Chromosomal length reference assembly for *Diaphorina citri* using single-molecule sequencing and Hi-C proximity ligation with manually curated genes in developmental, structural and immune pathways

Prashant S. Hosmani[1*], Mirella Flores-Gonzalez[1*], Teresa Shippy[2], Chad Vosburg[3,4], Crissy Massimino[4], Will Tank[2], Max Reynolds[4], Blessy Tamayo[4], Sherry Miller[2], Jordan Norus[4], Kyle Kercher[4], Bec Grace[4], Margaryta Jernigan[4], Doug Harper[4], Sam Adkins[4], Yesmarie DeLaFlor[4], Thomson Paris[5], Sara Vandervoort[2], Rebekah Adams[7], Seantel Norman[7], Jessica Ventura[7], Michael Perry[7], Matthew Weirauch[6], Josh Benoit[7], Wayne B. Hunter[5], Helen Wiersma-Koch[4], Tom D'elia[4], Susan Brown[2], Lukas A. Mueller[1] and Surya Saha[1]

[1]Boyce Thompson Institute, Ithaca, NY 14853; [2] Division of Biology, Kansas State University, Manhattan, KS 66506; [3] Department of Plant Pathology and Environmental Microbiology, University Park, PA 16802; [4]Indian River State College. Department of Biological Sciences. 3209 Virginia Ave, Fort Pierce, Fl, 34981; [5]USDA-ARS, U.S. Horticultural Research Laboratory, Fort Pierce, FL 34945; [6]The Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA; [7]Department of Biological Sciences, University of Cincinnati, Cincinnati, OH, USA

* Both authors contributed equally to this publication

Abstract

Hemipterans include some of the most important insect pests of agricultural systems and vectors of plant pathogens. The vector, *Diaphorina citri* (Asian citrus psyllid) belonging to the Psylloidae superfamily, is the primary target of approaches to stop the spread of the pathogen *Ca.* Liberibacter asiaticus that causes Huanglongbing or citrus greening disease. High quality genomic resources enable rapid functional discovery that can target disease transmission and control. The previous psyllid genome (Diaci v1.1) available in NCBI is missing 25% of the single copy markers conserved in other Hemipterans. Manual genome curation helped to identify a significant number of genome anomalies including misassemblies and missing genes. We present an improved and highly contiguous *de novo* assembly based on PacBio long reads followed by Dovetail Chicago and Hi-C based scaffolding. The current assembly (Diaci v3) has 13 chromosomal length scaffolds with a genome size of 475 Mb. This is the first report of a chromosomal length assembly in the Hemiptera order according to our knowledge. Full-length cDNA transcripts were sequenced with PacBio Iso-Seq technology from diseased and healthy tissue at multiple life stages. Iso-Seq along with diverse Illumina RNA-Seq expression data were used to predict 19,049 protein-coding genes in psyllid using MAKER annotation pipeline. We also generated a genome independent transcriptome with a comprehensive catalog of all genes in the psyllid.

Gene-targeting technologies like RNAi, antisense oligos and CRISPR require accurate annotation of genes. Lack of closely related and well characterized model organisms coupled with the diversity of insect genomes impacts the quality of predicted gene models. We used the improved genomic resources to create a high-quality manually curated gene set for developmental, structural and immune pathways. All

resources are available on https://citrusgreening.org/, a portal for all omics resources for the citrus greening disease research community. The high quality ACP genome assembly, annotation based on transcriptomics evidence, manual curation of critical pathways and a genome independent *de novo* transcriptome will provide a foundation for comparative analysis among genomes of agricultural pests and vectors in the Hemiptera.

1. Introduction

Citrusgreening or Huanglongbing (HLB) is a tritrophic disease complex involving citrus host, the Asian citrus psyllid (ACP, *Diaphorina citri* Kuwayama) vector and a phloem restricted bacterial pathogen *Ca.* Liberibacter asiaticus (CLas). HLB is the most devastating of all citrus diseases, and there is currently no effective control strategy.

The psyllid is a highly effective vector for the bacterial pathogen, CLas. The bacteria are transmitted through a circulatory mechanism through the psyllid during which they are multiplying within the insect [1]. When an infected female oviposits on a citrus leaf or flush, the nymphs feed on the phloem in a part of the leaf that already contains CLas as a result of previous feeding by the infected female. The infected nymphs feed and rapidly spread the pathogen as they move to other parts of the tree and CLas eventually moves to the roots which leads to a systemic infection in the tree typically before any symptoms are manifested [2]. By the time a tree is identified as infected, it has already been the source of inoculum for the surrounding trees in the grove.

High throughput methods in genomics, transcriptomics and proteomics coupled with recent developments in single cell technologies have allowed us to better understand processes at the cellular, organismal and ecological levels. However, a high quality genome assembly and annotation of both coding and non-coding genes is critical to the success of omics assays as well as interdiction methods based on gene targets such as RNAi and CRISPR. Genome sequencing of insects is challenging for species that are sexually reproducing and have low biomass [3]. This requires the pooling of multiple individuals to get sufficient tissue for both genomic and transcriptomic sequencing. This increases the heterozygosity in the DNA sample and increases the complexity of the genome assembly besides the issues introduced by the repetitive regions found in arthropod genomes [4].

The short read based assembly for the psyllid (Diaci v1.1) with a contig N50 of 34.4Kb and scaffold N50 of 109.8Kb has been an important resource for the citrus greening research community [5]. A BUSCO [6] analysis identified a significant number of conserved single-copy markers either missing (24.9%) in this assembly. A community-driven manual annotation project [7] has also identified a number of misassemblies such as tandem duplications and a high degree of fragmentation in the genome.

