PROJECT DELIVERABLE REPORT

Deliverable D3.5: manual multi-step work flow and decision protocol available to NPPO

Fruit Flies In-silico Prevention & Management

FF•IPM

Project Title:

In-silico boosted, pest prevention and off-season focused IPM against new and emerging fruit flies ('OFF-Season' FF-IPM)

SFS-2018-2

"This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 818184– H2020-SFS-2018-2

Document Information

Grant Agreement Number	818184	Acronym	FF-IPM
Full Title	In-silico boosted, pest prevention and off-season focused IPM against new and emerging fruit flies ('OFF-Season' FF-IPM)		
Topic	SFS-05-2018-2019-2020 New and emerging risks to plant health		
Funding scheme	RIA - Research and Innovation action		
Start Date	1st September 2019	**Duration**	48 months
Project URL	http://fruitflies-ipm.eu/		
EU Project Officer	George PREDOIU		
Project Coordinator	UNIVERSITY OF THESSALY - UTH		

Deliverable				
Work Package	WP3 – Development and enhancement of tools and methods for FF prevention			
Date of Delivery	**Contractual**	M42	**Actual**	M54
Nature	R - Report	**Dissemination Level**	PUBLIC	
Lead Beneficiary	RMCA			
Responsible Researcher	Marc De Meyer	Email	demeyer@africamuseum.be	
		Phone	+32 2 7695360	
Reviewer(s):	FF-IPM Consortium			
Keywords	Identification, barcodes, genome sequences			

Revision History

Version	Date	Responsible	Description/Remarks/Reason for changes
0.10	26/12/2023	RMCA, P. Deschepper & M. De Meyer	First draft
0.20	28/12/2023	RMCA, P. Deschepper	Second draft
0.30	16/01/2024	UTH, N. Papadopoulos	Revision
0.40	17/01/2024	RMCA, P. Deschepper & M. De Meyer	Third draft
0.50	18/01/2024	BPI, P. Mylonas	Revision of the third draft
0.60	23/01/2024	RMCA, P. Deschepper & M. De Meyer	Final version
0.70	24/01/2024	UTH, N. Papadopoulos	Review of the final draft
0.80	30/01/2024	EB	Approval of the final draft
0.90	08/02/2024	UTH, G. Pahlitzanakis	Submission of deliverable

Contents

1. Summary

This deliverable brings together the different identification tools developed or enhanced within the framework of the FF-IPM project and combines them into an applicable decision-making workflow and protocol suitable for individuals or agencies involved in identification services for the EU horticultural industry.
The tools cover both morphological and molecular approaches and address different needs and queries with regard to intercepted or detected fruit fly species and specimens of the target species *Ceratitis capitata (Cc)*, *Bactrocera dorsalis* (*Bd*) and *Bactrocera zonata (Bz)*. The following components were developed:

1. Mobile application for 23 adult fruit flies of economic importance to the EU: Morphological identification of adult intact specimens of target fruit fly species and closely related congeners
2. Mobile application for 13 larval fruit flies of economic importance to the EU: Morphological identification of juvenile intact specimens of target fruit fly species and closely related congeners
3. Loop mediated iso-thermal Amplification (LAMP): Molecular identification of all life stages in a limited time frame and with limited infrastructure of target fruit fly species and closely related congeners
4. DNA-barcodes: Molecular identification of all life stages of target fruit fly species and an extensive range of other fruit infesting tephritid taxa
5. Diagnostic Single Nucleotide Polymorphisms (SNPs): Molecular identification of all life stages of different geographical populations of target fruit fly species.

Regarding the latter we also present a machine learning approach to find diagnostic SNPs markers in the three target species, which are characterized by a highly complex demographic history and high levels of gene flow. We report genomic locations of 100 and 130 genomic SNPs in *C. capitata* and *B. dorsalis* respectively and 80 for *B. zonata* that allow for a prediction of the origin of a given samples with an accuracy between 75 and 90%. We show that the success of a random forest approach in tackling the origin tracing question is highly dependent on the grouping of the samples in coherent genetic clusters and that more specificity in terms of geographic location translates into a higher number of SNPs needed. Compared to other methods, the approach described here has the advantage in that it does not need additional reference material for direct comparison, but rather predicts the origin with a reported confidence by feeding SNP data directly into a statistical model. Trained models can easily be shared due to their small size (typically less than 10Mb) and in contrast to other methods (DAPC, admixture analysis or inference of phylogenetic clustering) do not require high performing infrastructures such as HPCs or expensive high computation setups.

2. Introduction

2.1. Scope and purpose

The family Tephritidae has more than 5000 species distributed globally (White & Elson-Harris, 1992, EFSA 2020). The larvae of about 35% of the species attack fruit that include fruit crops of economic importance (White & Elson-Harris, 1992). Some of these species are among the most destructive pests of fruit and vegetables and are of quarantine importance for the export market (Ekesi *et al.,* 2016). The larvae feed in fruit and this is the life stage detected during inspection of fruit for import or export, but are difficult to identify (Frías et al. 2008, Dutra et al. 2012). Trapping activities on the other hand will always yield adult specimens but their condition can be deteriorated because of the poor preservation in the trapping device. When any biological material (specimen or parts thereof) is intercepted or detected and considered of quarantine importance, a correct identification is essential to comply with international biosecurity measures, since not all species are of quarantine importance in all countries (Boykin et al., 2012). Such identification should be unambiguous in order to deploy the necessary measures to mitigate the potential danger of such unwanted introduction. The methodology used in the identification will differ according to the kind of biological material detected (life stage, physical condition), the time constraints in which the identification needs to be conducted, the expertise and infrastructure available to conduct the identification process and the actual objective of the identification (i.e. at species level or population level).
Within the framework of FF-IPM, one of the main objectives was the development of novel tools that would aid in the identification process at different level and for different intentions, and to combine this suite of identification tools in a decision protocol that will help the end-user in the actual process. The following major tools were developed and included in the protocol, with the different objectives for each tool:

1. Mobile application for 23 adult fruit flies of economic importance to the EU: Morphological identification of adult intact specimens of target fruit fly species and closely related congeners
2. Mobile application for 13 larval fruit flies of economic importance to the EU: Morphological identification of juvenile intact specimens of target fruit fly species and closely related congeners
3. Loop mediated iso-thermal Amplification (LAMP): Molecular identification of all life stages in a limited time frame and with limited infrastructure of target fruit fly species and closely related congeners
4. DNA-barcodes: Molecular identification of all life stages of target fruit fly species and an extensive range of other fruit infesting tephritid taxa
5. Diagnostic Single Nucleotide Polymorphisms (SNPs): Molecular identification of all life stages of different geographical populations of target fruit fly species.

All these tools are incorporated into a multi-step workflow and decision protocol to facilitate the user. Tools 1 and 2 were presented and explained in Deliverable 3.3., while tool 4 and the genomic sequence production used to develop the diagnostic SNPs are presented in Deliverable 3.4. Detailed description of these steps is thus not presented in the current deliverable but can be consulted in the above-mentioned deliverables.

2.2. Arrangements

The structure below is organized following the division into the five different elements as listed above, but grouped in four main groups:

1. mobile applications
2. LAMPs
3. DNA-barcodes
4. Diagnostic SNPs

This is followed by a final chapter incorporating the above in the final multi-step workflow and decision protocol. For those items covered in previously presented deliverable, a reference to the relevant part is included.

3. Mobile Applications

3.1. Material & methods

ID keys already exist for adult African fruit flies developed by the Royal Museum for Central Africa (RMCA, see Identification tools > Multi-entry key | Fruit flies (africamuseum.be)) and Australian and SE Asian fruit flies developed by Plant Health Australia (PHA; see https://fruitflyidentification.org.au/) which were used as the basis for the development of the restricted key within the FF-IPM project. After the list of selected fruit fly species was made, the existing character states were verified. Available images, illustrating the different character states were selected from existing image databases and new images were made where applicable.

For the larval identification app, the only useable features on a fruit fly larva are the cephalopharyngeal skeleton containing the mouth hooks and spiracles, which were therefore used to compile the larval key, and included various angle and distance measurements (mouth hooks) and position of front spiracles, as well as number of tubules. A total of 13 larval species were assessed, based largely on availability of material from colonies in various countries. Fact sheets for each of the shortlisted adult fruit fly species were compiled, along the outline of existing fact sheets for African fruit flies developed previously by the RMCA, and include information on morphology, identification through DNA barcoding, biology, host plants, impact and management, attractants and trapping, distribution, and quarantine regulations. Datasheets with images and measurements of each species were prepared to be used with the key. The LUCID software (https://www.lucidcentral.org/) was used for the development of two multi-entry keys, including the conversion: into mobile applications.

The development of the adult and larvae key has been presented in more detail in Deliverable 3.3. point 3.2 and 3.3 respectively.

3.2. Results

Based on both keys a selection of 47 morphological characters including 106-character states, were incorporated in the LUCID key. This comprises 22 characters states for generic recognition (i.e., *Bactrocera/Zeugodacus/Dacus* (BZD) versus *Ceratitis*), 63-character states for recognition within the BZD group, and 21 characters states for recognition within the *Ceratitis* group. The scoring of these character states for selected species was verified, their diagnostic power evaluated (i.e., the ability to differentiate all taxa from each other), and the general key functioning tested by having non-specialists going through the key and conducting identification of blind fruit fly samples. This allowed testing of possible erroneous scoring of character states and of positive and negative dependencies between character states (i.e., showing up of particular characters after prior ones were selected or not). The mobile App for a selection of adult fruit flies was tested and is now freely available on android and apple. As with the key to the adults described above, a mobile APP was developed for third instar larvae using the LUCID Builder software for a total of 13 species of economic significance: *Bactrocera dorsalis, B. zonata, B. oleae, B. tryoni, B. tau, B. correcta, B. minax, Ceratitis capitata, C. rosa, C. quilicii, C. cosyra, Zeugodacus cucurbitae* and *Dacus ciliatus*. The multi-entry key is based on 4 characters of the secondary tooth and front spiracle, and 5 characters involving distance and angle measurements. Factsheets with all information as listed under 3.1. are included in the mobile App.
It is the first time that a mobile application for identification specifically of fruit flies has been developed. The mobile App for larvae is currently available as a beta version and is undergoing testing to determine

any changes needed. Supporting information, such as the factsheets and videos are on freely available platforms to assist the non-expert with making Ids.

Detailed results are presented in Deliverable 3.3.

4. LAMP

4.1. LAMP for Ceratitis species

4.1.1. Material & Methods

The aim was to design a LAMP assay to detect *C. capitata* from among related taxa. The test includes the following species: *Ceratitis capitata* (target fruit fly species) and congeners *C. fasciventris, C. anonae, C. rosa, C. quilicii* (i.a. all members of the *Ceratitis* FARQ complex, see Zhang et al., 2021), *C. rubivora, C. cosyra* and *C. quinaria*. Subsequently genomic DNA was extracted from single samples using a TIANamp Micro DNA Kit (Tiangen, Beijing, China), and used a UV–Vis Spectrophotometer Q5000 (Quawell, USA) to test DNA quality.

Focus for primer design started from *cox1* mitochondria DNA given the abundant data available for this gene. Twelve additional protein-coding genes in the mitochondrial genome were included in case ideal specific primers cannot be obtained based on *cox1*. The sequences of all available *Ceratitis* homologous genes were multiple sequence aligned in MEGA 7 (Kumar et al., 2016). The intraspecific conserved and specific loci of the target fruit fly species were manually screened and specially marked in Geneious 10.1.3 (Kearse et al., 2012). Then the target species sequences were submitted with special markers at specific sites to Primer Explore V4 (http://primerexplorer.jp/elamp4.0.0/index.html) [26] and automatically designed as LAMP primers. (F1c/B1c = 20-28 bp, F2/B2 = 18-25 bp, F3/B3 = 20-28 bp; Tm: F1c/B1c = 59-61°C, F2/B2 = 54-56°C, F3/B3 = 54-56°C; GC rate = 40-50%) (Zhao et al. 2007). The LAMP primers that satisfy primers F2 and B2 that contain specific sites at the 3' ends and the conditions(|5'dG|>4, |3'dG|>4, |dimer dG|< 2) could then be used as alternative primers for specificity test. The primers were synthesized by Tsingke Biological Technology (Beijing, China).

The specificity of LAMP primers was tested by 2% agarose gel electrophoresis, including genomic DNA of a single *Ceratitis* fruit fly from each geographic region as positive control, and negative controls were DNA samples from other *Ceratitis* species and ddH_2O. The primers are specific only if the samples of all the geographical populations of the target species show evident trapezoidal strips and no such strips appear in the control species or ddH_2O through agarose gel electrophoresis. The total volume of LAMP amplification is 25.0 μL, including Warm Start LAMP 2× Master Mix E1700S (New England Biolabs Inc., MA, US). 12.5 μL, 0.5 μL each of 10 μM F3 and B3 primers, 4.0 μL each of FIP and BIP primers, template DNA 1.0 μL, ddH_2O 2.5 μL. After incubation at 63 ℃ for 40 min, the reaction was terminated at 85 ℃ for 3 min. After the amplification, 5 μL products were added into agarose gel electrophoresis channels for 30 min at 220 V. Then, examine the results under UV light using Gel Logic 212 PRO (Carestream Kodak, USA) to determine whether the primers were specific or not.

Referred to the LAMP primer ratio and amplification temperatures set by Li et al. (2020), using the WarmStart LAMP Kit (DNA and RNA) E1700S (New England Biolabs Inc., MA, US) to optimize the *Ceratitis* LAMP reaction systems. The optimal concentration ratio was screened from the primer concentration ratio (F3/B3:FIP/BIP = 1:4, 1:6, 1:8, 1:10), and the optimal reaction temperature was screened from four reaction temperature conditions, including 61, 63, 65 and 67℃ for 40 min, the reaction ended at 85℃ for 3 min.

After screening LAMP-specific primers and determining the optimal primer concentration ratio and reaction temperature, the concentration of template DNA was changed by gradient dilution of template DNA so as to test the sensitivity of LAMP primers. Here, ddH_2O was used as a negative control to detect whether the LAMP reaction could effectively amplify template DNA at 10.0 ng/μL, 1.0 ng/μL, 0.1 ng/μL and 0.01 ng/μL.

In the LAMP visual detection, the total reaction volume is 40.0 μL, including WarmStart Colorimetric LAMP 2× Master Mix M1800S (New England Biolabs Inc., MA, US) 25.0 μL, 0.5 μL each of 10 μM F3 and B3 primers, 4.0 μL each of FIP and BIP primers, template DNA 2.0 μL, ddH_2O 4 μL. The mixture was incubated at 63 ℃ for 40 min, and after the amplification, we observed the color change to judge whether the visual detection can be effectively realized.

In order to make the LAMP identification technology better serve the on-site detection at the port and improve the timeliness, the rapidly rough DNA extraction method of Kitano and Takakura (2020) was adopted. The extraction steps protocol is formulated as follows: Take part of the thoracic muscle tissue of the fruit fly samples into a 0.2 mL centrifuge tube, add 30 μL of TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0), and place the 0.2 mL centrifuge tube in a water bath. The reaction will be carried out at 95 °C for 3 min, and the supernatant will contain the roughly extracted DNA. Take 2.0 μL of the supernatant from the above reaction as a DNA template for LAMP visual detection.

Further details on material and methods are provided in Zhang et al. (2023).

4.1.2. Results

Primer design

LAMP-specific primers of the *Ceratitis* FARQ complex were designed and screened on the *cob* gene, and the specific primers of *C. cosyra* and *C. capitata* were designed and screened on the *cox1* gene. All primer sequences and length information are given in Table 4.1.

