*Database*, 2017, 1–20
doi: 10.1093/database/bax032
Original article

Original article

# Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community

**Surya Saha,[α,β,1,*] Prashant S. Hosmani,[α,β,1,*] Krystal Villalobos-Ayala,[α,β,2] Sherry Miller,[α,β,3] Teresa Shippy,[α,β,3] Mirella Flores,[α,β,1] Andrew Rosendale,[α,β,4] Chris Cordola,[β,2] Tracey Bell,[β,2] Hannah Mann,[β,2] Gabe DeAvila,[β,2]Daniel DeAvila,[β,2] Zachary Moore,[β,4] Kyle Buller,[β,4] Kathryn Ciolkevich,[β,4] Samantha Nandyal,[β,4] Robert Mahoney,[β,4] Joshua Van Voorhis,[β,4] Megan Dunlevy,[β,4] David Farrow,[β,4] David Hunter,[β,3] Taylar Morgan,[β,3] Kayla Shore,[β,3] Victoria Guzman,[β,3] Allison Izsak,[β,5] Danielle E. Dixon,[β,1,6] Andrew Cridge,[β,7] Liliana Cano,[β,7] Xiaolong Cao,[ρ,14] Haobo Jiang,[ρ,15] Nan Leng,[ϑ,13] Shannon Johnson,[ϑ,9] Brandi L. Cantarel,[ϑ,10] Stephen Richards,[ϑ,11,12] Adam English,[ϑ,11,12] Robert G. Shatters,[ϑ,γ,16] Chris Childers,[α,16] Mei-Ju Chen,[α,17] Wayne Hunter,[ϑ,β,γ,18] Michelle Cilia,[γ,19,20] Lukas A. Mueller,[γ,1,21] Monica Munoz-Torres,[α,β,22] David Nelson,[β,23] Monica F. Poelchau,[α,16] Joshua B. Benoit,[β,γ,4] Helen Wiersma-Koch,[α,β,2] Tom D'Elia[β,γ,2] and Susan J. Brown[β,γ,3]**

[1]Boyce Thompson Institute, Ithaca, NY 14853, [2]Indian River State College, Fort Pierce, FL 34981, [3]Division of Biology, Kansas State University, Manhattan, KS 66506, [4]University of Cincinnati, Cincinnati, OH 45220, [5]Cornell University, Ithaca, NY 14853, [6]University of Puget Sound, Tacoma, WA 98416, USA, [7]University of Otago, North Dunedin, Dunedin 9016, New Zealand, [8]Plant Pathology, University of Florida/IFAS Indian River Research and Education Center, Ft. Pierce, FL 34945, [9]Department of Biochemistry and Molecular Biology, [10]Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74074, [11]Illumina Inc., San Diego, CA 92122, [12]Los Alamos National Laboratory, Los Alamos, NM 87544, [13]Department of Bioinformatics, UT Southwestern Medical Center, Bioinformatics Core Facility, Dallas, TX 75390, [14]i5K Arthropod Genomics, [15]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, [16]USDA ARS, U.S. Horticultural Research Laboratory, Ft. Pierce, FL 34945, [17]USDA Agricultural Research Service, National Agricultural Library, Beltsville, MD 20705, USA, [18]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan, [19]USDA ARS, Emerging Pests and Pathogens Research Unit, Ithaca, NY 14853, [20]Plant Pathology and Plant-Microbe Biology Section, [21]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, [22]Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology, Berkeley, CA 94720 and [23]Department of Microbiology, Immunology and Biochemistry, The University of Tennessee Health Science Center, Memphis, TN 38163, USA

*Corresponding author: Tel.: +(607) 254-5480; Fax: +(607) 254-1242; Email: ss2489@cornell.edu

Roles: $^{\alpha}$Key personnel; $^{\beta}$Annotator; $^{\vartheta}$Genome Sequencing; $^{\rho}$Transcriptome; $^{\gamma}$Principal Investigator

[†]These authors contributed equally to this work.

Saha,S., Hosmani,P.S., Villalobos-Ayala,K. *et al.* Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community. *Database* (2017) Vol. 2017: article ID bax032; doi:10.1093/database/bax032

Received 5 January 2017; Revised 3 March 2017; Accepted 25 March 2017

## Abstract

The Asian citrus psyllid (*Diaphorina citri* Kuwayama) is the insect vector of the bacterium *Candidatus* Liberibacter asiaticus (CLas), the pathogen associated with citrus Huanglongbing (HLB, citrus greening). HLB threatens citrus production worldwide. Suppression or reduction of the insect vector using chemical insecticides has been the primary method to inhibit the spread of citrus greening disease. Accurate structural and functional annotation of the Asian citrus psyllid genome, as well as a clear understanding of the interactions between the insect and CLas, are required for development of new molecular-based HLB control methods. A draft assembly of the *D. citri* genome has been generated and annotated with automated pipelines. However, knowledge transfer from well-curated reference genomes such as that of *Drosophila melanogaster* to newly sequenced ones is challenging due to the complexity and diversity of insect genomes. To identify and improve gene models as potential targets for pest control, we manually curated several gene families with a focus on genes that have key functional roles in *D. citri* biology and CLas interactions. This community effort produced 530 manually curated gene models across developmental, physiological, RNAi regulatory and immunity-related pathways. As previously shown in the pea aphid, RNAi machinery genes putatively involved in the microRNA pathway have been specifically duplicated. A comprehensive transcriptome enabled us to identify a number of gene families that are either missing or misassembled in the draft genome. In order to develop biocuration as a training experience, we included undergraduate and graduate students from multiple institutions, as well as experienced annotators from the insect genomics research community. The resulting gene set (OGS v1.0) combines both automatically predicted and manually curated gene models.

**Database URL**: https://citrusgreening.org/

## Introduction

The Asian citrus psyllid (ACP), *Diaphorina citri* Kuwayama (Hemiptera:Liviidae), is a phloem-feeding insect native to Southeastern and Southwestern Asia with a host range limited to plants in the citrus genus and related Rutaceae spp. (1). Accidental anthropogenic introductions of psyllid-infested citrus combined with the ability of psyllids to disperse rapidly have allowed *D. citri* to extend its distribution to most of southern and eastern Asia, the Arabian Peninsula, the Caribbean, and South, Central and North America (1–6). For years, ACP has been classified as a global pest that is capable of devastating citrus crops through transmission of the bacterial agent, *Candidatus* Liberibacter asiaticus, CLas, which is associated with Huanglongbing (HLB) or citrus greening disease. The psyllid alone has little economic importance and causes only minor plant damage while feeding (7, 8).

HLB is the most destructive and economically important disease of citrus, with practically all commercial citrus species and cultivars susceptible to CLas infection (9). Infected trees yield premature, bitter and misshapen fruit that is unmarketable. In addition, tree death follows 5–10 years after initial infection (2, 9, 10). Furthermore, HLB drastically suppresses economic progress in southern and eastern Asia by impeding viable commercial citrus agriculture within those regions (11). Florida is one of the top citrus-producing regions in the world and the largest in the USA, with nearly double the output of California, the second largest citrus-producing state (12). HLB has severely impacted the 8.91 billion dollar Florida citrus industry with 23% reduction in yield, $1.7 billion in lost revenue and the loss of 8,257 jobs from 2006 to 2011 (13). In 2008, the HLB infection rate within central Florida was low (1.4% to 3.6%), but reaching 100% in the southern and eastern portions of the state (14, 15). In 2005, when HLB was first detected in Florida, 9.3 million tons of oranges were harvested, but production has declined steadily to 5.3 million tons in 2016 as ACP and HLB have spread (12).

Primary management strategies focus on disrupting the HLB transmission pathway by suppressing psyllid populations and impeding interactions between CLas and psyllids. These strategies currently rely on extensive chemical application, which has broad environmental impact and high costs, and are ultimately unsustainable. To develop molecular methods that exploit current gene-targeting technologies, detailed genetic and genomic knowledge, including a high-quality official gene set (OGS), is required (16, 17). Early efforts focused on *D. citri* transcript expression (3, 18–20), analysis of the full transcriptome (21, 22) and, more recently, analysis of the *D. citri* proteome (16). Arp et al. (23) performed a BLAST-based inventory of NCBI-predicted immune genes in *D. citri* (v100, see Materials and methods). In contrast, we have conducted broad structural and functional annotation with the aid of a comprehensive transcriptome and created an OGS for *D. citri* with a focus upon completing the repertoire of immune genes.

Manual curation improves the quality of gene annotation and establishing a 'version controlled' OGS provides a set of high quality, well-documented genes for the entire research community. Although ACP is a significant agricultural pest, it is not a model organism and the size of the research community does not warrant 'museum' or 'jamboree' annotation strategies (24). To maximize the number of genes annotated in a relatively short time, we augmented the 'cottage industry' strategy (25) by training undergraduates to perform basic annotation tasks. The dispersed ACP annotation community agreed on a set of standard operating procedures and defined a set of primary gene targets. Starting with automated gene predictions, we used several additional types of evidence including RNAseq and proteomics data including comparisons to other insects to generate the first official gene set (OGS v1.0). Using an independent transcriptome enabled us to identify a number of gene families that are either missing or misassembled in the draft genome.

Our manual curation efforts focused on genes of potential use in vector control including immunity-related genes and pathways, RNAi machinery genes, multiple clans of cytochrome P450 (CYP) genes and other genes relevant to insect development and physiology. We speculate that targeted analysis of these genes in *D. citri* will provide the foundation for a better understanding of the interactions between psyllid host and CLas pathogen, and will open the possibilities for research that can eventually find solutions to manage the dispersion of this very destructive pest and HLB.

## Materials and methods

### DNA extraction and library preparation

High-molecular weight DNA was extracted using the BioRad AquaPure Genomic DNA isolation kit from fresh intact *D. citri* collected from a citrus grove in Ft. Pierce, FL and reared at the USDA, ARS, US Horticultural Research Laboratory, Ft. Pierce, FL. To generate PacBio libraries, DNA was sheared using a Covaris g-Tube and SMRT-bell library was prepared using the 10 kb protocol (PacBio DNA template prep kit 2.0; 3–10 kb), cat #001-540-835.

### Genome sequencing and assembly

Samples were prepared for Illumina sequencing using the TruSeq DNA library preparation kits for paired-end as well as long-insert mate-pair libraries. Thirty-nine SMRTcells of the library were sequenced, all with 2 × 45 min movies. A total of 2 750 690 post-filter reads were generated, with an average of 70 530 reads per SMRTcell. The post-filter mean read length was 2504 bp with an error rate of 15%.

Velvet (26) was used with kmer 59 for generating the Diaci1.0 draft assembly. PacBio long reads were mapped to the draft assembly using blasr (27) with the following parameters: -minMatch 8 -minPctIdentity 70 -bestn 5 -nCandidates 30 -maxScore -500 -nproc 8 -no SplitSubreads. These alignments were parsed using PBJelly (28) with default parameters to scaffold the draft assembly and create the final Diaci1.1 reference genome (45). More details about the history of ACP genome sequencing can be found at citrusgreening.org (https://citrusgreening.org/organism/Diaphorina_citri/genome). The Diaci1.1 assembly was evaluated with BUSCO version 2 and the Hemipteran marker set with default parameters.

### Maker and NCBI annotation

The maker control files are included in Supplementary Data. BLAST 2.2.27+ was used with augustus version 2.5.5 (29) and exonerate version 2.2.0 (30). Only contigs longer than 10 kb were selected for annotation with psyllid-specific RNAseq data from Reese et al. (21). Details of ACP genome annotation by the NCBI Eukaryotic Genome Annotation pipeline are available at https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Diaphorina_citri/100/.

### MCOT transcriptome assembly

Reads were assembled with Trinity in two runs, one used reads as single-end reads, and the other used them as paired-end reads. Reads were trimmed based on fastq quality score (the -trimmomatic option was enabled and run under the default setting of Trinity). The transcripts from both runs were compiled together to create the final Trinity assembly.

Velvet-Oases (31) assemblies were performed for trimmed reads (trimmed in Trinity run, the read quality control step) from the egg, nymph and adult separately, with kmer length of 23, 25, 27 and 29 as single-end reads, and kmer length 25 as paired-end reads. The 15 outputs were combined using the Oases merge function (-long, kmer 27 and -min_trans_lgth 200) to generate the final assembly.

