

We also annotated four TALE-class homeobox genes encoding proteins that frequently serve as Hox cofactors in other organisms [79]. *D. citri* has two members of the MEINOX class, which is consistent with other insects. However, while most insects have only one copy of the PBC class gene *extradenticle* (*exd*), *D. citri* has two. One of these *exd* genes has no introns and appears to be a retrogene. In *D. citri,* both *exd* genes are expressed in a wide range of stages and tissues, suggesting they could both be functional. Somewhat surprisingly, comparison to other insect Exd proteins shows that the protein encoded by the retrogene is more conserved than the protein encoded by the *exd* copy with a more typical gene structure, raising the possibility that the retrogene has taken over the function of the original gene. Functional studies of both genes could help determine which of the genes has retained the original function of *exd* and whether either has acquired any new functions.

### 2.5.2.2. Immune Response

### Melanization

The melanization pathway is important for wound healing and defense against pathogens, in addition to its role in pigmentation [80]. We annotated 12 genes from the melanization pathway including two laccases and two tyrosinase prophenoloxidases. We also annotated the members of the *yellow* gene family, several of which have been implicated in melanization in other insects, although their precise role is still not well understood [81]. *D. citri* has nine *yellow* genes, including an apparent duplication of *yellow-y* [55]. Interestingly, one of these paralogs is expressed primarily in the egg and nymph stages, while the other is expressed mainly in adults. We also noted apparent differences in expression of some *yellow* genes between *C*Las-infected and uninfected psyllids, warranting further investigation into their potential role in *C*Las transmission.

#### 2.5.2.3. Metabolic and Cellular Functions

Housekeeping genes involved in metabolic pathways and essential cellular functions are excellent targets for RNAi-based pest control, since knockdown is often lethal. We have annotated genes from several essential pathways including carbohydrate metabolism, chitin metabolism, chromatin remodeling, protein degradation and organelle acidification.

##### Carbohydrate Metabolism

Although trehalose is the primary blood sugar in insects, glucose metabolism is also important [82]. Breakdown of trehalose produces glucose, which is then further metabolized by glycolysis [83]. The synthesis of glucose and trehalose (gluconeogenesis and trehaloneogenesis, respectively) follow the same pathway up to the production of glucose-6-phosphate, before diverging. Some insects synthesize both sugars, with glucose synthesis occurring primarily in neural cells [82]. We annotated 32 genes involved in carbohydrate metabolism [52]. The genes in these pathways are highly conserved, although there are a few differences in gene copy number. Most notably, *D. citri* has two copies each of *phosphoenolpyruvate carboxykinase* and *fructose bisphosphate-aldolase* instead of a single copy as found in many other insects. We did not find a *glucose-6-phosphatase* gene in *D. citri*, suggesting that trehalose is the end product of this pathway in *D. citri.*

##### Chitin Metabolism

Chitin is a major component of the insect cuticle and properly coordinated synthesis and breakdown of chitin are essential for growth and survival [84]. We annotated 19 orthologs of genes involved in chitin metabolism: three genes for enzymes involved in chitin synthesis [68], four chitin deacetylase genes [61] and 12 genes whose products are involved in chitin degradation [62]. Like most hemipterans and other hemimetabolous insects, *D. citri* has fewer chitin metabolism genes than do most holometabolous insects. This reduction is likely due to the fact that hemipterans do not undergo complete metamorphosis and their guts lack a true peritrophic membrane (another structure containing chitin) [85]. Consistent with the absence of a peritrophic membrane, *D. citri*, like other hemipterans, lacks the chitin synthase 2 gene (CHS2), which is specifically expressed in the peritrophic membrane in holometabolous insects [86,87]. *D. citri* and *A. pisum* [88] both lack a group VII chitinase, although one is present in some hemipteran species [89,90]. *D. citri* has apparent duplications of two genes involved in chitin metabolism: UDP-N-acetylglucosamine pyrophosphorylase (UAP) and Chitinase 10 (a group II chitinase). We also identified a chitinase gene that seems to have arisen by horizontal gene transfer [62]**.**

##### V-ATPase

Vacuolar ATP synthase (V-ATPase) regulates the acidity of various organelles by using energy from ATP to translocate protons across a membrane [91]. We annotated 14 genes encoding subunits of V-ATPase [53]. Gene copy number for each subunit varies between insects. The ACP genome has two copies of the $V_0$-α subunit gene and one copy of each of the others. We also annotated one gene encoding an accessory V-ATPase subunit that may help assemble the V-ATPase complex.

##### Ubiquitin-proteasome pathway

Ubiquitination is an ATP-dependent process that targets proteins for degradation. We annotated 15 genes of the ubiquitin-proteasome pathway in the ACP genome [51]. These genes were chosen based on studies by [15], who identified several *D. citri* ubiquitin-proteasome pathway proteins as being differentially expressed during *C*Las infection, and [64], who found

multiple ubiquitin-proteasome pathway genes among high lethality targets in a *Tribolium castaneum* RNAi screen.

#### Chromatin Remodeling

Chromatin remodeling proteins modify the positioning of nucleosomes along the chromosome to control the accessibility of particular regions of genomic DNA to protein binding. This is particularly important for regulating the transition between transcriptionally active and inactive chromatin states. We annotated 27 *D. citri* genes encoding chromatin remodeling proteins. Overall, the gene content of this family closely matches that of *Drosophila*. However, there is a duplication of *Mi-2* that has not been reported in other insects.

#### 2.5.2.4. Light sensing

Light sensing is essential for vision and maintenance of day-night rhythms. Both processes have a strong effect on the behavior of insects. To gain insight into these pathways in the psyllid, we searched for genes known to be involved in circadian rhythm and phototransduction.

#### Circadian Rhythm

We annotated 27 orthologs of genes putatively involved in circadian rhythm in other insects [50]. All components of the pathway are present in the *D. citri* genome. Notably, *D. citri* has two cryptochrome photoreceptor genes (*cry1* and *cry2*), which are believed to be the ancestral state of this pathway in insects [92–94].

#### Phototransduction

We annotated 21 genes in the phototransduction pathway. Opsin genes are important for the ability of an organism to detect light [95]. *D. citri* has four copies of opsin, the same number as the honeybee but fewer than the seven found in *Drosophila* [96]. Single gene copies of *UV-sensitive opsin, short-wavelength opsin, long-wavelength opsin*, and *rhodopsin 7* were annotated.

#### 2.5.2.5. Reproduction

Genes involved in reproduction, particularly spermatogenesis, have been targeted for use in a method of pest control called Sterile Insect Technique (SIT). Male insects sterilized by radiation, chemicals, or in this case, RNAi-based knockdown of spermatogenesis genes are released into the environment with the hope of out-competing fertile males in the natural population [97,98]. We have annotated 20 *D. citri* orthologs of genes that have been identified as potential SIT targets in other insects [99–101].

### 2.6. Improving OGSv1 curated gene models

A subset of genes that had been manually annotated in the Diaci v1.1 genome [6] were examined in the improved Diaci v3.0 genome. Of 159 previously annotated genes, 27 were removed from the Official Gene Set (OGS) version 3 (Table 5). All but one of these were artifactual duplicates resulting from misassemblies in Diaci v1.1. Most of these duplications are no longer present in Diaci v3.0 due to long-read based assembly and removal of duplicate scaffolds as previously described. One heat shock gene was removed because it was located in a region of Diaci v1.1 determined to be of endosymbiont origin and was shown by BLAST to be

identical to the molecular chaperone HtpG from "*Ca.* P. armatura". Of 24 annotated gene pathways or families examined, eight have a reduced and more accurate gene count after validation in Diaci v3.0 (Table 5).

We were able to improve almost all re-examined models. In many cases, the old models had only partial open reading frames (ORFs) and we were able to create models with complete ORFs. We were also able to add 5' and 3' UTRs to many models. The Iso-Seq transcripts were extremely helpful in creating longer models with a high level of confidence.

Table 5. Comparing Diaci v1.1 and Diaci v3.0 gene counts by pathway or gene family.

| **Diaci v1.1 Pathway / Gene Family** | **Number of genes in Diaci v1.1** | **Number of genes in Diaci v3.0** |
|---|---|---|
| Toll Receptors | 5 | 5 |
| Toll Receptor related genes | 3 | 3 |
| JAK/STAT pathway | 3 | 3 |
| C-type Lectins | 10 | 7 |
| Lysozymes | 5 | 5* |
| Superoxide dismutases | 4 | 4 |
| CLIP | 11 | 10 |
| Autophagy | 15 | 15 |
| Dicer | 4 | 2 |
| Drosha | 4 | 1 |
| Pasha | 2 | 2 |
| Loquacious and R2D2 proteins | 3 | 3 |
| Argonaute and PIWI proteins | 4 | 4 |
| Tudor Staphylococcal Nuclease protein | 2 | 1 |
| Vasa intronic gene | 1 | 1 |

| | | |
|---|---|---|
| Armitage | 1 | 1 |
| Fragile X Mental Retardation Protein | 1 | 1 |
| Spindle_E (homeless) | 2 | 1 |
| Rm62 protein | 1 | 1 |
| Ras-related nuclear protein | 1 | 1 |
| Aquaporin | 6 | 6 |
| Cathepsins and Cysteine Proteases | 34 | 23 |
| Heat shock proteins | 18 | 13 |
| Rab genes | 19 | 19 |
| **Total** | **159** | **132** |

Table 5 Legend: Gene counts of manually curated gene families or pathways in Diaci v1.1 versus Diaci v3.0. The asterisk (*) indicates a case where one lysozyme gene was missing from Diaci v3.0 but was found in the ALT contigs.

We were also able to improve the accuracy of annotated transcriptional isoforms in Diaci v3.0. We discarded 12 models previously annotated as differentially spliced isoforms. In most cases, these errors stemmed from falsely duplicated exons in the Diaci v1.1 genome being interpreted as alternative exons. Although our efforts were focused on assessing existing models, we did identify and annotate two previously unrecognized isoforms in the Diaci v3.0 genome during the validation process.

## 3. Conclusions and Potential Implications

The hemipteran *Diaphorina citri* (Asian citrus psyllid) is a primary target of approaches to stop the spread of the bacterial pathogen *C*Las that causes Huanglongbing (citrus greening disease) (reviewed in the book by Qureshi and Stansly (2020) [102], Chapters 11-17). In support of the global effort to control this insect vector, we provide a significantly improved, chromosomal-length genome assembly of the *D. citri* genome (Diaci v3.0), along with the corresponding Official Gene Set (OGSv3) which includes 18,947 protein coding genes, of which 1036 have been manually curated. The hologenome of this species is provided by direct sequencing of the two microbial symbionts "*Ca.* Profftella armatura" and "*Ca.* Carsonella armatura", and the *D. citri* strain of *Wolbachia* (wDi).

