Journal of Pest Science (2019) 92:1397–1407
https://doi.org/10.1007/s10340-019-01116-6

ORIGINAL PAPER

Sequencing of *Tuta absoluta* genome to develop SNP genotyping assays for species identification

Christine A. Tabuloc[1] · Kyle M. Lewald[1] · William R. Conner[1] · Yoosook Lee[2] · Ernest K. Lee[1] · Alexander B. Cain[3] · Kristine E. Godfrey[3] · Judit Arnó[4] · Nuria Agustí[4] · Clerison R. Perini[5] · Jerson C. Guedes[5] · Frank G. Zalom[1] · Joanna C. Chiu[1]

Received: 15 December 2018 / Revised: 23 March 2019 / Accepted: 6 April 2019 / Published online: 11 April 2019

Abstract

Tuta absoluta is one of the most devastating pests of fresh market and processing tomatoes. Native to South America, its detection was confined to that continent until 2006 when it was identified in Spain. It has now spread to almost every continent, threatening countries whose economies rely heavily on tomatoes. This insect causes damage to all developmental stages of its host plant, leading to crop losses as high as 80–100%. Although *T. absoluta* has yet to be found in the USA and China, which makes up a large portion of the tomato production in the world, computer models project a high likelihood of invasion. To halt the continued spread of *T. absoluta* and limit economic loss associated with tomato supply chain, it is necessary to develop accurate and efficient methods to identify *T. absoluta* and strengthen surveillance programs. Current identification of *T. absoluta* relies on examination of morphology and assessment of host plant damage, which are difficult to differentiate from that of native tomato pests. To address this need, we sequenced the genomes of *T. absoluta* and two closely related Gelechiidae, *Keiferia lycopersicella* and *Phthorimaea operculella*, and developed a bioinformatic pipeline to design a panel of 21-SNP markers for species identification. The accuracy of the SNP panel was validated in a multiplex format using the iPLEX chemistry of Agena MassARRAY system. Finally, the new *T. absoluta* genomic resources we generated can be leveraged to study *T. absoluta* biology and develop species-specific management strategies.

Keywords South American tomato pinworm · Tomato leaf miner · Tomato borer · Gelechiidae · Invasive pest · Single-nucleotide polymorphism

Key message

- Surveillance and identification of *Tuta absoluta* are challenging because it is morphologically similar to closely related species, e.g., *Keiferia lycopersicella* and *Phthorimaea operculella.*
- We generated new genomic sequences for these three species and identified single-nucleotide polymorphisms (SNPs) to facilitate species identification.

Communicated by A. Biondi.

Special Issue on Advances in the Management of *Tuta absoluta*.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10340-019-01116-6) contains supplementary material, which is available to authorized users.

✉ Joanna C. Chiu
jcchiu@ucdavis.edu

[1] Department of Entomology and Nematology, University of California, Davis, Davis, USA

[2] Department of Pathology, Microbiology and Immunology, School of Veterinary Medicine, University of California, Davis, Davis, USA

[3] Contained Research Facility, University of California, Davis, Davis, USA

[4] IRTA (Institute of Agrifood Research and Technology), Cabrils, Barcelona, Spain

[5] Department of Crop Protection, Universidade Federal de Santa Maria (UFSM), Santa Maria, Rio Grande do Sul, Brazil

- We validated a multiplex genotyping panel of 21 SNPs using the iPLEX MassARRAY platform and confirmed its accuracy for species identification.
- We generated new molecular tools and genome resources to aid *T. absoluta* management.

Introduction

Tuta absoluta (Meyrick 1917) (Lepidoptera: Gelechiidae), commonly referred to as tomato leaf miner, tomato borer, or the South American tomato pinworm, is a moth species that devastates fresh market and processing tomatoes. *Tuta absoluta* was originally identified from samples collected in Peru by Meyrick in 1917 (Povolny 1994; Đurić et al. 2014; Biondi et al. 2018). It was not recognized as a serious pest until it was found damaging tomatoes in Argentina in the 1960s (Bahamondes and Mallea 1969), where it caused substantial crop losses. *Tuta absoluta* causes crop losses as high as 80–100% because it will damage all developmental stages of its host plant. Adult females oviposit on the leaves, where the larvae will emerge from eggs and begin mining host tissues. Larvae can also enter the stems through the buds and feed within the tomato fruit, leaving them unmarketable. Its distribution was largely confined to South America until it was first detected in Spain in 2006 (Desneux et al. 2010; Guillemaud et al. 2015). Since then, *T. absoluta* has spread rapidly and is now established in Europe, northern, southern, and eastern parts of Africa, southern Central America, the Middle East, and in parts of South Asia (CABI 2016; Campos et al. 2017; Mutamiswa et al. 2017; Biondi et al. 2018; Mansour et al. 2018; Han et al. 2018; 2019).

Although *T. absoluta* has not been reported in North America, Australia, New Zealand, and some parts of Asia, CLIMEX computer modeling taking into account pest life history, climate data, and host plant availability predicts that it has a moderate to high likelihood of establishing in the commercial tomato-growing regions around the globe including California and Arizona in the southern United States (USDA 2011; Tonnang et al. 2015; Biondi et al. 2018). Although the primary host of *T. absoluta* is tomato, *Solanum lycopersicum* L., it can also colonize other solanaceous host plants such as potato, black nightshade eggplant, sweet pepper, jimsonweed, and deadly nightshade (Pereyra and Sanchez 2006; Desneux et al. 2010, 2011; Bawin et al. 2015; Mohamed et al. 2015; Negi et al. 2018). The ability of *T. absoluta* to inhabit a wide variety of host plants is expected to greatly facilitate its range expansion.

Early detection of invasion and timely response are instrumental in halting the continued spread of *T. absoluta,* especially into the USA, Mexico, and China, which together account for roughly 45% of tomato production in the world (FAOSTAT 2017). Unfortunately, *T. absoluta* identification and monitoring remain a challenge; *T. absoluta* larvae and adults are morphologically similar to many other gelechiid species. The tomato pinworm, *K. lycopersicella* (Walsingham 1897), and the potato tuber moth, *P. operculella* (Zeller 1873), are two primary gelechiids already occupying tomato-growing regions in the USA in which *T. absoluta* will likely invade (Michalak 2011). Although less commonly observed compared to *K. lycopersicella* and *P. operculella,* other gelechiids such as *Sinoe capsana* (Lee and Brambila 2012) and *Tuta* sp. *near chiquitella* (Gaskill 2013) have also been reported in the USA and have the potential to be misidentified as *T. absoluta*. The Guatemalan potato tuber moth, *Tecia solanivora* (Povolny 1973), presents risks of future introductions into many tomato-growing regions (EPPO Global Database 2019) and could be misidentified as *T. absoluta*. This creates a serious problem for early detection. Current identification requires the dissection and examination of male genitalia (Povolny 1975; Michalak 2011; Đurić et al. 2014) by highly practiced experts. Furthermore, rearing to adulthood in order for male genitalia to fully develop is not always practical if marketability of a shipment is to be maintained. While host plant damage caused by *T. absoluta* at immature stages could potentially be leveraged for identification and detection, damage is essentially indistinguishable from the damage caused by other morphologically similar pests occupying the same niche, e.g., *K. lycopersicella* and *P. operculella*.