Reads from the egg, nymph and adult were first aligned to the genome with Tophat (32) with the insert length parameter based on each library (53, 24 and 90). The parameter -read-realign-edit-dist was set to 0 to ensure better alignment results. Gene models were generated by Cufflinks (33) with default settings, with the -frag-bias-correct and -multi-read-correct function (-b, -u) enabled to give the most accurate gene models.

Transcripts from Maker (34, 35), Cufflinks (36), Oases (31) and Trinity (37) were translated to proteins with Transdecoder version 2.0.1 (38) and only unique proteins were retained. Furthermore, protein sequences from each program (Maker, Cufflinks, Oases, Trinity) were compared using BLASTP with a special scoring matrix (matching score of non-identical amino acids set to -100 of the BLOSUM62 matrix) to proteins from other arthropod species. The best protein models from each source were selected to create the final MCOT v1.0 protein set, and the corresponding transcript set. MCOT v1.0 set has 30 562 genes and is available at ftp://ftp.citrusgreening.org/annotation/MCOT/ and Ag Data Commons (39).

MCOT v1.0 proteins were analyzed using Interproscan 5 (40) based on InterPro databases with the options -goterms to get GO terms, -iprlookup to switch on look-up of corresponding InterPro annotation and -pa option to switch on lookup of corresponding pathway annotation. We also performed BLASTP comparison (parameters: -e 0.0001 -v 200 -b 200 -m 0) of the MCOT proteins to insect proteins from Uniprot and NCBI nr databases. These BLAST results were used as input for AHRD (Automated assignment of Human Readable Descriptions) (41), to assign functional descriptions to each MCOT protein. This functional annotation was performed using a filter of bit score of > 50 and *e*-value less than e-10. Pfam domains and gene ontology terms were also assigned from the Interproscan analysis.

### Annotation edit distance

We mapped transcripts from MCOT v1.0 to the genome using GMAP (42). Out of 30 562 transcripts, 19 744 were mapped with at least 90% query coverage and 90% identity.

A genome-guided transcriptome was generated to validate all annotation sets using all available RNAseq data (Supplementary Table S4) and insect proteins (NCBI taxonomy: 'Hexapoda [6960]') from Swiss-Prot were used as sources of evidence. In total, 622 million paired-end reads were mapped to NCBI-Diaci1.1 assembly using hisat2 (43) with a mapping rate of 81.78%. Mapped files were sorted with samtools rocksort. After sorting, a genome-guided transcriptome assembly was performed using StringTie (44) and the resulting transcriptome contained 210 890 transcripts (N50: 1691 bp).

Annotation edit distance (AED) was calculated for all gene models from NCBI v100, mapped MCOT, Maker v1.1 and curated gene sets. The Maker genome annotation (v2.31.8) pipeline was used for calculating AED (35, 45).

## Results and discussion

### NCBI-Diaci 1.1 draft genome assembly

The Diaci1.1 draft genome assembly was generated using Illumina paired-end and mate-pair data with low coverage Pacbio for scaffolding and uploaded to NCBI (PRJNA251515) and Ag Data Commons (46) after filtering out bacterial contamination. Illumina sequencing was performed on the HiSeq2000 using 100 bp or longer reads. Seven libraries were sequenced, with inserts ranging from 'short' (ca. 275 bp) to 10 kb. These are available in NCBI SRA and include 99.7 million paired-end reads (NCBI SRA: SRX057205), 35.1 million 2-kb mate-pair reads (NCBI SRA: SRX057204), 30 million 5-kb mate-pair reads (NCBI SRA: SRX058250) and 30 million 10-kb mate-pair reads (NCBI SRA: SRX216330). A second round of DNA sequencing was performed with PacBio at 12× coverage (NCBI SRA: SRX218985) for scaffolding the Diaci1.0 Illumina assembly to create the Diaci1.1 version of the *D. citri* genome (146). The Illumina data were assembled with the velvet (26) assembler followed by scaffolding with Pacbio long reads using the PBJelly (28) pipeline (see Materials and methods). The *D. citri* genome has an estimated size of 400–450 Mb (19) (47). This genome has a length of 485 Mb with 19.3 Mb of N's. It contains 161 988 scaffolds with an N50 of 109.8 kb. Given the high degree of fragmentation, we performed a Benchmarking of Universal Single-Copy Orthologs (BUSCO) (48) analysis with a set of conserved single-copy markers. A BUSCO analysis identifies the proportion of known single-copy genes correctly assembled in a genome. The accuracy and resolution of this analysis is enhanced by using a set of markers specific to a phylogenetic clade so we used 3550 markers from nine insects in the Hemipteran order based on orthologous groups defined in the ORTHODB v9 (49) database. We found a significant number of these genes to be missing (35.7%, See Supplementary Table S1a) as confirmed in the curation section below.

## MCOT transcriptome

To generate a more comprehensive set of gene models, we created the *D. citri* MCOT v1.0 transcriptome assembly. The MCOT pipeline (50) merges the output of multiple gene prediction and transcriptome assembly tools by clustering similar transcripts and selecting the best predicted protein for each cluster (see Materials and methods). By supplementing genome-based transcript models with *de novo*-assembled transcripts, we hoped to obtain a more complete collection of *D. citri* transcripts. Maker v1.1 (34, 35) gene models were predicted on the Diaci1.1 genome using RNAseq data from adult, nymph and egg tissue (21). The RNAseq data sets were also used to generate a genome-based transcriptome assembly using Cufflinks (36). *De novo* transcriptome assemblies of the adult, nymph and egg RNAseq data were performed with Oases (31) and Trinity (37). These data sets are all available at ftp://ftp.citrusgreening.org/annotation/MCOT/. *Diaphorina citri* MCOT v1.0 contains 30 562 proteins and is also available at Ag Data Commons (39).

The completeness of *D. citri* MCOT v1.0 was assessed with the BUSCO version 2 (48) tool with Hemipteran specific markers as described above. *Diaphorina citri* MCOT v1.0 contains 3114/3350 (92.9%) complete BUSCO orthologs, 2239 (66.8%) of which are single copy and 875 (26.1%) appear to be duplicated. Four additional BUSCO orthologs (0.1%) are fragmented, and only 232 (7%) were not found. BUSCO analysis was also performed on the other resources used for annotation including three stage-specific *de novo*-assembled transcriptomes from egg, nymph and adult tissue (21), the NCBI-Diaci1.1 genome assembly itself and the Maker v1.1 as well as the NCBI v100 predicted gene models (Figure 1, Supplementary Tables S1a and b). The *D. citri* MCOT v1.0 proteins that could be mapped to the genome assembly were also included in this analysis. The Maker v1.1 annotation set contains 18 205 protein-coding gene models. NCBI *D. citri* Annotation Release 100 (NCBI v100, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Diaphorina_citri/100/) contains a total of 19 311 protein-coding gene models along with non-coding RNAs (776) and pseudogenes (207). As shown in Figure 1 and Supplementary Table S1b, none of these data sets proved to be as complete as *D. citri* MCOT v1.0, which contains a higher percentage of complete BUSCO orthologs and fewer fragmented or missing markers. Proteins representing transcripts assembled by Trinity (37) and Oases (31) exclusively from RNAseq data improve the completeness of MCOT v1.0 in comparison to the NCBI v100 annotation based on the fragmented and incomplete Diaci1.1 draft genome.

The functional annotation for MCOT v1.0 proteins was generated using InterproScan5, BLAST and AHRD (41), which assigned descriptions to 23 098 genes (75.6%), GO annotations to 15 314 genes (50%) and Pfam domains to 18 170 genes (59%). MCOT is available at ftp://ftp.citrusgreening.org/annotation/MCOT/ and Ag Data Commons (39).

## Manual curation workflow

Manual curation of the *D. citri* (NCBI-Diaci1.1) draft genome assembly was undertaken to improve the quality of automated annotations (Supplementary Table S2) produced by the Maker (Maker v1.1) and NCBI pipelines (NCBI v100). This community-based manual annotation was focused on immunity-related genes as targets for ACP control.

The Apollo Genome Annotation Editor hosted at i5k Workspace@NAL (https://i5k.nal.usda.gov/) was implemented for community-based curation of gene models (51). Multiple evidence tracks (Supplementary Table S3) were added to Apollo to assist in manual curation. Standard operating procedures were outlined at the beginning and refined based on feedback from annotators and availability of evidence resources. A typical workflow for manual gene curation involved selection of orthologs from related species, search of the ACP genome and MCOT v1.0 transcriptome, followed by assessment and correction of the gene models based on evidence tracks to generate the final model. Any exceptions to this general workflow are specified in the respective gene reports (Supplementary Notes 1–39). We used ImmunoDB (52) as a primary source of orthologs for curation of immune genes (Supplementary Notes 1–31). However, we also report other gene families of functional and evolutionary importance (Supplementary Notes 32–39) including aquaporins, cuticle proteins and secretory proteins. Following correction, final gene models were verified using reciprocal BLAST analysis. With this community-based curation effort, we annotated a total of 530 genes, the majority of which include genes predicted to function in immunity, development and physiology.

### AED as a measure of quality for different annotations

We employed the AED (35, 45) metric to evaluate the quality of the different annotation data sets based on evidence from expression data. AED measures congruence of a predicted gene model with the RNAseq evidence supporting it. AED scores range between 0 and 1, where an AED score of 0 denotes perfect concordance and an AED score of 1 denotes lack of any supporting evidence. The AED

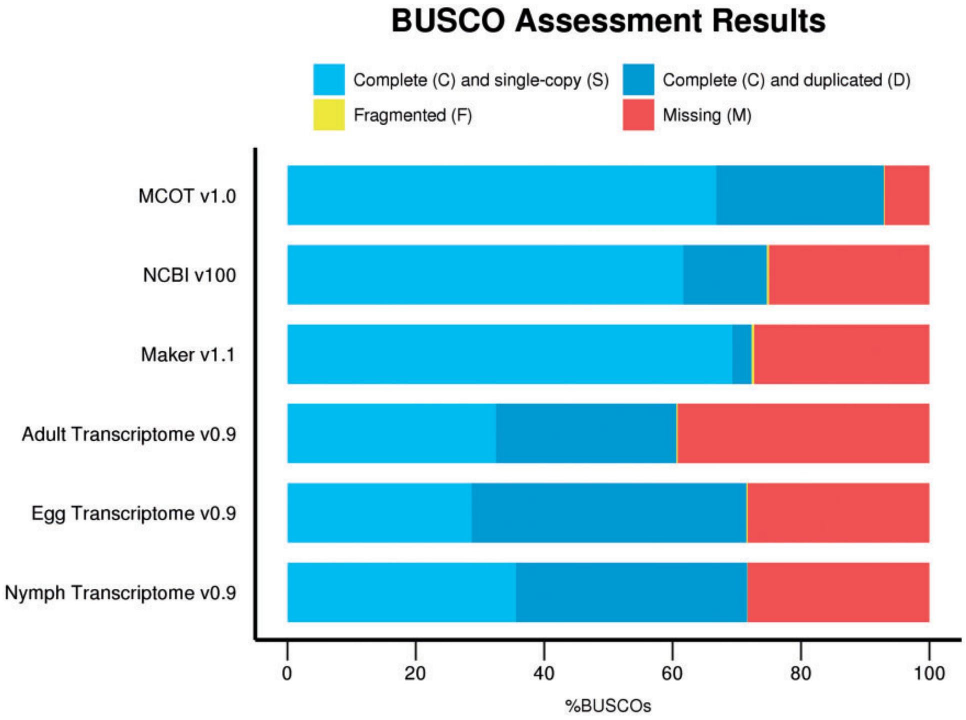

**Figure 1.** BUSCO completeness comparison for *D. citri* datasets using the Hemiptera markers based on ORTHODBv9.1 orthologs (nine species and $n = 3350$ genes). *Diaphorina citri* MCOT v1.0 has the most complete single-copy orthologs (92.9%, largest blue bar) and fewer missing orthologs (7%, smaller red bar) compared to Adult, Egg and Nymph Transcriptomes v0.9, NCBI v100 and Maker v1.1 sets.

had a 2-fold application here, selection of the best predicted gene set and quantification of improvement after manual curation. In order to generate AED scores, we compared the annotations to known insect proteins (NCBI taxonomy: 'Hexapoda [6960]') and to a comprehensive genome-guided transcriptome assembled from the latest data (see Materials and methods and Supplementary Table S4) and the NCBI-Diaci1.1 draft assembly using the Maker pipeline (34). Most of the RNAseq data used to create this transcriptome were not used in the prediction of genes by other pipelines (NCBI v100, Maker v1.1 and MCOT v1.0) as it had not yet been produced and therefore provides independent validation. RNAseq data used for building the transcriptome included adult, nymph, egg, CLas exposed and healthy (adult and nymph tissue) as well as gut-specific expression data (see Supplementary Table S4). We calculated AED scores for NCBI v100, Maker v1.1 and mapped MCOT v1.0 annotation sets.