Table 4.1. Three *Ceratitis* economically important fruit flies LAMP specific primers

Species	Primers	Sequences (5'-3')	Length
Ceratitis FARQ complex	FARQ-F3	GGATTATTCCTTGCTATACATTAC	24bp
	FARQ-B3	GTTGGGGTGAATAAATAAGATC	22bp
	FARQ-FIP	TCATCCATAATTTACATCTCGACAACA GCAGATATTAATTTAGCTTTC	48bp
	FARQ-BIP	ATTACGAACTATACACGCTAACGCATA ATACATTCCACGTCCTAC	45bp
C. cosyra	Ccos-F3	GGGTTTGGAATAATTTCCCATA	22bp
	Ccos-B3	AGTGTAGCAAGTCAACTGAA	20bp
	Ccos-FIP	CCAAGTAAACCGATTGCTAATATAGCA CAAGAATCAGGGAAAAAGGAGA	49bp
	Ccos-BIP	TTGTTTGAGCCCACCATATATTTACAG GCAATAATTATTGTTGCTGATGT	50bp
C. capitata	Ccap-F3	AGGAGCTGTAAATTTTATCACA	22bp
	Ccap-B3	GTTGGTATAAAATAGGGTCTCC	22bp
	Ccap-FIP	TGCAGTAAGAACTACTGCTCAAACAAC AGTAATTAATATACGATCTACCG	50bp
	Ccap-BIP	ACCAGTTTTAGCAGGAGCTATTACCTC CAGCTGGGTCAAAGA	42bp

The specificities of the three groups of primers in Table 4.1. were detected, respectively, by taking the *Ceratitis* FARQ complex, *C. cosyra*, and *C. capitata* as positive controls, and taking seven other different species and geographical populations of *Ceratitis* fruit flies as negative controls. The results of agarose gel electrophoresis show evident trapezoidal strips, *Ceratitis* specific LAMP primers sets showed strong specificity to the target samples (Figure 4.1).

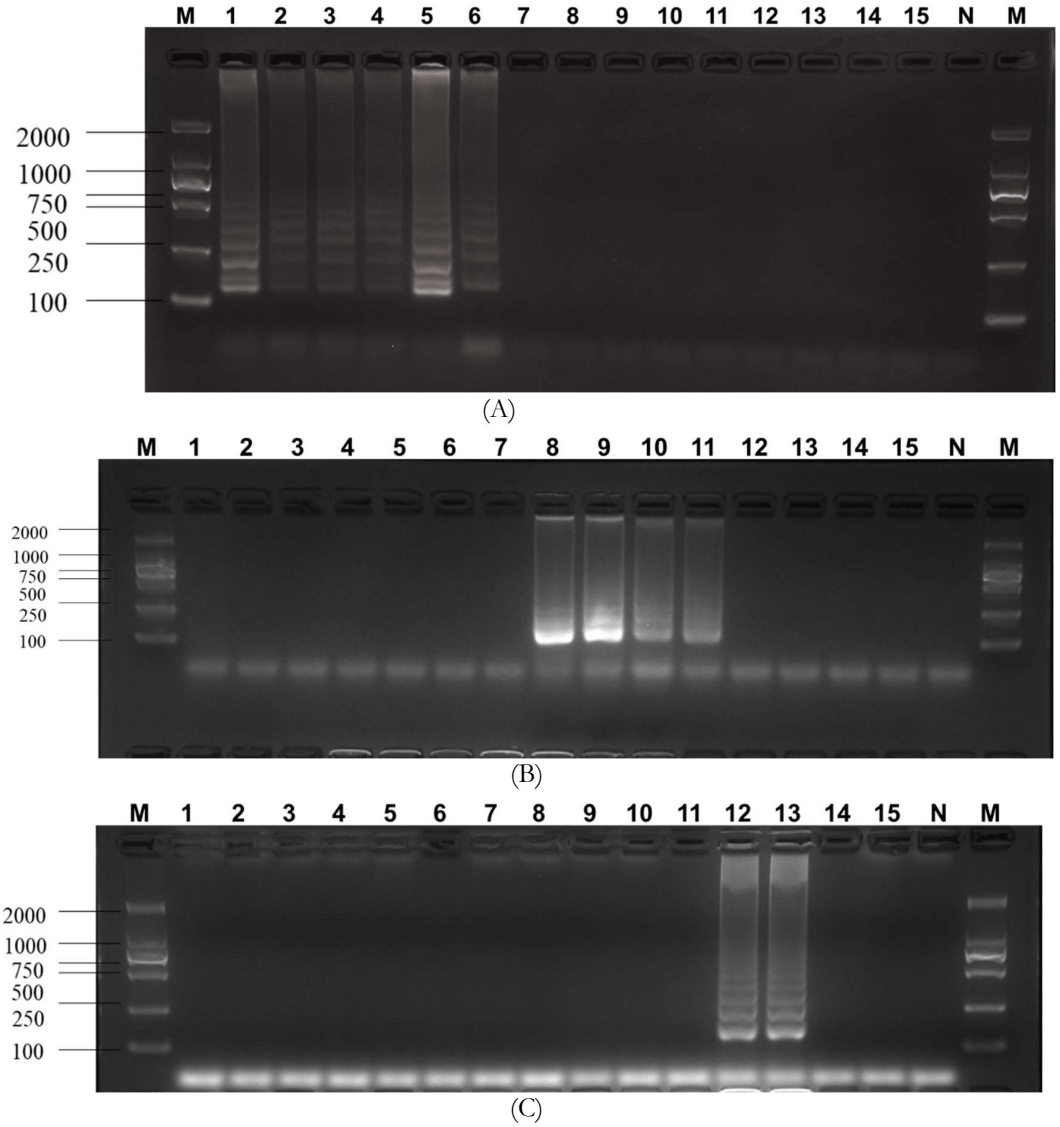


Figure 4.1. *Ceratitis* LAMP primers sets specificity verification by agarose gel electrophoresis analysis. Lane M. D2000 Marker, Lane 1. *C. fasciventris* Burundi, Lane 2. *C. anonae* Cameroon, Lane 3. *C. rosa* Tanzania, Lane 4. *C. quilicii* Tanzania, Lane 5. *C. quilicii* Kenya, Lane 6. *C. quilicii* South Africa, Lane 7. *C. rubivora* Tanzania, Lane 8. *C. cosyra* South Africa, Lane 9. *C. cosyra* Mozambique, Lane 10. *C. cosyra* Nigeria, Lane 11. *C. cosyra* Kenya, Lane 12. *C. capitata* South Africa, Lane 13. *C. capitata* Kenya, Lane 14. *C. quinaria* Sudan, Lane 15. *C. quinaria* Sudan, Lane N. negative control. (A) *Ceratitis* FARQ complex. (B) *C. cosyra.* (C) *C. capitata.*

Visual detection

Visual detection was developed to make it easier to observe LAMP outcomes in realistic applications. After amplifying the corresponding positive samples with specific LAMP primers, the color of the 25 µL LAMP reaction system single-tube, changed from pink to khaki, while the negative control remained pink. The results showed that the three groups of specific primers all achieved visual detection, thus making the practical application of LAMP possible.

LAMP workflow

After immersing the muscle tissue of *Ceratitis* samples in TE buffer and treating them at 95 °C for 3 min, the supernatant was used as a DNA template. All geographic population samples are subjected to LAMP reactions using the specific primers, the amplification results can be directly observed with naked eyes. The color of the positive control obviously changed from pink to khaki, while the color of the negative control remained pink. The LAMP detection system, which is based on rapid rough DNA extraction, greatly reduces the time required for a series of identification processes, beginning with DNA extraction, to less than 1 hour.

Additional information on the LAMP methodology for *Ceratitis* species can be found in Zhang et al. (2023).

4.2. LAMP for Bactrocera species

4.2.1. Material & Methods

The aim was to design a LAMP assay to detect *B. zonata* from among related taxa. A specific test for the other target fruit fly species, *B. dorsalis*, the Oriental fruit fly has not been developed as such LAMP was published by the time the FF-IPM project started (Blaser et al., 2018). The test includes the following species: *B. zonata* (target fruit fly species) and congeners: *B. dorsalis*, *B. correcta*, *B. tryoni*, *B. latifrons*, *B. oleae*, *B. cucurbitae* and *B. tau* (the latter two are currently placed in the genus *Zeugodacus* but this higher taxon is considered to be closely related to *Bactrocera*). DNA was extracted using the Qiagen Dneasy Blood and Tissue Kit with optimized protocol for insect tissue. In addition, to assess more field-friendly DNA extraction method larval posterior and anterior halves were incubated in 30 µL TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) at 95 °C for 5 min and used as-is for LAMP assays as the supernatant would contain the roughly extracted DNA.

Focus for primer design start from *cox1* mitochondria DNA given the abundant data available for this gene. No other protein-coding genes were included. LAMP primer specificity was assessed using DNA extracted from all *Bactrocera* species. Assay success was evaluated by running products on a 1.5% Agarose gel at 90V for 1hr and visualizing the resultant products under UV light. A negative control was run in every instance. In order to optimize the protocol for a more field-friendly nature a more rapid assay confirmation method than the Agarose gel was assessed.

A LAMP assay was conducted and instead of loading the resultant product on an agarose gel, 10 µL of the product was mixed with 0.5 µL Ethidium Bromide (Promega) in the 0.2 mL vial it was amplified in and directly visualized under UV light. Primer specificity would be confirmed if fluorescence was observed only for the *B. zonata* sample and all other samples resembled the negative control.

Further details on material and methods are provided in Bierman et al. (submitted); included as Annex IV.

4.2.2. Results

Primer design

LAMP-specific primers for *B. zonata* were designed and screened on the *cox1* gene. All primer sequences and length information are given in Table 4.2.

Table 4.2.: Primer sequences of LAMP primers designed for specific amplification of *Bactrocera zonata* DNA.

Primer	Sequence (5' - 3')
F3	GGA GGA TTT GGA AAT TGA CTT
B3	AAA TAG CTA GAT CAA CTG AAG C
FIP	GCG TAA GGG AAG GAG GTA ATA ATC AGT TCC CCT AAT ATT AGG AGC AC
BIP	AAG TAT AGT AGA AAA CGG AGC TGG TCC GTG AGC AAT AAC AGA TGA
LF	ATT CAT TCG TGG GAA TGC TAT GTC G
LB	AGG TTG AAC AGT TTA TCC TCC CCT A

Primer specificity was experimentally confirmed by performing the LAMP assay with the selected primers (Table 4.2.) using DNA extracted with the Qiagen DNeasy Blood and Tissue kit (Qiagen) from all *Bactrocera* specimens in this study. Negative controls for the assay was run in every instance where all components of the assay were added except for DNA. The specificity was visually confirmed on agarose gel electrophoresis where the ladder-like amplification of *B. zonata* only can be seen around 500 bp, with all other samples resembling the negative control (Figure 4.2.).

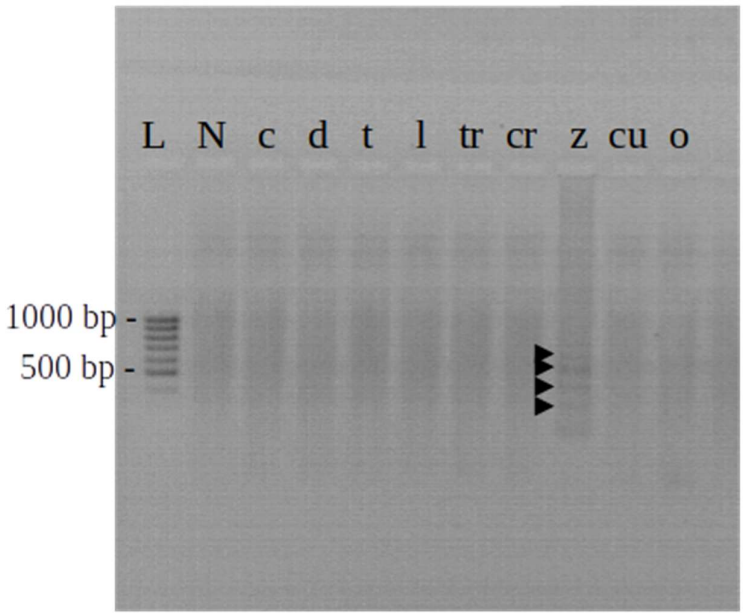


Figure 4.2.: LAMP assay products on 1.5% Agarose gel. From left to right lanes contain: L - 100 bp ladder; N - negative control; c - *Ceratitis capitata*; d - *Bactrocera dorsalis*; t - *Bactrocera tau*; l - *Bactrocera latifrons*; tr - *Bactrocera tryoni*; cr - *Bactrocera correcta*; z - *Bactrocera zonata*; cu - *Bactrocera cucurbitae*; o - *Bactrocera oleae*. The ladder-like bands around the 500 bp mark for *B. zonata* (highlighted by black arrows) indicate the successful amplification of the LAMP primers while all other samples resemble the negative control.

LAMP workflow

LAMP assay workflow, including the steps and duration in time for each part of the workflow is depicted below (Figure 4.3.): from dissection of larval samples on the left, through to the DNA extraction in TE buffer and transfer of 1µL crude extract to the LAMP assay mixture followed by addition of 0.5 µL of a UV-fluorescent dye such as Ethidium Bromide and immediate visualization under UV light. The complete work-flow can be accomplished in under 1 hour if sample material is at hand.

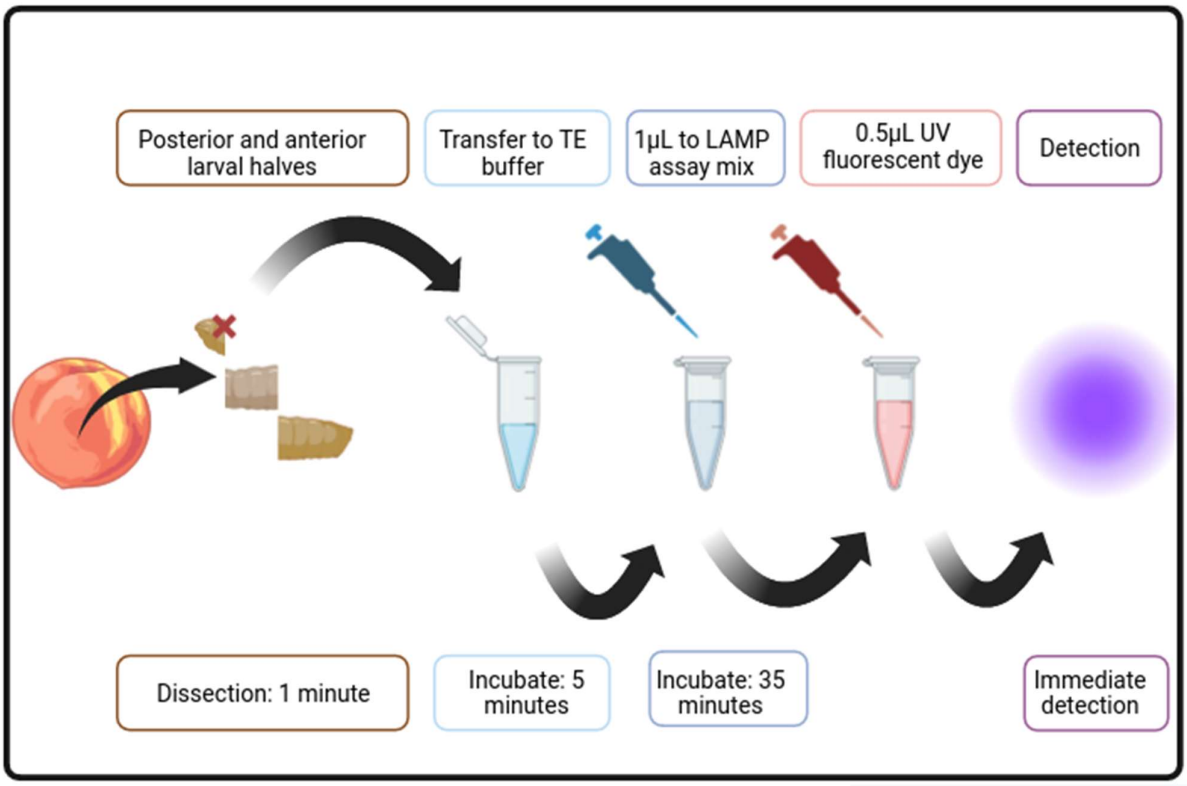


Figure 4.3.: Workflow for the LAMP assay.

Additional information on the LAMP methodology for *Bactrocera* species can be found in Bierman et al. (submitted, included as Annex IV).

5. DNA-barcodes

5.1. Material and methods

DNA barcodes were generated in two ways: either developed directly through Sanger Sequencing using the standard COX1 barcoding primers or extracted from genomic data obtained through Next Generation Sequencing. In the latter case, a consensus sequence was obtained for the COI barcode region from the variant called sequence alignment file.
Methodology for the production of DNA barcodes has been presented in more detail in Deliverable 3.4. point 3.2.