Alternatively, DNA barcoding via PCR amplification of mitochondrial cytochrome oxidase subunit I (COI) (Cifuentes et al. 2011) as well as RAPD-PCR (RAPD, Random Amplified Polymorphic DNA) (Bettaibi et al. 2012) has been utilized as molecular diagnostics to identify *T. absoluta* and to examine genetic variations between different geographical populations. However, these molecular diagnostics have not been tested for or utilized to differentiate between *T. absoluta* and morphologically similar species such as *K. lycopersicella* and *P. operculella*. Sint et al. (2016) developed species-specific primers from COI sequences of *T. absoluta, P. operculella*, and *Symmetrischema tangolias* (Gyen 1913) (Lepidoptera: Gelechiidae) and established multiplex PCR assays to enable the identification of these three species and their parasitoids, but did not include *K. lycopersicella* in their analysis.

In this study, we constructed a draft genome assembly for *T. absoluta* using Linked-Read library preparation by 10× Genomics Chromium platform and performed genome sequencing for *K. lycopersicella* and the *P. operculella*. We then designed and implemented a custom bioinformatic pipeline with the goal of identifying single-nucleotide polymorphisms (SNPs) to design a multiplex SNP genotyping assay for robust molecular species diagnostics. SNP genotyping was performed on the Agena MassARRAY system using iPLEX (Locus-specific primer extension reaction)

chemistry (Gabriel et al. 2009), allowing us to perform multiplex reactions to detect over 20 SNPs simultaneously. We validated the accuracy of this SNP panel to differentiate *T. absoluta* from *K. lycopersicella* and *P. operculella* using specimens from multiple life stages and determined the accuracy of species identification to be 100%.

Materials and methods

Origins of the Gelechiidae specimens

Tuta absoluta adults and larvae came from the laboratory colony maintained at IRTA in Cabrils (Barcelona), Spain, Costa Rican field collections (by Y. G. Bonilla), as well as collections from greenhouses and the field in eleven geographical locations in Argentina, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, and Uruguay (by J. C. Guedes and C. R. Perini) (Supplemental Table 1). The colony in Spain was initiated from individuals collected from several locations in the Barcelona province, as reported in Arnó et al. (2018). Live samples were collected, preserved in 90–95% ethanol, subsequently shipped to UC Davis, and stored at 4 °C prior to genomic DNA (gDNA) extraction.

Keiferia lycopersicella colonies were established from specimens collected in Immokalee, Florida, in the Fall of 2015 (by P. Stansly). Pupae were shipped to UC Davis in January 2016. *Keiferia lycopersicella* individuals were then reared on tomato seedlings or small plants (cv Patio Princess, W. Atlee Burpee and Company, Warminster, PA, USA) that were about 3 months old. Rearing was performed in a Bugdorm cage (MegaView Science Education Services Co., Ltd., Taichung, Taiwan) and held at 23–24 °C with overhead lights 24 h a day. Humidity was not controlled. Each cage consisted of 6–8 tomato plants in UC Mix soil. The plants were watered as needed with a fertilizer solution (Miracle-Gro mixed according to the manufacturer's recipe for indoor plants). Adults were introduced into a new cage, and a generation lasts about 30 days on average. Larvae, pupae, and adults were collected on dry ice, stored in − 80 °C, and subjected to gDNA extraction.

Phthorimaea operculella colonies were established from specimens collected from a commercial potato field near Arvin, Kern County, California (CA) (by D. Haviland), and shipped to UC Davis. At UC Davis, *P. operculella* individuals were reared on yellow or russet potato tubers. Four to six small tubers were placed on a 1.25-cm bed of autoclaved sand in a tray covered with paper towels and placed into a Bugdorm cage. Thirty to forty adult tuber moths were introduced into the cages. The cages were held in the same environmental conditions as the *K. lycopersicella* colonies. Over the course of their 45-day life cycle, larvae, pupae, and adults were collected on dry ice, stored in − 80 °C, and subjected to genomic gDNA extraction.

Genomic DNA extraction for *Tuta absoluta* reference genome sequencing

A single adult *T. absoluta* collected from Spain and preserved in 95% ethanol was first placed into nuclease-free water in a 1.5-ml tube for rehydration at room temperature for 15 min. After removing water, the specimen was subsequently homogenized in a 2% CTAB solution (100 mM Tris–HCl (pH 8.0), 10 mM EDTA, 1.4 M NaCl, and 2% CTAB). The sample was incubated at 65 °C for 5 min, and 200 µl of chloroform was added to the tubes and then inverted slowly 10 times to mix. To isolate nucleic acids, samples were centrifuged at 13,000 rpm for 10 min at 4 °C. The aqueous layer was transferred to a new tube and mixed with an equal volume of 100% isopropanol and left in − 20 °C overnight for gDNA to precipitate. The DNA was then pelleted at 13,000 rpm for 15 min at 4 °C. The DNA pellet was washed with 70% ethanol and spun down at 13,000 rpm for 5 min at 4 °C. After the pellet was air-dried, the gDNA was re-suspended in nuclease-free water. DNA was quantified using the Qubit dsDNA high sensitivity kit (Thermo Fisher Scientific, Pleasanton, CA, USA) in combination with Qubit fluorometer (Thermo Fisher Scientific, Pleasanton, CA, USA).