We have generated 36.2Gb of PacBio long reads from 41 SMRT cells with an 80X coverage of the 480Mb psyllid genome. The Canu assembler was used to create an PacBio only *de novo* assembly with 38,263 unitigs with a conitg N50 of 115.8kb. The assembly was followed by Dovetail Chicago and chromatin

interaction based Hi-C scaffolding to get 13 chromosomes (Diaci v3) with a total length of 442Mb and a N50 of 40,584kb. High-resolution short read expression data from CLas infected and healthy psyllids across multiple life stages, tissues and sexes and a variety of citrus hosts was used as evidence to annotate this assembly. PacBio Iso-Seq long read data from CLas infected and healthy adult and nymphs was also used in conjunction with Illumina data in the MAKER [8] pipeline to create the Official Gene Set v3 (OGSv3). This version of the annotation contains 19,049 protein coding genes with 21,346 isoforms with an average of 7.2 exons per transcript. In addition to the genome based OGSv3, we also generated genome independent transcriptome using all the transcript evidence to create a comprehensive catalog of all genes in the psyllid.

Gene sets involved in chitin metabolism, cuticle formation, segmentation and segmental identity, signal transduction, chromatin remodeling, phototransduction, circadian rhythm, carbohydrate metabolism, melanization and spermatogenesis were manually curated using Apollo annotation editor [9].

2. Results and Discussion

2.1. Genome assembly

Genome assembly of organisms such as ACP with low biomass is challenging due to the additional heterozygosity introduced by pooling together multiple individuals to extract sufficient material for library preparation. This is exacerbated by non-clonal and sexual reproduction even when colonies have been maintained for many generations to encourage inbreeding. These were the primary reasons for the fragmentation and misassemblies in version 1.1 of the ACP genome besides the fact that it was assembled with Illumina short reads even though mate pair libraries with an insert size of up to 20Kb were used for improving the contiguity.

In order to improve the assembly of this important insect vector, we have used PacBio long read sequencing and *de novo* assembly followed by two rounds of scaffolding. Intermediate range scaffolding was performed with Dovetail Chicago [10] method followed by long range scaffolding with Hi-C [11]. Scaffolding with the HiRise assembler [10] based on paired-end reads from Chicago libraries with a 120X coverage of the ACP genome increased the N50 from 28Kb to 383Kb while reducing the L50 from 6700 scaffolds to 383 scaffolds. This step resulted in 12,369 joins and corrected 48 misassemblies in the unitig set. The insert size distribution of the Chicago paired-end reads extended up to 250Kb. In comparison, the insert size distribution of the paired-end reads from the Hi-C library stretched up to 3Mb with a coverage of 232X of the genome. The HiRise assembler made 1003 joins in the Chicago assembly and increased the scaffold N50 from 29Kb to 26.6Mb.

13,320 scaffolds from the Chicago assembly ordered and oriented in 12 chromosomal length scaffolds representing the 12 putative chromosomes [12] in the ACP genome. We expected a high rate of duplication in the 24,916 unplaced scaffolds in this assembly due to the multiple individuals pooled to obtain the long read data. So we used Redundans [13]] to split the unplaced scaffolds into 1244 non-redundant and 23,672 duplicated scaffolds. The 1244 unique scaffolds were joined with 1000 Ns

separating adjacent scaffolds to create chromosome 00 in the Diaci version 3.0 reference assembly. The 23,672 duplicated scaffolds may represent alternative loci in the ACP population and are reported as alternate (ALT) contigs.

Table 1: Genome assembly statistics

Genome assembly version	**Diaci v1.1**	**Diaci v2**	**Diaci v3**
Size	485.7Mb	498.8Mb	473.9Mb
Scaffold N50	109.8Kb	749.5Kb	40.5Mb
Max scaffold length	1.0Mb	4.2Mb	50.3Mb
Contig N50	34.4Kb	127.6Kb	38.6Kb
Max contig length	431.2Kb	2.1Mb	409.7Kb
Count of scaffolds	161,988	1,906	13 chromosomes (1244 unplaced)
Number of N's	19.3Mb	4.5Mb	13.4Mb
% complete BUSCO	65.9	75.9	88.3
Repeat % (RM)	26.37%	31.96%	30.23%

2.2. Annotation

We have identified more repeats in the v3 genome assembly compared to previous published v1.1 assembly (Table 1). This can largely be attributed to the use of long-read PacBio sequences in the v3 assembly. Annotation of genes was carried out after masking repetitive elements. In total, we have identified 19,049 protein-coding genes with 21,345 alternatively spliced isoforms with MAKER annotation pipeline (Table 2). Use of PacBio Iso-Seq enabled accurate predictions of complete gene models and their isoforms. Predicted protein-coding genes were longer compared to OGSv1 gene models with more exons predicted per gene. Curated genes represent manually refined models based on evidence from orthology

and expression data using Apollo annotation editor [9]. Curation was performed by distributed groups based on curated workflow previously described in Hosmani et al. [14]. Detailed reports of curated gene families in pathways of interest are described in the following sections and supplementary notes. Overall, curated genes are longer with more exons compared to predicted genes. We also observe fewer non-canonical 5' and 3' splice sites in the curated genes and OGSv3 gene models, indicating higher quality of genes models compared to previous annotations.

Table 2: Genome annotation metrics for official gene sets (OGS) from each ACP genome assembly along with subset of curated genes.

Annotation set	**OGSv1**	**OGSv2**	**OGSv3**	**Curated**
No. of genes	19,311	20,793	19,049	811
No. of transcripts	20,966	25,292	21,345	916
No. of Exons Per transcript	5.42	7.06	7.29	7.87
Avg. transcript length (bp)	1,317	1,944	2,034	2,503
Avg. exon length (bp)	243	275	279	318
non-canonical splice sites	6.05%	3.13%	2.47%	1.91%

2.4 *De novo* transcriptome

In order to identify a comprehensive set of transcripts, we built a genome independent *D. citri de novo* transcriptome with long and short read data from from a range of experimental conditions (Supp. Table 3). It resulted in 40,637 genes and 60,261 transcripts with a contig N50 of 3,657bp and an average length of 1736.1 bp. Identifiers were determined according the evidence used to create the transcript (DcDTr for RNA-Seq transcripts and DcDTi for Iso-Seq transcripts) with 41,457 and 18,804 transcripts respectively. BUSCO results (Supplementary Table 1) indicate 79.9% complete single-copy orthologs and 20% missing single-copy orthologs for the hemiptera data set showing that the gene space is retained in the *de novo* transcriptome.