The plot for the AED cumulative fraction of transcripts (Figure 2) shows higher expression evidence support for NCBI v100 genes compared to the Maker v1.1 and mapped MCOT v1.0 annotation sets. Therefore, the NCBI v100 annotations were selected to create the official gene set v1 (OGS v1.0). Although the MCOT v1.0 scores higher than NCBI v100 in BUSCO results (Figure 1 and Supplementary Table S1b), the mapped-MCOT set scores lower in the AED analysis as a large number of MCOT proteins could not be mapped on the draft Diaci1.1 genome. The AED plot also shows that there are many gene models in all the annotation sets that do not have any expression evidence support (AED = 1.0). This could be due either to lack of RNAseq data for some of the loci or incorrect annotations. The predicted annotations may also have been affected by misassemblies in the NCBI-Diaci1.1 draft genome.

To quantify improvements in the gene structure by manual curation, we calculated AED scores for only the curated genes (530 genes). Cumulative fractions of transcripts of AED for curated genes show improvements indicating that the intron–exon structures in the curated genes have been corrected by manual curation (Figure 2).

## Training and curation strategy

### Harnessing the crowd

It is not possible for a single individual or computer system to fully curate a genome with precise biological fidelity. A growing number of genome sequencing projects have come to fruition thanks to the combined efforts of global consortia [e.g. parasitoid wasps (53), centipede (54) and bed bug (55)]. This indicates that a centralized model of genome annotation, the design used in earlier sequencing projects for the fruit fly (56) and the human genome (57),

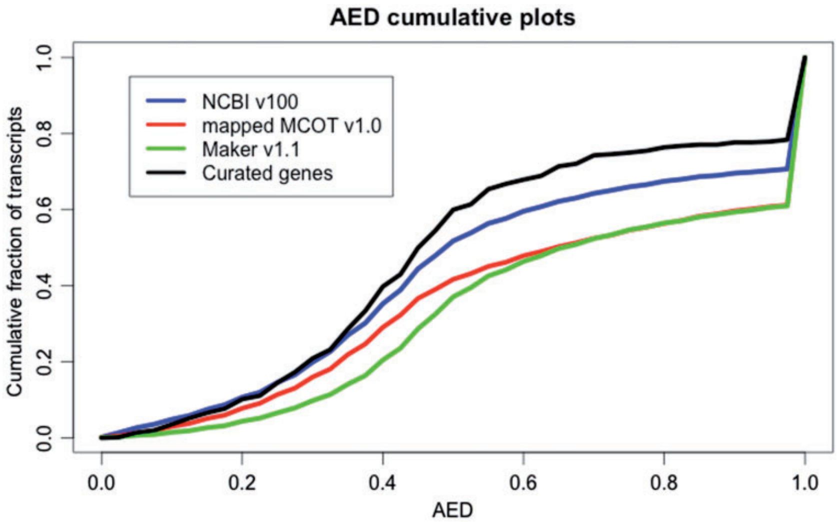

**Figure 2.** Annotation edit distance (AED) plot comparing NCBI v100 annotation (20 996), the mapped MCOT v1.0 annotation (19 744), Maker v1.1 (18 205) and curated genes (530) generated from a genome-guided transcriptome assembly based upon the NCBI-Diaci1.1 genome. AED cumulative fraction of transcripts plot shows that genes in the NCBI v100 annotation have more expression evidence support compared to MCOT v1.0 mapped set. Curated genes show better evidence support compared to any other automated prediction pipelines.

is giving way to the global and collaborative communities to generate high-quality genome annotation. Beyond the problem of scale, curators require insights from others with expertise in specific gene families, which makes the process of curation inherently collaborative. Mobilizing groups of researchers to focus on these specific and manageable areas is more likely to distill the most pertinent and valuable knowledge from genome analysis.

We formed a team of 36 curators by enlisting collaborators primarily distributed across seven academic institutions (Indian River State College, Cornell University/Boyce Thompson Institute, Kansas State University, University of Cincinnati, University of Florida, University of Otago and University of Tennessee Health Center), including undergraduate and graduate students, postdocs, staff researchers and faculty. The i5k Workspace@NAL (51) was selected as the platform for collaborative gene model curation as it is used widely by expert annotators from the insect genomics community. Training was provided at in-person workshop sessions and video conferences. The initial training workshop was structured to provide a review of principles associated with identification of specific genes by comparison to orthologs, gene structural aspects, how to search the genome for targets of interest, and using the combination of gene predictions and additional evidence tracks (Supplementary Table S3) to correct gene models with the Apollo platform. Finally, we demonstrated secondary analyses, such as BLAST comparison to other insects and phylogenetic analyses, to the annotators to examine specific genes or gene sets. The training materials used for the workshop are included in the Supplementary Data and also available online (58–60).

After initial in-person training, we continued to organize annotation via bi-weekly video conferences, documents on Google Drive and an online project management website (https://basecamp.com/2923904/projects/9184795). The project management website was used for coordinating all annotation activities and storage of working documents as well as all presentations and tutorials. The entire annotation workflow is shown in Figure 3 and a detailed training tutorial for the *D. citri* genome is included in the Supplementary Data. The online forum facilitated discussions outside of the video conferences and allowed annotators to interact at their convenience. We established standard operating procedures for minimum evidence required for annotating a gene model, gene naming conventions and quality control checks (Supplementary Tables S5a and b). The default annotation workflow (Figure 3) was followed by majority of student annotators, whereas more experienced annotators customized the protocol as described in the curation workflow section.

The undergraduate annotators at each site were mentored by a local faculty who coordinated separate in-person meetings. A few experienced annotators from the i5k community also participated in this curation effort. The video conferences facilitated a healthy discussion about curation methods among different groups of annotators. The combination of video conferences and online forum provided curators with the tools required to efficiently share data and information as they worked in teams across institutional boundaries. Moreover, as expert community curators volunteer their time to multiple projects, this model allowed them to contribute according to their availability.

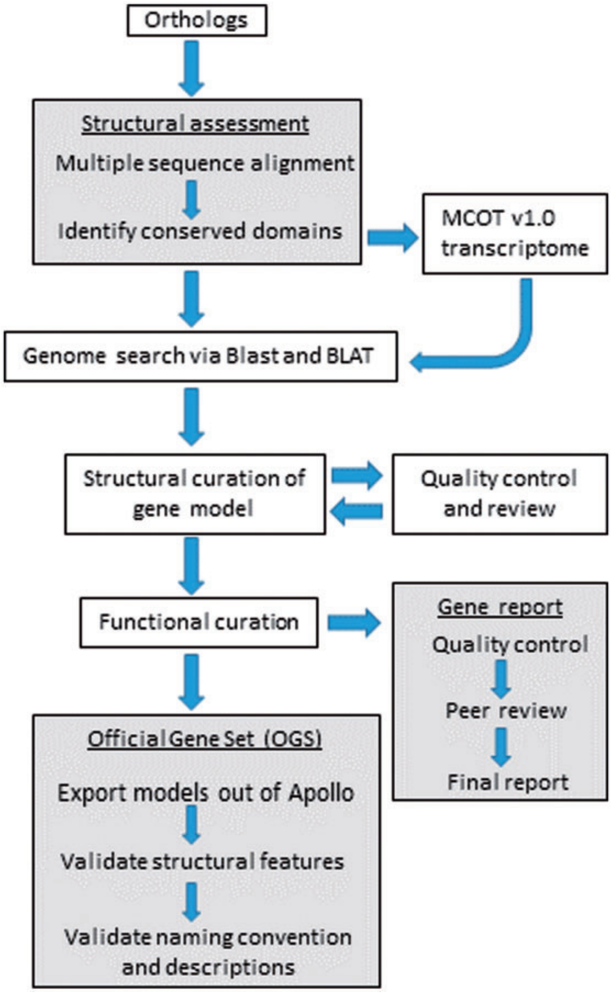
Orthologs
Structural assessment
Multiple sequence alignment
Identify conserved domains
MCOT v1.0 transcriptome
Genome search via Blast and BLAT
Structural curation of gene model
Quality control and review
Functional curation
Gene report
Quality control
Peer review
Final report
Official Gene Set (OGS)
Export models out of Apollo
Validate structural features
Validate naming convention and descriptions

**Figure 3**. Workflow followed by student annotators for the structural curation of gene models using i5k Apollo. The OGS was created at the end of the annotation cycle along with gene reports included as Supplementary Notes 1–39. The dark colored boxes denote processes with multiple steps.

### Curation workflow

Gene lists and orthologs were made available via the project management website (https://basecamp.com/2923904/projects/9184795) so that the annotators could volunteer to curate genes of interest. We used ImmunoDB (51) as a primary source of immune gene orthologs, which provided expert-curated immunity genes for *Aedes aegypti*, *Anopheles gambiae*, *Drosophila melanogaster* and *Culex quinquefasciatus*. Other closely related organisms used as sources of annotated orthologs included bed bug (*Cimex lectularius*) (55), pea aphid (*Acyrthosiphon pisum*) (61) and milkweed bug (*Oncopeltus fasciatus*) (62). All communication was done via emails and the project management website, which was also used to store presentations, gene reports, meeting minutes, presentations, working documents and data files. Annotation updates are shared with the community through the citrusgreening.org website (https://citrusgreening.org/annotation/index).

Candidate gene models were identified on the *D. citri* genome by using orthologous proteins as query in Apollo blat and i5k BLAST. The NCBI conserved domains database (63) was used to identify the conserved domains in the orthologs and candidate genes. Multiple sequence alignments were generated using MUSCLE (64), tcoffee (65) and clustal (66) to compare the ACP gene model to the query gene set. The final model was refined in Apollo using homology, RNAseq and proteomics evidence tracks. MEGA7 (67) was used to construct phylogenetic trees. Please see individual gene reports in Supplementary Notes 1–39 for detailed methods. When available, published literature was used to putatively assign molecular functions, participation in biological processes, and cellular localization for an annotated gene, associate term identifiers from the Gene Ontology Database (68) and PubMed identifiers from NCBI. Curated genes were assigned names and descriptions based on the function and domain structures available in published literature or NCBI.

Another strategy for identification of candidate genes followed by more experienced curators involved preprocessing the query before searching the ACP genome on Apollo. A set of representative genes was BLASTN searched against a database of all contigs in the ACP genome. These contigs were BLASTX searched against all the database of named insect genes. Analysis of the match helped to identify missing exons and extend partial exons to achieve the best possible manually curated version of the ACP gene. Once gene models were reconstructed from the genomic DNA, they were located on the *D. citri* genome using the i5k BLAST server (51). The models were then manually edited based on evidence tracks to produce the final gene model.

We performed multiple cycles of internal review of curated gene models to identify errors and suggest improvements to annotators. This was valuable in ensuring that annotators followed consistent guidelines throughout annotation. Curation of a gene family by an annotator was followed by a presentation summarizing their results during the video conference and peer-review. Annotations were ranked on a scale of A–D (Supplementary Table S5a) during review depending on completeness and support from various evidence tracks. Standard gene naming conventions were defined and agreed upon to ensure consistency (Supplementary Table S5b).