Under this project, we have also advanced the education and expertise of undergraduate and graduate students from multiple institutions who received training in gene curation as part of the community curation effort. Many trainees were also involved in the production of concrete deliverables in the form of research publications and presentations, including previous genome releases, this paper, and other direct studies completed as part of this project [10,50–55,61,62,66–68]. Standard operating procedures have been reported as part of a guide for other annotation communities [23,39] to implement similar programs.

All resources are available online at CitrusGreening.org, a portal for 'omics resources for the citrus greening disease research community [1]. The chromosomal-level *D. citri* genome assembly, automated annotation based on transcriptomics evidence, manual curation of critical pathways and a genome-independent *de novo* transcriptome will provide a foundation for comparative analysis among genomes of agricultural pests and plant-feeding hemipteran vectors of plant pathogens. Moreover, these resources will facilitate genome-wide association studies to identify the *D. citri* genes involved in *C*Las acquisition and transmission.

## Methods

### 1. *Diaphorina citri* Genome assembly

The DNA for PacBio sequencing and the Dovetail Chicago and Hi-C libraries was sourced from a *D. citri* colony originating from psyllids collected in Indian River County, Florida and maintained at the U.S. Horticultural Research Laboratory, USDA, Fort Pierce, Florida. High molecular weight DNA for PacBio sequencing was extracted from pooled *D. citri* adults using previously described methods [103]. DNA from a single male *D. citri* was used to generate the Dovetail Chicago library which was sequenced and used for both scaffolding and error correction.

PacBio sequencing was done on an RSII instrument to produce 36.2 Gb of long reads with an average length of 7.2 Kb. The Canu [104] assembler software was used to correct 40 X of the longest continuous long reads which were trimmed and used for the final assembly (-utgOvlErrorRate=0.013). The 38,263 unitigs produced by the Canu assembly with a contig N50 of 29 Kb were selected for scaffolding. 207 million Dovetail Chicago paired-end reads from a single male psyllid, with an insert size distribution of up to 250 Kb, were used to perform 12,369 joins and connect these unitigs into 25,942 scaffolds. This round of scaffolding added 12.3 Mb of Ns to the assembly. The Chicago scaffolded assembly was passed through another round of scaffolding with 388 million paired-end Hi-C reads [105]. The insert size distribution of the paired-end reads from the Hi-C library stretched up to 3 Mb with genome coverage of 819.14 X. The Hi-C scaffolding and validation reduced the number of scaffolds to 24,943 containing 12.4 Mb of Ns and a scaffold N50 of 26.7 Mb. This assembly consisted of 13 chromosomal-length scaffolds comprising 441 Mb of the genome. There were also 24,930 "unplaced scaffolds" that were not placed into the putative chromosomes. This unplaced set consisted of short scaffolds (scaffold N50 15 Kb). We reduced the duplication within the unplaced sequences by applying Redundans [32] at a threshold of 80 % identity and coverage. This split the unplaced set into

1271 unique scaffolds (33.2 Mb, N50 28.1 Kb) and 23,672 duplicated scaffolds (201 Mb, scaffold N50 13.7 Kb). The duplicated scaffolds are reported as alternate (ALT) contigs for the Diaci v3.0 assembly. Unique scaffolds were ordered based on length and joined with 1000 Ns separating adjacent scaffolds to create chromosome 00 of length 34.5 Mb. We performed two rounds of error corrections with Pilon [106] using Illumina reads from the single male psyllid individual. We opted to use this data set instead of short read data from multiple individuals to avoid introducing artificial heterozygosity into the genome assembly. Pilon was optimized to only correct regions of the genome where the change was supported by more than 90 % of the aligned bases at that position (--fix bases --diploid --mindepth 0.9). We also performed one round of error correction with Pilon and Illumina RNA-Seq to polish the genic regions of the assembly (--unpaired --fix bases --diploid) based on unspliced alignments to the genome. To minimize microbial contamination, we removed regions of the assembly with very high read coverage (>1000x) and BLAST hits (>200 bp and >90% coverage) to an endosymbiont genome. We used BUSCO [7] Arthropoda (arthropoda_odb10 with 1013 markers) and Hemiptera (hemiptera_odb10 with 2510 markers) marker sets to evaluate the completeness of all the data sets.

## 1.1. Endosymbiont assembly, validation and annotation

To identify endosymbionts present in *D. citri*, we identified contigs from the Diaci v3.0 assembly with sequence similarity to published "*Ca.* P. armatura", "*Ca.* C. ruddii", and *w*Di genomes. Endosymbiont contigs were manually validated using the Integrative Genomics Viewer (IGV) [107] to assess coverage of PacBio and Illumina DNA reads. Any mis-assembled contigs were split into multiple contigs and regions that had no support were removed from further analysis. We applied the SSPACE-LongRead [108] scaffolding algorithm to contigs verified in IGV in order to join any contigs for which there was PacBio evidence of overlap. When running SSPACE we required 20 PacBio reads to link contig-pairs for scaffolding (-l 20). We annotated scaffolds with Prokka [109] and analyzed them using Orthofinder [110] and BLAST Ring Image Generator (BRIG) [110,111]. We used Orthofinder to judge the completeness of the genome assemblies by comparing the orthogroups present in our assemblies to those present in published genomes. We defined a conserved orthogroup for a given endosymbiont to be one that contains at least one protein from each reference genome (Supplemental Figure 1). A conserved orthogroup then indicates protein functionality shared among all known genomes. We defined a unique orthogroup for a given endosymbiont to be one that is present in our assembly, and is not present in at least 50 % of the reference genomes.

### 1.2. Comparative genomic analysis

We used MCScanX [112] to identify syntenic blocks between the chromosomal-length scaffolds of the Diaci v3.0 genome and other publicly available psyllid genomes [22,35]. Simplified GFF files of predicted genes from each genome were prepared by using gffread v0.12.1 [113] to extract scaffold names, gene names and coordinates from available GFF files. For each genome pair, BLAST+ [114] was used to compare predicted proteins in all possible

pairwise combinations. The BLAST results and simplified GFF files were then used as input for MCScanX with default parameters. Parallel plots were created from the MCScanX results using SynVisio [115].

2. Iso-Seq

In order to detect alternative splicing we performed isoform sequencing (Iso-Seq) using PacBio to generate high quality (HQ), full length transcripts. Samples were collected from healthy and *C*Las-infected adult *D. citri*, and healthy and *C*Las-infected nymph *D. citri*. The sequences were processed using the PacBio SMRTlink v4.0 software [116]. The pipeline runs circular consensus sequencing (CCS) using the raw reads followed by a classification of full length and non-full length transcripts. The resulting transcripts were polished to generate HQ consensus isoforms and polished low quality (LQ) consensus isoforms. A comprehensive set of genome independent HQ Iso-seq isoforms from adult and nymph tissue was achieved with three rounds of error corrections with Pilon [106] using Illumina RNA-Seq reads.

3. *De novo* transcriptome

The *de novo* transcriptome was generated using the Iso-Seq data described above and publicly available short read RNA-Seq data from different experimental conditions and tissues, obtained from the NCBI SRA database (Supplemental Table 1).

Trinity [117] was used to perform *de novo* assembly of both RNAseq and Iso-Seq datasets using default parameters. Transcripts that did not encode at least one Pfam domain [118] or did not contain a coding region (as determined by TransDecoder [119]) were removed from the dataset. Sequences were then clustered with CD-HIT (cd-hit-est) [120] using a threshold of 75% sequence identity, and the longest transcript in each cluster was selected. We used BLAST [121] to identify contamination from archaeal, viral and bacterial sequences and removed those transcripts. We also removed transcripts that did not have any matches to the Trembl Insecta subset in the Uniprot database [122]. Finally, the remaining transcripts were mapped to Diaci v3.0 using GMAP [123] and the Cupcake ToFU [124] pipeline was used to collapse redundant isoforms at each locus. We assigned transcript identifiers according to the source of the raw data: DcDTr for RNA-Seq transcripts and DcDTi for Iso-Seq transcripts.

4. Automated predictions of protein-coding genes

A repeat library for the Diaci v3.0 genome was constructed using RepeatModeler [125]. The repeat library was screened for known protein association with ProtExcluder [126] based on similarity with proteins obtained from Swiss-Prot (Arthropoda) [122]. The resulting repeat library was used to mask the Diaci v3.0 genome with RepeatMasker [125]. Repeat annotation is available on citrusgreening.org FTP.

Protein-coding genes were predicted on the repeat-masked Diaci v3.0 genome through iterative processing within the MAKER (v3) annotation pipeline [127]. For homology evidence, manually annotated proteins of Arthropoda were downloaded from Swiss-Prot [122]. Expression evidence was obtained through multiple sources. RNA-Seq data generated as part of this work is listed in Supplemental Table 1. Publicly available RNA-Seq datasets were obtained from NCBI SRA database (Supplemental Table 1). All the RNA-Seq data was mapped to the genome using HISAT2 [128] and the transcriptome was assembled with StringTie [129]. Independently, high-quality PacBio Iso-Seq transcripts were mapped to the genome and clustered through Cupcake-ToFU clustering [124]. Transcriptomes obtained through RNA-Seq and Iso-Seq were processed with Mikado pipeline [130] for refining the transcriptome.

*Ab initio* gene predictions were performed with Augustus [131] and SNAP [132] gene predictors. Augustus was trained with available RNA-Seq data within BRAKER1 [133]. SNAP was trained iteratively within the MAKER pipeline based on MAKER guidelines. The gene predictions were supplied to MAKER along with expression and homology evidence which was run with default parameters. The Mikado-refined transcriptome was passed as a predictor (pred_gff). Gene identifiers were assigned based on their genomic location. For example, the Dcitr01g01000.1.1 gene consists of a five letter species identifier (Dcitr), a two digit scaffold/chromosome number (01), an abbreviation for gene (g), a unique five digit ID for the gene (01000) followed by version number (.1) and isoform number (.1). Consecutive genes were assigned identifiers in increments of 10 to allow addition of genes in the future.