Library preparation, sequencing, and assembly of *Tuta absoluta* reference genome

Genomic DNA from a single *T. absoluta* adult was submitted to the UC Davis DNA Technologies Core for Linked-Read library preparation using a Chromium Controller and the Chromium Genome Reagent Kit (10× Genomics, Pleasanton, CA, USA) according to manufacturer's protocols for v1 chemistry. The barcoded library was sequenced on one lane of an Illumina HiSeq 4000 sequencer (Illumina, San Diego, CA, USA) to produce 2 × 150 paired-end reads. A "pseudohap" assembly was generated from raw reads with Supernova 2.1.1 using 40 cpu cores. The only optional arguments used in *supernova run* were localcores and localmem, which were set to the aforementioned values. This assembly was used as the *T. absoluta* reference in subsequent analysis. Genome size estimate was obtained from Supernova as well as GenomeScope (Vurture et al. 2017) using k-mer length = 21, read length = 150, max k-mer coverage = 1000. For GenomeScope, the input histrogram of k-mer frequencies was generated using Jellyfish v2.2.5 (Marçais and Kingsford 2011) with k-mer length = 21. The completeness of the *T. absoluta* assembly was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) v3.0.2 (Simao et al. 2015) in genome mode with the insecta_odb9 lineage

data and by mapping RNA-seq reads (NCBI SRA accession number SRX1134908) from a published *T. absoluta* transcriptome (Camargo et al. 2015) to our assembly using STAR v2.6.1a (Dobin et al. 2013) with default parameters. Ribosomal RNA sequences were removed from the raw RNA-seq reads downloaded from NCBI using SortMeRNA v2.1 (Kopylova et al. 2012). The remaining reads were trimmed for quality and adapter sequences using Trimmomatic v0.35 (Bolger et al. 2014) with LEADING = 10, TRAILING = 10, ILLUMINACLIP = TrueSeq 3-PE. fa:2:30:10, and MINLEN = 36 prior to mapping onto the *T. absoluta* genome assembly.

Library preparation and genome sequencing of *Tuta absoluta*, *Keiferia lycopersicella* and *Phthorimaea operculella* replicates for comparative sequence analysis and identification of species-specific SNPs

Eight replicate libraries, each represents a single adult insect, were prepared for each of the three species. Instead of separately sequencing an individual at high depth similar to *T. absoluta*, we used one of the replicates as reference for *K. lycopesicella* and *P. operculella,* respectively. Genomic DNA was extracted as described in Nieman et al. (2015) and Yamasaki et al. (2016) using the Qiagen BioSprint 96 Automated Nucleic Acid Purification System and reagents (Qiagen Sciences, Germantown, MD, USA). DNA libraries were then prepared with 50-ng input DNA per library using the Kapa HyperPlus Kit (Kapa Biosystems, Wilmington, MA, USA). Libraries were quantified using the Qubit dsDNA high sensitivity kit in combination with Qubit fluorometer (Thermo Fisher Scientific, Pleasanton, CA, USA) and subjected to quality control using an Agilent 2100 Bioanalyzer with a High Sensitivity DNA chip (Agilent Technologies, Santa Clara, CA, USA). Each library had an 8-bp-long barcode. The multiple barcoded libraries were pooled and subjected to a two-tailed size selection, 0.35× and 0.7×, using AMPure XP beads (Beckman Coulter Life Sciences, Indianapolis, IN, USA). The final pooled sample was eluted in 22 µl of 10 mM Tris–HCl, pH8.0, and submitted to Novogene (Sacramento, CA, USA) for sequencing on a HiSeq 4000 platform (Illumina, San Diego, CA, USA). Raw reads from one replicate each of *K. lycopersicella* and *P. operculella* were assembled using SOAPdenovo2 r240 (Luo et al. 2012) with k-mer size of 63 to generate low coverage references for subsequent analysis.

Bioinformatic pipeline for comparative genomic analysis and SNP identification for iPLEX primer design

We developed a custom program *snp-id* (available in GitHub; https://github.com/ClockLabX/snp-id) that can identify SNPs suitable for iPLEX or other genotyping assays. The complete bioinformatics pipeline for our analysis is illustrated in Fig. 1. First, reads from each of the 8 replicates for each species were mapped back to the respective reference using BWA (BWA-MEM) v0.7.9a (Li and Durbin 2009). Reference genomes for all three species were aligned using the multiple genome alignment tool, Mauve (progressiveMauve) (Darling et al. 2010). The SNP identification script of *snp-id*, *search_iplex.py*, was then invoked with the following input: (i) the reference genome sequence for each species, (ii) alignment of each replicate to the corresponding reference genome for each species, and (iii) the multiple genome alignment of the three species. High-quality SNPs that are more likely to be invariant within species are chosen by requiring that SNPs be homozygous and uniform across all replicates within a species with no less than 3 replicates with coverage at that position. To satisfy the more stringent requirements for the iPLEX assay, only segments of 81–141 bases with non-polymorphic regions flanking the diagnostic SNP are chosen (Fig. 2). This selection criterion also satisfies the requirements of other SNP identification assays. Finally, results were searched against the NCBI nucleotide database using *blast_iplex.py*, which uses MegaBLAST (Zhang et al. 2000) with an e-value cutoff of 1e−10, to identify common contaminants to be excluded for iPLEX assay design.

The list of SNPs that were identified using *snp-id* (Supplemental File 1) were then used as the input for the MassARRAY Typer 4.0 Assay Designer Software (Agena Bioscience, San Diego, CA, USA) to design iPLEX PCR and extension primers (Table 1). The markers were named by the SNP location on the genome assembly, except in the case where the region clearly mapped to an annotated gene when queried in BLAST (as in the case of Eif-4a).

MassARRAY system combined with iPLEX chemistry for species identification

Genomic DNA from *K. lycopersicella*, *P. operculella*, and *T. absoluta* was extracted using the method as described in Nieman et al. (2015) and Yamasaki et al. (2016) using the Qiagen BioSprint 96 Automated Nucleic Acid Purification System and reagents (Qiagen Sciences, Germantown, MD, USA). Samples at different life stages were analyzed (Supplemental Table 1). Primer cocktails for multiplex PCR of 21 loci were prepared as described in Gabriel et al. (2009). DNA samples, primer cocktails for multiplex PCR, and primers for iPLEX extension reactions were then sent to the Veterinary Genetics Laboratory at UC Davis for MassARRAY iPLEX genotyping assay (Agena Bioscience, San Diego, CA, USA). MassARRAY 4.0 Typer Analyzer Software was used for genotype calling and species identification.