2.3. Transcription factor prediction

We identified a total of 1,015 putative transcription factors (TFs) in the *D. citri* genome (Figure 1). This value is similar to some beetles, such as *Anoplophora glabripennis* (1,397) and *Hypothenemus hampei*

(1,148), but substantially greater than others, such as *Tribolium castaneum* (788), *Nicrophorus vespilloides* (744), and *Dendroctonus ponderosae* (683). Of the 1,015 *D. citri* TFs, we could infer motifs for 337 (33%) (Supplementary data), mostly based on DNA binding specificity data from *D. melanogaster* (228 TFs), but also from species as distant as human (77 TFs) and mouse (12 TFs). Many of the largest TF families have inferred motifs for a substantial proportion of their TFs, including Homeodomain (93 of 104, 89%), bHLH (57 of 63, 90%) and Forkhead box (24 of 31, 77%). As expected, the largest gap in binding specificity knowledge is for C2H2 zinc fingers (only 39 of 377, ~10%), which evolve quickly by shuffling their many zinc finger arrays.

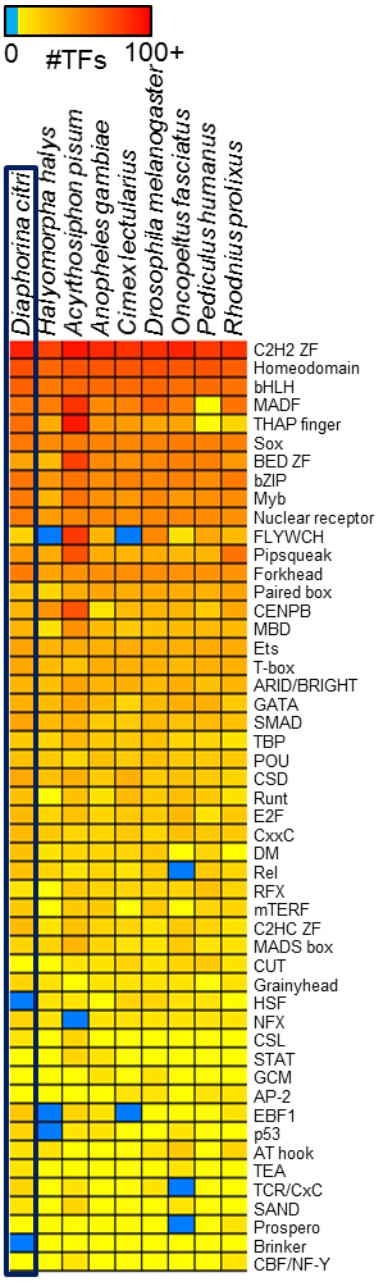

Figure 1. Distribution of transcription factor families across insect genomes. Heatmap depicting the abundance of transcription factor (TF) families across a collection of insect genomes. Each entry indicates the number of TF genes for the given family in the given genome, based on presence of DNA binding domains. Color key is depicted at the top (light blue means the TF family is completely absent). Color scale is log (base 2). *Diaphorina citri* is boxed.

2.5. Pathway-based manual curation

We also manually annotated several new sets of genes in genome v3, focusing our efforts on gene families or pathways of interest to the *D. citri* research community or that have been targeted for RNAi-based pest control in other insects [15–17].

2.5.1. Development

We annotated *D. citri* orthologs of genes important for development in other insects. This includes genes involved in segmentation, signal transduction and segmental identity.

Segmentation

During embryonic development, the insect embryo is divided into segments. The process of segmentation differs somewhat between insects, but the genes involved are fairly well conserved [18]. Out of 32 genes involved in Drosophila segmentation, we identified and annotated 23 in *D. citri* (Supplementary Report 1). Most of the differences from Drosophila were expected based on observations in other hemipterans.

The segmentation genes are traditionally grouped into categories based on their function in Drosophila. *D. citri* has all of the maternal effect genes expected for a hemipteran, including *caudal*, *nanos*, *dorsal*, and one *TGFalpha*. As expected, *bicoid* and *oskar*, were not found. Several gap genes are absent in *D. citri*, including *giant*, *buttonhead* and *otd1*. These genes are also missing in the pea aphid genome. However, *D. citri* does have a copy of *huckebein*, which was reported absent in the pea aphid [19]. All of the pair-rule genes are present in the ACP as one-to-one orthologs. The segment polarity genes include members of the Wnt and Hedgehog signal transduction pathways as described below, as well as the transcription factor genes *engrailed* (*en*) and *gooseberry* (*gsb*). *en* is present in *D. citri*, but we were unable to find the related gene *invected*. As in other insects, *gsb* is found in a cluster with its homolog *gooseberry-neuro*.

Wnt and other signaling pathways

The Wnt signaling pathway is critical for segment polarity determination during segmentation and also plays many other roles during development. There are multiple Wnt ligands that activate the downstream signaling components in specific contexts. Particular ligand genes have been lost in various insect lineages [20] and *D. citri* is no exception. We were unable to find orthologs of *Wnt8/D* or *Wnt9* (Supplementary Report 2), which are also absent in pea aphids. We also did not find *Wnt16*, which has been lost from the Holometabola but is found in several other hemipterans [19,21]. Conversely, *Wnt6* (contrary to a previous report [22] and *Wnt10* are present in the ACP genome but apparently missing in pea aphid [19]. The key downstream components of Wnt signaling are all present in the *D. citri* genome.

We also annotated 13 genes from the Hedgehog pathway, 16 from the Notch pathway, and six members of the insulin signaling pathway. All of these pathways seem to be highly conserved.