The curated gene models were exported from Apollo and all functional annotations were manually checked for consistency with community standards (Supplementary Table S5b). The gene set was validated using the i5k quality control pipeline (https://github.com/NAL-i5K/I5KNAL_OGS/wiki/QC-phase) which identifies intra-model, inter-model and single-feature errors. The cleaned manual annotations were then merged with the protein-coding genes from the NCBI *Diaphorina citri* Annotation Release 100 (NCBI v100, ftp://ftp.ncbi.nlm.nih.gov/genomes/Diaphorina_citri/ARCHIVE/ANNOTATION_RELEASE.100/) using the NAL's Merge prototype software (described at https://github.com/NAL-i5K/I5KNAL_OGS/wiki/Merge-phase; software is available on request). Non-coding RNAs from the NCBI *Diaphorina citri* Annotation Release 100 were added to the gene set after this merge, resulting in the Official Gene Set Dcitr_OGSv1.0 (69).

### Education

Manual curation was incorporated into a bioinformatics class in 2016 by one of the authors (Benoit) at the University of Cincinnati. Specifically, 1 week in the class was utilized for genome annotation through the i5k genome annotation workspace for 22 students. This focused on how RNAseq datasets contribute to gene prediction and why these models need to be corrected manually before genome publication. Each student was responsible for correcting two gene models for *D. citri* in Apollo. As part of this course, students were given an assessment test before and after the class. Importantly, the rubric for this survey was not specifically designed for this project, rather the ACP genome was selected as the focus organism to annotate after the initial assessment test. Three questions (Supplementary Table S6, Details necessary for correct answers are included within this table) of this test focused on gene prediction and genome annotation, which were the major educational focus of the genome annotation week. The average score on these questions before the class was only ∼38%, which improved to ∼90% at the completion of this course. These scores indicated that the students had a much improved knowledge of genome assembly and annotation following this class.

Two of the participating institutions (Benoit, University of Cincinnati; D'Elia, Indian River State College) integrated the *D. citri* genome annotation into senior capstone courses. Students that participated in capstone projects at UC covered eight topic areas; aquaporins, acidic amino acid transporters, glycosphingolipid metabolism, glycolysis, histone binding, vitamin metabolism, vitamin transport and Hox genes. With the addition of the capstone course, a total of 28 UC students participated in the manual curation process. Six students participated in capstone projects at IRSC during which they annotated 15 gene families. In total, 25 students directly participated in the manual curation process, contributing to 39 gene reports. This strategy reinforces the use of undergraduate students in gene annotation, as students show an increase in learning and produce scientific reports which contribute to peer-reviewed publications.

## Immune pathway in *D. citri*

Identification of the pathogen-induced immune components in ACP is critical for understanding and influencing the interaction between *D. citri* and CLas. The repertoire of immune genes is known to be a very diverse functional group and includes proteins that recognize infectious agents and initiate a signal, members of signal transduction pathways that relay the message to the nucleus, and the genes that are transcribed in response to infection. Below, we have briefly summarized our findings from manual structural and functional curation of immunity-related genes (Table 1). Additional details are provided in gene reports for specific gene families and in the Supplementary Notes 1–31.

### Pathogen recognition molecules

Recognition of infectious agents, the first step in immune defense, relies on a variety of cellular receptors including C-type lectins, Galectins, fibrinogen-related proteins (FREPs), peptidoglycan recognition proteins (PGRPs), beta-1, 3-Glucan recognition proteins ($\beta$GRPs) and thioester-containing proteins (TEPs) (Table 1). These proteins recognize pathogen-associated molecular patterns which are associated with microbial cells (70). Most of these recognition molecules are widely conserved in insects, although the copy number often varies.

*C-type lectins.* Ten C-type lectins (CTLs), carbohydrate-binding receptors (71, 72), were identified in *D. citri* (Supplementary Note 1). These include three oxidized low-density lipoprotein receptor genes and genes encoding C-type Lectin 3, C-type Lectin 5, C-type Lectin 8, E selectin, Perlucin, Agglucetin subunit alpha and a selectin-like osteoblast-derived protein. The number of CTLs in *D. citri* is comparable to the number found in bed bugs (11), pea aphids (6) and honeybees (10).

*Galactoside-binding lectins.* A total of three galactoside-binding lectins (galectins) and one partial galectin were identified and manually curated within the ACP genome (Supplementary Note 2). Galectins bind β-galactose with their structurally similar carbohydrate-recognition domains (73), which can function alone or in clusters creating a β-sandwich structure without $Ca^{2+}$-binding sites (74–76). Although we found more galectins in ACP than has been previously reported in the pea aphid [two genes (77)] and bed bug [one gene (55)], we did not observe a substantial lineage-specific expansion as seen in Dipterans.

**Table 1.** Immune gene pathways and gene counts in ACP, pea aphid and whitefly

| | Pathway/Genes | ACP | Pea aphid | Whitefly |
|---|---|---|---|---|
| Pathogen recognition molecules | | | | |
| | CTLs: C-type lectins | 10 | 5 | 5 |
| | GALEs: Galactoside-binding lectins | 4 | 1 | 4 |
| | FREPs: Fibrinogen-related proteins | 3 | 1 | 5 |
| | PGRPs: Peptidoglycan recognition proteins | 1 | 0 | 1 |
| | BGBPs: 1,3-beta-D Glucan-binding proteins | 0 | 2 | 5 |
| | TEPs: Thioester-containing proteins | 2 | 1 | 1 |
| Signaling cascades associated with pathogenesis | | | | |
| Toll pathway and receptors | Toll receptors | 5 | 7 | 5 |
| | Spaetzle | 5 | 6 | 9 |
| | Tube | 1 | 1 | 1 |
| | Pelle | 1 | 1 | 1 |
| | MyD88 | 1 | 1 | 1 |
| | CACT | 1 | 1 | 0 |
| | TRAF6 | 1 | 2 | 1 |
| IMD pathway members | CASPAR | 3 | 1 | 1 |
| | FADD | 0 | 0 | 0 |
| | IKKB/ird5 | 1 | 0 | 0 |
| | IMD | 0 | 0 | 0 |
| | TAK1 | 1 | 1 | 1 |
| | TAB | 1 | 1 | 1 |
| JAKSTAT pathway members | DOME | 1 | 3 | 1 |
| | HOP | 1 | 1 | 1 |
| | STAT | 1 | 3 | 1 |
| Response to pathogens and pathogen-associated stress | | | | |
| | AMPs: Antimicrobial peptides | 0 | 7 | 4 |
| | LYSs: Lysozymes | 5 | 3 | 5 |
| | SODs: Superoxide dismutases | 4 | 4 | 5 |
| | CLIPs: Clip-domain serine proteases | 14 | 3 | 6 |
| | Autophagy | 15 (2) | 8 | 16 |
| | PPOs: Prophenoloxidases | 2 | 2 | 4 |
| | IAPs: Inhibitors of apoptosis | 4 | 7 | 4 |

ACP genes identified only in MCOT v1.0 are in parentheses.

*Fibrinogen-related proteins.* Similar to other hemipterans (pea aphid and bed bug), few FREPs have been identified in ACP. Three complete FREPs were manually annotated (Scabrous, Angiopoietin and Tenascin) although partial un-annotatable FREP gene models were also detected. Like several of the other recognition molecule classes, the FREPs appear to have expanded in mosquitoes (78) (Supplementary Note 3). The suggestion that this expansion is related to blood feeding is consistent with the apparent absence in ACP of ficolin, tachylectins and aslectin, which are likely involved in detecting blood-borne parasites (78).

*PGRP and βGRP.* We identified one PGRP gene in *D. citri*. Insects have two classes of PGRPs: large (L) and small (S) (79). PRGP-L proteins recognize Gram-negative bacteria and activate the Imd pathway. PRGP-S proteins interact with $\beta$GRPs such as GNBP to recognize components of Gram-positive bacteria and then activate the Toll pathway. Based on sequence similarity to other insect proteins, the *D. citri* PGRP protein seems to belong to the S class. We did not find any GNBP genes in *D. citri* (Supplementary Note 4). This is somewhat surprising, as GNBPs have been found in several hemipterans including pea aphids (77), bed bugs (55) and brown planthoppers (80).

*Thioester-containing proteins.* Only two TEPs were identified in ACP (Supplementary Note 5), which is comparable to the number found in *A. pisum* and *Nasonia vitripennis*. TEPs are members of an ancient protein family that includes vertebrate C complement and alpha-2-macroglobulin proteins (81). Insect TEPs seem to play a similar role to their vertebrate homologs, binding to invaders such as parasites or microbes, marking them for degradation, and they are upregulated by the Janus kinase/signal transducer of activators of transcription (JAK/STAT) pathway during innate immune response (82).

**Signaling cascades associated with pathogenesis**

Once a potential infection has been detected, a cellular response is initiated by signaling cascade (Table 1). Typically, Gram-positive bacteria and fungi cause activation of the Toll pathway, whereas the Imd pathway responds to Gram-negative bacteria (83, 84). The JAK/STAT pathway plays a role in several immune functions, including antiviral defense (85).

*Toll pathway.* We identified four Toll receptors in *D. citri* (Supplementary Note 6). Comparison of the Toll receptors found in various insects suggests that there were six ancestral Toll receptors: *Toll-1*, *Toll-6*, *Toll-2/7*, *Toll-8*, *Toll-9* and *Toll-10* (86, 87). Phylogenetic analysis indicates that the *D. citri* genes are orthologs of *Toll-1*, *Toll-6*, *Toll-7* and *Toll-8*, but orthologs of *Toll-9* and *Toll-10* were not found. Pea aphids and bed bugs have Toll receptors from every class but *Toll-9* (55, 77). We found orthologs of five of the six Spätzle (Spz) ligand classes, including Spz1, Spz3, Spz4, Spz5 and Spz6 (Supplementary Note 7). The lack of Spz2 is not surprising as it has only been reported in Diptera and Hymenoptera. The downstream Toll pathway components are represented by single genes in most insects (88). Consistent with this, we identified single-copy orthologs of *tube*, *pelle*, *MyD88*, TRAF6, *cactus* and *dorsal* (Supplementary Notes 8–13). Taken together, our findings suggest that the Toll pathway is largely conserved in *D. citri*, as it is in other insects.

*Imd pathway.* As has been observed for several other hemipterans (77, 89–92), many components of the Imd pathway appear to be missing in *D. citri* (Supplementary Note 14). We were unable to identify orthologs of Dredd, FADD, Imd, IKKG and Relish in either the assembled genome or the MCOT transcriptome. We did, however, find orthologs of pathway components IKKB, TAK1 and TAB, as well as FAF1/Caspar, a negative regulator of the pathway. The apparent loss of Imd pathway genes in many hemipterans has led to speculation that association with Gram-negative endosymbionts may have favored the loss of these genes (61, 92), although it should be noted that organisms such as Wolbachia are also found in many insects with intact Imd pathways (93). Several Gram-negative bacteria have been identified as *D. citri* symbionts, including *Wolbachia*, *Candidatus* Carsonella, *Candidatus* Profftella armaturae, and an as-yet-unidentified enteric bacteria closely related to *Klebsiella variicola* and *Salmonella enterica* (94–98). Given this information, it is tempting to speculate that the loss of many Imd pathway genes may, in fact, be associated with the ability of *D. citri* to acquire and harbor at least some of these Gram-negative symbionts and might also be important for its ability to act as a carrier of CLas.

*JAK/STAT pathway.* The JAK/STAT pathway is a signaling pathway that provides direct communication between the membrane and nucleus (99). We identified genes encoding the major components of the JAK/STAT pathway, namely the orthologs of *domeless, hopscotch* and *marelle/Stat92E* (Supplementary Note 15). The JAK/STAT pathway is involved in many developmental processes, in addition to its role in immunity, and has been found in all sequenced insects to date, including other hemipterans.

**Response to pathogens and pathogen-associated stress**

In response to infection, insect cells employ microbicidal compounds such as antimicrobial peptides (AMPs), lysozymes and reactive oxygen species (ROS) to destroy invading cells and also activate tissue repair, wound healing and hematopoiesis processes. We searched the ACP genome for antimicrobial compounds (AMPs and lysozymes), the melanization-inducing Clip-domain serine proteases (CLIPs), the protective superoxide dismutases (SODs) and autophagy-related genes (Table 1).