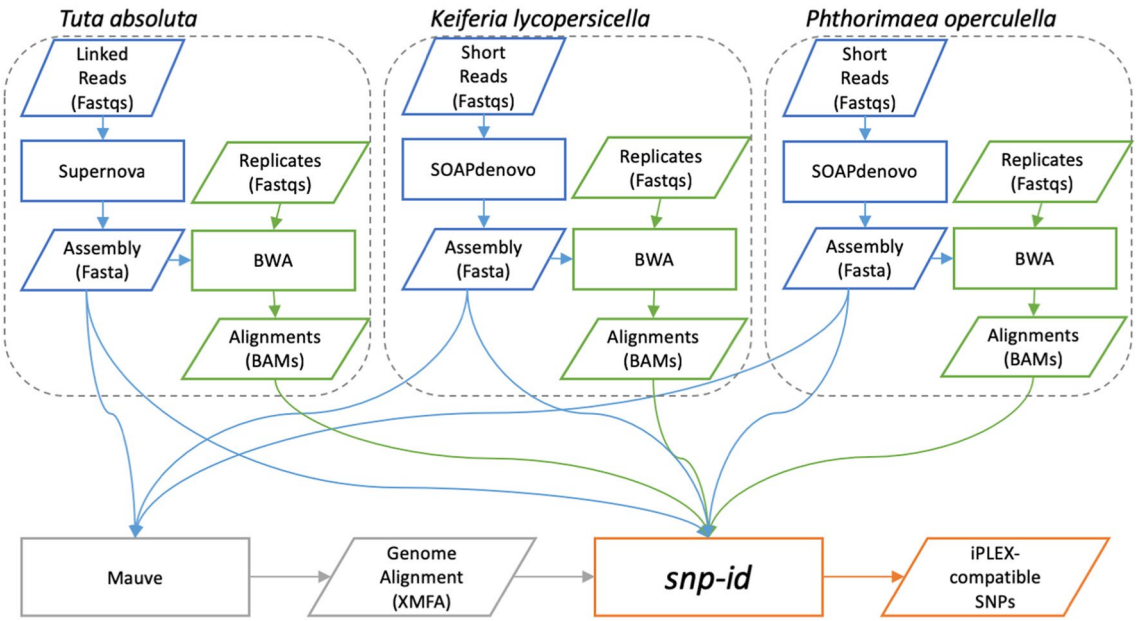

Fig. 1 Schematic illustrating the bioinformatic workflow for genome assembly and comparative genomic analysis. Raw reads from one sample each of *T. absoluta*, *K. lycopersicella*, and *P. operculella* were separately assembled to create reference genomes. Eight replicates for each species were then aligned back to their respective references. The three reference genomes were also aligned to each other to create a multi-genome alignment. All the reference genomes and alignments were passed to *snp-id*, which identified and generated iPLEX-compatible SNPs and sequences for assay design

The iPLEX workflow starts with a multiplex PCR reaction to amplify specific gene regions containing the polymorphic SNPs between species. The PCR products are then treated with shrimp alkaline phosphatase (SAP) to neutralize any free nucleotides. This is followed by a second round of SNP extension reaction that utilizes end terminating nucleotides. The extension primers for the SNP extension step are shown in Table 1. Because the amplicons from this reaction are identical in sequence for all samples except at the last nucleotide, i.e., location of the SNP, the mass of the extension primer plus one base of species-specific allele will produce variable spectra readings when analyzed by a mass spectrophotometer (Gabriel et al. 2009).

Phylogenetic analysis of gelechiid species COI sequences

COI sequences were identified from the genomes for *K. lycopersicella*, *P. operculella*, and *T. absoluta* and from NCBI for *Sinoe robiniella* (Fitch 1859) (accession no. MG365151.1) and *T. solanivora* (accession no. NC_029386.1). Alignment was performed with MAFFT v7.3.10 (Katoh and Standley 2013) using the L-INS-I algorithm. Maximum likelihood analysis was performed with RAxML v8.2.12 (Stamatakis 2014) using the GTRGAMMA model with 1000 rapid bootstrap searches.

Results

Tuta absoluta reference genome

A reference genome assembly of *T. absoluta* was generated from 638.8 million paired-end reads representing roughly 72× raw coverage. Counting only scaffolds greater than 10 kb, the assembly has a total size of 677.2 Mb. The contig N50 is 26.36 Kb and the scaffold N50 is 112.89 Kb as reported by Supernova. GC content of the assembly is 38.11%. The genome size estimated by Supernova varies widely from 674 Mb when 252 million reads were used for a raw coverage of 56× to 1.34 Gb when all reads were used, whereas GenomeScope (Vurture et al. 2017) produced an estimate of only 492 Mb. Two metrics reported by Supernova may explain the lower than expected scaffold sizes: (i) weighted mean molecule size was reported to be 24.55 Kb, which may reflect challenges in extracting long DNA from *T. absoluta*, and (ii) the repeat content index, which is the percent of read kmers with twice the expected depth, is 37.91%. However, our SNP identification method is not sensitive to scaffold size.

To assess the completeness of our *T. absoluta* assembly, we compared it to the Insecta set of universal single-copy orthologs with BUSCO v3.0.2 (Simao et al. 2015). Of the 1658 total BUSCO groups searched, 1532 (92.4%) were identified as complete in the assembly. Summarized

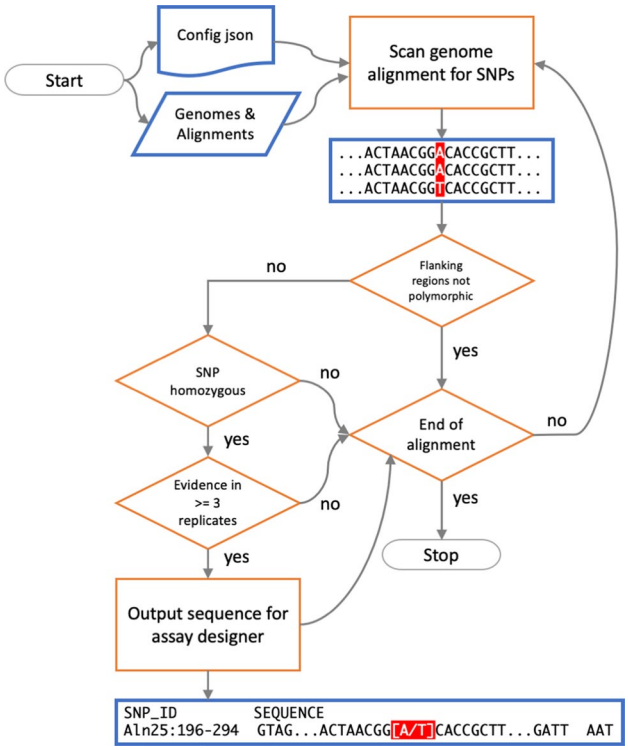

Fig. 2 Flowchart describing the algorithm of the *snp-id* program. The search_iplex.py script of the *snp-id* program requires an input file (json) that specifies all the reference genomes (Fasta), replicate alignments (BAM), and multi-genome alignment (XMFA). It scans the multi-genome alignment for candidate SNPs and tests for (i) polymorphisms in flanking regions, (ii) homozygosity, and (iii) evidence in other replicates. SNPs that satisfy all selection criteria are printed out in a format suitable for MassARRAY Typer 4.0 Assay Designer Software

benchmarking in BUSCO notation is as follows: *C*:92.4% [*S*:66.0%, *D*:26.4%], *F*:4.4%, *M*:3.2%, *n*:1658 (*C* = Complete BUSCOs, *S* = Complete and single copy, *D* = Complete and duplicated, *F* = Fragmented, *M* = Missing, *n* = Total BUSCO groups searched).