5. Transcription factor prediction

We identified likely transcription factors (TFs) by scanning the amino acid sequences of predicted protein coding genes for putative DNA binding domains (DBDs), and when possible, we predicted the DNA binding specificity of each TF using the procedures described in Weirauch *et al.* [134]. Briefly, we scanned all protein sequences for putative DBDs using the 81 Pfam [118] models listed in Weirauch and Hughes [135] and the HMMER tool [136], with the recommended detection thresholds of per-sequence Eval < 0.01 and per-domain conditional Eval < 0.01. Each protein was classified into a family based on its DBDs and their order in the protein sequence (e.g., bZIPx1, AP2x2, Homeodomain+Pou). We then aligned the resulting DBD sequences within each family using ClustalOmega [137], with default settings. For protein pairs with multiple DBDs, each DBD was aligned separately. From these alignments, we calculated the sequence identity of all DBD sequence pairs (i.e. the percent of AA residues that are exactly the same across all positions in the alignment). Using previously established sequence identify thresholds for each family [134], we mapped the predicted DNA binding specificities by simple transfer. For example, the DBD of Dcitr04g16960.1.1 is 98 % identical to the *Drosophila melanogaster* ‘oc’ (FBgn0004102) TF. Since the DNA binding specificity of ‘oc’ has already been experimentally determined, and the cutoff for the Homeodomain family of TFs is 70 %, we can infer that Dcitr04g16960.1.1 will have the same binding specificity as ‘oc’ (Supplementary File 1). WGCNA [138] was conducted according to methods previously adapted to examine gene expression patterns in relation to TF binding predictions [44,139,140]. Enriched TF binding motifs were identified in the 500 and 2000 bp regions upstream of the

putative transcription start site using the HOMER tool as in previous studies examining the dynamics between TFs and expression patterns in insects [45].

## 6. Manual curation and OGSv3

### 6.1. Manual curation

Manual curation of automated predictions was carried out using the Apollo annotation editor [21] plugin for JBrowse genome browser [141]. Among other genomic resources, the Apollo instance for the Diaci v3.0 genome is hosted at citrusgreening.org [1]. All the evidence used for automated annotation was added as tracks on JBrowse/Apollo. Other tracks to assist in accurate evidence-based curation included proteins from related insects mapped with Exonerate [142]. All publicly available RNA-Seq datasets were mapped as quantitative tracks. Manual curation was performed following the workflow previously described in Hosmani *et al., [39]* and modifications to the workflow, if any, are described in the individual pathway reports. A more detailed version of the annotation workflow is available from Shippy *et al.*, [143].

### 6.2. Official Gene Set v3 (OGSv3)

The Official gene set version 3 (OGSv3) consists of predicted gene models from the MAKER annotation pipeline and manually curated genes from the Apollo annotation editor. Before merging curated genes to create OGSv3, overlapping genes from the automatic predictions were removed. We also removed 258 predicted genes that were flagged as incorrect by annotators during the manual curation process. These included incorrect predictions, false duplications, and genes not supported by transcriptome evidence.

### 6.3. Orthology analysis

Orthology analysis of 12 species (*Bemisia tabaci, Acyrthosiphon pisum, Rhopalosiphum maidis, Diaphorina citri, Pachypsylla venusta, Drosophila melanogaster, Anopheles gambiae, Homalodisca vitripennis, Gerris buenoi, Cimex lectularius, Halyomorpha halys, Oncopeltus fasciatus*) was performed using a total of 240,242 genes. Orthofinder (v2.3.3) [110,144] was run to get orthogroups using the -S blast parameter as a sequence search program. Orthogroups annotation was performed with Interproscan (v5.32-71.0) [145] to assign GO terms and Pfam domains.

Figure 5: A visual representation of methods and pipelines for assembly, annotation, orthology, synteny, prediction and transcriptomics.

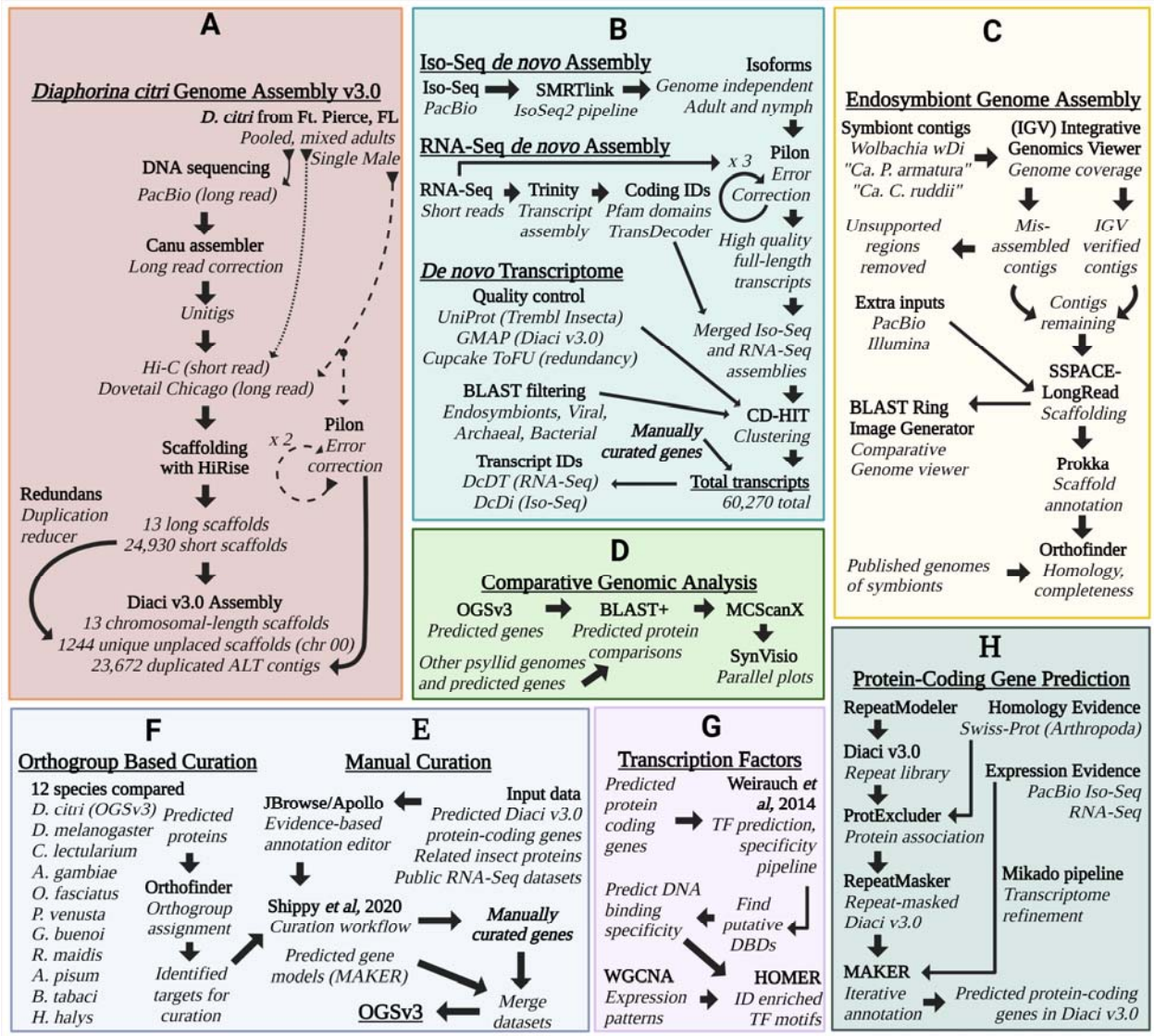



Figure 5 legend: A) The *Diaphorina citri* Diaci v3.0 genome assembly began with input of long read DNA sequencing data, followed by multiple rounds of assembly and scaffolding , duplication reduction, error correction and exclusion of non-target organism reads (see C). B) Iso-Sequencing used long read RNA sequencing data and an established computational pipeline (SMRTLink) to assemble gene isoforms independent of any genome. Additionally, a *de novo D. citri* transcriptome was generated using public short read RNA sequencing data collected from online repositories combined with the high quality Iso-seq isoforms. The transcriptome was assembled using Trinity, then cleaned and error corrected, subjected to BLAST to remove contaminants and clustered by locus to remove redundant isoforms. C) Genome assemblies of the *D. citri* endosymbionts including "*Candidatus* Profftella armatura", "*Candidatus* Carsonella ruddii" and *Wolbachia* were generated using reads excluded during the Diaci v3.0 assembly process, then cleaned and verified using Orthofinder and BLAST. D) We performed synteny analysis between the Diaci v3.0 genome and other full length psyllid genome assemblies. E) Manual curation was a major part of this genome project and involved teams of student annotators, graduate students and faculty across multiple institutions. We have

previously published the annotation pipeline used [23]. F) We compared the predicted proteins of 12 different insect species to generate orthogroups and assign GO terms to *D. citri* genes. G) Transcription factor (TF) prediction was performed following previously published pipelines and weighted gene co-expression network analysis (WGCNA) and HOMER were used to explore the effect of *C*Las infection on TFs and their target genes. H) Protein-coding genes were predicted and annotated following the MAKER annotation pipeline and results were informed using Mikado and RNAseq datasets.

## Declarations

### List of abbreviations

ALT = Alternate
BLAST = Basic Local Alignment Search Tool
*C*Las = "*Candidatus* Liberibacter asiaticus"
DBD = DNA Binding Domains
*D. citri* = *Diaphorina citri,* Asian citrus psyllid
FPKM = Fragments per kilobase of exon per million mapped
FTP = File Transfer Protocol
Gb/Mb/Kb/bp = Gigabyte, Megabyte, Kilobyte, base pair
GFF = General Features Format
GO = Gene Ontology
HLB = Huanglongbing or Citrus Greening disease
HQ = High Quality/Low Quality
HSI = Hispanic Serving Institute
OGS = Official Gene Set
ORF = Open Reading Frame
NCBI SRA = National Center for Biotechnological Information Sequence Read Archive
NIFA = National Institute of Food and Agriculture
TF = Transcription Factor
USDA = United States Department of Agriculture
WCGNA = Weighted Gene Co-expression Network Analysis

### Ethics statement

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Funding

All open-access fees, student annotators and post-docs were funded through USDA-NIFA grant 2015-70016-23028, USDA-NIFA HSI 2020-38422-32252 and 2020-70029-33199. Additional funding was provided by the USDA ARS CRIS Project 8062-22410-007-000-D, Kansas INBRE through an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103418.

Author contributions

SJB, TD, MH, WBH, RGS and LAM conceptualized the project and obtained funding. SJB, TD, SS and LAM contributed to project administration. PSH, MF-G, SS, TDS, MTW, and JB developed methods. PSH, MF-G, and SS curated data and contributed to software development. TDS, SM, MTR, CV, CM, WT, LD, CC, SH, MR, JN, KK, MRJ, DH, SA, YD, TP, SV, RA, SN, JV, MP, AL, STB, NP, YO, BT, RG, AN, DLH, AK, AT, BK, XC, ME, ND, RM, RH, KG, MK, MH, MMS, KMRP, AM, NW, JB, TW, SL, SMi, EM, SK and SS conducted the investigation. TD; SS; TDS; JB, and SJB provided leadership and oversight for the project. TDS, CV, CM, LD, and SM validated manual annotations. MAM, TDS; JB; MTR, and SS created visualizations of the data. PSH; MF-G; TDS; JB; TD; and SS wrote the initial draft. MAM; TDS; WBH, MH and SS reviewed and edited the manuscript. All authors approved the submitted version.