Hox genes

Hox genes specify the identity of regions along the body axis. Notably, they are usually arranged in a cluster in an order paralleling that of their functional domains along the anterior-posterior body axis, although breaks in the cluster have occurred in several lineages [23]. *D. citri* has the full complement of ten Hox genes, although *labial*, which was present in previous versions of the genome, is missing in v3.0 (Supplementary Report 3). The *D. citri* Hox cluster is split into two parts with the breakpoint between *Dfd* and *Scr*. Both clusters are on the same chromosome about 6 Mb apart.

2.5.2. Immune Response

Immune system genes were a primary focus of manual annotation in the first version of the *D. citri genome* (Saha et al 2017). Several groups of immune genes have been re-annotated in genome v3 (see Improving OGSv1 curated gene models) with substantial improvements in gene count and completeness (Table 3).

Table 3. Gene counts of manually curated immune system genes in genome v1.1 versus genome v3. The asterisk (*) indicates a case where one lysozyme gene was missing from genome v3 but was found in the ALT contigs.

Pathway / Gene Family	Genes in v1.1	Genes in v3.0
Toll Receptors	5	5
Toll Receptor related genes	3	3
JAK/STAT pathway	3	3
C-type Lectins	10	7
Lysozymes	5	5*
Superoxide dismutases	4	4
CLIP	11	10
Autophagy	15	15
Cathepsins and Cysteine Proteases	34	23
Total	90	75

In addition to improving previously annotated immune genes, we have also annotated genes involved in melanization. Besides its role in pigmentation, the melanization pathway plays important roles in wound healing and defense against pathogens [24]. We annotated 12 genes from the melanization pathway including two laccases and two tyrosinase prophenoloxidases (PPO). We also annotated the members of the yellow gene family, several of which have been implicated in melanization in other insects although their precise role is still not well understood [25]. *D. citri* has nine yellow genes, including an apparent duplication of yellow-y (Supplementary Report 4). Interestingly, one of these paralogs is expressed primarily in the egg and nymph stages, while the other is expressed mainly in adults. We also noted apparent differences in expression of some of the yellow genes between CLas-infected and uninfected psyllids that might warrant further investigation.

2.5.3. Metabolic and Cellular Functions

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

Carbohydrate Metabolism

Although trehalose is the primary blood sugar in insects, glucose metabolism is also important [26]. Breakdown of trehalose produces glucose, which is then further metabolized by glycolysis [27]. The synthesis of glucose and trehalose (gluconeogenesis and trehaloneogenesis, respectively) follow the same pathway through the production of glucose-6-phosphate, before diverging. Some insects synthesize both sugars, with glucose synthesis occurring primarily in neural cells [26]. We annotated 21 genes involved in glycolysis and gluconeogenesis (Supplementary Report 5). The genes in both pathways are highly conserved, although there are few differences in gene copy number. ACP genome has only one copy of the hexokinase gene, whereas most other insects have multiple copies. Conversely, it has two copies of phosphoenolpyruvate carboxykinase instead of a single copy as found in many other insects. We did not find a glucose-6-phosphatase gene in *D. citri*, suggesting that trehalose is the end product of this pathway in the ACP.

Chitin Metabolism

Chitin is a major component of the insect cuticle and properly coordinated synthesis and breakdown of chitin are essential for growth and survival. We annotated 18 orthologs of genes involved in chitin metabolism (Supplementary Report 6). Like most hemipterans, ACP has fewer chitin metabolism genes than do holometabolous insects. This reduction is likely due to the fact that hemipterans do not undergo metamorphosis and lack a gut peritrophic membrane (another structure containing chitin) [28]. Consistent with the absence of a peritrophic membrane, *D. citri* lacks the chitin synthase gene (CHS2) that is specifically expressed in the peritrophic membrane in holometabolous insects [29,30]. *D. citri* does have apparent duplications of two genes involved in chitin metabolism: UDP-N-acetylglucosamine pyrophosphorylase (UAP) and Chitinase 10 (a group II chitinase).

vATPase

Vacuolar ATP synthase (V-ATPase) regulates the acidity of various organelles by using energy from ATP to translocating protons across a membrane [31]. We annotated the genes encoding all 13 subunits of V-ATPase (Supplementary Report 7). Gene copy number for each subunit varies between insects. The ACP genome has two copies of the V0-a, V1-D and V1-G subunit genes and one copy of each of the others, which is consistent with the gene copy number reported in pea aphids.

Chromatin Remodeling

We annotated 27 genes from six chromatin remodeling complexes (Supplementary Report 8). Overall, the gene content of this family closely matches that of Drosophila. However, there is a duplication of Mi-2 that has not been reported in other insects.

2.5.4. Environmental/Sensory

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

Circadian Rhythm

We annotated 29 orthologs of genes involved in circadian rhythm in other insects (Supplementary Report 9). All components of the pathway are present in the genome. ACP has both types of cryptochrome photoreceptor genes (cry1 and cry2), consistent with the apparent ancestral state of this pathway [32–34].

Phototransduction

We annotated 18 genes in the phototransduction pathway (Supplementary Report 10). *D. citri* has four copies of opsin, the same number as the honeybee but fewer than the seven found in Drosophila [35]. Interestingly, ACP has a gene apparently encoding rhodopsin kinase, an enzyme found in Drosophila [36] but not Tribolium, honeybee or pea aphid.

2.5.5. Reproduction

Genes involved in reproduction, particularly spermatogenesis, have been targeted for use in a method of pest control called Sterile Insect Technique (SIT). Male insects sterilized by radiation, chemicals, or in this case, RNAi-based knockdown of spermatogenesis genes, are released into the environment to hopefully out-compete fertile males in the natural population [37,38]. We annotated 20 *D. citri* orthologs of genes (Supplementary Report 11) that have been used in other insects for SIT [39–41].