We also examined the coverage of coding regions by mapping a published *T. absoluta* transcriptome (Camargo et al. 2015) to our assembly. After removing ribosomal RNA sequences and performing adapter and quality trimming, 17,345,874 read pairs were mapped to our *T. absoluta* assembly using STAR (Dobin et al. 2013). There are 75.89% of uniquely mapped reads, 11.84% of multi-mapped reads, and 12.1% of reads that are too short to map. Only 0.17% of reads are unmapped for other reasons.

We observed the presence of *Wolbachia* sequences in the *T. absoluta* genome assembly. A total of 1.198 Mb in 148 scaffolds have significant BLAST matches to *Wolbachia* strains in GenBank. We also identified *Wolbachia* sequences in other *T. absoluta, K. lycopersicella,* and *P. operculella* genome replicates analyzed in this study, suggesting *Wolbachia* infection is prevalent in these species.

Bioinformatic analysis enables SNP identification and genotyping primer design

The bioinformatic workflow for genome assembly and comparative genomic analysis of *T. absoluta*, *K. lycopersicella*, and *P. operculella* is outlined in Fig. 1, and a flow diagram charting the steps of the *snp-id* program to select gene regions suitable for SNP genotyping using Agena MassARRAY platform in combination with iPLEX chemistry (Gabriel et al. 2009) is presented in Fig. 2. Due to the stringent requirements used in identifying SNPs, the output of *snp-id* (Supplemental File 1) can be readily adopted to be used for other SNP genotyping assays. The stringency can also be tuned in the script by adjusting (1) the minimum number of genome replicates required with the same SNP to allow for genetic variability, (2) the maximum number of other polymorphisms within the amplicon, and (3) the length of nucleotides flanking the target SNP in each amplicon. Since all replicates, including the ones used to construct draft genome references for *K. lycopersicella* and *P. operculella*, have relatively low (~ 10×) sequencing depth, we showed that low coverage genomes are sufficient in identifying SNPs for species identification with our workflow.

Multiplex SNP genotyping assay is successful in differentiating *Tuta absoluta, Keiferia lycopersicella* and *Phthorimaea operculella*

We extracted gDNA from at least 24 individuals at different life stages (adults or larvae) of each species (Supplemental Table 1) to validate our panel of 21 species-specific markers. The markers were designed such that each SNP specifically identifies *T. absoluta*, *K. lycopersicella*, or *P. operculella* from the other two species (Table 2). All markers performed as expected, and all 137 specimens, 85 *T. absoluta*, 24 *K. lycopersicella*, and 28 *P. operculella*, were correctly classified (Table 2). These include 5 adult specimens from Costa Rica (CRA1-5) that were previously suspected to be *T. absoluta* based on morphological characters but were not identified with certainty due to poor conditions of the specimens (personal comm. Y. G. Bonilla). The positive identification of the Costa Rican *T. absoluta* specimens speaks to the utility of the SNP markers to enable identification of less than perfect specimens as well as immature stages.

Phylogenetic analysis suggests that the likelihood of misidentifying USA gelechiids as *Tuta absoluta* using the SNP panel is low

Although our SNP panel was designed to differentiate *T. absoluta* from *K. lycopersicella* and *P. operculella*, two

Table 1 Amplification and extension primers for iPLEX MassARRAY SNP genotyping assay

SNP Loci	PCR amplification forward primer[a]	PCR amplification reverse primer	Extension primer
LOC110375043	ACGTTGGATGGGTAAAATGTG CTCAGCTGG	ACGTTGGATGAACTCGTTCCC GAACCAGC	CCGAACCAGCTGATCTCG
LOC106138973	ACGTTGGATGGGAACTAAATG CTTGGTCGC	ACGTTGGATGAGCAACTGGCA CAAGTAGTC	GGGTGCACAAGTAGTCATGTCAG
LOC110369709	ACGTTGGATGCACTAGTGTGTATT ATATTAC	ACGTTGGATGGTCGATTCACA GTTAGACGG	CGGTCAACGTCGAAAT
LOC110369696_T[b]	ACGTTGGATGGCAAGTTCCGT GTACTTGAG	ACGTTGGATGTCCCCTCAGGA CCCGAGAA	CCCCTCGCAGTCGTT
LOC101743970	ACGTTGGATGCTTATTACAACATG TCATC	ACGTTGGATGCAGTAAACATT TTGCTGTTAG	TCAGTTTGCTGTTAGCTTTTATTTT
LOC105392331	ACGTTGGATGTGCATCAAGCA CGTCGACTG	ACGTTGGATGTTCAACGAAAG TAAACAGC	CAAGTAAACAGCTGCTATTG
LOC106110340	ACGTTGGATGTCCCCTTTCGTATC ATCCTG	ACGTTGGATGTAAGTTATGCA TGCTGTTC	CCTGTTCTAAATTATTTTTCCAAA TTAT
Eif-4a	ACGTTGGATGTGTGGCCAACC TCATCTAAG	ACGTTGGATGATCGCAATAGA ACGCAGGAC	GGGGCACGCAGGACACAAAGT ATAAA
LOC101746640	ACGTTGGATGTCATCAGCAGC ATGGCTGTC	ACGTTGGATGTTCGTCAAAAT TCACAAAC	CTCCAAATTCACAAACAAAAA AAACC
LOC106131324	ACGTTGGATGGAAAATTATTT TCACTTACC	ACGTTGGATGATCAAAATTCG GCCCTTCAC	TCACTTCTTCTTCGTCC
LOC110375524	ACGTTGGATGTAGAGACTTGT TGAAACAG	ACGTTGGATGCAATCCAGACT ACAAGATCA	CTAAAAAATTCAACAACACTTATT TTA
LOC105701312	ACGTTGGATGTTCCACTGCAG TGAAACAAG	ACGTTGGATGATTATTGATGG CTCTGGCTC	TCCTTTCTCATAGTCCACAT
LOC110384087	ACGTTGGATGTCACCGAGTCT ACTAGTTAC	ACGTTGGATGTCAGCCTTTTC GAATAGGAG	GAGCATTTTCGAATAGGAGAATCA TAAT
LOC106136952	ACGTTGGATGTAGTAGGAATC AAAATGCC	ACGTTGGATGGGAAGACTACA TTTTGCCTC	CCATTTTGCCTCTAAATTTAACTTA
LOC110378105	ACGTTGGATGTCGACGCCGGA AAAAATCAC	ACGTTGGATGTTGCCACACTTTTT TCCCCG	GACGGTGAATGAAAAGTTCGG
LOC105393522	ACGTTGGATGCTGCAGTTAAT TTGAATGGC	ACGTTGGATGCATACATAAAG ATTAAGCATC	TGAGCATCTCTGCACCTTA
LOC110369696_K[c]	ACGTTGGATGGGAATTACAAC GTAGTGTGG	ACGTTGGATGGTTGATTGGTT GTTGTGCGG	AAACTGTTGTGCGGTGAGCG
LOC110999038	ACGTTGGATGTCGGCATTATA ATGTCCATC	ACGTTGGATGAAAGAGTTTAG TAGGTCGAG	TCTAGAAACAAATTACTTGATTGC
LOC106137173	ACGTTGGATGGTGCCATAAAT ATCGATCCG	ACGTTGGATGCCAACAGTTTC TTTCACCCC	CATATCACTGTACACGTTTTCAG
LOC110371221	ACGTTGGATGTCTCCTTCGAG AAACAGACG	ACGTTGGATGTTTGTGTGGCC GTATTACCG	CGATGTAAATGCAGAAACC
LOC106110944	ACGTTGGATGAGTTCAACCCT TTTGAGGGC	ACGTTGGATGCGTATTTATTCAAC AATCAA	GTTATTCAACAATCAAATAAATAT TACG