*Antimicrobial peptides.* Although >250 AMPs have been identified in insects, we searched the ACP genome and the MCOT transcriptome for 10 classes of known AMPs without success. The AMPs investigated included attacin, cecropin, defensin, diptericin, drosocin, drosomycin, gambicin, holotricin, metchnikowin and thaumatin. Some of these AMPs appear to be widely conserved, whereas others have only been identified in a limited number of species. Although defensins are one of the most widely conserved, ancient groups of AMPs (100), the absence of defensin in ACP is not unprecedented as its absence has also been reported the hemipteran *A. pisum* (77). Although the pea aphid is lacking defensin (as well as most other previously identified insect AMPs) it does contain six thaumatin (antifungal) homologs. Despite its presence in the closely related pea aphid, we were unable to identify thaumatin in the ACP genome. It must be noted that absence of previously identified AMPs does not necessarily suggest absence of all AMPs. AMPs are an extremely large, diverse group of molecules often defined by structure and function rather than conserved motifs, making identification through comparative sequence analysis difficult. Additionally, most AMPs are probably yet to be identified and will need to be discovered through experimental work as opposed to orthologous searches based upon currently available sequence information.

*Lysozymes*. Five genes encoding lysozymes were found in the *D. citri* genome (Supplementary Note 16). Lysozymes hydrolyze bacterial peptidoglycan, disrupting cell walls and causing cell lysis. Many insects produce lysozymes, particularly c-type lysozymes, and secrete them into the hemolymph following bacterial infection. C-type lysozymes that commonly defend against Gram-positive bacteria have been reported in many different insect orders including Diptera, Hemiptera and Lepidoptera (101). Although c-type lysozymes were not found in the initial search of the ACP genome, two c-type lysozyme transcripts were found in *D. citri* MCOT v1.0 and were subsequently used to identify these genes in the ACP genome. Additionally, three i-type lysozymes were identified in the ACP assembly.

*Superoxide dismutases*. Insect hemocytes can produce a burst of ROS to kill pathogens (102). As ROS are also damaging to host cells, SODs are necessary to detoxify ROS. We found a total of four SOD genes in *D. citri* (Supplementary Note 17). Similar to other insects (103–105), *D. citri* contains both CuZn and Mn SODs. One of the *D. citri* genes is an Mn SOD and the other three are CuZn SODs.

*Clip-domain serine proteases*. Eleven CLIPs, from four distinct evolutionary clades (CLIPA, CLIPB, CLIPC and CLIPD), were manually annotated in the *D. citri* genome and corresponding models identified in MCOT v1.0 (Supplementary Note 18). These clades are present as multigene families in insect genomes and function in the hemolymph in innate immune responses (106). In *Drosophila*, CLIPs are involved in melanization and the activation of the Toll pathway (107).

*Autophagy*. Using the *D. citri* genome and the MCOT gene set, we identified *D. citri* orthologs of autophagy-related genes known in *Drosophila* (Supplementary Note 19). Autophagy is the regulated breakdown of unnecessary or dysfunctional components of the cell. This process is highly conserved among all animals and is critical to the regulation of cell degradation and recycling of cellular components. The main pathway is macroautophagy, where specific cytoplasmic components are isolated from the remaining cell in a double-membraned vesicle called the autophagosome (108–110). We identified 17 out of 20 autophagy-related genes (Supplementary Note 19). There is only a single *Autophagy-related 8* gene in ACP gene sets compared to two for the *Drosophila* gene set, but this is common for non-Dipteran insects (109). Thus, as expected, psyllids have the required repertoire of autophagy-related genes to undergo macroautophagy.

In summary, there is a reduction in the number of immune recognition, signaling and response genes in *D. citri* compared to insects from Diptera. The reduction in the immunity genes is also observed in other hemipteran insect genomes such as *A. pisum* (77), *Pediculus humanus* (89), *Bactericera cockerelli* (90), *Rhodnius prolixus* (91) and *Bemisia tabaci* (92). The reduction of immunity-related genes in these insects has been attributed to association with their endosymbionts, which sometimes complement the immunity of the insects (111, 112). In addition, this reduction in immune genes may be associated with insects that feed on nutritionally poor and relatively sterile food sources, such as blood and fluid from the xylem/phloem (89, 111). However, blood-feeding mosquitoes actually show an increase in immune genes and this expansion has been attributed to the likelihood of encountering pathogens in their food source. Arp et al. (23) pointed out additional inconsistencies with the diet hypothesis, including the presence of a full immune system in an insect species that develops in a sterile environment.

## RNA interference pathway in *D. citri*

The RNA interference (RNAi) pathway is a highly conserved, complex method of endogenous gene regulation and viral control mediated through short interfering RNAs (siRNAs), microRNAs (miRNAs) and piwiRNAs (piRNAs). Although all of these small RNA molecules function to modulate or silence gene expression, the method of gene silencing and the biogenesis differs (113). In *Drosophila*, it appears that genes in the RNAi machinery have subfunctionalized to have roles in specific small RNA silencing pathways (114–120). Although the RNAi machinery genes have been shown to be conserved across major taxa, functional studies in insects have been limited to a handful of Diptera. Investigating the complement of RNAi genes in *D. citri* (Table 2) may provide insight into the role that RNAi has on the immune response of phloem-feeding insects and could aid in better use of RNAi as a tool for pest management (121–123).

### Core machinery

Class II (Drosha type) and class III (Dicer type) RNase III enzymes play an essential role in the biogenesis of small RNA molecules with Drosha and Dicer1 functioning to produce miRNAs and Dicer2 functioning to produce siRNAs

**Table 2.** Homolog number of core machinery and auxiliary RNAi components in insects

| | Dcr1 | Dcr2 | Drosha | Loqs | R2D2 | Pasha | AGO1 | AGO2 | AGO3 | PIWI/Aub | Armi | TSN | VIG-1 | Spn-E | Rm62 | Ran | FMR1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *D. melanogaster* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *A. gambiae* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 6 | 1* | 0 |
| *A. aegypti* | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 7 | 2 | 1 | 1 | 1 | 9 | 1* | 1 |
| *C. quinquefasciatus* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 6 | 2 | 1 | 0* | 1 | 10 | 1 | 1 |
| *T. castaneum* | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 3* | 1* | 2* | 1* | 2* | 1* | 2* |
| *C. lectularius* | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 4 | 2* | 1* | 1* | 1* | 0* | 1* | 1* |
| *D. citri* | 1 | 1 | 1 or 2 | 2 | 1† | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |

Proteins highlighted in orange are RNase Type III enzymes. Proteins highlighted in blue are dsRNA-binding proteins. Proteins highlighted in green are AGO family proteins. Proteins highlighted in yellow have been implicated in RISC or in small RNA biogenesis. *Drosophila melanogaster* has six other RNA helicase genes with homology to Rm62. Some of the mosquito Rm62 proteins could be orthologous to these related RNA helicases. * indicates homolog number was determined by BLAST and reciprocal BLAST analysis using NCBI's non redundant databases. Homolog number with no asterisk were determined by publications or reported in ImmunoDB. † indicates homolog is likely present but was unable to be annotated in the current genome assembly.

(124–126). Our analysis of the *D. citri* genome revealed four possible loci with identity to insect Dicer proteins (Supplementary Note 20). However, further analysis of the MCOT transcriptome suggests that *D. citri* contains only one gene orthologous to *Dicer1* (MCOT05108.0.CO) and one gene orthologous to *Dicer2* (MCOT13562.0.CO). The remaining two loci are likely the result of genome fragmentation and misassembles. BLAST analysis of Drosha also identified multiple loci with homology to other insect Drosha proteins. MCOT transcriptome analysis indicates at least one or possibly two *drosha* homologs are present in *D. citri* (Supplementary Note 21).

dsRNA-binding proteins act in concert with RNase III-type enzymes to bind and process precursor dsRNA molecules into small effector molecules (116, 120, 127, 128). In some cases, these dsRNA-binding proteins also function to load small RNA molecules into the RNA-induced silencing complex (RISC) (129–131). In *Drosophila*, Pasha partners with Drosha (128), Loquacious (Loqs) partners with Dicer1 (120, 127) and R2D2 partners with Dicer2 (116). In the *D. citri* genome, we identified two *pasha* homologs (Supplementary Note 22) and two *loqs* homologs but were initially unable to identify a true *r2d2* ortholog (Supplementary Note 23) in the genome. The apparent absence of *r2d2* was consistent with previous reports (22, 131). However, a search of the MCOT (MCOT18647.0.CO) transcriptome identified a gene with similarity to R2D2 orthologs from bed bug, *Tribolium castaneum* and mosquitoes. Although *r2d2* is likely to be present in the *D. citri* genome, it is not annotatable given the limitations of the current assembly. Alternatively, if *r2d2* is missing from in *D. citri*, it is possible that one of the Loqs proteins identified functions in the RNAi pathway (132), as Loqs has been shown to associate with Dicer2 in both *Drosophila* and *Aedes aegypti* (133, 134).

Argonaute (AGO) proteins present small RNA guide molecules to their complementary targets through silencing complexes and provide the 'Slicer' catalytic activity that is required for mRNA cleavage in some RNA silencing pathways (135–138). In *Drosophila* AGO1 is involved in the miRNA pathway, AGO2 is involved in silencing by siRNAs (114, 119) and AGO3, PIWI and Aubergine (Aub) function in the piRNA pathway (115, 117, 139). In the *D. citri* genome, we have identified four *AGO* genes, *AGO1*, *AGO2*, *AGO3* and one gene corresponding to the PIWI/Aub class of proteins (Supplementary Note 24).

### Auxiliary (RISC and other) factors

A subset of other genes known to be involved in the function or regulation of the RISC in *Drosophila* and other organisms have been identified and annotated in the *D. citri* genome. The genes identified include two *Tudor Staphylococcal Nucleases* (TSN, Supplementary Note 25), one *vasa-intronic gene* (*vig-1*, Supplementary Note 25), one *armitage* (*Armi*) gene (Supplementary Note 27) and one *Fragile X Mental Retardation 1* (*FMR1*, Supplementary Note 28) gene. Additionally, several more genes known to be involved in the biogenesis or function of small RNA molecules have been identified. These include two *spindle-E* genes (Supplementary Note 29), one *Rm62* gene (Supplementary Note 30) and one *Ran* gene (Supplementary Note 31).

In summary, the *D. citri* genome has a full complement of RNAi machinery genes. Duplications are more frequent in genes that have previously been associated with the miRNA pathway (*drosha*, *pasha* and *loqs*) as opposed to the RNAi or piRNA pathways. This is an interesting finding as the same result was found upon analysis of the pea aphid genome (140) but was not seen in the whitefly genome (92).

## Building the foundation for P450/Halloween genes targeted to reduce insect pests

CYPs in eukaryotes are heme-containing membrane bound enzymes that activate molecular oxygen via a mechanism involving a thiolate ligand to the heme iron. Usually this requires an electron donor protein, the NADPH CYP reductase, in the ER or ferredoxin and ferredoxin reductase in the mitochondria (141, 142). Insects have four deep branching clades on phylogenetic trees and this represents some losses during evolution as up to 11 clades are found in other animals. These are termed CYP2, CYP3, CYP4 and mitochondrial clans in P450 nomenclature (143). Most species have a tendency to expand P450s in one or more clans via tandem duplications. One interpretation of these P450 'blooms' is diversification to handle many related compounds from the environment that may be toxic or potential carbon sources.

*Diaphorina citri* in its current assembly has 60 P450 genes that are identified and named as distinct P450s (Table 3). There are also numerous fragments named as partials. *Diaphorina* has a P450 bloom in the clusters CYP3172, CYP3174, CYP3175, CYP3176, CYP3178 in the CYP4 clan. There is another in the CYP3167 family in the CYP2 clan and a third that includes CYP6KA, CYP6KC and CYP6KD in the CYP3 clan. *Diaphorina citri* has three CYP4G genes.