5.2. Results

Regarding DNA barcodes: 1,854 specimens were DNA barcoded and made available through the main online library and BOLD (Barcoding of Life Database) repository, as well as GENBANK. In addition, upon request of the European and Mediterranean Plant Protection Organization (EPPO), newly generated DNA barcodes as well as existing libraries generated by the partners involved, were made available to the EPPO-Q-Bank, which is a reference database for DNA barcodes specifically geared towards pest species

of relevance to the European Union. In total, DNA barcodes for 173 species were made available through the different open access libraries. More detailed results have been presented in Deliverable 3.4. point 4.2.

6. Diagnostic SNPs

The preliminary extended DNA sequencing reads, based on whole-genome sequencing, for the three target species have been presented in Deliverable 3.4. point 3.1.

6.1. Ceratitis capitata

The Mediterranean fruit fly, Medfly, *C. capitata* , is one of the three target fruit flies in the FF-IPM project. Of African origin, it has been introduced at various locations outside the species' natural range on many occasions and successful establishment has been reported. During the past century it has spread to other tropical and more temperate regions worldwide and is now established in northern Africa, the Mediterranean, Latin America, Australia and the Hawaiian islands and is annually detected in California. Under a warming climate, incursions of this warmth loving species is increasing annually. In many cases, intercepted flies are of unknown origin which make it hard to pinpoint trade routes that facilitate transfer of this pest species. For this reason, tools that aid in tracing the origin of intercepted specimens are increasingly important to tackle further spread of *C. capitata*.

6.1.1. Material and Methods

Sample collection and genotyping

A total of 134 samples of *C. capitata* were collected from 25 locations in 20 countries. Samples originating from Africa were provided by the RMCA while samples from the Mediterranean were field collected in 2020-2021. Additional samples from Central and South America were provided by Konstantinos Bourtzis (International Atomic Energy Agency, IAEA). Sampled populations are located in following main geographical regions where *C. capitata* is well established: West Africa, East Africa, the Mediterranean, Central America and South America (Table 6.1.). Geographic coverage is depicted in Figure 6.1.

Table 6.1.: Overview of sampled locations used for detection of diagnostics SNPs

Region	Location	Nr. Samples	Nr. Samples in region
West Africa	Benin	5	9
	Senegal	4	
East Africa	Burundi	3	22
	Kenya	2	
	Mozambique	5	
	South Africa	4	
	Tanzania	5	
	Zambia	3	
Mediterranean	Croatia - Opuzen	6	57
	Croatia - Split	7	
	Greece - Thessaloniki	7	
	Greece - Volos	5	
	Italy - Galatina	5	
	Italy - Lazio	7	
	Italy - Molise	7	
	Spain - Ibiza	7	
	Spain - Pobla del Duc	6	
Central America	El Salvador	4	16
	Guatemala	6	
	Nicaragua	2	
	Panama	4	
South America	Argentina	5	13
	Bolivia	2	
	Brazil	6	

DNA extraction was performed using the Qiagen Blood & Tissue extraction kit following the manufacturer's recommendations. Subsequently, DNA extracts were submitted for whole genome sequencing (150 bp paired-end) on the NovaSeq 6000 platform with a minimum output of 6 Gb. Reads were quality trimmed with 'fastp v0.23.2' (https://github.com/OpenGene/fastp) using -q 20, which omits reads when more than 40% of bases have a phred-quality score of below 20. Adapter sequences were trimmed using overlap analysis (--correction), which also performed base correction using default parameters. Only reads with length of 100 bases or more were retained. Finally, fastp the default was used for options not mentioned here. Trimmed reads were then aligned to a newly available *C. capitata* reference genome by RMCA and CIRAD available at genbank upon publication of the manuscript (in preparation) using BWA 0.7.17 'bwa-mem' (https://github.com/lh3/bwa). Briefly, a *de novo* chromosome-level genome assembly was constructed using 38.8 Gb of PacBio CCS reads resulting in a 78x coverage of the *C. capitata* genome and was processed using Hifiasm (https://github.com/chhylp123/hifiasm). In order to capture chromosome conformation, a Dovetail™ Omni-C™ library read library was constructed and downstream analysis was performed using the HiRise pipeline (https://github.com/DovetailGenomics/HiRise_July2015_GR).

Next, bam files were processed, and variants were called using Elprep 5.1.1 'elprep sfm' (https://github.com/ExaScience/elprep) with marking duplicates, coordinate sorting and –reference-confidence BP_resolution enabled. The resulting g.vcf files were then combined and jointly called using GATK's *CombineGVCFs* and *GenotypeGVCFs,* respectively (https://gatk.broadinstitute.org/hc/en-us). Additionally, Following GATK hard-filters were applied to the variants:

gatk VariantFiltration
-V $jointvariants
-filter "QD < 2.0"
-filter "QUAL < 30.0"
-filter "SOR > 3.0"
-filter "FS > 60.0"
-filter "MQ < 40.0"
-filter "MQRankSum < -12.5"
-filter "ReadPosRankSum < -8.0"

Next, variants were filtered in bcftools 1.14 to be biallelic, and have a minor allele frequency of 5%. Genotypes with depth of below three were set to missing data. Only variants without missing data were retained. This dataset was then pruned to rule out linkage between variants.

Finally, to avoid the effect of linkage disequilibrium (LD) on the results (especially for recently admixed populations), SNPs were pruned using the SNPrelate package in R with the following parameters: a window size of 10 kb, step size of 10, and a LD coefficient (r2) threshold of 0.2.

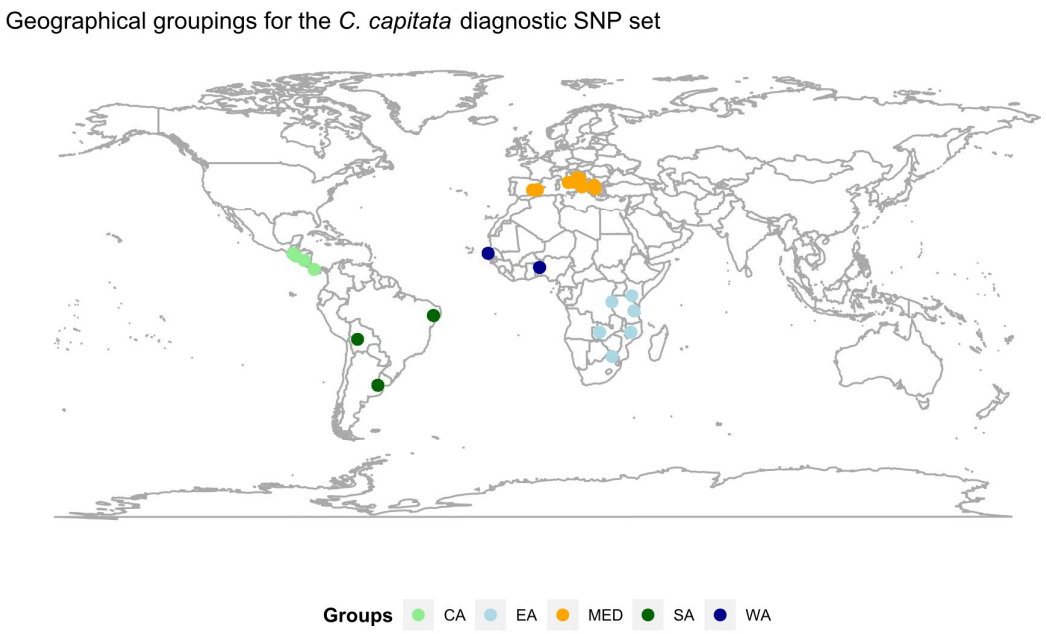


Figure 6.1.: Sampling locations of *C. capitata*. See table 6.1. for an overview of sample sizes per location.

Worldwide population structure analysis

In the interest of extracting diagnostic SNPs from the genome, it is essential to first assess the genetic structure that is present in the dataset. This is necessary to correctly identify the number of genetic groups present in the data and label every individual according to sensible regions of origin. For this, we performed a PCA as implemented in the package "SNPRelate" and visualized the first three PC axes.

Identification of diagnostic SNPs

Over the last years, several approaches have been applied to detect and extract diagnostic SNP loci. Methods relying on F_{ST} pruning have effectively been used to create a diagnostic SNP panel with the goal of origin tracing for a screwworm (*Cochliomyia hominivorax*; Tietjen, Arp & Lohmeyer, 2023), the Mexican fruit fly (*Anastrepha ludens,* Dupuis et al., 2019) and the Spanish Cedar (*Cedrela odorata*, Finch et al., 2020). In the Nile tilapia (*Oreochromis niloticus*), a DAPC discriminatory approach has been used by Kajungiro et al. (2019) to discern different populations from each other.

Here, we tested the power of Random Forest machine learning to reveal SNPs that are good predictors for population origin. Random Forest analysis can perform classification tasks very efficiently on large datasets even when diagnostic features are sparse.

First, we split the genotype matrix into a "test set" (70%) and a "validation set" (30%) and ran the random forest model implemented in the *rfe* function ("caret" package, R) using the test set only. The random forest algorithm randomly subdivides the test dataset into a training dataset, which is used to train the algorithm, and an 'out-of-bag' (OOB) set that is used to perform internal validation of the classification success of a tree. For our analysis, we fed the random forest model with a set of 35,591 unlinked SNPs and all of our 117 individuals, coded by their genetic cluster (EA, WA, MED, CA, SA). With the aim of extracting a minimum set of diagnostic SNPs, the *rfe* function allows "recursive feature elimination" using predefined sizes of a set of most important features of which predictive accuracy can be explored. Here

assessed accuracy while only including the 5, 10, 15, 20, 50, 80, 100, 150 and 200 topmost important SNPs and then evaluated each model.

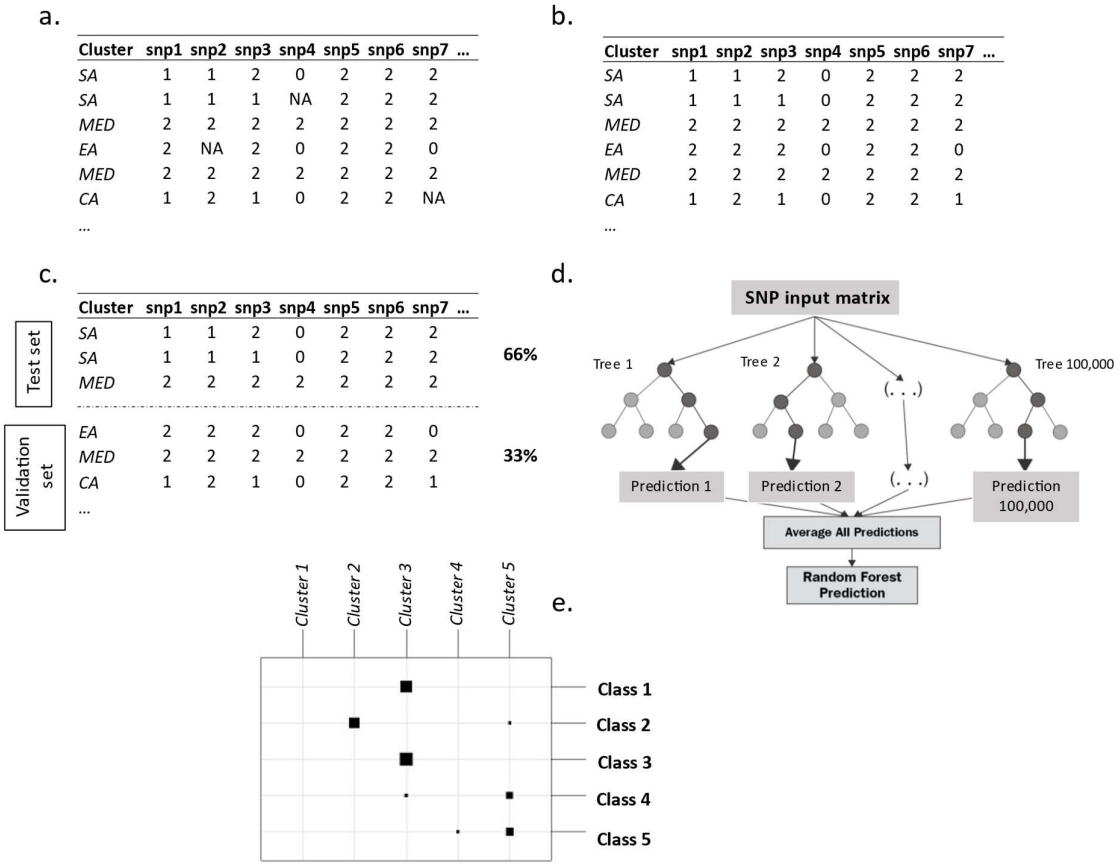


Figure 6.2.: A conceptualized overview of the different steps performed in a random forest approach to identify diagnostic SNPs. A) A genotype matrix is created with alleles coded as 0 (two reference alleles), 1 (one alternative allele), 2 (two alternative alleles) for each individual. B) Missing genotypes are imputed by looking at the best represented allele within the group. This step is performed when a dataset comprises a significant amount of missing data which would otherwise compromise the number of SNPs that can be retained for further analysis. C) The genotype dataset is split in a test and a training set. The test set is used to train the random forest algorithm, while the validation set is used to validate how well the final set of diagnostic SNPs perform in the classification task. D) The random forest model is run using 1000,000 decision trees. E) The approach is validated. A confusion matrix gives information on how well we can predict the origin of a the validation set.

Secondly, once a minimum set of putative diagnostic SNPs is attained, we validated the precision of classification using the validation dataset, which contains samples that the model was not trained upon. For this we used the *predict* function from the 'stats' package implemented in R. See Figure 6.2. for conceptualized representation of a random forest approach for detecting diagnostics SNPs. Additionally, we interpreted results of a discriminant analysis of principal components (DAPC) when using the full SNP set compared to the diagnostic SNPs only. DAPC, as implemented in the package "adegenet", is a clustering

method and first applies a principal components analysis (PCA) on the SNP data which is then subjected to a discriminant analysis. In the latter, variance in the dataset is partitioned into a between-group and within- group component, in an effort to maximize contrasts between groups.

6.1.2. Results

Genetic structure of Ceratitis capitata using the full dataset

After applying all quality filtering and linkage pruning, our dataset comprised a total of 35,591 SNPs. Results of a PCA are visualized in Figure 6.3. and point towards the presence of five genetic clusters in our data that align with the following geographical regions: East Africa (EA), West Africa (WA), the Mediterranean (MED), Central America (CA) and South America (SA). This subdivision is chosen because it provides sensible grouping that still contain a sufficient number of individuals that permit robust modelling.

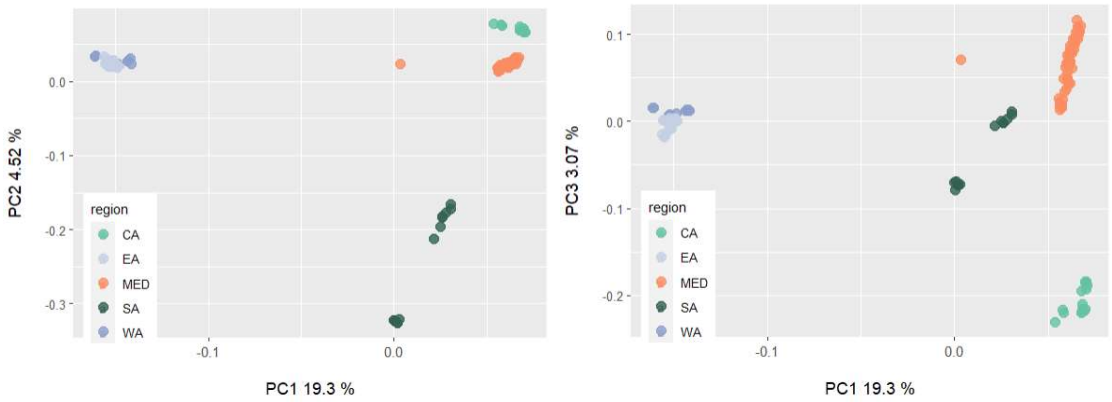


Figure 6.3.: PCA performed on the full SNP dataset in order to detect genetic structure in the data.

Genetic structure of Ceratitis capitata using a set of diagnostic SNPs

Backwards feature selection using 5, 10, 15, 20, 50, 80, 100, 150 and 200 SNPs indicated that using 100 SNPs delivered sufficient accuracy while keeping the number of SNPs to a minimum. It must be said however, that a smaller set of only 50 diagnostic SNPs is still powerful enough to detect differences between East and West Africa, but error rate increases when predicting the origin of Central American and South American samples. Scaffold name and base position of our diagnostic SNP panel can be consulted in Annex I.