[a]All sequences are written 5′–3′

[b]This SNP genotyping assay identifies *T. absoluta*

[c]This SNP genotyping assay identifies *K. lycopersicella*

gelechiids that are morphologically similar to *T. absoluta* and are the primary gelechiids that are found in commercial tomato fields in the USA, inclusion of other gelechiids in SNP design in future studies will further improve the resolution and utility of our diagnostic markers for species identification. Other gelechiids that are occasionally encountered in traps for monitoring *T. absoluta* in the USA include *S. capsana* and *Tuta sp. near chiquitella*. Another gelechiid species that is morphologically similar to *T. absoluta* and presents a risk of invasion into North America, Africa, and Asia is *T. solanivora* (Guatemalan potato tuber moth) (EPPO Global Database 2019). There are no sequences available for *S. capsana and Tuta sp. near chiquitella* in NCBI. However, COI sequences are available for *S. robiniella* and *T.*

Table 2 iPLEX MassARRAY SNP genotyping assays to differentiate *T. absoluta* (*Ta*), *K. lycopersicella* (*Kl*), and *P. operculella* (*Po*)

Species comparison (species 1 vs. 2/3)	SNP Loci[a]	SNP species 1	SNP species 2/3
Ta versus *Kl/Po*	Loc110375043	C	T
	Loc106138973	T	C
	Loc110369709	A	G
	Loc110369696_T	T	C
	Loc101743970	A	G
	Loc105392331	A	T
	Loc106110340	A	C
	Eif-4a	A	G
	Loc101746640	G	A
	Loc106131324	A	T
	Loc110375524	C	T
	Loc105701312	C	T
Kl versus *Ta/Po*	Loc110384087	T	A
	Loc106136952	C	A
	Loc110378105	A	G
	Loc105393522	A	G
	Loc110369696_K	A	G
Po versus *Ta/Kl*	Loc110999038	G	A
	Loc106137173	G	A
	Loc110371221	A	G
	Loc106110944	T	C

[a]Number of specimens tested for this 21-SNP panel: 85 *T. absoluta*, 24 *K. lycopersicella,* and 28 *P. operculella*

solanivora. Together with COI sequences from our genome data for *T. absoluta*, *K. lycopersicella,* and *P. operculella*, we used maximum likelihood tree estimation to generate a phylogram to determine the genetic distances between these gelechiids (Fig. 3). We reasoned that if *K. lycopersicella* and *P. operculella* are more closely related to *T. absoluta* as compared to *Sinoe* species and *T. solanivora,* then it is less likely that our 21-SNP panel will misidentify *S. capsana* and *T. solanivora* as *T. absoluta*. Indeed, this is what we observed (Fig. 3).

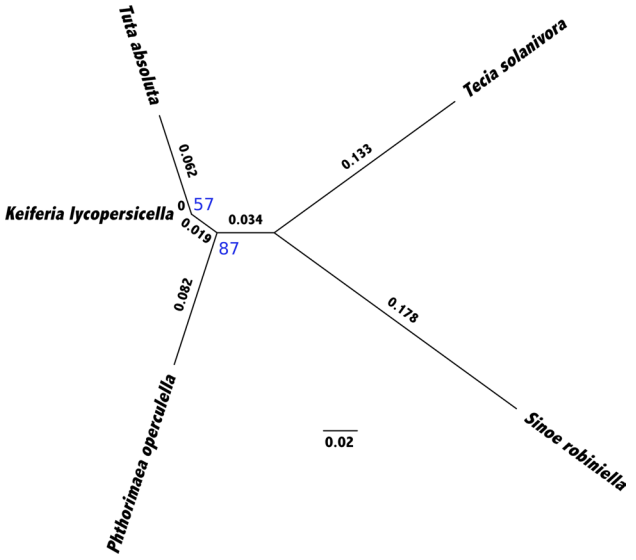

Fig. 3 Phylogram describing the genetic distances between *T. absoluta* and morphologically similar gelechiids. Maximum likelihood tree of COI nucleotide sequences showing the phylogenetic relationship between *T. absoluta*, *K. lycopersicella*, *P. operculella*, *S. robiniella*, and *T. solanivora*. The GTRGAMMA model was used in the tree search. Branch lengths are number of substitutions per site. Numbers in blue are bootstrap values from 1000 rapid bootstrap searches

Discussion

In this study, we generated a draft genome assembly for the devastating tomato pest *T. absoluta* and genomic sequences for two other Gelechiidae, *K. lycopersicella* and *P. operculella*, that show high levels of similarity in morphology. Through the development and use of a custom bioinformatic pipeline, we identified a large number of species-specific SNP markers (Supplemental File 1) and designed a multiplex panel of 21 SNPs that can be used to differentiate these three species at all life stages efficiently and accurately with minimal DNA input. In addition to species identification, these SNP markers will facilitate detection of hybridization among morphologically similar species that colocalize and may impact the spread of undesirable traits such as insecticide resistance (Teeter et al. 2010; Lee et al. 2013, 2014).

Each SNP is selected based on the criteria that it is homozygous and is invariant among the replicate species genomes we used for SNP identification. These criteria were imposed to increase the chance that the SNP alleles are conserved within each of the three species of interest, even for populations from diverse geographical regions. Our SNP validation experiments using *T. absoluta* specimens collected from 13 geographical locations in South America, Central America, and Europe confirmed the utility of the high-quality SNPs designed using our selection criteria to process samples from diverse geographical populations. Although we were not able to collect different geographical populations of *K. lycopersicella* and *P. operculella* for SNP validation, the fact that the SNP alleles for identifying those two species were isolated using the same criteria suggests that it is likely our SNP panel will be able to handle *K. lycopersicella* and *P. operculella* specimens from diverse populations. This can be confirmed in future studies when specimens from diverse locations become available.