CYP2 and mito clans have many 1:1 orthologs but these are rare in the CYP3 and CYP4 clan (Figure 4). One exception in the CYP3 clan is CYP3087A1 and two neighbors CYP6DB1 and CYP6KB1 as they may be orthologs and probably should be in the same family. *Rhodnius prolixus* and *A. pisum* CYP3 clan genes have undergone gene blooms that were not found in *D. citri*. This may be interpreted as the common ancestor having few CYP3 clan P450s. The cluster at arc A (Figure 4) on the tree consisting of CYP6KB1, CYP6DB1 and CYP3087A1 may be evidence that these three are orthologs and should be in the same family. The large aphid clade of CYP6CY at arc C has no members from *Rhodnius* or *Diaphorina*, so it seems to be aphid specific. At arc B (Figure 4) the CYP395 family has four subfamilies C, D, E, F. There are many CYP395 genes in other hemipteran species, including *C. lectularius* (bedbug CYP395A,B), *Apolygus lucorum* (Hemiptera, a Mirid bug, CYP395G, H, J, K, L, M) and *Cyrtorhinus lividipennis* (Hemiptera, green mirid bug, CYP395H, J, N). The fact that they have been placed in different subfamilies suggests they are diverging from their common ancestor. The CYP3084 family with subfamilies A, B, C, D is only found in *Rhodnius* so far (144). The families CYP3088, CYP3089, CYP3090 and CYP3091 are also Hemiptera specific with some members in the same species noted earlier. The number of P450s varies with arthropod species from a low of 25 in the mite *Aculops lyoperscii* and 36 in *P. humanus* (body louse) to over 200 in *Ixodes scapularis* (black-legged tick) and up to 158 in some mosquitos (145).

**Table 3.** Number of P450 genes reported in 10 Hemipteran species

| Species | Number of P450s |
|---|---|
| *Apolygus lucorum* (mirid bug) | 46 |
| *Acyrthosiphon pisum* (pea aphid) | 58 |
| *Cyrtorhinus lividipennis* (green miridbug) | 59 |
| *Cimex lectularius* (bedbug) | 60 |
| ***Diaphorina citri* (Asian citrus psyllid)** | 60 |
| *Laodelphax striatellus* (small brown planthopper) | 63 |
| *Nilaparvata lugens* (brown planthopper) | 66 |
| *Rhodnius prolixus* (kissing bug) | 87 |
| *Bemisia tabaci* (whitefly) | 128 |
| *Homalodisca vitripennis* (glassy winged sharpshooter) | 142 |

## Conclusion

We report the first draft assembly for the *D. citri* genome and the corresponding official gene set (OGS v1.0) which includes 530 manually curated genes and about 20 000 genes predicted by the NCBI Eukaryotic Genome Annotation Pipeline (NCBI v100). The community curation effort involved undergraduate students at multiple locations who were trained, individually or in a class setting, in gene curation as a part of this initiative. These students were supported by contributions from expert annotators in the insect genomics community. The major advantage of having both expert curators and undergraduate students work together in the annotation project was the training, exchange of ideas and community building. We also present standard operating procedures that can be used to guide and coordinate annotation by large virtual teams. We would like to note that implementing consistent annotation practices across a highly diverse and virtual team of annotators required regular discussions backed up by extensive documentation that was updated in response to user feedback. However, multiple rounds of manual review by senior annotators were still required to confirm that all annotations conform to certain basic criterion. A number of evidence sources (Supplementary Table S3) were added during the course of the project based upon availability and utility to ongoing annotation. One of the major decisions taken by the ACP community as a result of this annotation effort and subsequent detailed evaluation of the Diaci1.1 genome was to generate an improved reference genome for ACP using the latest methods. An interim but improved assembly based on long-read sequencing technology is available at Ag Data Commons (146).

This community annotation process will be continued by recruiting the next cohort of student annotators and scientists to improve structural and functional characterization of the next version of the ACP genome (146). The MCOT transcriptome reported in this article is available at Ag Data

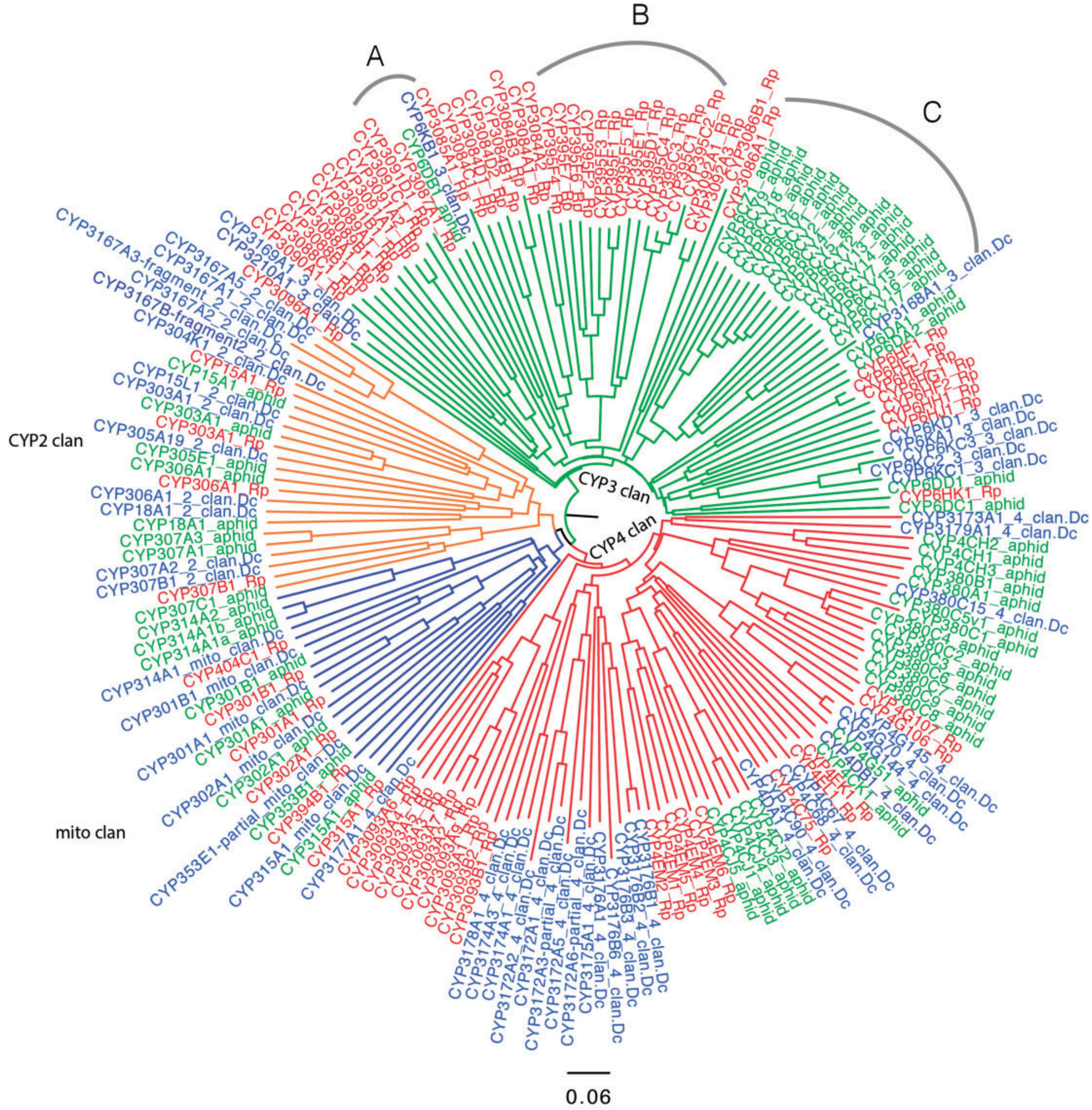

**Figure 4.** Phylogeny of P450 genes with insect orthologs. Neighbor-joining midpoint rooted tree of P450s from *R. prolixus* (red), *A. pisum* (green) and *D. citri* (blue) was generated using CLUSTAL Omega and drawn in Figtree. Four clans of P450s, CYP2 (orange), CYP3 (green), CYP4 (red) and mito (blue) clan are shown in the phylogenetic tree. Arc A, B and C are described in the main body of text. Excluding some partial genes, a total of 189 P450s were used to generate this phylogenetic tree.

Commons (39) and offers a genome-independent and comprehensive representation of the gene repertoire of *D. citri* that was used to improve the genome annotation. The MCOT transcriptome allowed us to identify lineage specific-gene models in *D. citri* and curate them. This gene set will support efforts in other hemipteran species that transmit bacterial pathogens.

In summary, we curated and described genes related to immunity and the RNAi pathway in addition to the CYP genes. We report blooms in P450 genes in the CYP4, CYP2 and CYP3 clans which may be an evolutionary response to environmental stresses. Other important gene families that were curated as a part of the OGS include aquaporins, cathepsins, cuticle and secretory proteins. We found the number of immunity-related genes to be reduced, even after direct targeting for improvement, in the *D. citri* genome similar to pea aphid and whitefly, which may reflect the association with microbial symbionts that have coevolved in both insects and the consumption of relatively sterile plant-derived fluids. The genomic resources from this project will provide critical information underlying ACP biology that can be used to improve control of this pest.

## Supplementary data

Supplementary data are available at *Database* Online.

## Acknowledgements

Kascha Bohnenblust from Kansas State University assisted with the organization of meetings for the community curation effort. The following personnel at Los Alamos National Laboratory, Los Alamos, NM contributed to the Illumina and Pacbio sequencing: Kimberly McMurry, Cheryl D. Gleasner, Krista Reitenga, Shunsheng (Cliff) Han and Goutam Gupta. Christian Haudenschild from lllumina, Inc. supported the genome assembly. The following personnel at USDA-ARS, U.S. Horticultural Research Laboratory, Fort Pierce, FL contributed to the sample preparation for genome sequencing: Maria T. Gonzalez, Belkis Diego and Kathy Moulton. We would like to thank Felipe Simão for providing the Hemipteran marker set for BUSCO analysis.

## AUTHOR CONTRIBUTIONS

SS, PSH, KVA, SM, TS, AR, CC, TB, HM, GD, DD, ZM, KB, KC, SN, RM, JVV, MD, DF, DH, TM, KS, VG, AI, DED, LC, AC, CC, MC, MF, WH, MMT, DN, MFP, JB, HKW, TD and SJB contributed to the community curation. XC, HJ and MF generated the MCOT v1.0 transcriptome. WH, SJ, BLC, SR, AE and NL were involved in the genome sequencing. WH, MC, LAM, JB, TD and SJB were the Principal Investigators involved in this work. SS, PSH, SM, TS, MF, WH, MC, MMT, DN, MFP, JB, TD and SJB wrote the final manuscript.

## Funding

All open-access fees, student annotators and post-docs were funded through USDA-NIFA grant 2015-70016-23028. The genome sequencing was supported by grants from the NIH National Center for Research Resources (5P20RR016469). USDA-ARS, Ft. Pierce, FL and a grant from the Citrus Research Board, Inc. (217 North Encina, PO Box 230, Visalia, CA 93279).

*Conflict of interest*. None declared.