2.6. Improving OGSv1 curated gene models

Gene models which had been manually annotated on the v1.1 genome [5] were mapped to improved v3 genome and manually corrected as needed. Of 159 genes examined, 27 were removed from the Official Gene Set. All but one of these were artifactual duplicates resulting from misassemblies in genome v1.1. Most of these duplications are no longer present in genome v3.0, due to long-read based assembly and removal of duplicate scaffolds as previously described. One heat shock gene was removed because it was located in a region of genome v1.1 determined to be of endosymbiont origin and was shown by BLAST to be identical to the molecular chaperone HtpG from Candidatus Profftella armatura. Of 23 annotated gene pathways or families examined, 8 have a reduced, more accurate gene count after validation in genome v3.0 (Supplementary Table 2). We were also able to more accurately annotate isoforms in genome v3.0. Most of the 12 isoforms that were discarded had mapped differently to the v1.1 genome because of assembly-related exon duplications, thus giving the false appearance of isoforms. Although our efforts were focused on assessing existing models, we did identify and annotate two previously unrecognized isoforms in genome v3.

We were able to improve almost all of the v1.1 models in some way. In many cases, the old models had only partial open reading frames (ORFs) and we were able to create models with complete ORFs. We were also able to add 5' and 3' UTRs to many models. The Iso-Seq transcripts were extremely helpful in creating longer models with a high level of confidence.

3. Conclusion

Undergraduate and in some cases, graduate students from various institutions were involved in this community curation effort. They were trained, individually or in groups and remotely or in-person, in gene curation as a part of this initiative. The students were mentored by senior peer student annotators as well as expert annotators. We had regular interactions with scientists from the insect genomics community working on genes and pathways under curation. Implementation of rigorous and consistent annotation practices across a virtual team of highly diverse annotators required project management tools and regular video conferences followed by extensive documentation that was continuously updated in response to user feedback. Standard operating procedures have been reported in as a guide for other annotation communities [14] to implement similar programs.

Hemipterans include some of the most important insect pests of agricultural systems and vectors of plant pathogens. The vector, *Diaphorina citri* (Asian citrus psyllid) belonging to the Psylloidae superfamily, is the primary target of approaches to stop the spread of the pathogen *Ca.* Liberibacter asiaticus that causes Huanglongbing or citrus greening disease. We report a significantly improved and chromosomal length genome assembly for the *D. citri* genome and the corresponding official gene set (OGSv3) which includes 811 curated genes and 19,049 protein coding genes. This is the first report of a chromosomal length assembly in the Hemiptera order according to our knowledge. All resources are available on https://citrusgreening.org/, a portal for all omics resources for the citrus greening disease research community [42]. The high quality ACP genome assembly, annotation based on transcriptomics evidence,

manual curation of critical pathways and a genome independent *de novo* transcriptome will provide a foundation for comparative analysis among genomes of agricultural pests and vectors in the Hemiptera.

Funding

All open-access fees, student annotators and post-docs were funded through USDA-NIFA grant 2015-70016-23028.

Acknowledgements

Kascha Bohnenblust from Kansas State University assisted with the organization of meetings for the community curation effort. Michelle Coleman at Kansas State University extracted the DNA used for PacBio sequencing.

Methods

1. Genome assembly

The DNA for PacBio sequencing and Dovetail Chicago libraries as well as tissue for Hi-C was sourced from a *D. citri* colony at the U. S. Horticultural Research Laboratory, USDA, Ft. Pierce, Florida. High molecular weight DNA was extracted from ACP adults using the BioNano extraction kit [43] at Kansas State University for PacBio sequencing. DNA from a single male ACP was used to generate the Dovetail Chicago library which was sequenced and used for both scaffolding and error correction.

The PacBio sequencing was done on the RSII instrument to produce 36.1Gb of long reads with an average length of 7.2Kb. The Canu [44] assembler was used to correct 40X of the longest CLR reads which were trimmed and used for the final assembly (-utgOvlErrorRate=0.013). The 38,263 unitigs produced by the Canu assembly with a contig N50 of 28Kb were selected for scaffolding. Dovetail Chicago paired-end reads (207 million) from a single male psyllid were used to perform 12,369 joins and scaffold these unitigs into 25,942 scaffolds. This round of scaffolding added 12.3Mb of Ns to the assembly. The Chicago scaffolded assembly was passed through another round of scaffolding with 388 million paired-end Hi-C reads. The Hi-C scaffolding reduced the number of scaffolds to 24,943 with a scaffold N50 of 26.6Mb containing 12.4Mb of Ns. This assembly consisted of 13 scaffolds with 441.Mb of the genome representing the 13 chromosomes [12]. There were also 24,930 unplaced scaffolds that were not scaffolded into the putative chromosomes. This unplaced set was comprised of short scaffolds (scaffold N50 15Kb). We reduced the duplication within the unplaced sequences by applying redundans [13] at a threshold of 80% identity and coverage. This split the unplaced set into 1244 unique scaffolds (33.2Mb, N50 28.1Kb) and 23,672 duplicated scaffolds (201Mb, scaffold N50 13.7Kb). The duplicated scaffolds are reported as alternate (ALT) contigs for the Diaci version 3.0 assembly. The 1244 unique scaffolds were ordered based on length and joined with 1000 Ns separating adjacent scaffolds to create chromosome 00 of length 344.5Mb. We performed two rounds of error corrections with pilon [45] using Illumina reads from the single male psyllid individual. We opted to use this data set instead of short read data from multiple individuals to avoid introducing artificial heterozygosity into the genome assembly. Pilon was optimized to only correct regions of the genome where the change was supported by more than 90% of the aligned bases at that position (--fix bases --diploid --mindepth 0.9). We also performed one round of error correction with pilon and Illumina RNA-Seq to polish the genic regions of the assembly (--unpaired --fix bases –diploid) based on unspliced alignments to the genome.

2. Iso-Seq

PacBio Iso-Seq was used to generate high-quality full-length transcripts and characterize the transcriptome with 4 SMRT cells of long reads each from healthy adults, CLas infected adults and mixed healthy and CLas infected nymphs. Using the Iso-Seq2 bioinformatics pipeline provided by SMRTlink v4.0

software, we generated a comprehensive set of genome independent isoforms from adult and nymph tissue. We also performed three rounds of error corrections with pilon [45] using Illumina reads.