A new random forest model was trained using only the final set of 100 diagnostic SNPs. Validation of the model using genotype information of samples that were not included in the training set showed that our panel of diagnostic SNPs was highly accurate and could thus successfully differentiate between the five geographic groups (WA, EA, MED, CA and SA, see Figure 6.4.). Our model could accurately estimate the correct origin of every sample within the validation set.

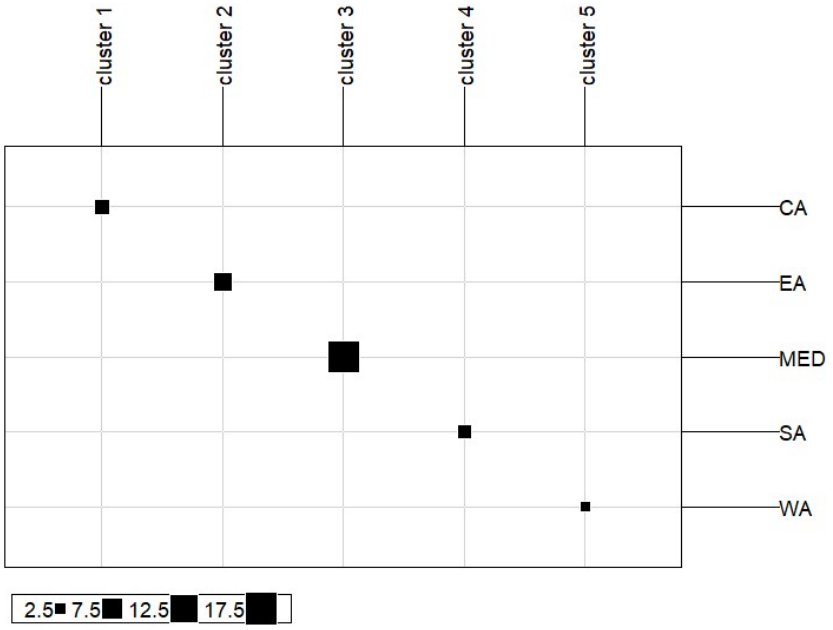


Figure 6.4.: Confusion matrix comparing the predicted and actual group identity of the validation samples. The size of the rectangles is representative of the number of samples grouped in each cluster.

Next, we assessed whether the diagnostic SNPs can be used in a DAPC to discriminate between the geographical groups (Figure 6.5.).

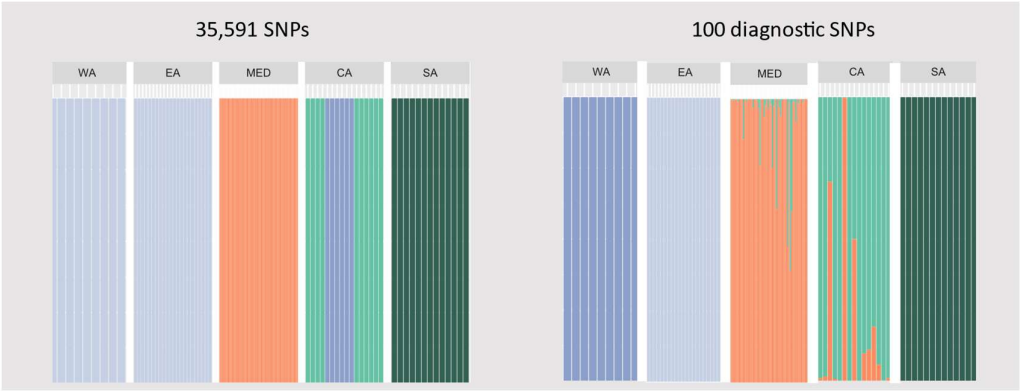


Figure 6.5.: Posterior membership assignment resulting from DAPC analysis using 35,591 unlinked SNPS (left) and the subset of 100 diagnostic SNP markers (right). Colours refer to the genetic cluster an individual is assigned to.

In the DAPC analysis, five genetic clusters were chosen based on the BIC score. In our full SNP dataset, African samples, both from East (EA) and West Africa (WA) form a single distinct cluster while they are separated using diagnostic SNPs only. Likely the recurrent feature elimination algorithm was able to extract informative SNPs to separate EA from WA while the extent of admixture observed when using the full SNP set caused the DAPC to create one cluster for both EA and WA. Within the invasive range, samples from the Mediterranean all belong to a single genetic cluster. In Central America, we detect two genetic groups, private to the region, while in South America, all individuals share the same unique south

American lineage. In earlier work, differentiating the Mediterranean genotypes from the Central American genotypes proved to be challenging and is likely the result of the higher levels of admixture between the two regions as shown in Deschepper *et al.* (2020). However, using whole genome sequencing in concert with DAPC seems to resolve the uncertainty of Central American samples for a large part.

Using the 100 diagnostics SNPs directly as input for a DAPC analysis did not prove to be as powerful as feeding the information for the diagnostic SNPs directly into the trained random forest model but achieved very comparable results (Figure 6.6). Likely the hierarchical nature of a decision tree can harness the power of highly divergent SNPs more efficiently which results in the better performance compared to a DAPC analysis. Additionally, using the random forest model for origin prediction does not require a reference dataset which is crucial when performing a DAPC analysis.

6.2. Bactrocera dorsalis

6.2.1. Material and Methods

In the context of Global change and economic integration, the spread of *B. dorsalis* has become increasingly serious. Attacking >250 species of fruits and vegetables, the oriental fruit fly, *B. dorsalis* (Hendel) (Diptera: Tephritidae), poses a serious threat to agricultural products and represents one of the most detrimental invasive pests worldwide. It presently spreads to 75 countries across Asia, Africa, and Oceania. Facilitated by the biological features of high prolificacy, short life history, broad host range, and adaptability, *B. dorsalis* is classified as the top member in the competitive hierarchy of fruit flies and could replace and drive various fruit fly species to extinction, including other highly invasive fruit flies such as *C. capitata*, *C. cosyra*, *Bactrocera tryoni*, and *B. zonata*.

Bactrocera dorsalis has been introduced to different places outside the natural distribution range of the species for many times and has received high attention from international plant quarantine and invasive biology. In many cases, the origin and transmission path of *B. dorsalis* intercepted by port quarantine departments and monitored in pest free areas are often unclear, making it difficult to develop invasion prevention and control strategies. Therefore, it has become an important scientific challenge for the prevention and control of *B. dorsalis* invasion to find more accurate traceability molecular markers and clarify the source of intercepted and monitored samples.

To meet the traceability needs of port quarantine and plant protection departments for *B. dorsalis*, this study conducted in-depth analysis and surveyed the whole genome of *B. dorsalis* for "diagnostic SNPs".

Sample collection and genotyping

We collected 429 samples of *B. dorsalis* using methyl eugenol traps in 50 populations from 29 countries, roughly covering the entire distribution range of *B. dorsalis*. The 50 populations were divided into six geographical groups: China (CN) (N = 141), Northern Southeast Asia (NSA) (N = 38), Southern Southeast Asia (SSA) (N = 57), South Asia (SA) (N = 78), Africa (AF) (N = 105), and Hawaii (HW) (N = 10). All samples were preserved in an 95% ethanol solution after collection and stored at -80 °C. DNA was extracted from the thoracic muscle of each fly using a Promega Wizard SV Genomic DNA Purification System. For each sample, a library with an average insert size of 350 bp was constructed using the Illumina TruSeq Nano DNA Library Prep Kit and sequenced on the Illumina NovaSeq platform with PE150 bp reads. Every sample was sequenced at a capacity of 6 Gb data to guarantee a sequencing depth of at least 10 × coverage.

The raw sequencing data were filtered using fastp with default parameters before mapping. The filtered data were mapped to most recent *B. dorsalis* chromosome-level genome assembly using the Burrows-

Wheeler Aligner (BWA) - mem algorithm with default parameters (GCA_030710565.1). Samtools was used to sort the bam output and calculate the sequencing coverage and depth for each sample. Duplicates were removed using Picard (https://sourceforge.net/ projects/picard/). Variants were called using the GATK. The following steps were executed in consecutive order: HaplotypeCaller (calling per sample single nucleotide polymorphisms (SNPs) and InDels), CombineGVCFs (combining per-sample gVCF files), GenotypeGVCFs (joint genotyping of all samples), SelectVariants (extracting SNPs and InDels), and VariantFiltration (hardfiltering variant calls based on the criteria: quality-by-depth ratio (QD) < 2.0 || read mapping quality (MQ) < 40.0 || probability of strand bias (FS) > 60.0 || symmetric odds ratio (SOR) > 3.0 || MQRankSum < –12.5 || ReadPosRankSum < –8.0).

Before the analysis, VCFtools was used to remove SNPs with a missing genotype rate > 0 and minor allele frequency < 0.05, and variants were filtered to be biallelic. To avoid the effect of linkage disequilibrium (LD) on the results (especially for recently admixed populations), SNPs were pruned using PLINK with the following parameters: a window size of 50 kb, step size of 10, and a LD coefficient (r2) threshold of 0.2.

An overview of sampling locations with their sample size can be found in Figure 6.6. and Table 6.2.

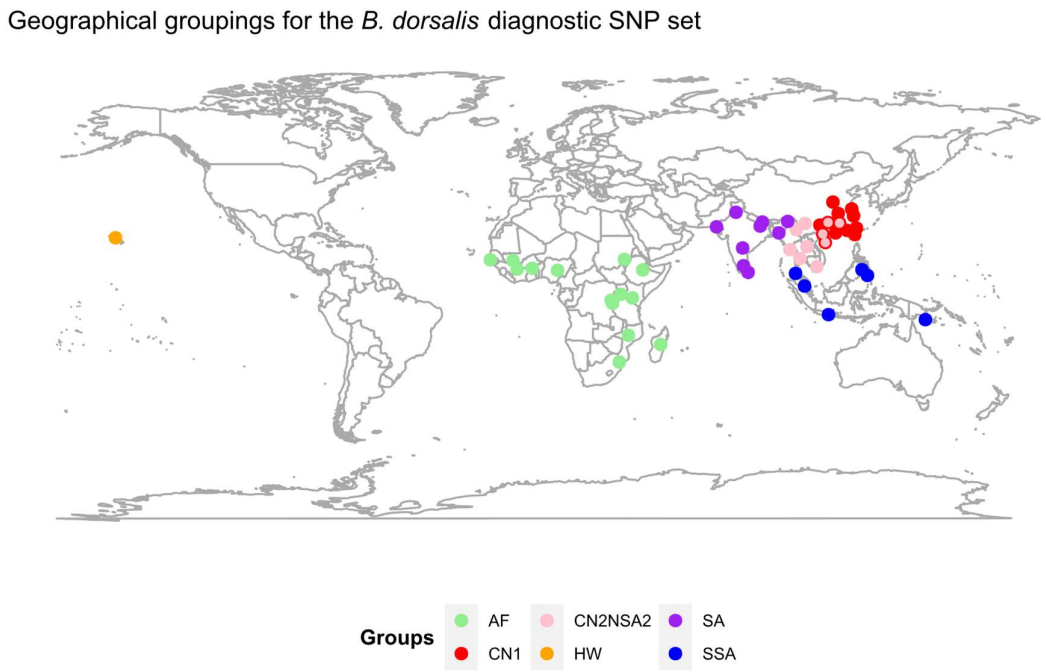


Figure 6.6.: Overview of sampled locations. Colors indicate to which genetic group each population was ascertained to for the identification of diagnostics SNPs. See Table 6.2. for sample sizes per location.

Table 6.2.: Overview of sampled locations used for detection of diagnostics SNP

Region/Group	Location	Nr. Samples	Nr. Samples in region
Africa, AF	SDSG	5	105
	SNBG	4	
	MLBM	8	
	GHND	5	
	CIBD	10	
	NGNS	2	
	ETAA	10	
	UGEB	10	
	KENB	10	
	CDKB	8	
	BIBJ	6	
	MWZB	10	
	ZAMP	7	
	MG	10	
South Asia, SA	PKMR	9	73
	INPJ	9	
	INAS	10	
	INBH	7	
	INTG	5	
	INTN	10	
	LKAD	9	
	BDKG	8	
	NPSH	6	
Southern Southeast Asia, SSA	THPK	10	55
	PHDM	7	
	PHDV	10	
	MYKL	10	
	IDJI	8	
	PGPM	10	
Northern Southeast Asia1, NSA1	MMYG	8	8
Hawaii, HW	USHW	10	10
China1, CN1	HNLY	10	110

	JSZJ	9	
	HBWH	10	
	ZJZJ	10	
	HNSY	9	
	JXNC	8	
	GZDY	8	
	GXNN	6	
	FJZZ	9	
	GDGZ	10	
	HNDZ	8	
	TWPD	6	
	TWTY	7	
China2+Northern Southeast Asia2, CN2+NSA2	GXNN	2	49
	HNDZ	2	
	JXNC	2	
	HNSY	1	
	YNBS	7	
	SCLS	10	
	LAVT	5	
	VNTG	10	
	MMYG	1	
	THBK	9	

Worldwide population structure analysis

In the interest of extracting diagnostic SNPs from the genome, it is essential to first assess the genetic structure that is present in the dataset. For this, we convert the VCF file to phylip format using the vcf2phylip.sh script, and then use FasTree to construct a phylogenetic tree based on the maximum likelihood method. Additionally, we performed a PCA as implemented in the package "SNPRelate" and visualized the first two PC axes.

6.2.2. Results

Geographical groups in Bactrocera dorsalis and diagnostics SNPs

Due to the close proximity between China and Northern Southeast Asia, *B. dorsalis* has more frequent exchange between these to regions. SNP characteristics are very similar, and the differences may be relatively small. From the phylogenetic trees (Figures 6.7 & 6.8), and PCA diagrams (Figures 6.9 & 6.10) it can be seen that most of the samples from China and northern Southeast Asia are clustered together. For this reason, we decided to divide samples from a number of Chinese populations (CN1) into a Chinese

cluster and a cluster representing a combination of Chinese and Northern South-East Asian samples (CN2+NSA2), which resulted in better overall predictions.

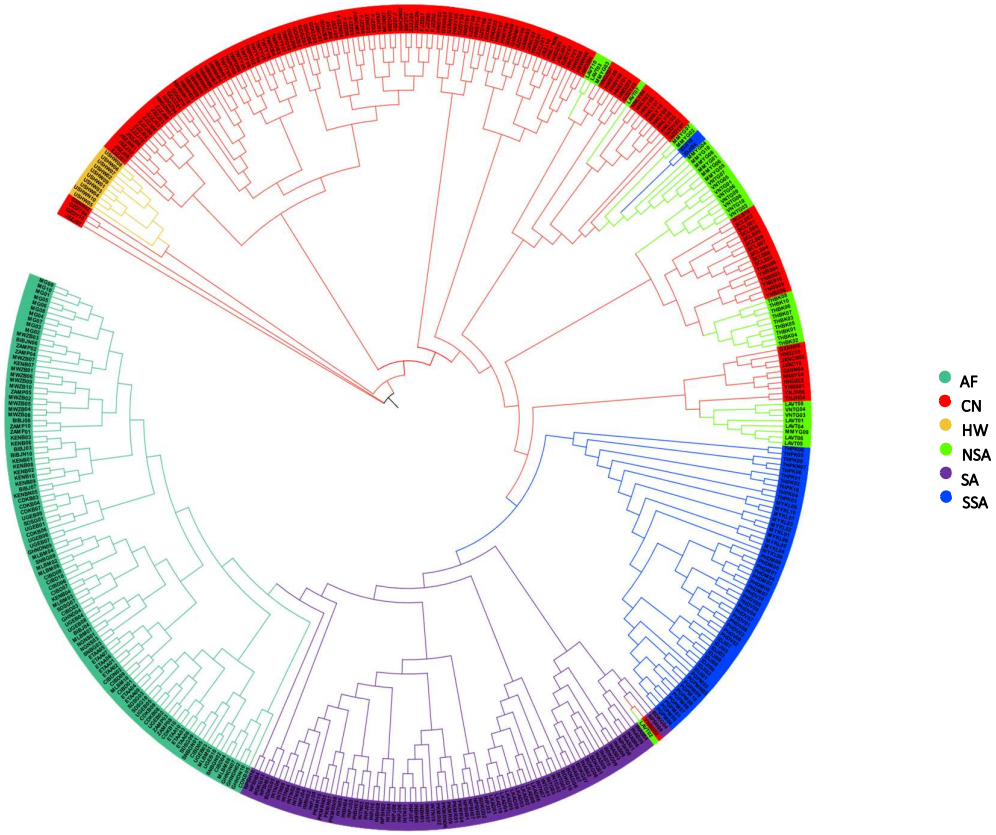


Figure 6.7.: The maximum likelihood phylogenetic tree of *B. dorsalis* based on whole-genome SNPs.