Acknowledgements

Kascha Bohnenblust Johnson from Kansas State University assisted with the organization of meetings for the community curation effort. Some of the computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, EPS-0919443, ACI-1440548, CHE-1726332, and NIH P20GM113109. Maria Gonzalez, Biological Science Technician, USDA, ARS, Ft. Pierce, FL, assisted with psyllid collection, DNA and RNA extractions, and production of high-molecular weight gDNA.

Disclaimers

Mention of trade names or commercial products herein is solely for the purpose of providing specific information and does not imply recommendation or endorsement, to the exclusion of other similar products or services by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References

1. Flores-Gonzalez M, Hosmani PS, Fernandez-Pozo N, Mann M, Humann JL, Main D, et al.. Citrusgreening.org: An open access and integrated systems biology portal for the Huanglongbing (HLB) disease complex. bioRxiv.

2. Singerman A, Useche P: Impact of citrus greening on citrus operations in Florida. https://edis.ifas.ufl.edu/publication/FE983 Accessed 2022 Nov 28.

3. Lee JA, Halbert SE, Dawson WO, Robertson CJ, Keesling JE, Singer BH. Asymptomatic spread of Huanglongbing and implications for disease control. *Proc Natl Acad Sci U S A*. 112:7605–102015;

4. Ammar E-D, Hall DG, Hosseinzadeh S, Heck M. The quest for a non-vector psyllid: Natural variation in acquisition and transmission of the huanglongbing pathogen “Candidatus Liberibacter asiaticus” by Asian citrus psyllid isofemale lines. *PLoS One*. Public Library of Science (PLoS); 13:e01958042018;

5. El-Desouky A, Shatters RG Jr, Heck M. Huanglongbing pathogens: acquisition, transmission and vector interactions. *Asian citrus psyllid: biology, ecology and management of the Huanglongbing vector*. :113–392020;

6. Saha S, Hosmani PS, Villalobos-Ayala K, Miller S, Shippy T, Flores M, et al.. Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community. *Database* . 2017; doi: 10.1093/database/bax032.

7. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210–22015;

8. : CitrusGreening.org. https://citrusgreening.org/organism/Diaphorina_citri/genome Accessed 2022 Nov 28.

9. Saha S, Hosmani P, Flores M, Hunter W, Brown S, Mueller LA. Using long reads, optical maps and long-range scaffolding to improve the Diaphorina citri genome.

10. Mann M, Saha S, Cicero JM, Pitino M, Moulton K, Hunter WB, et al.. Lessons learned about the biology and genomics of Diaphorina citri infection with “Candidatus Liberibacter asiaticus” by integrating new and archived organ-specific transcriptome data. *Gigascience*. 2022; doi: 10.1093/gigascience/giac035.

11. Chu C-C, Gill TA, Hoffmann M, Pelz-Stelinski KS. Inter-Population Variability of Endosymbiont Densities in the Asian Citrus Psyllid (Diaphorina citri Kuwayama). *Microb Ecol*. 71:999–10072016;

12. Tamames J, Gil R, Latorre A, Peretó J, Silva FJ, Moya A. The frontier between cell and organelle: genome analysis of Candidatus Carsonella ruddii. *BMC Evol Biol*. 7:1812007;

13. Morrow JL, Hall AAG, Riegler M. Symbionts in waiting: the dynamics of incipient endosymbiont complementation and replacement in minimal bacterial communities of psyllids. *Microbiome*. 5:582017;

14. Hosseinzadeh S, Ramsey J, Mann M, Bennett L, Hunter WB, Shams-Bakhsh M, et al.. Color morphology of Diaphorina citri influences interactions with its bacterial endosymbionts and “Candidatus Liberibacter asiaticus.” *PLoS One*. 14:e02165992019;

15. Ramsey JS, Johnson RS, Hoki JS, Kruse A, Mahoney J, Hilf ME, et al.. Metabolic Interplay between the Asian Citrus Psyllid and Its Profftella Symbiont: An Achilles’ Heel of the Citrus

Greening Insect Vector. *PLoS One*. 10:e01408262015;

16. Hunter WB, Cooper WR, Sandoval-Mojica AF, McCollum G, Aishwarya V, Pelz-Stelinski KS. Improving Suppression of Hemipteran Vectors and Bacterial Pathogens of Citrus and Solanaceous Plants: Advances in Antisense Oligonucleotides (FANA). *Frontiers in Agronomy*. 2021; doi: 10.3389/fagro.2021.675247.

17. Nakabachi A, Ueoka R, Oshima K, Teta R, Mangoni A, Gurgui M, et al.. Defensive bacteriome symbiont with a drastically reduced genome. *Curr Biol*. 23:1478–842013;

18. Hoffmann M, Coy MR, Kingdom Gibbard HN, Pelz-Stelinski KS. Wolbachia infection density in populations of the Asian citrus psyllid (Hemiptera: Liviidae). *Environ Entomol*. 43:1215–222014;

19. Neupane S, Bonilla SI, Manalo AM, Pelz-Stelinski KS. Complete de novo assembly of Wolbachia endosymbiont of Diaphorina citri Kuwayama (Hemiptera: Liviidae) using long-read genome sequencing. *Sci Rep*. 12:1252022;

20. Pelz-Stelinski KS, Brlansky RH, Ebert TA, Rogers ME. Transmission parameters for Candidatus liberibacter asiaticus by Asian citrus psyllid (Hemiptera: Psyllidae). *J Econ Entomol*. 103:1531–412010;

21. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al.. Apollo: Democratizing genome annotation. *PLoS Comput Biol*. 15:e10067902019;

22. Carlson CR, Ter Horst AM, Johnston JS, Henry E, Falk BW, Kuo Y-W. High-quality, chromosome-scale genome assemblies: comparisons of three Diaphorina citri (Asian citrus psyllid) geographic populations. *DNA Res*. 2022; doi: 10.1093/dnares/dsac027.

23. Shippy TD, Miller S, Massimino C, Vosburg C, Hosmani PS, Flores-Gonzalez M, et al.: Annotating genes in Diaphorina citri genome version 3. protocols.io. https://www.protocols.io/view/annotating-genes-in-diaphorina-citri-genome-versio-261ge44nwv47/v1 (2020). Accessed 2022 Nov 28.

24. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al.. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res*. 26:342–502016;

25. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*. 13:278–892015;

26. Eisenstein M. Startups use short-read data to expand long-read sequencing market. *Nat Biotechnol*. 33:433–52015;

27. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods*. 58:268–762012;

28. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al.. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*. 7:e477682012;

29. Lee J: PBJelly: A clone of https://sourceforge.net/projects/pb-jelly/. https://github.com/cgjosephlee/PBJelly Accessed 2022 Nov 28.

30. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al.. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 326:289–932009;

31. Labina ES. A chromosomal study of 11 species of Psyllinea (Insecta: Homoptera). *Comp Cytogenet*. 1:149–542007;

32. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res*. 44:e1132016;

33. Morrow JL, Om N, Beattie GAC, Chambers GA, Donovan NJ, Liefting LW, et al.. Characterization of the bacterial communities of psyllids associated with Rutaceae in Bhutan by high throughput sequencing. *BMC Microbiol*. 20:2152020;

34. : Diaphorina citri mitochondrion, complete genome - Nucleotide - NCBI. https://www.ncbi.nlm.nih.gov/nuccore/kx073968 Accessed 2022 Dec 21.

35. Li Y, Zhang B, Moran NA. The Aphid X Chromosome Is a Dangerous Place for Functionally Important Genes: Diverse Evolution of Hemipteran Genomes Based on Chromosome-Level Assemblies. *Mol Biol Evol*. 37:2357–682020;

36. Maryańska-Nadachowska A, Kuznetsova VG, Nokkala S. Standard and C-banded meiotic karyotypes of Psylloidea (Sternorrhyncha, Homoptera, Insecta). *Folia Biol*. 49:53–622001;

37. Labina ES, Maryańska-Nadachowska A, Kuznetsova VG. Meiotic karyotypes in males of nineteen species of Psylloidea (Hemiptera) in the families Psyllidae and Triozidae. *Folia Biol* . 55:27–342007;

38. Maryanska-Nadachowska A, Yang M-M. Karyotypes and morphology of testes in some species of Pachypsyllinae Becker-Migdisova (Psylloidea, Spondyliaspididae). *Folia Biologica*. Polish Scientific Journals Database; 45:21–61997;

39. Hosmani PS, Shippy T, Miller S, Benoit JB, Munoz-Torres M, Flores-Gonzalez M, et al.. A quick guide for student-driven community genome annotation. *PLoS Comput Biol*. 15:e10066822019;

40. Saha S, Cooksey AM, Childers AK, Poelchau MF, McCarthy FM. Workflows for Rapid Functional Annotation of Diverse Arthropod Genomes. *Insects*. 2021; doi: 10.3390/insects12080748.

41. Panfilio KA, Vargas Jentzsch IM, Benoit JB, Erezyilmaz D, Suzuki Y, Colella S, et al.. Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome. *Genome Biol*. 20:642019;

42. Benoit JB, Adelman ZN, Reinhardt K, Dolan A, Poelchau M, Jennings EC, et al.. Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. *Nat Commun*. Springer Nature America, Inc; 7:101652016;

43. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, et al.. Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *Genome Biol*. 17:2272016;

44. Olafson PU, Aksoy S, Attardo GM, Buckmeier G, Chen X, Coates CJ, et al.. The genome of the stable fly, Stomoxys calcitrans, reveals potential mechanisms underlying reproduction, host interactions, and novel targets for pest control. *BMC Biol*. 19:412021;

45. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al.. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 38:576–892010;

46. Hamaratoglu F, Affolter M, Pyrowolakis G. Dpp/BMP signaling in flies: from molecules to biology. *Semin Cell Dev Biol*. 32:128–362014;

47. Mann Marina, Fattah-Hosseini Somayeh, Ammar El-Desouky, Stange Richard, Warrick EricaRose, Sturgeon Kasie, et al.. Diaphorina citri Nymphs Are Resistant to Morphological Changes Induced by “Candidatus Liberibacter asiaticus” in Midgut Epithelial Cells. *Infect Immun*. American Society for Microbiology; 86:e00889–172018;

48. Kruse A, Fattah-Hosseini S, Saha S, Johnson R, Warwick E, Sturgeon K, et al.. Combining ’omics and microscopy to visualize interactions between the Asian citrus psyllid vector and the Huanglongbing pathogen Candidatus Liberibacter asiaticus in the insect gut. *PLoS One*. 12:e01795312017;

49. Rhee DY, Cho D-Y, Zhai B, Slattery M, Ma L, Mintseris J, et al.. Transcription factor networks in Drosophila melanogaster. *Cell Rep*. 8:2031–432014;

50. Reynolds M, de Oliveira L, Vosburg C, Paris T, Massimino C, Norus J, et al.. Annotation of Putative Circadian Rhythm-Associated Genes in Diaphorina citri (Hemiptera : Liviidae). bioRxiv.