## References

1. Halbert,S.E. and Núñez,C.A. (2004) Distribution of the Asian citrus psyllid, *Diaphorina citri* Kuwayama (Rhynchota: Psyllidae) in the Caribbean basin. *Fla. Entomol.*, 87, 401–402.
2. Halbert,S.E. and Manjunath,K.L. (2004) Asian citrus psyllids (Sternorrhyncha: Psyllidae) and greening disease of citrus: a literature review and assessment of risk in Florida. *Fla. Entomol.*, 87, 330–353.
3. Boykin,L.M., De Barro,P., Hall,D.G. *et al.* (2012) Overview of worldwide diversity of *Diaphorina citri* Kuwayama mitochondrial cytochrome oxidase 1 haplotypes: two Old World lineages and a New World invasion. *Bull. Entomol. Res.*, 102, 573–582.
4. French,J.V., Kahlke,C.J. and Da Graça,J.V. (2001) First record of the Asian citrus psylla, *Diaphorina citri* Kuwayama (Homoptera: Psyllidae) in Texas. *Subtrop. Plant Sci.*, 53, 14–15.
5. Pluke,R.W.H., Qureshi,J.A. and Stansly,P.A. (2008) Citrus flushing patterns, *Diaphorina citri* (Hemiptera: Psyllidae) populations and parasitism by *Tamarixia radiata* (Hymenoptera: Eulophidae) in Puerto Rico. *Fla. Entomol.*, 91, 36–42.
6. Tsai,J.H. and Liu,Y.H. (2000) Biology of *Diaphorina citri* (Homoptera: Psyllidae) on four host plants. *J. Econ. Entomol.*, 93, 1721–1725.
7. Teixeira,D.D.C., Saillard,C., Eveillard,S. *et al.* (2005) "*Candidatus Liberibacter americanus*", associated with citrus huanglongbing (greening disease) in São Paulo State, Brazil. *Int. J. Syst. Evol. Microbiol.*, 55, 1857–1862.
8. Capoor,S.P., Rao,D.G., Viswanath,S.M. *et al.* (1967) *Diaphorina citri* Kuway., a vector of the greening disease of citrus in India. *Indian J. Agric. Sci.*, 37, 572–575.
9. Bové,J.M. (2006) Invited review Huanglongbing: a destructive, newly-emerging, century-old disease of citrus 1. *J. Plant Pathol.*, 88, 7–37.
10. Manjunath,K.L., Halbert,S.E., Ramadugu,C. *et al.* (2008) Detection of '*Candidatus* Liberibacter asiaticus' in *Diaphorina citri* and its importance in the management of citrus huanglongbing in Florida. *Phytopathology*, 98, 387–396.
11. Leong,S.C.T., Abang,F., Beattie,A. *et al.* (2012) Impacts of horticultural mineral oils and two insecticide practices on population fluctuation of *Diaphorina citri* and spread of huanglongbing in a citrus orchard in Sarawak. *Sci. World J.*, 2012, 7. Doi:10.1100/2012/651416.
12. Honig,L. *October Crop Production Executive Summary*. https://www.nass.usda.gov/Newsroom/Executive_Briefings/2016/10_12_2016.pdf (9 May 2017, date last accessed).
13. Hodges,A.W. and Spreen,T.H. (2015) Univ. Florida IFAS Ext., 712, Economic Impacts of Citrus Greening -(HLB) in Florida, 2006/07-2010/11.
14. Tiwari,S., Lewis-Rosenblum,H., Pelz-Stelinski,K. *et al.* (2010) Incidence of *Candidatus* Liberibacter asiaticus infection in abandoned citrus occurring in proximity to commercially managed groves. *J. Econ. Entomol.*, 103, 1972–1978.
15. Tabachnick,W.J. (2015) *Diaphorina citri* (Hemiptera: Liviidae) vector competence for the citrus greening pathogen *Candidatus* Liberibacter asiaticus. *J. Econ. Entomol.*, 108, 839–848.
16. Ramsey,J.S., Johnson,R.S., Hoki,J.S. *et al.* (2015) Metabolic interplay between the Asian citrus psyllid and its Profftella symbiont: an Achilles' heel of the citrus greening insect vector. *PLoS One*, 10, e0140826.
17. Andrade,E.C. and Hunter,W.B. (2016) In: Abdurakhmonov,I.Y. (ed). *RNA Interference – Natural Gene-Based Technology for Highly Specific Pest Control (HiSPeC)*. Croatia, Rijeka: InTech. http://dx.doi.org/10.5772/61612.
18. Marutani-Hert,M., Hunter,W.B. and Hall,D.G. (2010) Gene response to stress in the Asian citrus psyllid (Hemiptera: Psyllidae). *Fla. Entomol.*, 93, 519–525.
19. Hunter,W.B., Dowd,S.E., Katsar,C.S. *et al.* (2009) Psyllid biology: expressed genes in adult Asian citrus psyllids, *Diaphorina citri* Kuwayama. *The Open Entomol. J.*, 3, 18–29.
20. Hunter,W.B., Hail,D., Tipping,C. et al. (2010) In: *Symposium Proceedings 2010 Pierce s Disease Research Symposium, California Department of Food and Agriculture, Sacramento, CA*. California Department of Food and Agriculture, San Diego, California, pp. 24–27.

21. Reese,J., Christenson,M.K., Leng,N. *et al.* (2014) Characterization of the Asian citrus psyllid transcriptome. *J. Genomics*, 2, 54–58.
22. Fisher,T., Vyas,M., He,R. *et al.* (2014) Comparison of potato and Asian citrus psyllid adult and nymph transcriptomes identified vector transcripts with potential involvement in circulative, propagative liberibacter transmission. *Pathogens*, 3, 875–907.
23. Arp,A.P., Pelz-Stelinski,K. and Hunter,W. (2016) Annotation of the Asian citrus psyllid genome reveals a reduced innate immune system. *Front. Physiol.*, 7, 570.
24. Stein,L. (2001) Genome annotation: from sequence to biology. *Nat. Rev. Genet.*, 2, 493–503.
25. Elsik,C.G., Worley,K.C., Zhang,L. *et al.* (2006) Community annotation: procedures, protocols, and supporting tools. *Genome Res.*, 16, 1329–1333.
26. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18, 821–829.
27. Chaisson,M.J. and Tesler,G. (2012) Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application. *BMC Bioinformatics*, 13, 238.
28. English,A.C., Richards,S., Han,Y. *et al.* (2012) Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. *PLoS One*, 7, e47768.
29. Stanke,M., Steinkamp,R., Waack,S. *et al.* (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.*, 32, W309–W312.
30. Slater,G.S.C. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 31.
31. Schulz,M.H., Zerbino,D.R., Vingron,M. *et al.* (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28, 1086–1092.
32. Kim,D., Pertea,G., Trapnell,C. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14, R36.
33. Trapnell,C., Roberts,A., Goff,L. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, 7, 562–578.
34. Cantarel,B.L., Korf,I., Robb,S.M.C. *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, 18, 188–196.
35. Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, 13, 329–342.
36. Trapnell,C., Williams,B.A., Pertea,G. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28, 511–515.
37. Grabherr,M.G., Haas,B.J., Yassour,M. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29, 644–652.
38. Haas,B.J., Papanicolaou,A., Yassour,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, 8, 1494–1512.
39. Saha,S., Cao,X., Flores,M. *et al.* (2017) *Diaphorina citri* MCOT transcriptome. Ag Data Commons. http://dx.doi.org/10.15482/USDA.ADC/1342726.
40. Jones,P., Binns,D., Chang,H.-Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30, 1236–1240.
41. Schoof,H. In: *Plant and Animal Genome XXIV Conference*; Scherago International: San Diego, CA, 2016.
42. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859–1875.
43. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, 12, 357–360.
44. Pertea,M., Pertea,G.M., Antonescu,C.M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, 33, 290–295.
45. Eilbeck,K., Moore,B., Holt,C. *et al.* (2009) *Quantitative Measures for the Management and Comparison of Annotated Genomes*, 10, 67.
46. Saha, S., Hunter, W., Mueller, L., *et al.* (2017) *Diaphorina citri* genome assembly Diaci 1.9. Ag Data Commons. http://dx.doi.org/10.15482/USDA.ADC/1342727.
47. Marutani-Hert,M., Hunter,W.B. and Hall,D.G. Establishment of Asian citrus psyllid (*Diaphorina citri*) primary cultures. *In Vitro Cell. Dev. Biol. Anim.*, 45, 317–320.
48. Simao,F.A., Waterhouse,R.M., Ioannidis,P. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212.
49. Waterhouse,R.M., Zdobnov,E.M., Tegenfeldt,F. *et al.* (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.*, 39, D283–D288.
50. Cao,X. and Jiang,H. (2015) Integrated modeling of protein-coding genes in the *Manduca sexta* genome using RNA-Seq data from the biochemical model insect. *Insect Biochem. Mol. Biol.*, 62, 2–10.
51. Poelchau,M., Childers,C., Moore,G. *et al.* (2014) The i5k Workspace@NAL–enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res.*, 43, D714–D719.
52. Waterhouse,R.M., Kriventseva,E.V., Meister,S. *et al.* (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, 316, 1738–1743.
53. Werren,J.H., Richards,S., Desjardins,C.A. *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science*, 327, 343–348.
54. Chipman,A.D., Ferrier,D.E.K., Brena,C. *et al.* (2014) The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biol.*, 12, e1002005.
55. Benoit,J.B., Adelman,Z.N., Reinhardt,K. *et al.* (2016) Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. *Nat. Commun.*, 7, 10165.
56. Adams,M.D., Celniker,S.E., Holt,R.A. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science (80-.)*, 287, 2185–2195.
57. Venter,J.C., Adams,M.D., Myers,E.W. *et al.* (2001) The sequence of the human genome. *Science (80-.)*, 291, 1304–1351.
58. Munoz-Torres,M. *Apollo Workshop at KSU 2015*. https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015 (1 January 2017, date last accessed).

59. Munoz-Torres,M. *Apollo Exercises Kansas State University 2015*. https://www.slideshare.net/MonicaMunozTorres/apollo-exercises-kansas-state-university-2015 (1 January 2017, date last accessed).
60. Munoz-Torres,M. *Apollo Annotation Guidelines for i5k projects Diaphorina citri*. https://www.slideshare.net/MonicaMunozTorres/apollo-annotation-guidelines-for-i5k-projects-diaphorina-citri (1 January 2017, date last accessed).
61. International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol.*, 8, e1000313.
62. Vargas Jentzsch,I.M., Hughes,D.S.T. and Poelchau,M.F.T. The *O. fasciatus* curation community, Richards S, Panfilio KA. 2015. *Oncopeltus fasciatus* official gene set v1. 1.
63. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, 43, D222–D226.
64. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32, 1792–1797.
65. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302, 205–217.
66. Sievers,F., Wilm,A., Dineen,D. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7, 539.
67. Kumar,S., Stecher,G. and Tamura,K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, 33, 1870–1874.
68. Harris *et al.*; Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 32, D258–D261.
69. Saha,S. (2017) *Diaphorina citri* Official Gene Set v1.0. Ag Data Commons. http://dx.doi.org/10.15482/USDA.ADC/1345524.
70. Christophides,G.K., Vlachou,D. and Kafatos,F.C. (2004) Comparative and functional genomics of the innate immune system in the malaria vector *Anopheles gambiae*. *Immunol. Rev.*, 198, 127–148.
71. Dodd,R.B. and Drickamer,K. (2001) Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. *Glycobiology*, 11, 71R–79R.
72. Cambi,A. and Figdor,C.G. (2003) Dual function of C-type lectin-like receptors in the immune system. *Curr. Opin. Cell Biol.*, 15, 539–546.
73. Cummings,R.D. and Liu,F.T. (2009) Essentials of Glycobiology. In: Varki,A., Cummings,R.D., Esko,J.D., Stanley,P., Hart,G., Aebi,M., Darvill,A., Kinoshit,T., Packer,N.H., Prestegard,J.J., Schnaar,R.L. and Seeberger,P.H. (eds). Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009.
74. Leffler,H., Carlsson,S., Hedlund,M. *et al.* (2002) Introduction to galectins. *Glycoconj. J.*, 19, 433–440.
75. Mitchell,D.A., Fadden,A.J. and Drickamer,K. (2001) A novel mechanism of carbohydrate recognition by the C-type lectins DC-SIGN and DC-SIGNR subunit organization and binding to multivalent ligands. *J. Biol. Chem.*, 276, 28939–28945.
76. Wang,L., Wang,L., Yang,J. *et al.* (2012) A multi-CRD C-type lectin with broad recognition spectrum and cellular adhesion from Argopectenirradians. *Dev. Comp. Immunol.*, 36, 591–601.
77. Gerardo,N.M., Altincicek,B., Anselme,C. *et al.* (2010) Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. *Genome Biol.*, 11, R21.
78. Wang,X., Zhao,Q. and Christensen,B.M. (2005) Identification and characterization of the fibrinogen-like domain of fibrinogen-related proteins in the mosquito, *Anopheles gambiae*, and the fruitfly, *Drosophila melanogaster*, genomes. *BMC Genomics*, 6, 1.
79. Dziarski,R. and Gupta,D. (2006) The peptidoglycan recognition proteins (PGRPs). *Genome Biol.*, 7, 1.
80. Bao,Y.-Y., Qu,L.-Y., Zhao,D. *et al.* (2013) The genome-and transcriptome-wide analysis of innate immunity in the brown planthopper, *Nilaparvata lugens*. *BMC Genomics*, 14, 1.
81. Blandin,S. and Levashina,E.A. (2004) Thioester-containing proteins and insect immunity. *Mol. Immunol.*, 40, 903–908.
82. Agaisse,H. and Perrimon,N. (2004) The roles of JAK/STAT signaling in *Drosophila* immune responses. *Immunol. Rev.*, 198, 72–82.
83. Lindsay,S.A. and Wasserman,S.A. (2014) Conventional and non-conventional *Drosophila* Toll signaling. *Dev. Comp. Immunol.*, 42, 16–24.
84. Myllymäki,H., Valanne,S. and Rämet,M. (2014) The *Drosophila* Imd signaling pathway. *J. Immunol.*, 192, 3455–3462.
85. Myllymäki,H. and Rämet,M. (2014) JAK/STAT pathway in Drosophila immunity. *Scand. J. Immunol.*, 79, 377–385.
86. Evans,J.D., Aronstein,K., Chen,Y.P. *et al.* (2006) Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol. Biol.*, 15, 645–656.
87. Benton,M.A., Pechmann,M., Frey,N. *et al.* (2016) *Toll* genes have an ancestral role in axis elongation. *Curr. Biol.*, 26, 1609–1615.
88. Viljakainen,L. (2015) Evolutionary genetics of insect innate immunity. *Brief. Funct. Genomics.*, 14, 407–412.
89. Kim,J.H., Min,J.S., Kang,J.S. *et al.* (2011) Comparison of the humoral and cellular immune responses between body and head lice following bacterial challenge. *Insect Biochem. Mol. Biol.*, 41, 332–339.
90. Nachappa,P., Levy,J. and Tamborindeguy,C. (2012) Transcriptome analyses of *Bactericera cockerelli* adults in response to *Candidatus Liberibacter solanacearum* infection. *Mol. Genet. Genomics*, 287, 803–817.
91. Mesquita,R.D., Vionette-Amaral,R.J., Lowenberger,C. *et al.* (2015) Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. *Proc. Natl Acad. Sci. U.S.A.*, 112, 14936–14941.
92. Chen,W., Hasegawa,D.K., Kaur,N. *et al.* (2016) 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.*, 14, 110.
93. Zug,R. and Hammerstein,P. (2012) Still a host of hosts for *Wolbachia*: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. *PLoS One*, 7, e38544.
94. Nakabachi,A., Yamashita,A., Toh,H. *et al.* (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science (80-.)*, 314, 267.