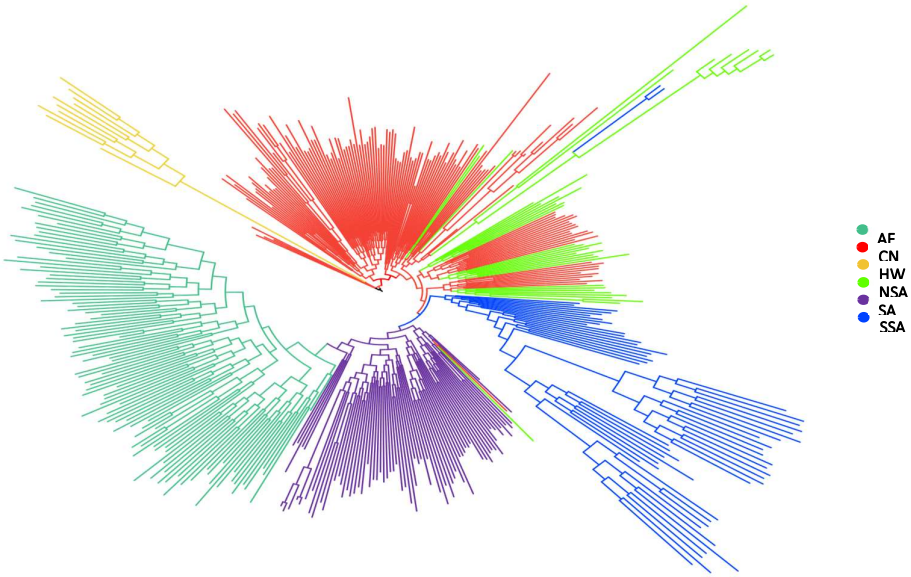


Figure 6.8.: The maximum likelihood phylogenetic tree of *Bactrocera dorsalis* based on whole-genome SNPs (With branch length).

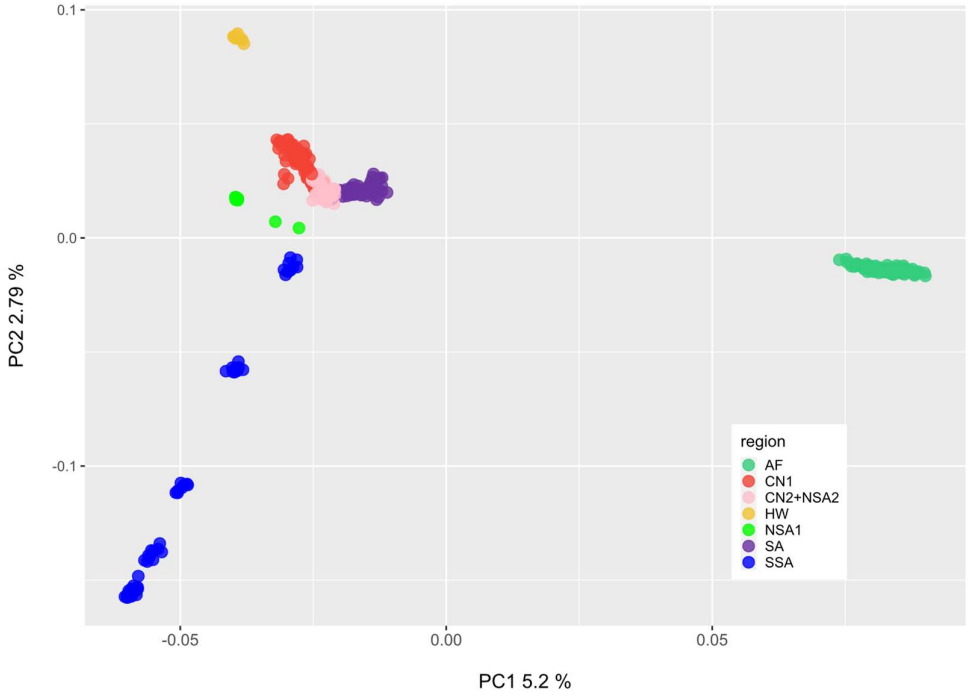


Figure 6.9.: PCA for *Bactrocera dorsalis* (the full SNP set)

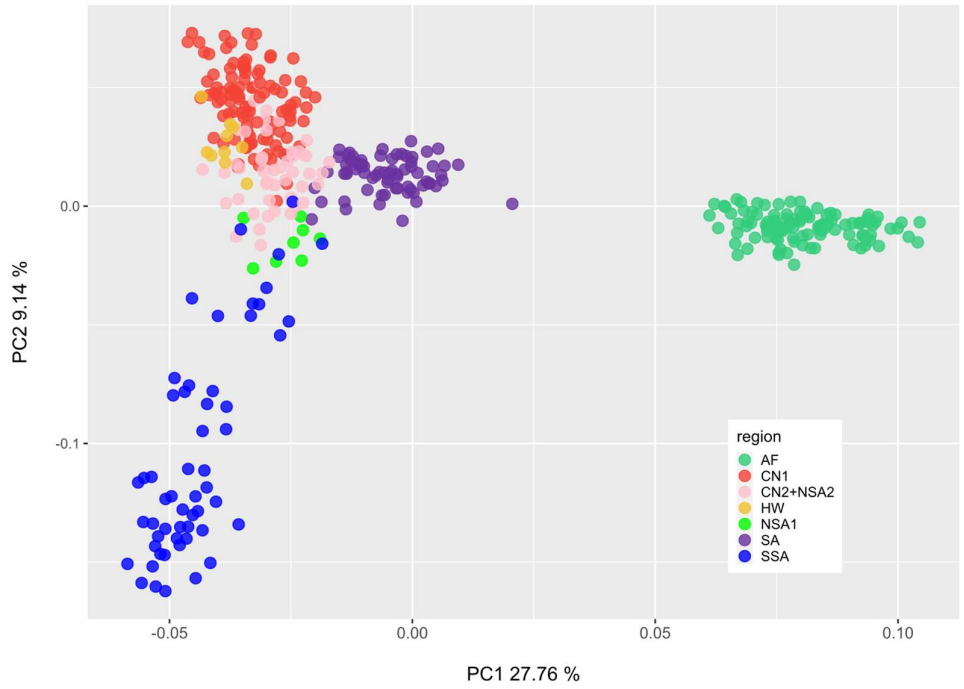


Figure 6.10.: PCA for *Bactrocera dorsalis* (the diagnostic SNPs only)

Backwards feature selection using 5, 10, 15, 20, 30, 40, 50, 80, 100, 130, 150, 180 or 200 SNPs indicated that using 130 SNPs delivered sufficient accuracy while keeping the number of SNPs to a minimum (Fig. 6.12.). The accuracy of the model with 130 SNPs was 91.08% when tested on the validation samples (Figure 6.11). Due to admixture between CN2+NSA2 and CN1, we can still observe samples that are mislabeled and thus leaves room for improvement of the model. However, more sampling in this region of high admixture would be necessary (Figure 6.12). Chromosome name and base position of our diagnostic SNP panel can be consulted in Annex II.

The PCA on the diagnostic SNPs (Figure 6.10) only lacks the fine structure of the PCA using the full SNP set (Figure 6.9). However, distinct clustering can still be observed for SSA, SA, AF and CN1.

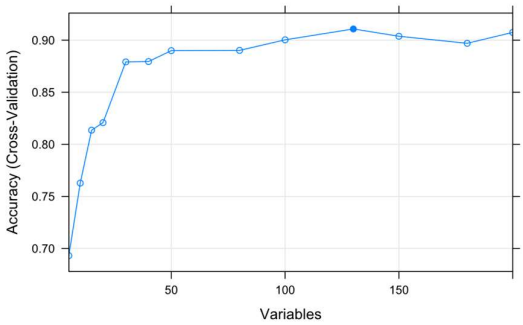


Figure 6.11.: Accuracy of the backwards feature selection analysis using different SNP group sizes.

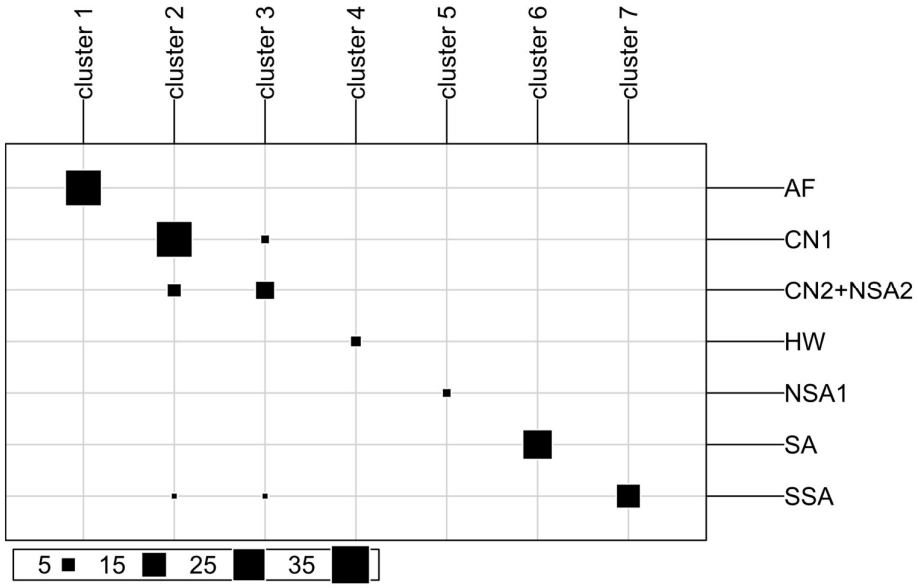


Figure 6.12.: Confusion matrix comparing the predicted and actual group identity of the validation samples. The size of the rectangles is representative of the number of samples grouped in each cluster.

6.3. Bactrocera zonata

6.3.1. Material & methods

The peach fruit fly (*B. zonata*) is native to South and Southeast Asia (White & Evenhuis, 1999). It as a highly polyphagous species that is assumed to adapt quickly to new host plants. Additionally, this species has a high reproductive potential and strong flight capacities (Qureshi *et al.*, 1975) making it ideally suited to invade new territories and of quarantine significance in many countries, including the whole of the EU (https://www.eppo.int/).

Sample collection, genotyping and genetic structure

A total of 159 samples were collected between 2011 and 2022 distributed in following countries: Reunion, Mauritius, Egypt, Iran, Oman, Sudan, Israel, India and Pakistan (lab colony). See Table 6.3 for an overview of the samples per location that were used for our study after curation for quality filtering and figure 6.13 for a map of the sampled locations.

Table 6.3. List of *Bactrocera zonata* samples used in development of diagnostic SNPs.

Country	Population	#samples
Egypt	Egy_Cai	6
India	Ind_Par	6
India	Ind_Pat	7
India	Ind_Ven	4
Iran	Ira_Cha	8
Iran	Ira_Kah	8
Iran	Ira_Min	8
Iran	Ira_Tis	6
Iran	Ira_Zar	8
Israel	Isr_Tel	7
Mauritius	Mauritius	7
Oman	Oma	6
Pakistan	Pak_Isl	7
Reunion	Reunion	2
Sudan	Sud_Gez	7

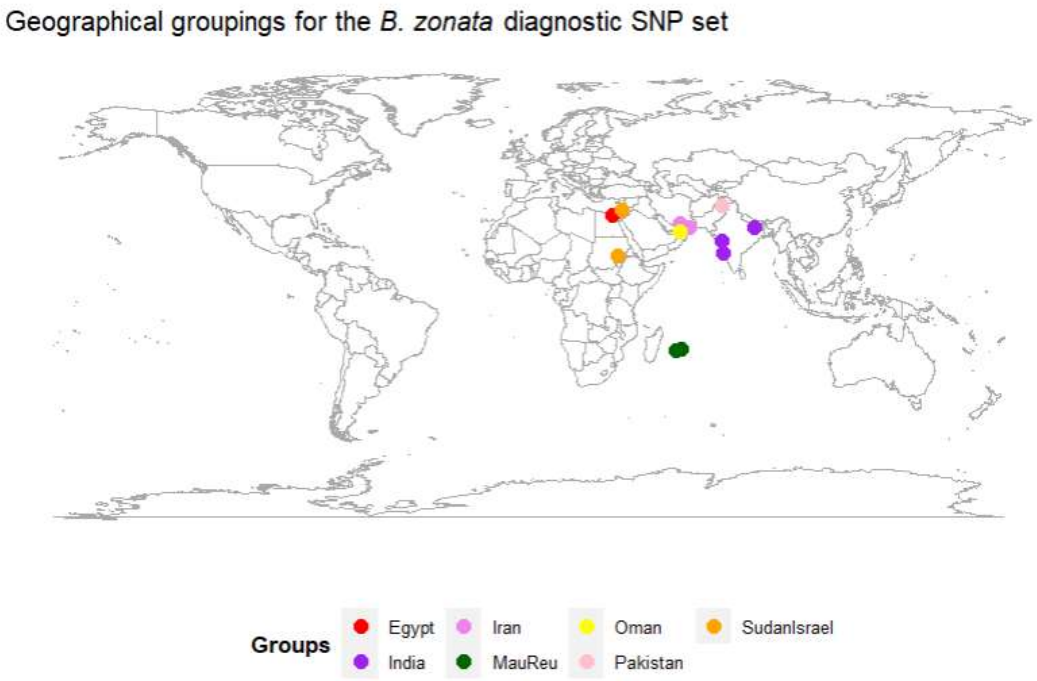


Figure 6.13.: Overview of sampled locations for *Bactrocera zonata*. Colors indicate to which genetic group each population was ascertained to for the identification of diagnostics SNPs. See Table 6.3. for sample sizes per location.

For detailed DNA extraction, sequencing and SNP calling protocol, see the materials and methods section under *C. capitata*. Some dataset specific filtering choices have been made for the *B. zonata* vcf file in order to retain samples with a missingness larger than 20% and increase the number of SNPs retained in the final vcf file. More specifically, we did not exclude loci with missing data but rather filtered on 95% missingness on the locus level. Since the methodology for identifying diagnostic SNPs does not allow any missing data we imputed the allelic states where missing using the *rfImpute* function from the randomForest package. Additionally, we have filtered on an average depth of 10, minor allele count of three and only retained biallelic loci.

To Assess the genetic structure of *B. zonata*, we carried out an analysis of principal components (PCA) as implemented in the package SNPrelate. This information was then used to assign every population to a genetic cluster/region and use this information downstream in the procedure to extract diagnostic SNPs.

6.3.2. Results

Geographical groups in Bactrocera zonata and diagnostics SNPs

After removing samples with an insufficient amount of read data and after quality filtering, 97 samples were retained for further analysis. PCA suggested the presence of multiple well defined genetic clusters (Figure 6.14). We can identify separate grouping for Egyptian and Pakistani samples while Omani, Indian and Iranian populations overlap to a minor extent. Sudanese and Israeli samples together form a distinct cluster an no clear discrimination can be between samples from both countries. Hence, Sudanese and Israeli samples were grouped together for further downstream analyses so that the accuracy of the diagnostic SNPs model is not affected by the large amounts of admixture observed between Israel and Sudan. Mauritian and Reunion samples were also grouped since the Reunion population only consists of two samples.