51. Tank W, Shippy T, Thate A, Massimino C, Hosmani PS, Flores-Gonzalez M, et al.. Ubiquitin Proteasome Pathway annotation in Diaphorina citri can reveal potential targets for RNAi based pest management. bioRxiv.

52. Tamayo B, Kercher K, Vosburg C, Massimino C, Jernigan MR, Hasan DL, et al.. Annotation of glycolysis, gluconeogenesis, and trehaloneogenesis pathways provide insight into carbohydrate metabolism in the Asian citrus psyllid. bioRxiv.

53. Grace R, Massimino C, Shippy TD, Tank W, Hosmani PS, Flores-Gonzalez M, et al.. Genomic identification, annotation, and comparative analysis of Vacuolar-type ATP synthase subunits in Diaphorina citri. bioRxiv.

54. Vosburg C, Reynolds M, Noel R, Shippy T, Hosmani PS, Flores-Gonzalez M, et al.. Utilizing a chromosomal-length genome assembly to annotate the Wnt signaling pathway in the Asian citrus psyllid, Diaphorina citri. Gigabyte.

55. Massimino C, Vosburg C, Shippy T, Hosmani PS, Flores-Gonzalez M, Mueller LA, et al.. Annotation of yellow genes in Diaphorina citri, the vector for Huanglongbing disease. *Gigabyte*. GigaScience Press; 2021:1–152021;

56. Saha S, Shippy TD, Brown SJ, Benoit JB, D’Elia T. Undergraduate Virtual Engagement in Community Genome Annotation Provides Flexibility to Overcome Course Disruptions. *J Microbiol Biol Educ*. 2021; doi: 10.1128/jmbe.v22i1.2395.

57. Adedokun OA, Parker LC, Childress A, Burgess W, Adams R, Agnew CR, et al.. Effect of

time on perceived gains from an undergraduate research program. *CBE Life Sci Educ*. 13:139–482014;

58. Thiry H, Laursen SL. The role of student-advisor interactions in apprenticing undergraduate researchers into a scientific community of practice. *J Sci Educ Technol*. Springer Science and Business Media LLC; 20:771–842011;

59. Daniels H, Grineski SE, Collins TW, Morales DX, Morera O, Echegoyen L. Factors Influencing Student Gains from Undergraduate Research Experiences at a Hispanic-Serving Institution. *CBE Life Sci Educ*. 2016; doi: 10.1187/cbe.15-07-0163.

60. Mueller L, Saha S, Chen C, Benoit J. Leveraging co-expression network analysis to establish transcriptional shifts in the gut of the Asian citrus psyllid during Candidatus Liberibacter asiaticus infection. *Proceedings of The 1st International Electronic Conference on Entomology*. Basel, Switzerland: MDPI;

61. Miller S, Shippy TD, Tamayo B, Hosmani PS, Flores-Gonzalez M, Mueller LA, et al.. In silico characterization of chitin deacetylase genes in the Diaphorina citri genome. *Gigabyte*. GigaScience Press; 2021:1–112021;

62. Shippy TD, Miller S, Tamayo B, Hosmani PS, Flores-Gonzalez M, Mueller LA, et al.. Manual curation and phylogenetic analysis of chitinase family genes in the Asian citrus psyllid, Diaphorina citri. bioRxiv.

63. Kola VSR, Renuka P, Madhav MS, Mangrauthia SK. Key enzymes and proteins of crop insects as candidate for RNAi based gene silencing. *Front Physiol*. frontiersin.org; 6:1192015;

64. Ulrich J, Dao VA, Majumdar U, Schmitt-Engel C, Schwirz J, Schultheis D, et al.. Large scale RNAi screen in Tribolium reveals novel target genes for pest control and the proteasome as prime target. *BMC Genomics*. bmcgenomics.biomedcentral.com; 16:6742015;

65. Baum JA, Bogaert T, Clinton W, Heck GR, Feldmann P, Ilagan O, et al.. Control of coleopteran insect pests through RNA interference. *Nat Biotechnol*. nature.com; 25:1322–62007;

66. Shippy T, Hosmani PS, Flores-Gonzalez M, Mueller LA, Hunter WB, Brown SJ, et al.. Annotation of Hox cluster and Hox cofactor genes in the Asian citrus psyllid, Diaphorina citri, reveals novel features.

67. Miller S, Shippy TD, Hosmani PS, Flores-Gonzalez M, Mueller LA, Hunter WB, et al.. Annotation of segmentation pathway genes in the Asian citrus psyllid, Diaphorina citri. *Gigabyte*. GigaScience Press; 2021:1–132021;

68. Miller S, Shippy TD, Tamayo B, Hosmani PS, Flores-Gonzalez M, Mueller LA, et al.. Annotation of chitin biosynthesis genes in Diaphorina citri, the Asian citrus psyllid. *Gigabyte*. GigaScience Press; 2021:1–122021;

69. Damen WGM. Evolutionary conservation and divergence of the segmentation process in arthropods. *Dev Dyn*. 236:1379–912007;

70. Shigenobu S, Bickel RD, Brisson JA, Butts T, Chang C-C, Christiaens O, et al.. Comprehensive survey of developmental genes in the pea aphid, Acyrthosiphon pisum:

frequent lineage-specific duplications and losses of developmental genes. *Insect Mol Biol*. 19 Suppl 2:47–622010;

71. Chen W, Hasegawa DK, Kaur N, Kliot A, Pinheiro PV, Luan J, et al.. The draft genome of whitefly Bemisia tabaci MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biol*. Springer Science and Business Media LLC; 2016; doi: 10.1186/s12915-016-0321-y.

72. Rosenberg MI, Lynch JA, Desplan C. Heads and tails: evolution of antero-posterior patterning in insects. *Biochim Biophys Acta*. 1789:333–422009;

73. Fu S-J, Zhang J-L, Xu H-J. A genome-wide identification and analysis of the homeobox genes in the brown planthopper, Nilaparvata lugens (Hemiptera: Delphacidae). *Arch Insect Biochem Physiol*. Wiley; 108:e218332021;

74. Holstein TW. The evolution of the Wnt pathway. *Cold Spring Harb Perspect Biol*. 4:a0079222012;

75. Swarup S, Verheyen EM. Wnt/Wingless signaling in Drosophila. *Cold Spring Harb Perspect Biol*. 2012; doi: 10.1101/cshperspect.a007930.

76. Wang L, Tang N, Gao X, Chang Z, Zhang L, Zhou G, et al.. Genome sequence of a rice pest, the white-backed planthopper (Sogatella furcifera). *Gigascience*. 6:1–92017;

77. Doumpas N, Jékely G, Teleman AA. Wnt6 is required for maxillary palp formation in Drosophila. *BMC Biol*. 11:1042013;

78. Lemons D, McGinnis W. Genomic evolution of Hox gene clusters. *Science*. 313:1918–222006;

79. Moens CB, Selleri L. Hox cofactors in vertebrate development. *Dev Biol*. 291:193–2062006;

80. Hillyer JF. Insect immunology and hematopoiesis. *Dev Comp Immunol*. 58:102–182016;

81. Ferguson LC, Green J, Surridge A, Jiggins CD. Evolution of the insect yellow gene family. *Mol Biol Evol*. 28:257–722011;

82. Miyamoto T, Amrein H. Gluconeogenesis: An ancient biochemical pathway with a new twist. *Fly* . 11:218–232017;

83. Thompson SN. Trehalose – The Insect “Blood” Sugar. *Advances in Insect Physiology*. Academic Press; p. 205–85.

84. Zhu KY, Merzendorfer H, Zhang W, Zhang J, Muthukrishnan S. Biosynthesis, Turnover, and Functions of Chitin in Insects. *Annu Rev Entomol*. 61:177–962016;

85. Silva CP, Silva JR, Vasconcelos FF, Petretski MDA, Damatta RA, Ribeiro AF, et al.. Occurrence of midgut perimicrovillar membranes in paraneopteran insect orders with comments on their function and evolutionary significance. *Arthropod Struct Dev*. 33:139–482004;

86. Muthukrishnan S, Merzendorfer H, Arakane Y, Kramer KJ. Chitin Metabolism in Insects. *Insect Molecular Biology and Biochemistry*. p. 193–235.

87. Arakane Y, Muthukrishnan S, Kramer KJ, Specht CA, Tomoyasu Y, Lorenzen MD, et al.. The Tribolium chitin synthase genes TcCHS1 and TcCHS2 are specialized for synthesis of epidermal cuticle and midgut peritrophic matrix. *Insect Mol Biol*. 14:453–632005;

88. Nakabachi A, Shigenobu S, Miyagishima S. Chitinase-like proteins encoded in the genome of the pea aphid, Acyrthosiphon pisum. *Insect Mol Biol*. Wiley; 19 Suppl 2:175–852010;

89. Xi Y, Pan P-L, Ye Y-X, Yu B, Xu H-J, Zhang C-X. Chitinase-like gene family in the brown planthopper, Nilaparvata lugens. *Insect Mol Biol*. 24:29–402015;

90. Yang X, Zhou C, Long G, Yang H, Chen C, Jin D. Characterization and functional analysis of chitinase family genes involved in nymph-adult transition of Sogatella furcifera. *Insect Sci*. Wiley; 28:901–162021;

91. Nelson N, Perzov N, Cohen A, Hagai K, Padler V, Nelson H. The cellular biology of proton-motive force generation by V-ATPases. *J Exp Biol*. 203:89–952000;

92. Zhu H, Yuan Q, Briscoe AD, Froy O, Casselman A, Reppert SM. The two CRYs of the butterfly. *Curr Biol*. 15:R953–42005;

93. Yuan Q, Metterville D, Briscoe AD, Reppert SM. Insect cryptochromes: gene duplication and loss define diverse ways to construct insect circadian clocks. *Mol Biol Evol*. 24:948–552007;

94. Cortés T, Ortiz-Rivas B, Martínez-Torres D. Identification and characterization of circadian clock genes in the pea aphid Acyrthosiphon pisum. *Insect Mol Biol*. 19 Suppl 2:123–392010;

95. Porter ML, Blasic JR, Bok MJ, Cameron EG, Pringle T, Cronin TW, et al.. Shedding new light on opsin evolution. *Proc Biol Sci*. 279:3–142012;

96. Terakita A. The opsins. *Genome Biol*. 6:2132005;

97. Whyard S, Erdelyan CNG, Partridge AL, Singh AD, Beebe NW, Capina R. Silencing the buzz: a new approach to population suppression of mosquitoes by feeding larvae double-stranded RNAs. *Parasit Vectors*. 8:962015;

98. Darrington M, Dalmay T, Morrison NI, Chapman T. Implementing the sterile insect technique with RNA interference - a review. *Entomol Exp Appl*. 164:155–752017;

99. Dong Y-C, Wang Z-J, Chen Z-Z, Clarke AR, Niu C-Y. Bactrocera dorsalis male sterilization by targeted RNA interference of spermatogenesis: empowering sterile insect technique programs. *Sci Rep*. 6:357502016;

100. Cruz C, Tayler A, Whyard S. RNA Interference-Mediated Knockdown of Male Fertility Genes in the Queensland Fruit Fly Bactrocera tryoni (Diptera: Tephritidae). *Insects*. 2018; doi: 10.3390/insects9030096.