There are a number of assays one can employ for SNP genotyping to facilitate species identification, e.g., TaqMan real-time PCR (Dhami et al. 2016; Zhang et al. 2016; Linck et al. 2017), High-Resolution Melt (HRM) real-time PCR

(Dhami and Kumarasinghe 2014; Ajamma et al. 2016), species-specific PCR (Sint et al. 2016), KASP genotyping (Middlesex, UK), and SNP microarrays. We chose to adopt the Agena MassARRAY platform in combination with iPLEX chemistry (Gabriel et al. 2009) to maximize the number of SNP markers we can multiplex in a single assay to reduce false positive rate and increase rigor of species identification. The iPLEX method allows the multiplex detection of up to 40 SNPs in a single reaction and can be completed within 5 h after gDNA extraction. The economical high multiplexing capacity of iPLEX assays provides increased diagnostic accuracy when compared with other PCR-based techniques (Lee et al. 2015).

We should point out that it is not necessary to use all 21 markers simultaneously in order to determine the species identity of a specimen. However, using a combination of the SNP markers will provide higher confidence for species identification by reducing false positives (Lee et al. 2015), given the presence of genetic variations in field populations. Other genotyping technologies mentioned above can be used in combination with the SNP markers generated in this study for *T. absoluta* species diagnostics, but the multiplexing capacity of some of these technologies, e.g., HRM and TaqMan, will not be as high as iPLEX.

We anticipate that increasing taxon sampling will continue to improve the utility and accuracy of the SNP diagnostics presented here. Nevertheless, we believe that the SNP panel in its current format is valuable for quick screening of adult and immature stages and complementary to morphological identifications to monitor early introduction of *T. absoluta* into the USA, given that the two primary gelechiids commonly found in tomato hosts in the USA, *K. lycopersicella* and *P. operculella,* can be distinguished from *T. absoluta* using our SNP panel.

Finally, the new genomic resources for *T. absoluta*, *K. lycopersicella*, and *P. operculella* can be leveraged for design of genetic pest control, e.g., RNA interference (Camargo et al. 2015, 2016), and for understanding various aspects of *T. absoluta* biology, e.g., *Wolbachia* infection, chemoreception, and insecticide resistance, to improve management.

Author contribution

JCC, FGZ, KEG, and CAT designed the research. CAT and KML conducted experiments. ABC, JA, NA, KEG, CRP, and JCG contributed to specimen collection and rearing. CAT, KML, WRC, YL, EKL, and JCC analyzed data and performed bioinformatic analysis. JCC and CAT wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements We would like to thank Dr. Yannery Gómez Bonilla (Investigator, Entomologia, INTA, Costa Rica) for the samples from Costa Rica, Dr. Phil Stansly (Professor, University of Florida, USA) for the *K. lycopersicella* samples, and David Haviland (Entomology and IPM Farm Advisor, University of California Cooperative Extension, Bakersfield, CA, USA) for the *P. operculella* samples.

Funding This project was supported by the Specialty Crop Block Program at the USA Department of Agriculture (USDA) through Grant 14-SCBGP-CA-006 to KG, FGZ, and JCC and USDA NIFA CA-D-ENM-2150-H awarded to JCC. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the USDA. JA and NA are funded by CERCA Programme (Generalitat de Catalunya).

Data availability All genome sequencing data generated from this study have been deposited to BioProject accession number PRJNA512383 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/). The Whole Genome Shotgun project of *T. absoluta* has been deposited at DDBJ/ENA/GenBank under the accession SNMR00000000. The version described in this paper is version SNMR01000000.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ajamma YU, Mararo E, Omondi D et al (2016) Rapid and high throughput molecular identification of diverse mosquito species by high resolution melting analysis. F1000Research 5:1949

Arnó J, Oveja MF, Gabarra R (2018) Selection of flowering plants to enhance the biological control of *Tuta absoluta* using parasitoids. Biol Control 122:41–50

Bahamondes LA, Mallea AL (1969) Biología en Mendoza de *Scrobipalpula absoluta* (Meyrick) Polvony (Lepidoptera: Gelechiidae), especie nueva para la República Argentina. Revista de la Facultad de Ciencias Agrarias 15:96–104

Bawin T, Dujeu D, DeBaker L et al (2015) Could alternative solanaceous hosts act as refuges for the tomato leafminer, *Tuta absoluta*? Arthropod Plant Interact 9:425–435

Bettaibi A, Mezghani-Khemakhem M, Bouktila D et al (2012) Genetic variability of the tomato leaf miner (*Tuta absoluta* Meyrick; Lepidoptera: Gelechiidae) in Tunisia, inferred from RAPD-PCR. Chilean J Agric Res 72:212–216

Biondi A, Narciso R, Guedes C et al (2018) Ecology, worldwide spread, and management of the invasive South American tomato pinworm, *Tuta absoluta*: past, present, and future. Ann Rev Entomol 63:239–258

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30:2114–2120

CABI (Centre for Agriculture and Bioscience International) (2016) *Tuta absoluta*. CABI website https://www.cabi.org/isc/datasheet/49260. Accessed 20 Oct 2018

Camargo RA, Herai RH, Santos LN et al (2015) *De novo* transcriptome assembly and analysis to identify potential gene targets for RNAi-mediated control of the tomato leafminer (*Tuta absoluta*). BMC Genom 16:635

Camargo RA, Barbosa GO, Possignolo IP et al (2016) RNA interference as a gene silencing tool to control *Tuta absoluta* in tomato (*Solanum lycopersicum*). Peer J 4:e2673

Campos MR, Biondi A, Adiga A (2017) From the Western Palaearctic region to beyond: *Tuta absoluta* ten years after invading Europe. J Pest Sci 90:787–796

Cifuentes D, Chynoweth R, Bielza P (2011) Genetic study of Mediterranean and South American populations of tomato leafminer *Tuta absoluta* (Povolny, 1994) (Lepidoptera: Gelechiidae) using ribosomal and mitochondrial markers. Pest Manag Sci 67:1155–1162

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5:e11147

Desneux N, Wajberg E, Wyckhuys K et al (2010) Biological invasion of European tomato crops by *Tuta absoluta*: ecology, geographic expansion and prospects for biological control. J Pest Sci 83:197–215