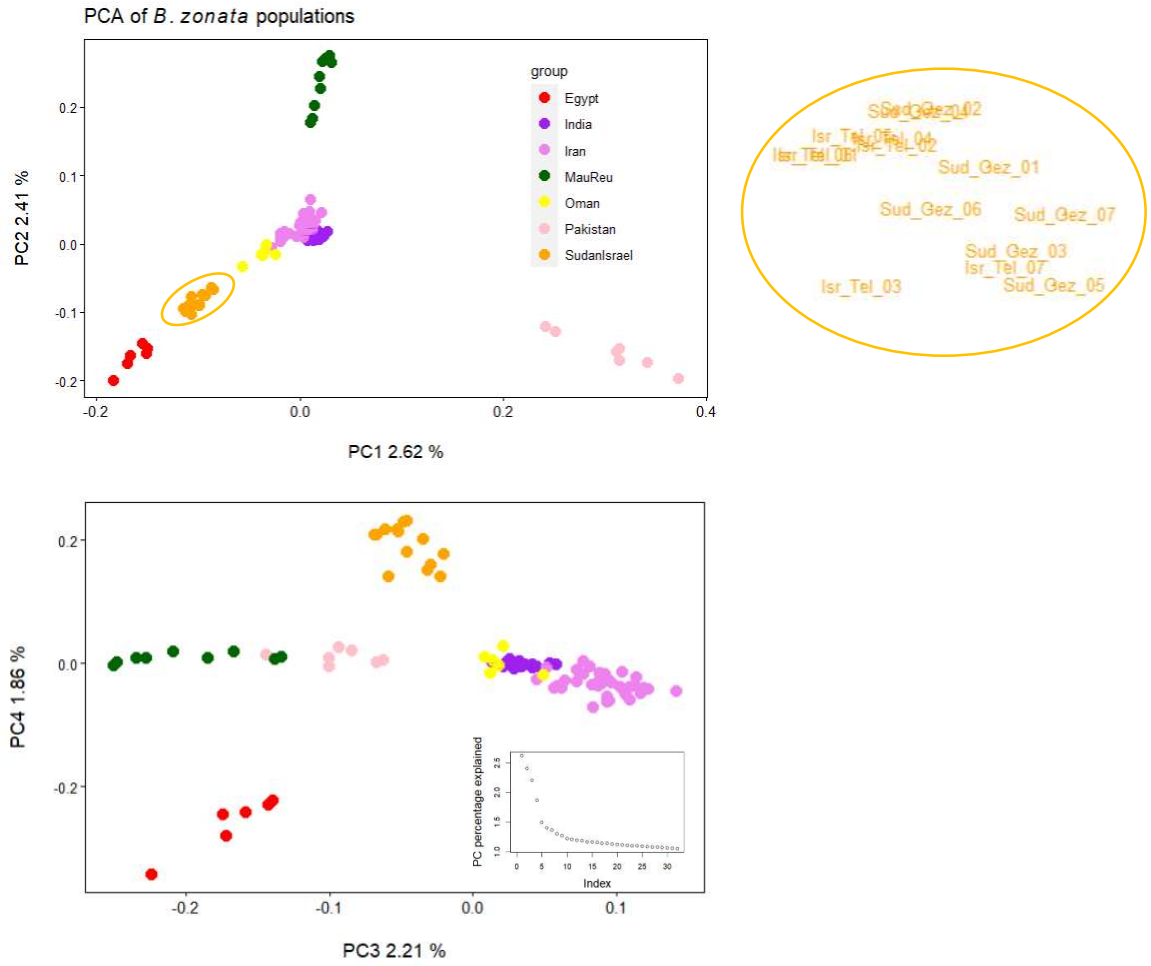


Figure 6.14.: PCA using the 97 *B. zonata* samples visualized using PC1, PC2 and PC3. The orange cluster is composed of Israeli and Sudanese samples without distinction between the two as represented in orange.

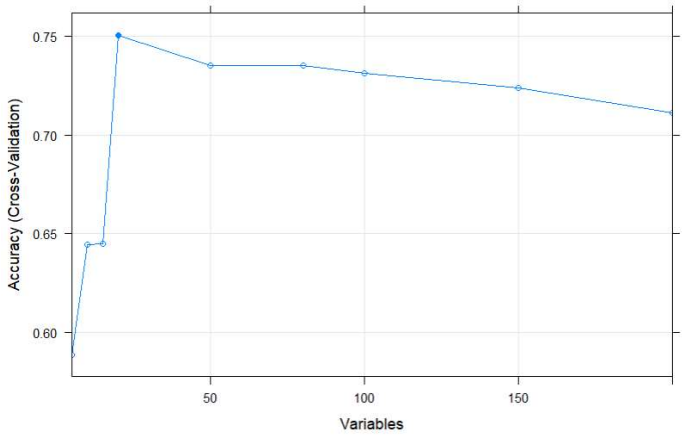


Figure 6.15.: Accuracy of the backwards feature selection analysis using different SNP group sizes.

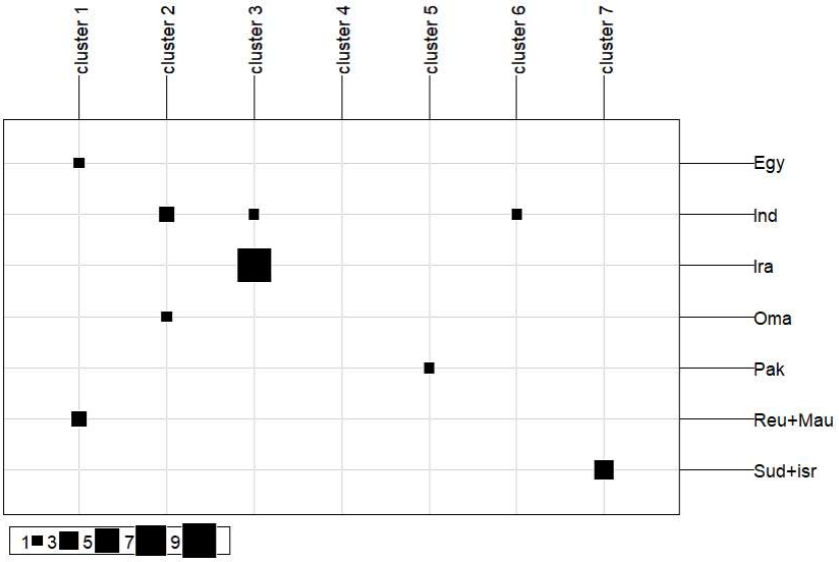


Figure 6.16.: Confusion matrix comparing the predicted and actual group identity of the validation samples using 20 diagnostics SNPs.

After running the backwards feature selection model to extract a minimum panel of diagnostics SNPs we can observe that the accuracy of the model quickly climbs up to 75.0 % accuracy (see figure 6.15), corresponding with a SNP number of 20. After that, accuracy is sloping downwards and thus positions of the 20 most predictive SNPs were extracted to build the final random forest model. From the confusion matrix we can see that only the Sudan+Israel and Pakistan samples can be traced to their origin with a 100% accuracy using the validation samples (Figure 6.16). For others, the pattern is more complex. For instance, validation samples from India have a predicted origin from India (2 samples), Iran (1 sample) and Reu+Mau (1 sample). This is likely caused by the demographic history of Reu+Mau and Iran populations which are assumed to be directly derived from Indian populations. These are listed in annex III.

6.4. Utilization of the trained models

The trained models can be used to trace the origin to a certain region as defined in this work. Optimally, the user of the model would want genotype information for as much diagnostics SNP loci as possible and thus it is recommended to sequence the genome of a query sample with sufficient depth (> 15X on average). Lacking genotypes will decrease the accuracy of the origin prediction.

After sequencing a query sample and having called variants using the correct reference genome, the VCF file can be subsetted in order to solely contain the diagnostic SNP loci. This can be done by the following line of code in linux terminal or WSL (Windows Subsystem for Linux):

```
vcfintersect -b diagnsnps.bed variants.g.vcf.gz > diagnsnpsonly.vcf
```

The random forest model can be used to make an estimate on the origin of a genotyped sample with only a few lines of R code.

Read in the diagnostic SNPs vcf file as a .gds file and open it in the R session

```
Install.packages(SNPRelate)

library(SNPRelate)

snpgdsVCF2GDS(query.fn, "query.gds", method="biallelic.only")

querygeno = snpgdsOpen("query.gds")
```

Get genotypes coded as 0,1,2

```
genotypes <- snpgdsGetGeno(querygeno, snp.id = NULL, snpfirstdim=TRUE)
```

Check how many times 0,1,2 and NA occur as genotypes (to check whether you don't have an excessive amount of NA)

```
table(c(genotypes), exclude=NULL)
```

Make a dataframe with all diagnostic SNPs in your VCF file

```
df_init <- cbind(read.gdsn(index.gdsn(querygeno, "sample.id")))

df <- data.frame(df_init, t(genotypes))

colnames_vec <- c(paste0(read.gdsn(index.gdsn(querygeno, "snp.chromosome")),"_",read.gdsn(index.gdsn(querygeno, "snp.position"))))
```

Load the trained model

```
load("path/to/model/model.rf")
```

Inspect the loaded model

```
rf_model_purged

plot(rf_model_purged)
```

Inspect the diagnostic SNP loci that are represented in the model.

```
predictors(rf_model_purged)
```

Make a datasets that contains all diagnostics SNPs that are represented in the model and populate it with NA. Afterwards, we add 0,1 or 2 where we have genotype info.

```
Install.packages("rqdatatable")

library("rqdatatable")

NA_df <- as.data.frame(matrix("NA", ncol = length(predictors(rf_model_purged)), nrow = nrow(df)))

NA_df <- cbind(df$sample, NA_df)

match(colnames(NA_df), colnames(df)) %>% summary()

colnames(NA_df) <- c(colnames(df)[1],predictors(rf_model_purged))

Q_df <- natural_join(df, NA_df,
```

```
            by = "sample",
            jointype = "INNER")
Q_df <- Q_df[, colnames(NA_df)]
```

Make the prediction and present results as a barplot

```
predict(rf_model_purged, as.factor(Q_df), type = "prob") %>% barplot(main = Q_df,cex.names = 0.7, col = "lightgrey")
```

Give the region of origin with the highest likelihood

```
print(paste(Q_df,predict(rf_model_purged, as.factor(Q_df), type = "response")))
```

7. Decision workflow and protocol

All of the above identification tools were incorporated in decision workflow and protocol for guiding the identification of the three target fruit flies, as well as any other fruit fly of EU quarantine importance. In addition to the tools developed or provided within the FF-IPM project, reference is being made to other tools which are publicly available and which can be either complementary or alternatives to the proposed methodologies.

This workflow and protocol is divided along three main factors influencing the identification process:

a) Speed at which an identification needs to be conducted
b) Condition of the material collected
c) Purpose of the identification

All tools were listed, indicating their advantages and shortcomings, the duration of the execution of the methodology and the interaction and/or chronology of deployment.
They were afterwards included in visual decision diagrams illustrating their interactions, as well as possible links to alternative methods.

The developed tools allow both morphological and molecular (i.e. DNA based) identification. Which tool to use will depend on three factors:

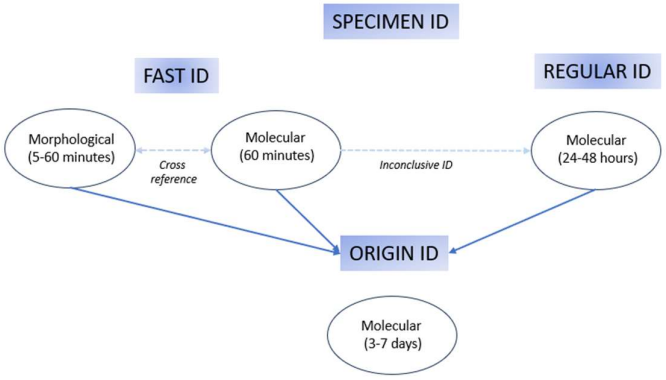
a) Speed at which an identification needs to be conducted
b) Condition of the material collected
c) Purpose of the identification

Speed: identification can be required within a short time framework or only necessary over a longer period. Fast identification is, for example, required when any fruit fly specimen is intercepted or trapped in a monitoring process and is suspected to be of quarantine importance. In such cases a speedy action (e.g. refusal or entrance of imported goods, eradication or mitigation programme to eliminate or reduce any transient population) is required. In a number of cases the proposed methodologies for a fast identification may not lead to an unambiguous identification and thus more detailed but more time-consuming methodologies may be required. Moreover, the speed of the identification will also depend on the infrastructure required and present at study site. As such an 'in situ' (i.e. at the point of interception or detection) can be possible or transport to a facility with more elaborate research infrastructure can be required.

Condition: Intercepted specimens can be of any life stage. While morphological identification of egg, first instars or pupal stages to species level is difficult, if not impossible for some groups, DNA based methods can be applied for any life stage but will require more time. Trapped material will normally be in the adult stage. However, trapping conditions (especially when exposed for a prolonged period) can damage the specimens. These factors will determine whether a swift identification using morphological characteristics is feasible or if assistance of molecular tools is required.

Purpose: In the majority of cases, the main purpose of the process is to define the taxon identification: "Does the specimen belong to species X?". Although EU regulation allows identification for particular levels above species level (i.e. usually genus level, except for priority pests), an unambiguous identification at species level is to be recommended as it may allow bespoke methodologies to be employed. Secondly, in addition to the species level identification, policy makers and end users may also be interested in the initial origin of the intercepted or detected material. While for interceptions metadata can provide some information on this (e.g. origin of imported goods that were infested, information obtained from passengers carrying infested fruits), this is more difficult for material detected in monitoring programmes like trapping surveys. Nevertheless, a known source can assist in mapping pathways and points of entry and subsequently in developing methods to prevent introduction of these organisms. Morphological characteristics will however be of little or no use in tracing origin but molecular data, in particular data obtained through whole genome sequencing, may assist in this.

The general workflow can thus be depicted as follows:



Below the different identification tools developed within the framework of the FF-IPM project are listed with indication of their scope and their advantages and shortcomings regarding the above mentioned factors:

Multi-entry key for morphological identification of adult fruit flies:
Scope: identification of the three target fruit flies and 20 closely related taxa of EU quarantine importance. Selection of the latter was based on economic significance and frequency of interception, after consultation of potential end users. Given the high diversity of taxa in the targeted taxa (i.e. *Ceratitis* and *Bactrocera* (for the latter also including *Zeugodacus* being formerly ranked as subgenus of *Bactrocera*), the coverage is limited.
Speed: experienced user can provide an identification within the timespan of a few minutes. For more difficult cases (e.g. differentiation between morphologically similar species) more extensive timespan could be necessary, in particular for comparison with virtual or physical reference material. Identification can be conducted in situ, at the point of entry. Only a magnifying lens or portable microscope is required.
Condition: The key works only for adult fruit flies, immature stages cannot be identified. Intact specimens will provide a faster identification as there is a larger suite of character states that can be scored. Partially damaged specimens may also still be identifiable but will depend on the structures available.

Purpose: only suitable for taxon identification at species level. Not suitable for tracing origin.

Multi-entry key for morphological identification of larval fruit flies:
Scope: identification of the three target fruit flies and 10 closely related taxa of EU quarantine importance. Selection of the latter was based on economic significance and frequency of interception, after consultation of potential end users but more limited in scope given the limited material available as reference material for development of the key. Given the high diversity of taxa in the targeted taxa (i.e. *Ceratitis* and *Bactrocera* (for the latter also including *Zeugodacus* being formerly ranked as subgenus of *Bactrocera*), the coverage is limited.
Speed: examination of the material requires more elaborate preparation of the specimen, compared to identification of an adult in addition to the need of taking measurements, hence resulting in a longer timespan required. An experienced user can provide an identification within the timespan of an hour. Access to a high profile microscope is required, making in situ identification less practical but not impossible.
Condition: The key works only for third instar larvae of fruit flies, other immature stages cannot be identified.
Purpose: only suitable for taxon identification at species level. Not suitable for tracing origin.

LAMP for target fruit flies
Scope: *identification of the target Ceratitis capitata* and resulting exclusion of a number of closely related Ceratitis species of economic significance, and identification of the target *Bactrocera zonata* and resulting exclusion of a number of closely related *Bactrocera* species of economic significance. LAMP for the target *Bactrocera dorsalis* was not developed as such a LAMP technique was made publicly available prior to the start-up of this activity (Blaser et al., 2018). However, reference to this technique is made in the protocol allowing to have it incorporated in the workflow.
Speed: full process (from dissection to actual visualization of presence/absence target DNA) requires about one hour. LAMP deployment can be conducted with minimum of apparatus required hence possible in situ.
Condition: LAMP technique works for any life stage.
Purpose: only suitable for taxon identification at species level. Only suitable for discrimination of target fruit flies against a number of related taxa. Not suitable for tracing origin.