3. Automated predictions of protein-coding genes

Repeat library for the Diaci v3 genome was constructed using RepeatModeler [46]. Repeat library was screened for known protein association with ProtExcluder [47] based on similarity with proteins obtained from Swiss-Prot (Arthropoda). Resulting repeat library was used to mask genome with RepeatMasker [46]. Repeat annotation is available on citrusgreening.org FTP. Repeat masked genome was used to predict protein-coding genes in the genome.

Protein-coding genes were predicted on v3 of *Diaphorina citri* genome through iterative process within MAKER (v3) annotation pipeline [8]. For homology evidence, manually annotated proteins of Arthropoda [6656] were downloaded from Swiss-Prot [48]. Expression evidence was obtained through multiple sources. RNA-Seq Data generated is listed in supplementary table 3. Publicly available RNA-Seq datasets were obtained from NCBI SRA database (Supplementary table 3). All the RNA-Seq data was mapped to the genome using HISAT2 [49] and transcriptome was assembled with StringTie [50]. Independently, high-quality PacBio Iso-Seq transcripts were mapped to the genome and clustered through Cupcake-ToFU clustering (https://github.com/Magdoll/cDNA_Cupcake). Transcriptomes obtained through RNA-Seq and Iso-Seq were processed with Mikado pipeline [51] for refining the transcriptome.

Ab initio gene predictions were performed with Augustus [52] and SNAP [53] gene predictors. Augustus was trained with available RNA-Seq data within BRAKER1 [54]. SNAP was trained iteratively within MAKER pipeline based on MAKER guidelines. The gene predictions were supplied to MAKER along with expression and homology evidence which was run with default parameters. The Mikado refined transcriptome was passed as a predictor (pred_gff). Gene identifiers were assigned based on the genomic location. For example, the Dcitr01g01000.1.1 gene consists of five letter species identifier (Dcitr), two digit scaffold/chromosome number (01), a spacer for protein-coding gene (g), a unique five digit ID for the gene (01000) followed by version number (.1) and isoform number (.1). Consecutive genes were added with spacing of 10 to allow addition of novel genes in the future.

2. Manual curation and OGSv3

Manual curation of automated predictions was carried out using the Apollo annotation editor [9] plugin for Jbrowse genome browser [55]. The Apollo instance for the v3 genome is hosted at citrusgreening.org [42] along with other genomic resources. All the evidence used for automated annotation was added as tracks on Jbrowse/Apollo. Other tracks to assist in accurate evidence-based curation included proteins from related insects mapped with exonerate. All publicly available RNA-Seq datasets were mapped as quantitative tracks. Manual curation was performed following the workflow previously described in [14] and modifications to the workflow, if any, are described in the individual pathway reports. Official gene set 3 (OGSv3) consists of predicted gene models from the MAKER annotation pipeline and manually

curated genes from the Apollo annotation editor. Before merging curated genes to create OGSv3, overlapping genes from the automatic predictions were removed.

5. *De novo* transcriptome

The protocol used to generate the *de novo* transcriptome included preparing the Iso-Seq data as described above, generating a *de novo* transcriptome using short read RNA-Seq data and finally merging both datasets. We used trinity [56] for all *de novo* assemblies. The resulting dataset was filtered using Pfam domains [57] and TransDecoder [58] to find coding regions within those transcripts that match to a Pfam domain. Resulting Iso-Seq dataset described previously was merged together with filtered short RNA-seq transcriptome. This merged set was clustered with CD-HIT (cd-hit-est) [59], which compared nucleotide datasets and identified the sequences in *de novo* short read Rna-Seq subset that are similar and to Iso-Seq subset above a threshold of 75% sequence identity. In order to reduce contamination, we removed endosymbionts, archaea, viral and bacteria sequences using BLAST [60]. Filtering was also done using Trembl insecta. Finally, the last round of filtering was done using *D. citri* genome V3.0 using GMAP [61] and cupcake (https://github.com/Magdoll/cDNA_Cupcake) in order to retain only *D. citri* transcripts and to remove redundancy. As a result of that process we obtained 60,261 transcripts. An identifier was chosen according the source raw data: DcDT for RNA-Seq transcripts and DcDi for Iso-Seq transcripts.

6. Transcription factor prediction

We identified likely transcription factors (TFs) by scanning the amino acid sequences of predicted protein coding genes for putative DNA binding domains (DBDs), and when possible, we predicted the DNA binding specificity of each TF using the procedures described in Weirauch *et al.* [62]. Briefly, we scanned all protein sequences for putative DBDs using the 81 Pfam [57] models listed in Weirauch and Hughes [63] and the HMMER tool [64], with the recommended detection thresholds of per-sequence Eval < 0.01 and per-domain conditional Eval < 0.01. Each protein was classified into a family based on its DBDs and their order in the protein sequence (e.g., bZIPx1, AP2x2, Homeodomain+Pou). We then aligned the resulting DBD sequences within each family using clustalOmega [65], with default settings. For protein pairs with multiple DBDs, each DBD was aligned separately. From these alignments, we calculated the sequence identity of all DBD sequence pairs (i.e. the percent of AA residues that are exactly the same across all positions in the alignment). Using previously established sequence identify thresholds for each family [62], we mapped the predicted DNA binding specificities by simple transfer. For example, the DBD of Dcitr04g16960.1.1 is 98% identical to the *Drosophila melanogaster* 'oc' (FBgn0004102) TF. Since the DNA binding specificity of 'oc' has already been experimentally determined, and the cutoff for the Homeodomain family of TFs is 70%, we can infer that Dcitr04g16960.1.1 will have the same binding specificity as 'oc' (Supplementary data).