101. Ali MW, Zheng W, Sohail S, Li Q, Zheng W, Zhang H. A genetically enhanced sterile insect technique against the fruit fly, Bactrocera dorsalis (Hendel) by feeding adult double-stranded RNAs. *Sci Rep*. 7:40632017;

102. Qureshi JA, Stansly PA. Asian Citrus Psyllid: Biology, Ecology and Management of the Huanglongbing Vector. CABI;

103. Brown SJ, Coleman M. Isolation of High Molecular Weight DNA from Insects. *Methods Mol Biol*. 1858:27–322019;

104. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 27:722–362017;

105. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*. 72:65–752015;

106. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al.. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9:e1129632014;

107. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al.. Integrative genomics viewer. *Nat Biotechnol*. 29:24–62011;

108. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*. 27:578–92011;

109. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics.

110. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 16:1572015;

111. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics.

112. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al.. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 40:e492012;

113. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Res*. 2020; doi: 10.12688/f1000research.23297.2.

114. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.. BLAST+: architecture and applications. *BMC Bioinformatics*. 10:4212009;

115. Bandi V, Gutwin C. Interactive Exploration of Genomic Conservation.

116. : Software downloads. PacBio. https://www.pacb.com/support/software-downloads/ (2015). Accessed 2022 Dec 22.

117. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 29:644–522011;

118. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al.. The Pfam protein families database. *Nucleic Acids Res*. 38:D211–222010;

119. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al.. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8:1494–5122013;

120. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22:1658–92006;

121. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 215:403–101990;

122. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 47:D506–152019;

123. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 21:1859–752005;

124. Tseng E. Cupcake ToFU: supporting scripts for Iso-Seq after clustering step.

125. Smit, AFA, Hubley, R & Green, P.: RepeatMasker Open-4.0. http://www.repeatmasker.org (2013-2015).

126. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al.. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol*. 164:513–242014;

127. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al.. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*. 18:188–962008;

128. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 37:907–152019;

129. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 33:290–52015;

130. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience*. 2018; doi: 10.1093/gigascience/giy093.

131. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*. 7:622006;

132. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 5:592004;

133. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. *Methods Mol Biol*. 1962:65–952019;

134. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al.. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 158:1431–432014;

135. Weirauch MT, Hughes TR. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. *Subcell Biochem*. 52:25–732011;

136. Eddy SR. A new generation of homology search tools based on probabilistic inference.

*Genome Inform*. 23:205–112009;

137. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al.. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 7:5392011;

138. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 9:5592008;

139. Attardo GM, Abd-Alla AMM, Acosta-Serrano A, Allen JE, Bateta R, Benoit JB, et al.. Comparative genomic analysis of six Glossina genomes, vectors of African trypanosomes. *Genome Biol*. 20:1872019;

140. Scott MJ, Benoit JB, Davis RJ, Bailey ST, Varga V, Martinson EO, et al.. Genomic analyses of a livestock pest, the New World screwworm, find potential targets for genetic control programs. *Commun Biol*. 3:4242020;

141. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al.. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol*. 17:662016;

142. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 6:312005;

143. Shippy T, Miller S, Massimino C, Vosburg C, Hosmani PS, Flores-Gonzalez M, et al.. Annotating genes in Diaphorina citri genome version 3 v1 (protocols.io.bniimcce). protocols.io.

144. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 20:2382019;

145. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al.. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 30:1236–402014;

Supplementary Information:

S0: Github Repo https://github.com/s-hoyt/endosymbiont-analysis
Contains scripts, prokka changes logs, unique protein names

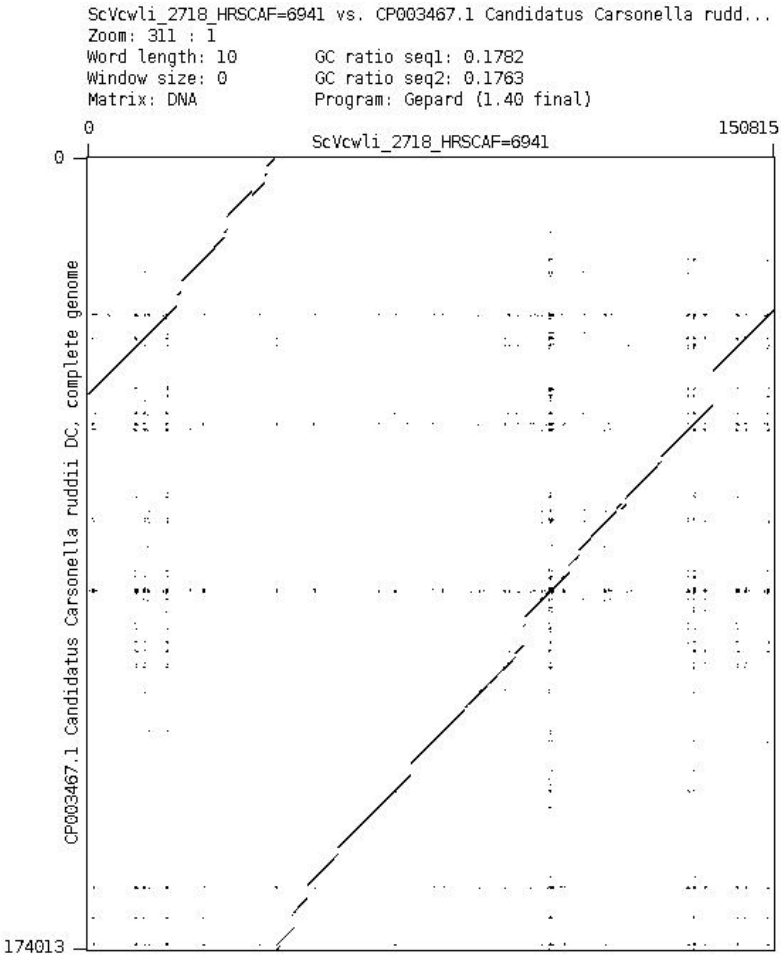
S1: Genomes in Custom Database

| **Wolbachia Strains** | **Accession Number** |
|---|---|
| Wolbachia endosymbiont of Drosophila melanogaster (a-proteobacteria) | ASM802v1 |
| Wolbachia endosymbiont strain TRS of Brugia malayi (a-proteobacteria) | ASM838v1 |
| Wolbachia endosymbiont of Onchocerca ochengi (a-proteobacteria) | ASM30688v1 |
| Wolbachia endosymbiont of Drosophila simulans wNo (a-proteobacteria) | ASM37658v1 |
| Wolbachia endosymbiont of Cimex lectularius (a-proteobacteria) | ASM82931v1 |
| Wolbachia sp. wRi (a-proteobacteria) | ASM2228v1 |
| Wolbachia endosymbiont of Drosophila simulans wHa (a-proteobacteria) | ASM37660v1 |
| Wolbachia endosymbiont of Onchocerca volvulus str. Cameroon (a-proteobacteria) | W_O_volvulus_Cameroon_v3 |
| Wolbachia endosymbiont of Drosophila simulans wAu (a-proteobacteria) | Wau001 |
| Wolbachia endosymbiont of Diaphorina citri | wACP3 |
| **Carsonella Strains** | **Accession Number** |
| Candidatus Carsonella ruddii PV (g-proteobacteria) | ASM1036v1 |
| Candidatus Carsonella ruddii CE isolate Thao2000 (g-proteobacteria) | ASM28723v1 |
| Candidatus Carsonella ruddii CS isolate Thao2000 (g-proteobacteria) | ASM28725v1 |
| Candidatus Carsonella ruddii HC isolate Thao2000 | ASM28727v1 |

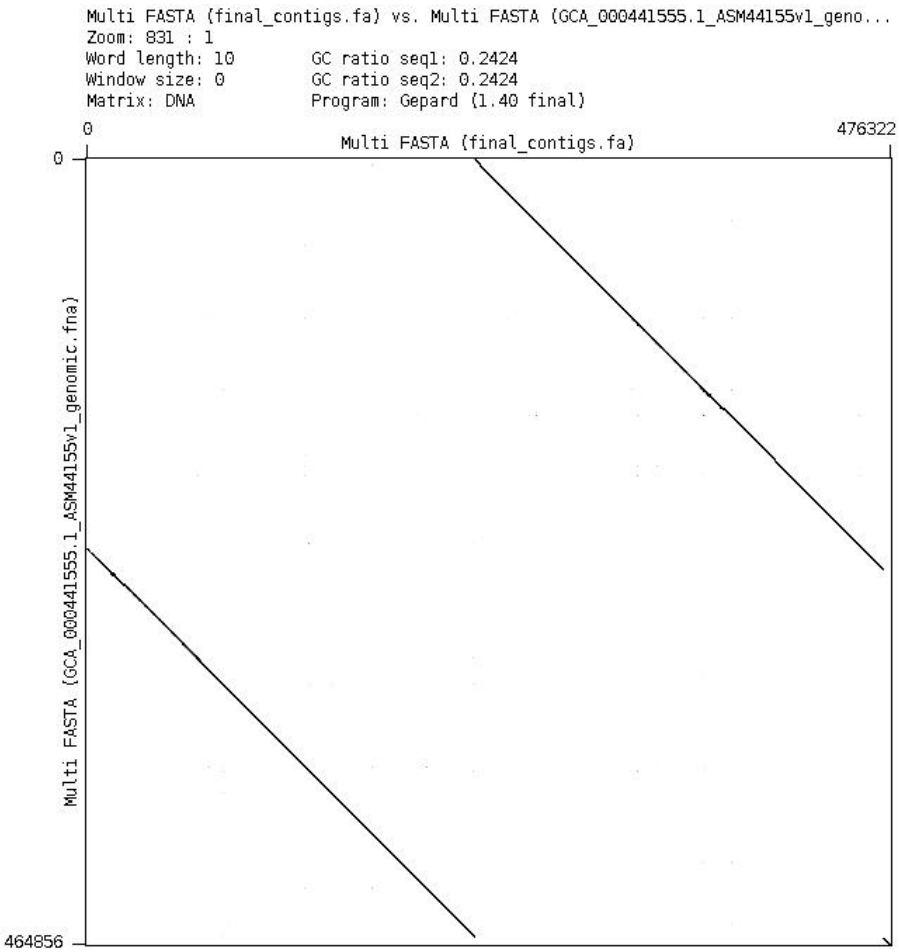
| | |
|---|---|
| (g-proteobacteria) | |
| Candidatus Carsonella ruddii HT isolate Thao2000 (g-proteobacteria) | ASM28729v1 |
| Candidatus Carsonella ruddii PC isolate NHV (g-proteobacteria) | ASM28731v1 |
| Candidatus Carsonella ruddii DC (g-proteobacteria) | ASM44157v1 |
| Candidatus Carsonella ruddii (g-proteobacteria) [YCCR] | ASM127451v1 |
| Candidatus Carsonella ruddii (g-proteobacteria) [BC] | ASM200935v1 |
| **Profftella Strains** | **Accession Number** |
| Candidatus Profftella armatura (b-proteobacteria) | ASM44155v1 |
| Candidatus Profftella armatura (b-proteobacteria) [YCPA] | ASM127878v1 |