Diagnostic SNPs for target fruit flies
Scope: identification of geographic populations of the three target fruit flies, based on diagnostic Single Nucleotide Polymorphisms that are considered characteristic for populations of a particular biogeographic region.
Speed: full process (from dissection to actual recognition and tracing of SNPs) requires several days (mainly because of the need for outsourcing to specialized companies. Actual time will depend on company and price that can be paid for the service).
Condition: diagnostic SNP scan be traced for any life stage.
Purpose: only suitable for identification of geographic population. An a priori taxon identification is required.

Complementary or alternative methods
In addition to the novel tools developed within the framework of FF-IPM there are a number of other identification tools, developed independently. The range is very wide and includes several other approaches. Listing them all is beyond the scope of this deliverable and we only present those suitable as subsequent steps in case the end result of the above mentioned tools leads to an indecisive identification.

Extended multi-entry keys for identification of a larger suite of fruit flies.

The key generated within the framework of the FF-IPM entails a limited number of fruit flies, selected on the basis of their economic significance and feedback from potential end-users. More extensive keys for fruit flies, comprising all representatives of a particular taxon or of a particular region, exist. These keys are generated in the same manner as the FF-IPM keys (i.e. multi-entry keys) which facilitates the use. However, as they comprise a much larger number of taxa, a more in-depth knowledge of morphological characteristics is required in order to discern the often subtle differences between taxa. Also these keys do not exist as mobile applications and thus can only be used on PCs (or online consultation through a web browser).

Virgilio et al. multi-entry keys for adult dacines from the Afrotropical Region. This key include a key for all existing *Ceratitis* species in addition to all Afrotropical *Bactrocera*, *Dacus* and *Zeugodacus* species. Link to open access or free download
: https://fruitflies.africamuseum.be/outputs/identification_tools_mobile_applications

Doorenweerd et al. multi-entry key for adults dacines native to Asia, Australasia and Oceania. This key includes all *Bactrocera*, *Dacus*, *Monacrostichus* and *Zeugodacus* species from these regions and described prior to 2021. Key can be consulted and used online at: https://idtools.org/tools/2103/

LAMP for Bactrocera dorsalis

As indicated above, a separate LAMP tool was developed for the target species *Batrocera dorsalis* during rhe initial start of the FF-IPM project. This is a suitable tool in case the developed LAMP protocols do not given an unambiguous answer. All details of the LAMP can be found at Blaser et al. (2018).

Alternative genetic markers

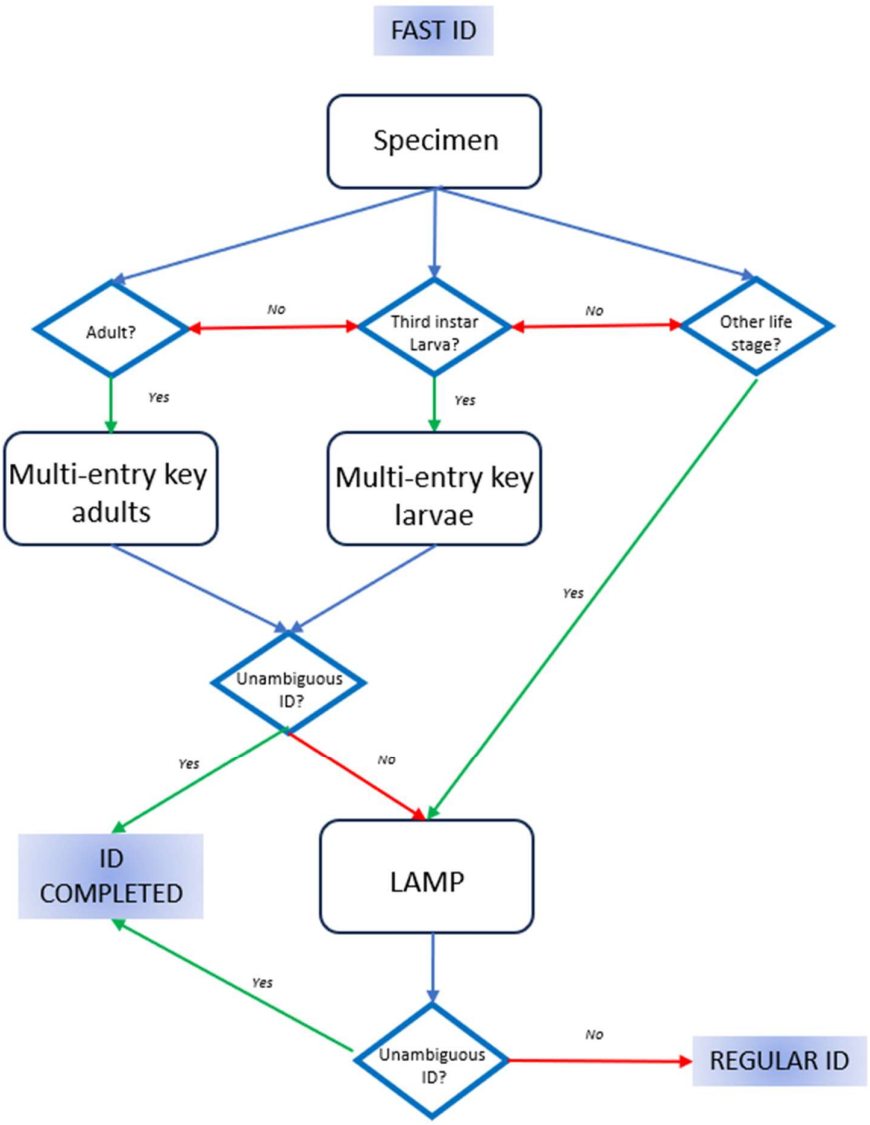
In some cases, DNA barcodes (i.e. using the mitochondrial COI locus as universal barcode) will not allow full differentiation between species (Virgilio et al. 2019), in particular for closely related organisms such as species complexes. Alternative markers or regions (e.g. Cytb, ITS) can be explored in that case. This option is included in the flowchart for regular ID.

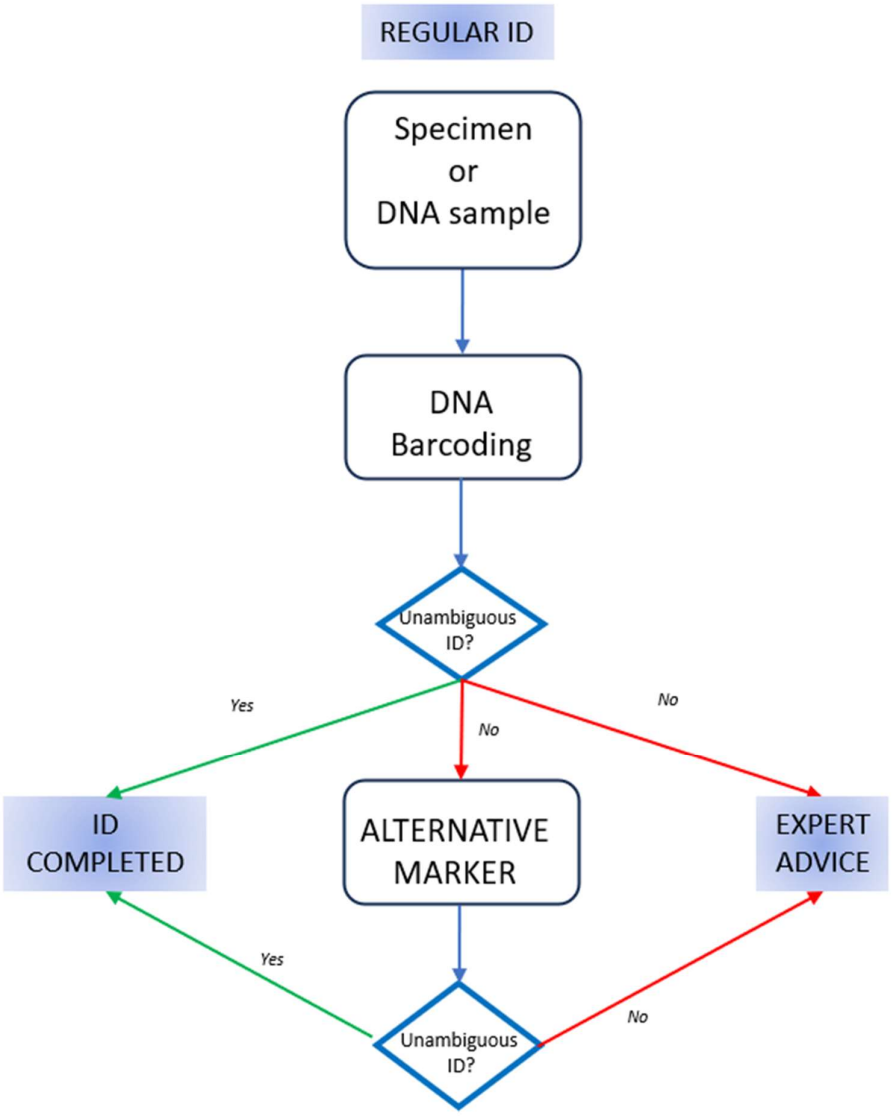
Expert advice

Occasionally no unambiguous identification can be obtained using the different techniques outlined above and in the flowchart. Expert advice of a taxonomic expert is essential in that case. Such advice is also recommended in particular cases where material that is detected or intercepted is considered of major significance, for example at the detection of invasive alien species.

Decision tree

The sequence of utilization, depending on the outcome of the previous step, is depicted in the flowcharts below:


FAST ID
Specimen
Adult?
No
Third instar Larva?
No
Other life stage?
Yes
Yes
Multi-entry key adults
Multi-entry key larvae
Yes
Unambiguous ID?
Yes
No
ID COMPLETED
LAMP
Yes
Unambiguous ID?
No
REGULAR ID


REGULAR ID
Specimen or DNA sample
DNA Barcoding
Unambiguous ID?
Yes
No
No
ID COMPLETED
ALTERNATIVE MARKER
EXPERT ADVICE
Yes
Unambiguous ID?
No

References

Bierman A, Karsten M, Smit A, De Meyer M, Papadopoulos N, Terblanche JS (submitted). Field friendly identification of the peach fruit fly, *Bactrocera zonata,* (Diptera: Tephritidae) using Loop-mediated Isothermal Amplification. *Curr. Res. Insect Sci.*

Blaser S, Diem H, von Felten A, Gueuning M, Andreou M, Boonham N, Tomlinson J, Müller P, Utzinger J, Frey JE, Bühlmann A 2018. From laboratory to point of entry: development and implementation of a loop-mediated isothermal amplification (LAMP)-based genetic identification system to prevent introduction of quarantine insect species. *Pest Manag. Sci.* 74, 1504-1512.

Boykin LM, Armstrong K, Kubatko L, De Barro P 2012. DNA barcoding invasive insects: database roadblocks. *Invertebrate Systematics* 26, 506-514.

Deschepper P, Todd NR, Virgilio M, De Meyer M, Barr NB, Ruiz-Arce R 2021. Looking at the big picture: worldwide population structure and range expansion of the cosmopolitan pest *Ceratitis capitata* (Diptera, Tephritidae). *Biol. Invasions,* 23, 3529–3543.

Dupuis JR, Ruiz-Arce R, Barr NB, Thomas DB, Geib SM 2019. Range-wide population genomics of the Mexican fruit fly: Toward development of pathway analysis tools. *Evol. Appl.*. doi.org/10.1111/eva.12824.

Dutra VS, Ronchi-Teles B, Steck GJ, Silva JG 2012. Description of larvae of *Anastrepha* spp.(Diptera: Tephritidae) in the *fraterculus* group. *Annls Entomol. Soc. Amer.*, 105(4), 529-538.

EFSA Panel on Plant Health (PLH), Bragard C, Dehnen-Schmutz K, Di Serio F, Gonthier P, Jacques M.A., Jaques Miret JA, Justesen AF, Magnusson CS, Milonas P, Navas-Cortes JA, Parnell S, Potting R, Reignault PL, Thulke HH, Van der Werf W, Civera AV, Yuen J, Zappalà L, Bali EM, Papadopoulos N, Papanastassiou S, Czwienczek E, MacLeod A. 2020. Pest categorisation of non-EU Tephritidae. EFSA Journal 18, e05931.

Ekesi S, De Meyer M, Mohamed SA, Virgilio M, Borgemeister C 2016. Taxonomy, ecology, and management of native and exotic fruit fly species in Africa. *Ann Rev. Entomol.*, 61, 219-238.

Finch KN., Cronn RC, Ayala Richter MC *et al.* 2020. Predicting the geographic origin of Spanish Cedar (*Cedrela odorata* L.) based on DNA variation. *Conserv. Genet.*, 21, 625–639.

Frías D, Selivon D, Hernández-Ortiz V 2008. Taxonomy of immature stages: new morphological characters for Tephritidae larvae identification. In: *Fruit Flies of Economic Importance from Basic to Applied Knowledge. Proceedings of the 7th International Symposium on Fruit Flies of Economic Importance, Salvador (Brazil)*, 29-44.

Kajungiro RA *et al.* 2019. Population Structure and Genetic Diversity of Nile Tilapia (*Oreochromis niloticus*) Strains Cultured in Tanzania. *Frontiers in Genetics*. doi.org/10.3389/fgene.2019.01269.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A 2012. Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data. *Bioinformatics,* 28, 1647–1649, doi:10.1093/bioinformatics/bts199.

Kitano D, Takakura KI 2020. Simple and on-site DNA purification for LAMP reaction applicable to non-adult tephritid fruit fly (Diptera: Tephiritidae). *J. Appl. Entomol.* 144, 824-829.

Kumar S, Stecher G, Tamura K 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol.* 33, 1870–1874, doi:10.1093/molbev/msw054.

Li XS, XM Huang, SX Wu, RR Huang, LY Wei, JL Hua. 2020. Rapid detection of plant bacterial wilt by loop-mediated isothermal amplification. *Biotechnol. Bull.* 37, 272–281, doi: 10.13560/j.cnki.biotech.bull.1985.2020-0389.

Qureshi ZA, Ashraf M, Bughio AR, Siddiqui QH. 1975. Population fluctuation and dispersal studies of the fruit fly, Dacus zonatus Saunders. International Atomic Energy Agency; Food and Agriculture Organization: Sterility principle for insect control 1974. Proceedings of the symposium on the sterility principle for insect control jointly organized by the IAEA and the FAO of the United Nations and held in Innsbruck, 22-26 July 1974. International Atomic Energy Agency. Vienna Austria, 201-206

Tietjen M, Arp AP, Lohmeyer KH 2023. Development of a diagnostic single nucleotide polymorphism (SNP) panel for identifying geographic origins of *Cochliomyia hominivorax*, the New World screwworm. *Veterinary Parasitology*. doi.org/10.1016/j.vetpar.2023.109884.

Virgilio M, Daneel JH, Manrakhan A, Delatte H, Meganck K, De Meyer M 2019. An integrated diagnostic setup for the morphological and molecular identification of the *Ceratitis* FAR complex (*C. anonae*, *C. fasciventris*, *C. rosa*, *C. quilicii*, Diptera, Tephritidae). *Bull. Entomol. Res*. 109, 376-382.

White IM, Elson-Harris MM 1992. Fruit Flies of Economic Significance; their Identification and Bionomics. CAB International, Wallingford.

White IM, Evenhuis NL. 1999. New species and records of Indo-Australasian Dacini (Diptera: Tephritidae). *Raffles Bulletin of Zoology*, 47(2):487-540; 31

Zhang Y, De Meyer M, Virgilio M, Feng S, Badji K, Li Z 2021. Phylogenomic resolution of the Ceratitis FARQ complex (Diptera: Tephritidae). *Mol. Phylogenet. Evol.* 161,107160.

Zhang Y, Li W, Virgilio M, De Meyer M, Li Z. 2023. Loop-mediated isothermal amplification (LAMP) identification technique for economically important fruit flies of the genus *Ceratitis* (Diptera: Tephritidae). *J. Econ. Entomol.*, 116: 1982-1989.

Zhao F, Zou WM 2007. Application of LAMP in the rapid detection of aquatic animal pathogens. *South Chin. Fisher. Sci.* 3, 71–75, doi: 10.3969/j.issn.2095-0780.2007.02.013.