Supplementary data

Table 1. Busco table (genome and annotation)

BUSCO	**Hemiptera complete**	**Duplicated**	**Fragmented**	**Missing**
Diaci v1.1	65.9	4.3	0.4	33.7
Diaci v2.0	75.9	20.2	0.2	23.9
Diaci v3.0	88.3	24.5	0.1	11.6
OGS1	74.5	13.0	0.3	25.2
OGS2	81.6	37.3	0.2	18.2
OGS3	80.2	29.4	0.1	19.7
Denovo transcriptome	79.9	26.7	0.1	20.0

Table 2. Gene counts for manually curated gene families in genome v1.1 compared to genome v3.0. *One lysozyme gene was missing from the v3.0 genome but was present in the ALT contigs.

v1.1 Pathway / Gene Family	Number of genes in v1.1	Number of genes in v3.0
Toll Receptors	5	5
Toll Receptor related genes	3	3
JAK/STAT pathway	3	3
C-type Lectins	10	7
Lysozymes	5	5*
Superoxide dismutases	4	4
CLIP	11	10
Autophagy	15	15
Dicer	4	2
Drosha	4	1
Pasha	2	2
Loquacious and R2D2 proteins	3	3
Argonaute and PIWI proteins	4	4
Tudor Staphylococcal Nuclease protein	2	1
Vasa intronic gene	1	1
Armitage	1	1
Fragile X Mental Retardation Protein	1	1
Spindle_E (homeless)	2	1
Rm62 protein	1	1
Ras-related nuclear protein	1	1
Aquaporin	6	6
Cathepsins and Cysteine Proteases	34	23
Heat shock proteins	18	13
Rab genes	19	19
Total	159	132

Table 3. RNA-Seq data

SRA	Data	Reference
SRR1259429	Citrus spp. CLas- Whole body Adult	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130328
SRR1259432	Citrus spp. CLas+ Whole body Adult	
SRR1259461	Citrus spp. CLas- Whole body Nymph	
SRR1259434	Citrus spp. CLas+ Whole body Nymph	
SRR2632316	C. reticulata CLas- Male antennae Adult	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159372
SRR2632319	C. reticulata CLas- Female antennae Adult	
SRR2632320	C. reticulata CLas- Male terminal abdomen Adult	
SRR2632321	C. reticulata CLas- Female terminal abdomen Adult	
SRR602249	C. macrophylla CLas- Whole body Adult	https://www.jgenomics.com/v02p0054.htm#headingA5
SRR610529	C. macrophylla CLas- Whole body Nymph	
SRR610530	C. macrophylla CLas- Whole body Egg	
SRR5514656	ACP-Gut-Healthy-rep1a	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179531#sec002
SRR5514657	ACP-Gut-Healthy-rep1b	
SRR5514655	ACP-Gut-Clas-rep4b	
SRR5514654	ACP-Gut-Clas-rep4a	
SRR5514651	ACP-Gut-Clas-rep2b	
SRR5514649	ACP-Gut-Healthy-rep4b	
SRR5514643	ACP-Gut-Clas-rep1b	

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

References

1. Ammar E-D, Shatters RG Jr, Hall DG. Localization of Candidatus Liberibacter asiaticus, Associated with Citrus Huanglongbing Disease, in its Psyllid Vector using Fluorescence in situ Hybridization: Huanglongbing Disease Bacterium in its Psyllid Vector. Journal of Phytopathology. 2011;159:726–34.

2. Louzada ES, Vazquez OE, Braswell WE, Yanev G, Devanaboina M, Kunta M. Distribution of "Candidatus Liberibacter asiaticus" Above and Below Ground in Texas Citrus. Phytopathology. 2016;106:702–9.

3. Richards S, Murali SC. Best Practices in Insect Genome Sequencing: What Works and What Doesn't. Curr Opin Insect Sci. 2015;7:1–7.

4. Saha S. Long Range Sequencing and Validation of Insect Genome Assemblies. In: Brown SJ, Pfrender ME, editors. Insect Genomics: Methods and Protocols. New York, NY: Springer New York; 2019. p. 33–44.

5. Saha S, Hosmani PS, Villalobos-Ayala K, Miller S, Shippy T, Flores M, et al. Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community. Database [Internet]. 2017;2017. Available from: http://dx.doi.org/10.1093/database/bax032

6. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

7. Saha S, Hosmani P, Flores M, Hunter W, Brown S, Mueller LA. Using long reads, optical maps and long-range scaffolding to improve the Diaphorina citri genome [Internet]. 2017. Available from: https://figshare.com/articles/Using_long_reads_optical_maps_and_long-range_scaffolding_to_improve_the_Diaphorina_citri_genome/5375116

8. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:188–96.

9. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol. 2019;15:e1006790.

10. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016;26:342–50.

11. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.

12. Labina ES. A chromosomal study of 11 species of Psyllinea (Insecta: Homoptera). Comp Cytogenet. ZOOLOGICAL INST ST PETERSBURG UNIV, UNIVERSITETSKAYA NAB., 1, C/O DR. ILYA …; 2007;1:149–54.

13. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 2016;44:e113.

14. Hosmani PS, Shippy T, Miller S, Benoit JB, Munoz-Torres M, Flores-Gonzalez M, et al. A quick guide for student-driven community genome annotation. PLoS Comput Biol. 2019;15:e1006682.

15. Kola VSR, Renuka P, Madhav MS, Mangrauthia SK. Key enzymes and proteins of crop insects as candidate for RNAi based gene silencing. Front Physiol. frontiersin.org; 2015;6:119.

16. Ulrich J, Dao VA, Majumdar U, Schmitt-Engel C, Schwirz J, Schultheis D, et al. Large scale RNAi screen in Tribolium reveals novel target genes for pest control and the proteasome as prime target. BMC Genomics. bmcgenomics.biomedcentral.com; 2015;16:674.

17. Baum JA, Bogaert T, Clinton W, Heck GR, Feldmann P, Ilagan O, et al. Control of coleopteran insect pests through RNA interference. Nat Biotechnol. nature.com; 2007;25:1322–6.

18. Damen WGM. Evolutionary conservation and divergence of the segmentation process in arthropods. Dev Dyn. 2007;236:1379–91.