S2: Dot Plots and Blast results showing Alignment between our genome assemblies and published genomes

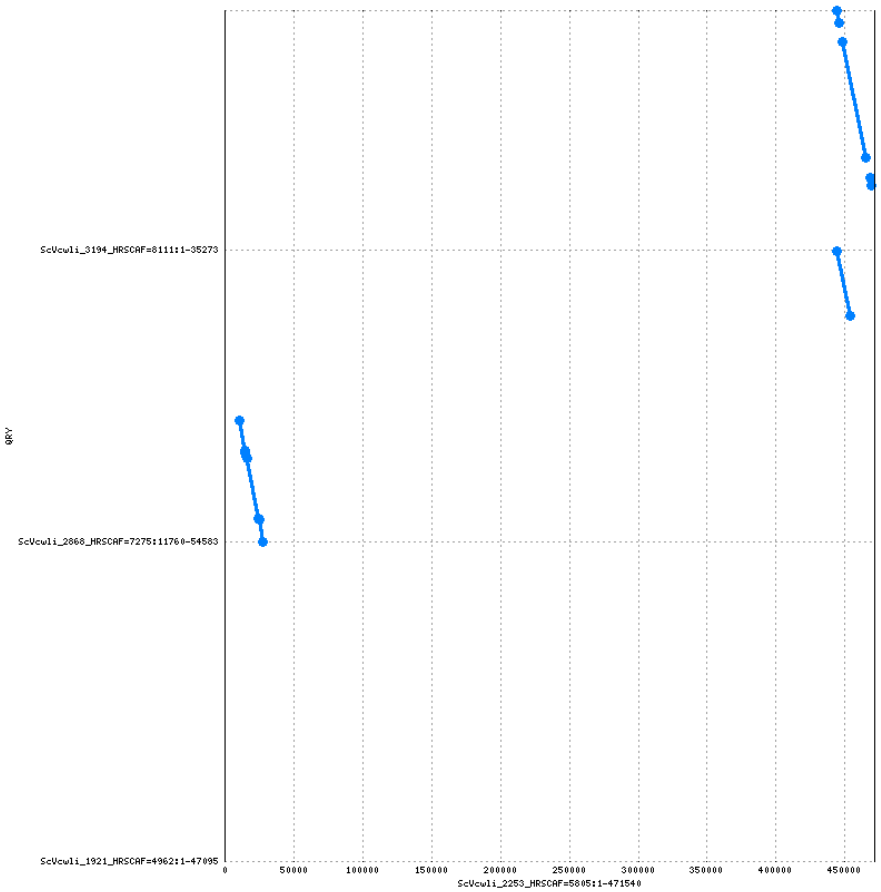
A: Carsonella aligned to Candidatus Carsonella ruddii DC




B: Proftella aligned to Candidatus Profftella armatura DC

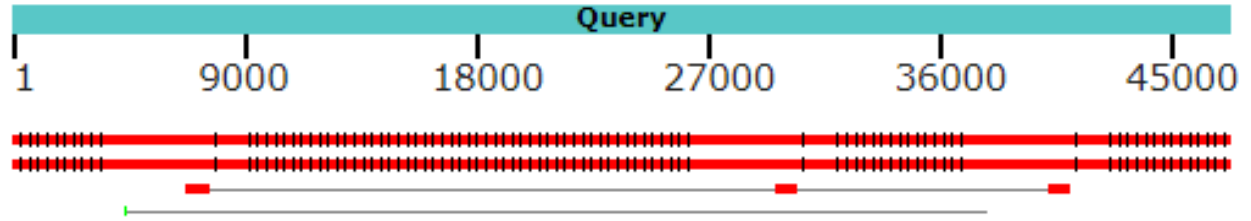



C: 471Kb Profftella chromosome to other three isolated contigs

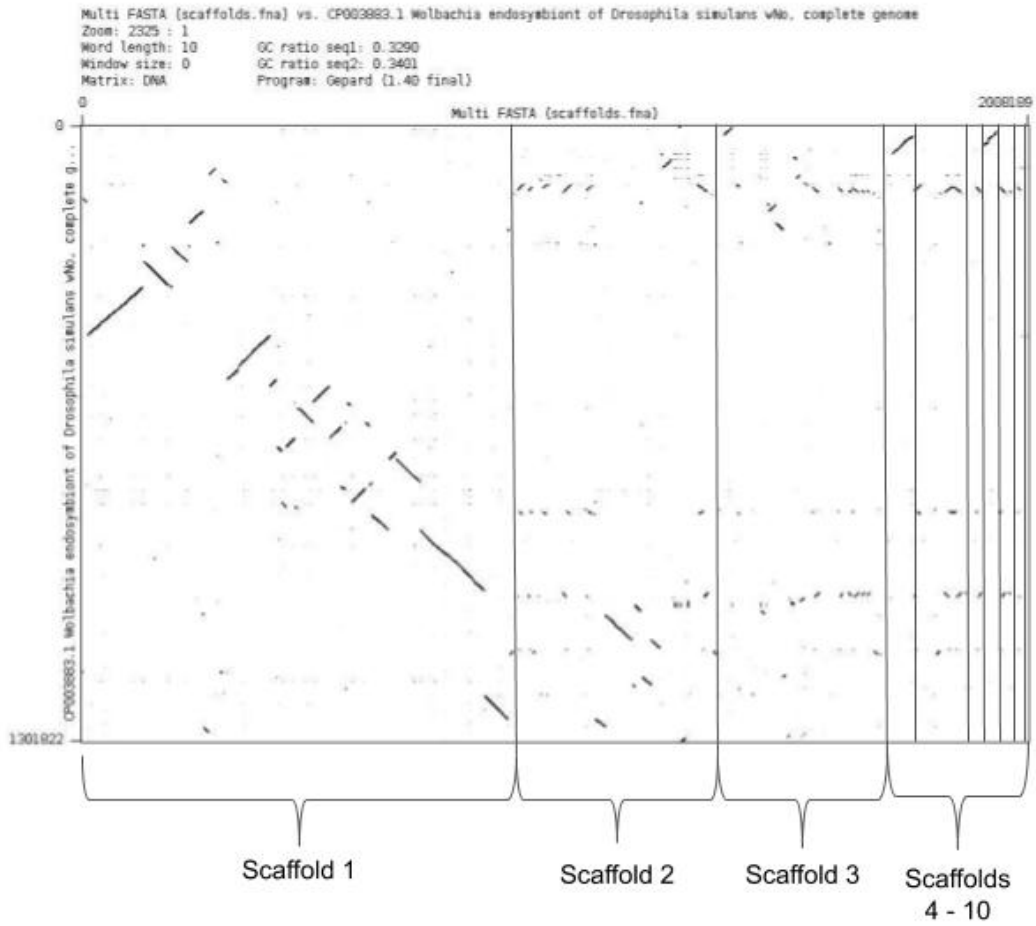



D: Profftella 47kB contig BLAST results

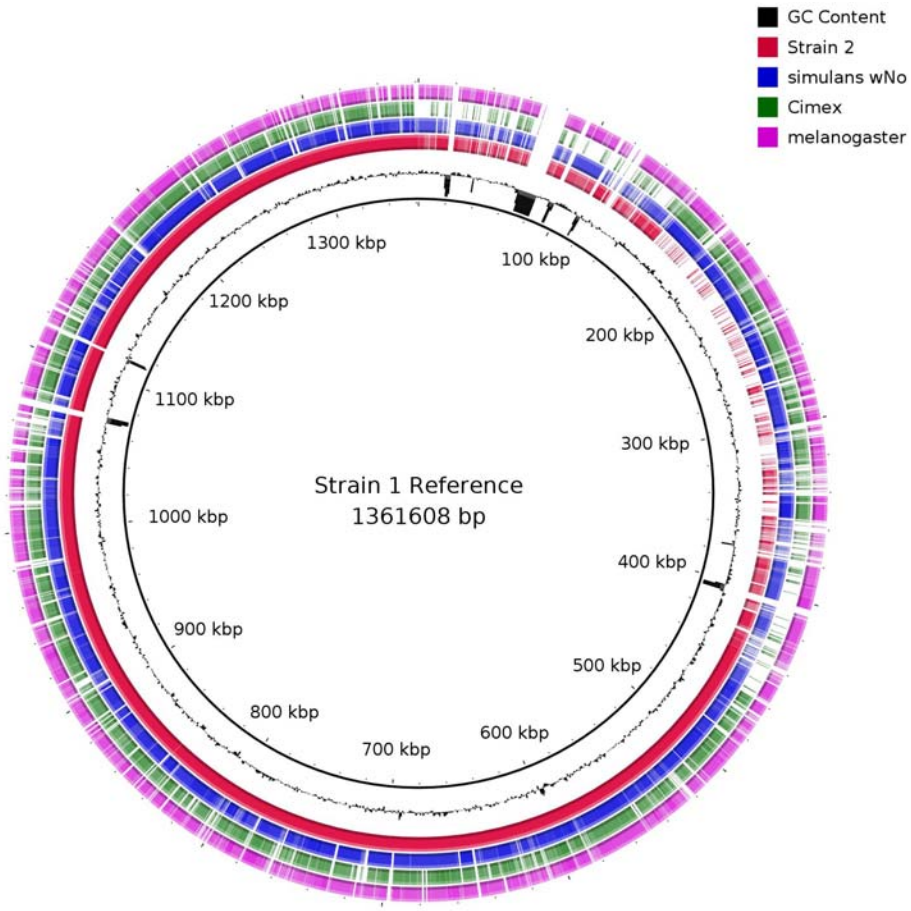
| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| Candidatus Profftella armatura strain YCPA plasmid, complete sequence | 8737 | 1.394e+05 | 99% | 0.0 | 99% | CP012592.1 |
| Candidatus Profftella armatura plasmid, complete sequence | 8737 | 1.394e+05 | 99% | 0.0 | 99% | CP003469.1 |
| PREDICTED: Diaphorina citri all-trans-phytoene synthase/15-cis-phytoene synthase-like (LOC103519589), partial mRNA | 1679 | 5039 | 5% | 0.0 | 100% | XM_017448245.1 |
| Schistocephalus solidus genome assembly S. solidus, NST_G2, scaffold SSLN_scaffold0001257 | 52.8 | 158 | 0% | 1.0 | 100% | LL901460.1 |