95. Hilgenboecker,K., Hammerstein,P., Schlattmann,P. *et al.* (2008) How many species are infected with *Wolbachia*?—a statistical analysis of current data. *FEMS Microbiol. Lett.*, 281, 215–220.
96. Saha,S., Hunter,W.B., Reese,J. *et al.* (2012) Survey of endosymbionts in the *Diaphorina citri* metagenome and assembly of a *Wolbachia* wDi draft genome. *PLoS One*, 7, e50067.
97. Sloan,D.B. and Moran,N.A. (2012) Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Mol. Biol. Evol.*, 29, 3781–3792.
98. Nakabachi,A., Ueoka,R., Oshima,K. *et al.* (2013) *Defensive Bacteriome Symbiont with a Drastically Reduced Genome*, 23, 1478–1484.
99. O'Shea,J.J., Schwartz,D.M., Villarino,A.V. *et al.* (2015) The JAK-STAT pathway: impact on human disease and therapeutic intervention*. *Annu. Rev. Med.*, 66, 311–328.
100. Zhang,L. and Gallo,R.L. (2016) Antimicrobial peptides. *Curr. Biol.*, 26, R14–R19.
101. Callewaert,L. and Michiels,C.W. (2010) Lysozymes in the animal kingdom. *J. Biosci.*, 35, 127–160.
102. Lavine,M.D. and Strand,M.R. (2002) Insect hemocytes and their role in immunity. *Insect Biochem. Mol. Biol.*, 32, 1295–1309.
103. Bordo,D., Djinovic,K. and Bolognesi,M. (1994) Conserved patterns in the Cu, Zn superoxide dismutase family. *J. Mol. Biol.*, 238, 366–386.
104. Parker,J.D., Parker,K.M., Sohal,B.H. *et al.* (2004) Decreased expression of Cu–Zn superoxide dismutase 1 in ants with extreme lifespan. *Proc. Natl Acad. Sci. U.S.A.*, 101, 3486–3489.
105. Colinet,D., Cazes,D., Belghazi,M. *et al.* (2011) Extracellular superoxide dismutase in insects characterization, function, and interspecific variation in parasitoid wasp venom. *J. Biol. Chem.*, 286, 40110–40121.
106. Kanost,M.R. and Jiang,H. (2015) Clip-domain serine proteases as immune factors in insect hemolymph. *Curr. Opin. Insect Sci.*, 11, 47–55.
107. Veillard,F., Troxler,L. and Reichhart,J.-M. (2016) *Drosophila melanogaster* clip-domain serine proteases: structure, function and regulation. *Biochimie*, 122, 255–269.
108. Chang,Y.-Y. and Neufeld,T.P. (2010) Autophagy takes flight in *Drosophila*. *FEBS Lett.*, 584, 1342–1349.
109. Malagoli,D., Abdalla,F.C., Cao,Y. *et al.* (2010) Autophagy and its physiological relevance in arthropods: current knowledge and perspectives. *Autophagy*, 6, 575–588.
110. Zirin,J. and Perrimon,N. (2010). Drosophila as a model system to study autophagy. *Semin. Immunopathol.*, 32, 363–372.
111. Altincicek,B., Gross,J. and Vilcinskas,A. (2008) Wounding-mediated gene expression and accelerated viviparous reproduction of the pea aphid *Acyrthosiphon pisum*. *Insect Mol. Biol.*, 17, 711–716.
112. Ratzka,C., Gross,R. and Feldhaar,H. (2012) Endosymbiont tolerance and control within insect hosts. *Insects*, 3, 553–572.
113. Ghildiyal,M. and Zamore,P.D. (2009) Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.*, 10, 94–108.
114. Hammond,S.M., Bernstein,E., Beach,D. *et al.* (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature*, 404, 293–296.
115. Pal-Bhadra,M., Bhadra,U. and Birchler,J.A. (2002) RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila*. *Mol. Cell*, 9, 315–327.
116. Liu,Q., Rand,T.A., Kalidas,S. *et al.* (2003) R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. *Science (80-.)*, 301, 1921–1925.
117. Aravin,A.A., Klenov,M.S., Vagin,V.V. *et al.* (2004) Dissection of a natural RNA silencing process in the *Drosophila melanogaster* germ line. *Mol. Cell. Biol.*, 24, 6742–6750.
118. Lee,Y.S., Nakahara,K., Pham,J.W. *et al.* (2004) Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, 117, 69–81.
119. Okamura,K., Ishizuka,A., Siomi,H. *et al.* (2004) Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev.*, 18, 1655–1666.
120. Saito,K., Ishizuka,A., Siomi,H. *et al.* (2005) Processing of pre-microRNAs by the Dicer-1-Loquacious complex in *Drosophila* cells. *PLoS Biol.*, 3, e235.
121. Scott,J.G., Michel,K., Bartholomay,L. *et al.* (2013) Towards the elements of successful insect RNAi. *J. Insect Physiol.*, 59, 1212–1221.
122. Christiaens,O. and Smagghe,G. (2014) The challenge of RNAi-mediated control of hemipterans. *Curr. Opin. Insect Sci.*, 6, 15–21.
123. Kola,V.S.R., Renuka,P., Madhav,M.S. *et al.* (2015) Key enzymes and proteins of crop insects as candidate for RNAi based gene silencing. *Front. Physiol.*, 6, 119.
124. Bernstein,E., Caudy,A.A., Hammond,S.M. *et al.* (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409, 363–366.
125. Knight,S.W. and Bass,B.L. (2001) A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science (80-.)*, 293, 2269–2271.
126. Lee,Y., Ahn,C., Han,J. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425, 415–419.
127. Leuschner,P.J., Obernosterer,G. and Martinez,J. (2005) MicroRNAs: *Loquacious* speaks out. *Curr. Biol.*, 15, R603–R605.
128. Yeom,K.H., Lee,Y., Han,J. *et al.* (2006) Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. *Nucleic Acids Res.*, 34, 4622–4629.
129. Liu,X., Jiang,F., Kalidas,S. *et al.* (2006) Dicer-2 and R2D2 coordinately bind siRNA to promote assembly of the siRISC complexes. *RNA*, 12, 1514–1520.
130. Liu,X., Park,J.K., Jiang,F. *et al.* (2007) Dicer-1, but not Loquacious, is critical for assembly of miRNA-induced silencing complexes. *RNA*, 13, 2324–2329.
131. Okamura,K., Robine,N., Liu,Y. *et al.* (2011) R2D2 organizes small regulatory RNA pathways in *Drosophila*. *Mol. Cell Biol.*, 31, 884–896.
132. Taning,C.N.T., Andrade,E.C., Hunter,W.B. *et al.* (2016) Asian citrus psyllid RNAi pathway – RNAi evidence. *Sci. Rep.*, 6, 38082.
133. Czech,B., Malone,C.D., Zhou,R. *et al.* (2008) An endogenous small interfering RNA pathway in *Drosophila*. *Nature*, 453, 798–802.
134. Haac,M.E., Anderson,M.A., Eggleston,H. *et al.* (2015) The hub protein loquacious connects the microRNA and short interfering RNA pathways in mosquitoes. *Nucleic Acids Res.*, 43, 3688–3700.

135. Hammond,S.M., Boettcher,S., Caudy,A.A. *et al*. (2001) Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science (80-.)*, 293, 1146–1150.
136. Liu,J., Carmell,M.A., Rivas,F.V. *et al*. (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science (80-.)*, 305, 1437–1441.
137. Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, 431, 343–349.
138. Verdel,A., Jia,S., Gerber,S. *et al*. (2004) RNAi-mediated targeting of heterochromatin by the RITS complex. *Science (80-.)*, 303, 672–676.
139. Kennerdell,J.R., Yamaguchi,S. and Carthew,R.W. (2002) RNAi is activated during *Drosophila* oocyte maturation in a manner dependent on *aubergine* and *spindle-E*. *Genes Dev*., 16, 1884–1889.
140. Jaubert-Possamai,S., Rispe,C., Tanguy,S. *et al*. (2010) Expansion of the miRNA pathway in the hemipteran insect *Acyrthosiphon pisum*. *Mol. Biol. Evol*., 27, 979–987.
141. Zhang,Y., Wang,Y., Wang,L. *et al*. (2016) Knockdown of NADPH-cytochrome P450 reductase results in reduced resistance to buprofezin in the small brown planthopper, *Laodelphax striatellus* (fall{é}n). *Pestic. Biochem. Physiol*., 127, 21–27.
142. McLean,K.J., Luciakova,D., Belcher,J. *et al*. (2015) *Monooxygenase, Peroxidase and Peroxygenase Properties and Mechanisms of Cytochrome P450*. In: Hrycay,E.G. and Bandiera,S.M. (eds). New York: Springer, pp. 299–317.
143. Feyereisen,R. (2006) Evolution of insect P450. *Biochem. Soc. Trans*., 34, 1252–1255.
144. Schama,R., Pedrini,N., Juárez,M.P. *et al*. (2016) *Rhodnius prolixus* supergene families of enzymes potentially associated with insecticide resistance. *Insect Biochem. Mol. Biol*., 69, 91–104.
145. Richards,S., Gibbs,R.A., Weinstock,G.M. *et al*. (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature*, 452, 949–955.
146. Saha,S., Hunter,W., Mueller,L. *et al*. (2017) *Diaphorina citri* genome assembly Diaci 1.9. Ag Data Commons. http://dx.doi.org/10.15482/USDA.ADC/1342727.