19. Shigenobu S, Bickel RD, Brisson JA, Butts T, Chang C-C, Christiaens O, et al. Comprehensive survey of developmental genes in the pea aphid, Acyrthosiphon pisum: frequent lineage-specific duplications and losses of developmental genes. Insect Mol Biol. 2010;19 Suppl 2:47–62.

20. Holstein TW. The evolution of the Wnt pathway. Cold Spring Harb Perspect Biol. 2012;4:a007922.

21. Wang L, Tang N, Gao X, Chang Z, Zhang L, Zhou G, et al. Genome sequence of a rice pest, the white-backed planthopper (Sogatella furcifera). Gigascience. 2017;6:1–9.

22. Doumpas N, Jékely G, Teleman AA. Wnt6 is required for maxillary palp formation in Drosophila. BMC Biol. 2013;11:104.

23. Lemons D, McGinnis W. Genomic evolution of Hox gene clusters. Science. 2006;313:1918–22.

24. Hillyer JF. Insect immunology and hematopoiesis. Dev Comp Immunol. 2016;58:102–18.

25. Ferguson LC, Green J, Surridge A, Jiggins CD. Evolution of the insect yellow gene family. Mol Biol Evol. 2011;28:257–72.

26. Miyamoto T, Amrein H. Gluconeogenesis: An ancient biochemical pathway with a new twist. Fly . 2017;11:218–23.

27. Thompson SN. Trehalose – The Insect “Blood” Sugar. Advances in Insect Physiology. Academic Press; 2003. p. 205–85.

28. Silva CP, Silva JR, Vasconcelos FF, Petretski MDA, Damatta RA, Ribeiro AF, et al. Occurrence of midgut perimicrovillar membranes in paraneopteran insect orders with comments on their function and evolutionary significance. Arthropod Struct Dev. 2004;33:139–48.

29. Muthukrishnan S, Merzendorfer H, Arakane Y, Kramer KJ. Chitin Metabolism in Insects. Insect

Molecular Biology and Biochemistry. 2012. p. 193–235.

30. Arakane Y, Muthukrishnan S, Kramer KJ, Specht CA, Tomoyasu Y, Lorenzen MD, et al. The Tribolium chitin synthase genes TcCHS1 and TcCHS2 are specialized for synthesis of epidermal cuticle and midgut peritrophic matrix. Insect Mol Biol. 2005;14:453–63.

31. Nelson N, Perzov N, Cohen A, Hagai K, Padler V, Nelson H. The cellular biology of proton-motive force generation by V-ATPases. J Exp Biol. 2000;203:89–95.

32. Zhu H, Yuan Q, Briscoe AD, Froy O, Casselman A, Reppert SM. The two CRYs of the butterfly. Curr Biol. 2005;15:R953–4.

33. Yuan Q, Metterville D, Briscoe AD, Reppert SM. Insect cryptochromes: gene duplication and loss define diverse ways to construct insect circadian clocks. Mol Biol Evol. 2007;24:948–55.

34. Cortés T, Ortiz-Rivas B, Martínez-Torres D. Identification and characterization of circadian clock genes in the pea aphid Acyrthosiphon pisum. Insect Mol Biol. 2010;19 Suppl 2:123–39.

35. Terakita A. The opsins. Genome Biol. 2005;6:213.

36. Lee S-J, Xu H, Montell C. Rhodopsin kinase activity modulates the amplitude of the visual response in Drosophila. Proc Natl Acad Sci U S A. 2004;101:11874–9.

37. Whyard S, Erdelyan CNG, Partridge AL, Singh AD, Beebe NW, Capina R. Silencing the buzz: a new approach to population suppression of mosquitoes by feeding larvae double-stranded RNAs. Parasit Vectors. 2015;8:96.

38. Darrington M, Dalmay T, Morrison NI, Chapman T. Implementing the sterile insect technique with RNA interference - a review. Entomol Exp Appl. 2017;164:155–75.

39. Dong Y-C, Wang Z-J, Chen Z-Z, Clarke AR, Niu C-Y. Bactrocera dorsalis male sterilization by targeted RNA interference of spermatogenesis: empowering sterile insect technique programs. Sci Rep. 2016;6:35750.

40. Cruz C, Tayler A, Whyard S. RNA Interference-Mediated Knockdown of Male Fertility Genes in the Queensland Fruit Fly Bactrocera tryoni (Diptera: Tephritidae). Insects [Internet]. 2018;9. Available from: http://dx.doi.org/10.3390/insects9030096

41. Ali MW, Zheng W, Sohail S, Li Q, Zheng W, Zhang H. A genetically enhanced sterile insect technique against the fruit fly, Bactrocera dorsalis (Hendel) by feeding adult double-stranded RNAs. Sci Rep. 2017;7:4063.

42. Flores-Gonzalez M, Hosmani PS, Fernandez-Pozo N, Mann M, Humann JL, Main D, et al. Citrusgreening.org: An open access and integrated systems biology portal for the Huanglongbing (HLB) disease complex [Internet]. bioRxiv. 2019 [cited 2019 Dec 9]. p. 868364. Available from:

https://www.biorxiv.org/content/10.1101/868364v1

43. Brown SJ, Coleman M. Isolation of High Molecular Weight DNA from Insects. Methods Mol Biol. 2019;1858:27–32.

44. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.

45. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.

46. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0 [Internet]. 2013-2015. Available from: http://www.repeatmasker.org

47. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164:513–24.

48. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15.

49. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.

50. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.

51. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience [Internet]. 2018;7. Available from: http://dx.doi.org/10.1093/gigascience/giy093

52. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;7:62.

53. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.

54. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. Methods Mol Biol. 2019;1962:65–95.

55. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17:66.

56. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.

57. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database.

Nucleic Acids Res. 2010;38:D211–22.

58. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.

59. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

61. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.

62. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.

63. Weirauch MT, Hughes TR. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell Biochem. 2011;52:25–73.

64. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23:205–11.

65. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.