E: Wolbachia scaffolds aligned to Wolbachia endosymbiont of Drosophila simulans wNo


Multi FASTA (scaffolds.fna) vs. CP003883.1 Wolbachia endosymbiont of Drosophila simulans wNo, complete genome
Zoom: 2325 : 1
Word length: 10
Window size: 0
Matrix: DNA
GC ratio seq1: 0.3290
GC ratio seq2: 0.3401
Program: Gepard (1.40 final)
0
2008189
Multi FASTA (scaffolds.fna)
0
1301922
CP003883.1 Wolbachia endosymbiont of Drosophila simulans wNo, complete g...
Scaffold 1
Scaffold 2
Scaffold 3
Scaffolds 4 - 10

Supplemental Figure 1: A comparison between the two strains of *Wolbachia* and three published genomes.




Supplemental Figure 1 Legend: Regions of dark red indicate where strain 1 and strain 2 share the same sequence; where there is more white indicates gaps where strains 1 and 2 differ. Regions of high GC content indicate regions of the reference genome that do not align to the other genomes at all. melanogaster = *Drosophila melanogaster Wolbachia* symbiont, simulans wNo = *Drosophila simulans Wolbachia* symbiont, cimex = *Wolbachia* symbiont of *Cimex lectularis*. Accession numbers and references for all *Wolbachia*, *Carsonella* and *Profftella* genomes used for comparison to *D. citri* endosymbionts are shown in the table.

| **Wolbachia Strains** | **Accession Number** | **Reference** |
|---|---|---|
| Wolbachia endosymbiont of Drosophila melanogaster (a-proteobacteria) | ASM802v1 | (Wu et al. 2004) |

| | | |
|---|---|---|
| Wolbachia endosymbiont strain TRS of Brugia malayi (a-proteobacteria) | ASM838v1 | (Foster et al. 2005) |
| Wolbachia endosymbiont of Onchocerca ochengi (a-proteobacteria) | ASM30688v1 | (Darby et al. 2012) |
| Wolbachia endosymbiont of Drosophila simulans wNo (a-proteobacteria) | ASM37658v1 | (Ellegaard et al. 2013) |
| Wolbachia endosymbiont of Cimex lectularius (a-proteobacteria) | ASM82931v1 | (Nikoh et al. 2014) |
| Wolbachia sp. wRi (a-proteobacteria) | ASM2228v1 | (Klasson et al. 2009) |
| Wolbachia endosymbiont of Drosophila simulans wHa (a-proteobacteria) | ASM37660v1 | (Ellegaard et al. 2013) |
| Wolbachia endosymbiont of Onchocerca volvulus str. Cameroon (a-proteobacteria) | W_O_volvulus_Cameroon_v3 | No publication listed* |
| Wolbachia endosymbiont of Drosophila simulans wAu (a-proteobacteria) | Wau001 | (Sutton et al. 2014)** |
| Wolbachia endosymbiont of Diaphorina citri | wACP3 | (Saha et al. 2012)^ |
| **Carsonella Strains** | **Accession Number** | |
| Candidatus Carsonella ruddii PV (g-proteobacteria) | ASM1036v1 | (Nakabachi et al. 2006) |
| Candidatus Carsonella ruddii CE isolate Thao2000 (g-proteobacteria) | ASM28723v1 | (Sloan and Moran 2012) |
| Candidatus Carsonella ruddii CS isolate Thao2000 (g-proteobacteria) | ASM28725v1 | |
| Candidatus Carsonella ruddii HC isolate Thao2000 (g-proteobacteria) | ASM28727v1 | |
| Candidatus Carsonella ruddii HT isolate Thao2000 (g-proteobacteria) | ASM28729v1 | |

| | | |
|---|---|---|
| Candidatus Carsonella ruddii PC isolate NHV (g-proteobacteria) | ASM28731v1 | |
| Candidatus Carsonella ruddii DC (g-proteobacteria) | ASM44157v1 | (Nakabachi et al. 2013) |
| Candidatus Carsonella ruddii (g-proteobacteria) [YCCR] | ASM127451v1 | (Wu et al. 2015) |
| Candidatus Carsonella ruddii (g-proteobacteria) [BC] | ASM200935v1 | No publication listed^^ |
| **Profftella Strains** | **Accession Number** | |
| Candidatus Profftella armatura (b-proteobacteria) | ASM44155v1 | (Nakabachi et al. 2013) |
| Candidatus Profftella armatura (b-proteobacteria) [YCPA] | ASM127878v1 | (Wu et al. 2015) |

*No publication found, but can find the data using this assembly ID and BioProject PRJEB4840 or GenBank GCA_000530755.1

**Can find the data using this assembly ID and BioProject PRJEB6321 or GenBank GCA_000953315.1

^Can find the data using this assembly ID. Could use BioProject PRJNA29451 or GenBank GCA_000331595.1

^^No publication or accession found

Supplemental Table 1: RNA-seq data used for genome annotation and de novo transcriptome assembly.

| SRA | Data | Reference |
|---|---|---|
| SRR1259429 | Citrus spp. CLas- Whole body Adult | [1] |
| SRR1259432 | Citrus spp. CLas+ Whole body Adult | |
| SRR1259461 | Citrus spp. CLas- Whole body Nymph | |
| SRR1259434 | Citrus spp. CLas+ Whole body Nymph | |
| SRR2632316 | C. reticulata CLas- Male antennae Adult | [2] |
| SRR2632319 | C. reticulata CLas- Female antennae Adult | |
| SRR2632320 | C. reticulata CLas- Male terminal abdomen Adult | |
| SRR2632321 | C. reticulata CLas- Female terminal abdomen Adult | |
| SRR602249 | C. macrophylla CLas- Whole body Adult | [3] |
| SRR610529 | C. macrophylla CLas- Whole body Nymph | |
| SRR610530 | C. macrophylla CLas- Whole body Egg | |
| SRR5514656 | ACP-Gut-Healthy-rep1a | [4] |
| SRR5514657 | ACP-Gut-Healthy-rep1b | |
| SRR5514655 | ACP-Gut-Clas-rep4b | |
| SRR5514654 | ACP-Gut-Clas-rep4a | |
| SRR5514651 | ACP-Gut-Clas-rep2b | |
| SRR5514649 | ACP-Gut-Healthy-rep4b | |
| SRR5514643 | ACP-Gut-Clas-rep1b | |

| | | |
|---|---|---|
| SRR5514642 | ACP-Gut-Clas-rep1a | |
| SRR5514653 | ACP-Gut-Clas-rep3b | |
| SRR5514652 | ACP-Gut-Clas-rep3a | |
| SRR5514650 | ACP-Gut-Clas-rep2a | |
| SRR5514648 | ACP-Gut-Healthy-rep4a | |
| SRR5514647 | ACP-Gut-Healthy-rep3b | |
| SRR5514646 | ACP-Gut-Healthy-rep3a | |
| SRR5514645 | ACP-Gut-Healthy-rep2b | |
| SRR5514644 | ACP-Gut-Healthy-rep2a | |

References

1. Vyas M, Fisher TW, He R, Nelson W, Yin G, Cicero JM, et al.. Asian Citrus Psyllid Expression Profiles Suggest Candidatus Liberibacter Asiaticus-Mediated Alteration of Adult Nutrition and Metabolism, and of Nymphal Development and Immunity. *PLoS One*. 2015; doi: 10.1371/journal.pone.0130328.

2. Wu Z, Zhang H, Bin S, Chen L, Han Q, Lin J. Antennal and Abdominal Transcriptomes Reveal Chemosensory Genes in the Asian Citrus Psyllid, Diaphorina citri. *PLoS One*. 2016; doi: 10.1371/journal.pone.0159372.

3. Reese J, Christenson MK, Leng N, Saha S, Cantarel B, Lindeberg M, et al.. Characterization of the Asian Citrus Psyllid Transcriptome. *J Genomics*. 2014; doi: 10.7150/jgen.7692.

4. Kruse A, Fattah-Hosseini S, Saha S, Johnson R, Warwick E, Sturgeon K, et al.. Combining ’omics and microscopy to visualize interactions between the Asian citrus psyllid vector and the Huanglongbing pathogen Candidatus Liberibacter asiaticus in the insect gut. *PLoS One*. 2017; doi: 10.1371/journal.pone.0179531.

Supplemental Table 2: BUSCO statistics.

| | **BUSCO Arthropoda 1013 Gene Set (%)** | | | |
|---|---|---|---|---|
| | **Complete** | **Duplicated** | **Fragmented** | **Missing** |
| **Diaci v3.0** | 93.5 | 26.4 | 3.3 | 3.2 |
| **Diaci v2.0** | 92.1 | 36.8 | 4.6 | 3.3 |
| **Diaci v1.1** | 82.6 | 4.6 | 11.4 | 6.0 |
| ***de novo* transcriptome** | 88 | 30.8 | 3.5 | 8.5 |

| | **BUSCO Hemiptera 2510 Gene Set (%)** | | | |
|---|---|---|---|---|
| | **Complete** | **Duplicated** | **Fragmented** | **Missing** |
| **Diaci v3.0** | 93.8 | 33.6 | 3.3 | 2.9 |
| **Diaci v2.0** | 93.8 | 33.6 | 3.3 | 2.9 |
| **Diaci v1.1** | 88.9 | 4.7 | 7.0 | 4.1 |
| ***de novo* transcriptome** | 94.2 | 24 | 2.6 | 3.2 |

Supplemental Table 2 legend: Percentages of the Arthropoda (arthropoda_odb10) and Hemiptera (hemiptera_odb10) BUSCO gene set that are complete, duplicated, fragmented or missing in Florida *Diaphorina. citri* genome versions, official gene sets and transcriptomes.