Desneux N, Luna M, Guillemaud T et al (2011) The invasive South America tomato pinworm, *Tuta absoluta*, continues to spread in Afro-Eurasia and beyond: the new threat to tomato world production. J Pest Sci 84:403–408

Dhami MK, Kumarasinghe L (2014) A HRM real-time PCR assay for rapid and specific identification of the emerging pest Spotted-Wing Drosophila (*Drosophila suzukii*). PLoS ONE 9:e98934

Dhami MK, Dsouza M, Waite DW et al (2016) Real-time PCR assay for the identification of the brown marmorated stink bug (*Halyomorpha halys*). Front Mol Biosci 3:5

Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21

Đurić Z, Delić D, Hrnčić S et al (2014) Distribution and molecular identification of *Tuta absoluta* (Meyrick, 1917) (Lepidoptera, Gelechiidae) populations in Bosnia and Herzegovina and Montenegro. Pol J Entomol 83:121–129

EPPO Global Database (2019) *Tecia solanivora* (TECASO). https://gd.eppo.int/taxon/TECASO/distribution. Accessed 15 Feb 2019

FAOSTAT (Food Agric. Org. U. N.) (2017) FAOSTAT statistics database. Rome, Italy. http://www.fao.org/faostat. Accessed 20 Oct 2018

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet 2:1–18

Gaskill DAR (2013) *Tuta absoluta* and the tomato commodity survey. Florida CAPS annual workshop and meeting presentation. https://www.freshfromflorida.com/Divisions-Offices/Plant-Industry/Bureaus-and-Services/CAPS/CAPS-Presentations-and-Outreach. Accessed 15 Feb 2019

Guillemaud T, Blin A, Le Goff I et al (2015) The tomato borer, *Tuta absoluta*, invading the Mediterranean Basin, originates from a single introduction from Central Chile. Sci Rep 5:8371

Han P, Zhang YN, Lu ZZ et al (2018) Are we ready for the invasion of *Tuta absoluta*? Unanswered key questions for elaborating an integrated pest management package. Entomol Gen 38:113–125

Han P, Bayram Y, Shaltiel-Harpez L et al (2019) *Tuta absoluta* continues to disperse in Asia: damage, ongoing management and future challenges. J Pest Sci. https://doi.org/10.1007/s10340-018-1062-1

Katoh K, Standley D (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Bio Evol 30:772–780

Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in meta-transcriptomic data. Bioinformatics 28:3211–3217

Lee S, Brambila J (2012) A new species of the genus *Sinoe* (Lepidoptera: Gelechiidae: Litini) from Florida. Fla Entomol 95:873–876

Lee Y, Marsden CD, Norris LC et al (2013) Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. Proc Natl Acad Sci USA 110:19854–19859

Lee Y, Marsden CD, Nieman CC et al (2014) A new multiplex SNP genotyping assay for detecting hybridization and introgression between the M and S molecular forms of *Anopheles gambiae*. Mol Ecol Resour 14:297–305

Lee Y, Weakley AM, Nieman CC et al (2015) A multi-detection assay for malaria transmitting mosquitoes. J Vis Exp 96:e52385

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760

Linck H, Kruger E, Reineke A (2017) A multiplex TaqMan aPCR assay for sensitive and rapid detection of phytoplasmas infecting Rubus species. PLoS ONE 12:e0177808

Luo R, Liu B, Xie Y et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18

Mansour R, Brevault T, Chailleux A et al (2018) Occurrence, biology, natural enemies and management of *Tuta absoluta* in Africa. Entomol Gen 38:83–112

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770

Michalak PS (2011) New pest guidelines for tomato leaf miner (*Tuta absoluta*). http://www.aphis.usda.gov/import_export/plants/manuals/emergency/index.shtml. Accessed 20 Oct 2018

Mohamed ESI, Mahmoud MEE, Elhaj MAM (2015) Host plants record for tomato leaf miner *Tuta absoluta* (Meyrick) in Sudan. Bull OEPP/EPPO 45:108–111

Mutamiswa R, Machekano H, Nyamukondiwa C (2017) First report of tomato leaf miner, *Tuta absoluta* (Meyrick) (Lepidoptera: Gelechiidae), in Botswana. Agric Food Secur 6:49

Negi S, Sharma PL, Sharma KC (2018) Effect of host plants on developmental and population parameters of invasive leafminer, *Tuta absoluta* (Meyrick) (Lepidoptera: Gelechiidae). Phytoparasitica 46:213–221

Nieman CC, Yamasaki Y, Collier TC et al (2015) A DNA extraction protocol for improved DNA yield from individual mosquitoes. F1000Research 4:1314

Pereyra P, Sanchez N (2006) Effect of two solanaceous host plants on development and population parameters of the tomato leaf miner, *Tuta absoluta* (Meyrick) (Lepidoptera: Gelechiidae). Neotrop Entomol 35:671–676

Povolny D (1975) On three neotropical species of Gnorimoschemini (Lepidoptera, Gelechiidae) mining Solanaceae. Acta Univ Agric 23:379–393

Povolny D (1994) Gnorimoschemini of southern South America VI: identification keys, checklist of Neotropical taxa and general considerations (Insecta, Lepidoptera, Gelechiidae). Steenstrupia 20:1–42

Simao FA, Waterhouse RM, Ioannidis P et al (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Sint D, Sporleder M, Wallinger C et al (2016) A two-dimensional pooling approach towards efficient detection of parasitoid and pathogen DNA at low infestation rates. Methods Ecol Evol 7:1548–1557

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Teeter KC, Thibodeau LM, Gompert Z et al (2010) The variable genomic architecture of isolation between hybridizing species of house mice. Evolution 64:472–485

Tonnang HEZ, Mohamed SF, Khamis F (2015) Identification and risk assessment for worldwide invasion and spread of *Tuta absoluta* with a focus on sub-Saharan Africa: implications for phytosanitary measures and management. PLoS ONE 10:e0135283

USDA (United States Department of Agriculture) (2011) New pest response guidelines: tomato leafminer (*Tuta absoluta*). USDA Animal and Plant Health and Inspection Service website. https://www.aphis.usda.gov/import_export/plants/manuals/emergency/downloads/Tuta-absoluta.pdf. Accessed 20 Oct 2018

Vurture GW, Sedlazeck FJ, Nattestad M et al (2017) GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33:2202–2220

Yamasaki YK, Nieman CC, Chang AN (2016) Improved tools for genomic DNA library construction of small insects. F1000Research 5:211

Zhang Z, Schwartz S, Wagner L et al (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214

Zhang T, Wang YJ, Guo W (2016) DNA barcoding, species–species PCR and real-time PCR techniques for the identification of six *Tribolium* pests of stored products. Sci Rep 6:28494