Annexes

Annex I *Genomic positions of C. capitata diagnostic* SNPs*, ranked in their importance for predicting the origin.*

Scaffold	Position
Scaffold_1__1_contigs__length_106982999	90482789
Scaffold_5__1_contigs__length_63380341	22824082
Scaffold_2__1_contigs__length_122778147	84437518
Scaffold_1__1_contigs__length_106982999	90063414
Scaffold_2__1_contigs__length_122778147	77500229
Scaffold_2__1_contigs__length_122778147	70379495
Scaffold_6__2_contigs__length_98771167	84590987
Scaffold_6__2_contigs__length_98771167	62192994
Scaffold_5__1_contigs__length_63380341	44007575
Scaffold_6__2_contigs__length_98771167	73314754
Scaffold_2__1_contigs__length_122778147	81443445
Scaffold_5__1_contigs__length_63380341	60557592
Scaffold_2__1_contigs__length_122778147	86760546
Scaffold_1__1_contigs__length_106982999	89755054
Scaffold_6__2_contigs__length_98771167	25015125
Scaffold_5__1_contigs__length_63380341	39328656
Scaffold_6__2_contigs__length_98771167	32417937
Scaffold_1__1_contigs__length_106982999	19070865
Scaffold_2__1_contigs__length_122778147	117890793
Scaffold_6__2_contigs__length_98771167	55373620
Scaffold_4__1_contigs__length_86735756	34017466
Scaffold_4__1_contigs__length_86735756	28623585
Scaffold_1__1_contigs__length_106982999	68737163
Scaffold_4__1_contigs__length_86735756	31276228
Scaffold_4__1_contigs__length_86735756	21042369
Scaffold_3__1_contigs__length_91678139	2650622
Scaffold_6__2_contigs__length_98771167	75127899
Scaffold_6__2_contigs__length_98771167	39841628
Scaffold_2__1_contigs__length_122778147	63940428
Scaffold_2__1_contigs__length_122778147	68495841
Scaffold_4__1_contigs__length_86735756	31465384
Scaffold_6__2_contigs__length_98771167	53011461
Scaffold_2__1_contigs__length_122778147	99601137
Scaffold_4__1_contigs__length_86735756	83307267
Scaffold_6__2_contigs__length_98771167	68343594
Scaffold_2__1_contigs__length_122778147	77124941
Scaffold_4__1_contigs__length_86735756	27876259
Scaffold_3__1_contigs__length_91678139	61630814
Scaffold_6__2_contigs__length_98771167	64882231

Scaffold_2__1_contigs__length_122778147	38287857
Scaffold_4__1_contigs__length_86735756	65206179
Scaffold_6__2_contigs__length_98771167	80893665
Scaffold_4__1_contigs__length_86735756	56982492
Scaffold_4__1_contigs__length_86735756	85672636
Scaffold_2__1_contigs__length_122778147	334007
Scaffold_6__2_contigs__length_98771167	73444895
Scaffold_6__2_contigs__length_98771167	70064085
Scaffold_3__1_contigs__length_91678139	79922116
Scaffold_4__1_contigs__length_86735756	28909280
Scaffold_1__1_contigs__length_106982999	96115326

Annex II *Genomic positions of B. dorsalis diagnostic SNPs ranked in their importance for predicting the origin.*

Scaffold	Position	Scaffold	Position
Chr01	16933966	Chr02	61592149
Chr01	17664922	Chr04	31398157
Chr03	38685089	Chr01	35372825
Chr03	66587020	Chr02	52874117
Chr03	61013985	Chr01	17345067
Chr05	3898491	Chr02	12664578
Chr01	17298781	Chr03	51156773
Chr03	57117418	Chr01	17902773
Chr02	71151073	Chr01	17595515
Chr01	14861237	Chr05	56926208
Chr02	83124110	Chr03	30772830
Chr05	47599909	Chr01	8283260
Chr01	17571312	Chr03	4309674
Chr03	45561050	Chr02	30704448
Chr01	94442904	Chr01	20577682
Chr03	64863712	Chr03	52353730
Chr01	17689383	Chr03	42176693
Chr01	20793498	Chr03	26872068
Chr01	2722892	Chr02	47828365
Chr01	14142858	Chr04	7354184
Chr02	47402798	Chr01	21276814
Chr03	53875241	Chr02	88272393
Chr01	19937898	Chr02	90052588
Chr01	21289436	Chr02	5124388
Chr01	21000942	Chr02	28512688
Chr03	46377173	Chr05	22822216
Chr01	16988652	Chr04	8424383
Chr03	9536996	Chr02	78828686
Chr02	38608893	Chr01	15005582
Chr02	39895429	Chr01	17174146
Chr01	17878840	Chr01	7041152
Chr03	36591546	Chr02	42619087
Chr01	15104053	Chr01	37249743
Chr03	27360675	Chr03	38262824
Chr02	21876148	Chr02	99559735
Chr02	93421016	Chr01	16980514
Chr01	16104184	Chr02	29809103
Chr05	55495816	Chr01	31920906
Chr01	20307965	Chr01	5578668
Chr01	14104957	Chr05	6453929
Chr03	57190037	Chr02	31861586

Chr03	40796565	Chr04	85706565
Chr03	53148038	Chr05	6515250
Chr01	13690929	Chr03	51349522
Chr05	9069198	Chr05	62492347
Chr01	13185885	Chr03	58658237
Chr01	19831754	Chr01	20014494
Chr02	31366978	Chr01	13793560
Chr01	3686983	Chr04	6471687
Chr01	19502956	Chr02	3494000
Chr02	102840382	Chr02	43243272
Chr01	27236918	Chr01	50259623
Chr03	60999272	Chr01	21411058
Chr01	12321398	Chr01	10816124
Chr03	69966713	Chr01	27353061
Chr02	42628198	Chr01	23970690
Chr01	16374825	Chr01	3661293
Chr01	43055729	Chr04	75505016
Chr01	882665	Chr02	23075710
Chr02	102324519	Chr05	21328223
Chr02	42810790	Chr01	11371349
Chr01	26955650	Chr03	97985353
Chr01	103523581	Chr03	52591435
Chr04	8856601	Chr01	9804530
Chr02	89210941	Chr03	37970481

Annex III *Genomic positions of B. zonata diagnostic SNPs ranked in their importance for predicting the origin.*

Scaffold	Position
ptg000006l	2449074
ptg000008l	15236349
ptg000006l	2438095
ptg000006l	2413750
ptg000002l	23699149
ptg000006l	2471063
ptg000006l	3162856
ptg000006l	649804
ptg000006l	3111999
ptg000006l	2524082
ptg000006l	2464444
ptg000006l	3174289
ptg000079l	1908950
ptg000006l	507583
ptg000006l	2546245
ptg000006l	2482875
ptg000006l	2458180
ptg000006l	5037544
ptg000006l	59204
ptg000006l	59202

Annex IV Bierman et al. manuscript (submitted)

Field friendly identification of the Peach Fruit fly, Bactrocera zonata, (Diptera: Tephritidae) using Loop-Mediated Isothermal Amplification

Anandi Bierman[1], Minette Karsten[2], Annelie Smit[1] , Marc de Meyer[3], Nikos Papadopoulos[4] & John S. Terblanche[1]

[1]Centre for Invasion Biology, Department of Conservation Ecology & Entomology, Stellenbosch University, South Africa, [2]Department of Conservation Ecology & Entomology, Stellenbosch University, South Africa [3]Department of Biology, Royal Museum for Central Africa; [4]Laboratory of Entomology & Agricultural Zoology, Department of Agriculture Crop Production and Rural Environment, University of Thessaly, Volos, Greece

Corresponding author: Anandi Bierman: anandie@sun.ac.za

Abstract

The rapid and accurate identification of invasive pest insects at quarantine stations or ports of entry is crucial for preventing their establishment and spread in new territories. Here, we present an effective and rapid genetic identification technique utilizing loop-mediated isothermal amplification (LAMP) to detect the presence of the Peach Fruit Fly, *Bactrocera zonata*. The LAMP assay was designed to target the Cytochrome Oxidase subunit I (COI) gene region of *B. zonata*, ensuring high specificity and sensitivity, and to discriminate it from a range of other economically important *Bactrocera* species. By utilizing isothermal conditions, LAMP eliminates the need for lengthy thermal cycling and enables rapid amplification of target DNA within a short period. Additionally, a rapid DNA extraction method combined with the use of an intercalating UV fluorescent dye allows for fast and reliable interpretation of results without the need for specialized equipment. Validation studies were conducted using known samples of *B. zonata*, as well as other *Bactrocera* species commonly encountered in quarantine settings, including *B. dorsalis, B. correcta, B. tryoni, B. latifrons, B. oleae, B. cucurbitae* and *B. tau*. The power of this LAMP assay lies in its specificity to accurately distinguish *B. zonata* from other closely-related *Bactrocera* species. Moreover, the assay exhibits a remarkably short turnaround time, with results obtained in under an hour, facilitating timely decision-making at quarantine stations.

Keywords: Fruit flies, interception, diagnostic testing, invasive species, quarantine pest, field-friendly, early detection, border bio-security

1. INTRODUCTION

The inadvertent spread of invasive insect species through global trade causes substantial economic losses in agriculture (Bacon et al., 2012; Horton et al., 2013; Szybiszewska et al., 2016). With the continuous growth of global trade, it is expected that the number and impact of successful invasions by plant pests will also increase (Levine and D'Antonio 2003; Essl et al., 2011). The European Union (EU) has established regulations that classify non-EU Tephritidae as quarantine organisms, prohibiting their introduction into the EU (EFSA Panel on Plant Health (PLH) 2020). Consequently, imported produce must undergo inspection by each EU member state to prevent the entry, establishment, and spread of these quarantine organisms. While experts can differentiate between adult tephritid species, identifying the larval stage, which is commonly found in intercepted produce, is challenging, even at the third larval instar (Balmès & Mouttet, 2017; EPPO, 2011). Larvae could be reared to the adult stage for morphological identification, however this results in considerable prolongation of the identification process and adult identifications can be complicated by species complexes.

DNA-based molecular tools, particularly those involving DNA sequence or gene region analysis, have been increasingly used alongside morphological characters for species identification in various insect pest species. The advantages of DNA-based tools are further amplified when polymerase chain reaction (PCR) methods are employed, as they can often be applied even with limited quantities of poorly preserved specimens. Several DNA amplification-based techniques have been utilized for identifying insect pest species, including PCR-restriction fragment length polymorphism (RFLP) (Armstrong et al., 1997; Barr et al., 2006), amplified fragment length polymorphism (AFLP) (Kakouli-Duarte et al., 2001), oligonucleotide array-based methods (Naeole and Haymer 2003) and, more recently, DNA barcoding (Barr et al., 2012; Van Houdt et al., 2010, Virgilio et al., 2017). However, PCR-based methods have certain drawbacks, such as the need for a relatively expensive precision thermal cycler and the time-consuming process of shipping samples from quarantine checkpoints to reference laboratories, leading to significant delays in species identification. This delay poses a challenge, especially for perishable plant imports like fruits, as it can result in substantial economic losses for importers. To overcome this issue, rapid molecular on-site tests conducted directly at the point of entry have shown promise (Blaser et al., 2018; Alon et al., 2023).

One such rapid molecular diagnostic tool is loop-mediated isothermal amplification (LAMP) (Notomi et al., 2000). This method offers rapid and robust amplification and operates under isothermal conditions, requiring only a simple thermostat-based instrument. The LAMP reaction employs specially designed primer pairs and a DNA polymerase, such as BstDNA polymerase, which possesses strand displacement activity. Recent studies have demonstrated that the inclusion of an additional pair of primers (loop primers) can enhance the efficiency of the LAMP method by reducing the amplification time by half (Nagamine et al., in 2002). Overall, LAMP represents a valuable molecular diagnostic tool due to its simplicity, rapidity, and reliable performance.

The loop-mediated isothermal amplification (LAMP) technique has found diverse applications across various fields. In disease studies LAMP has been utilized for diagnostic identification of bacterial, protozoan, and viral infectious agents (Iwamoto et al., 2003; Poon et al., 2006; Hong et al., 2004). Additionally, LAMP has been employed for pathogen quantification (Toriniwa and Komiya 2006), as well as for the detection of plant viruses (Fukuta et al., 2004), as a field method for identifying emerging viruses of biomedical significance (Parida 2008), and as a diagnostic tool for distinguishing between termite species (Itakura et al., 2006). To date, there have been several reported applications demonstrating the potential of LAMP for the rapid identification of insect pests, particularly tephritids, to enhance quarantine interception and indentification at ports of entry or other relevant settings. LAMP kits have been developed for rapid identification of the *Bactrocera dorsalis* complex of *B. cucurbitae*/*B. latifrons* and of *B. correcta*/*B. zonata* (Blaser et al., 2018) as well as for *B. tryoni* (Blacket et al., 2020). In addition to the *Bactrocera* species, LAMP assays are available for *Dacus ciliatus* (Sabahi et al., 2018) and *Zeugodacus scutellatus* (Kitano & Takakura, 2020). The other major tephritid pest species namely *Ceratitis*, have LAMP assays developed for *C. capitata* (Huang et al., 2009) and Dermauw et al., (2022) developed a user-friendly extraction method and LAMP assay for *C. capitata* and *C. cosyra* group1 (as defined by Virgilio et al., 2017) or *Ceratitis* species belonging to the FARQ complex which consists of *C. rosa*, *C. capitata*, *C.fasciventris* and *C. anonae* (as discussed by Virgilio et al., 2019, Zhang et al., 2021). Most recently, the isothermal amplification technique has been combined with detection through CRISPR-Cas12a

(Alon et al., 2023) to provide rapid and easy-to-use DNA-based detection of *Ceratitis capitata* and *Bactrocera zonata*.

The peach fruit fly, *Bactrocera zonata*, (Saunders) (Diptera: Tephritidae) is native to South and South-East Asia but it has invaded and become established in a number of countries in the Middle East, the Arabian Peninsula, North Africa and some of the Indian Ocean Islands (i.e. Mauritius and La Réunion) (CABI 2017; EPPO 2010; De Meyer et al., 2007). It is frequently detected in California, USA (Papadopoulos et al., 2013) and every year since 2010 in Vienna Austria (Egartner et al., 2019). It has recently established in Israel (EPPO 2002). In tropical and subtropical regions, where its host plants are available year-round, it causes significant economic losses (Stonehouse et al., 1998, OEPP/EPPO 2005). *Bactrocera zonata* is a species of agricultural interest due to its strong flying ability (Qureshi et al., 1975), short generation time, adaptability to different habitats, and wide range of host plants (CABI 2017). The recent expansion of its host range to olives following changes in its microbiome (Awad et al., 2023) highlight the threat posed by *B. zonata* to commercially grown crops.

In this study we describe a novel field-friendly, rapid protocol for the identification of *Bactrocera zonata* from adult and larval tissue. Using the LAMP method, we show that *B. zonata* can be discerned from other *Bactrocera* species in less than 1 hour, without the need for complicated laboratory equipment.

2. MATERIALS AND METHODS

2.1 Samples and sequence data

Physical specimens used for DNA extractions for testing of LAMP assays were obtained from a collection held at Stellenbosch University, Conservation Ecology and Entomology Department, and originating from the International Atomic Energy Agency (IAEA, Austria), and the Royal Museum of Central Africa (RMCA, Tervuren, Belgium). Table 1 provides a summary of the species used in the current study as well as the accession numbers and origins of sequence data used for LAMP primer design. Sequence data of the Cytochrome Oxidase subunit I gene region (COI) for LAMP primer design was obtained from GenBank (https://ncbi.nlml.nih.gov) and the Barcode of Life Database (BOLD; https://boldsystems.org/) in fasta format. All sequences were deposited either by the Royal Museum of Central Africa (RMCA, Tervuren, Belgium) or the International Atomic Energy Agency (IAEA, Austria).

Table 1: Species list of samples used in the current study including origin of physical samples used for testing LAMP assays and origin and accession number of sequence data used for LAMP primer design.

Species	GenBank Accession No.	Specimen origin	Sequence origin
Bactrocera dorsalis	KM023410.1	CRI, Nelspruit	RMCA (GenBank)
Bactrocera correcta	FFIPM488-22 & FFIPM489-22	Vietnam (lab strain)	IAEA (BOLD)
Bactrocera zonata	KM023423.1	Pakistan (lab strain)	RMCA (GenBank)
Bactrocera tryoni	AB720890.1, FFIPM508-22, FFIPM509-22	Queensland, Brisbane, Australia (lab strain)	IAEA (NCBI) and IAEA (BOLD)
Bactrocera latifrons	GQ154148.1, GQ154147.1, GQ154146.1, GQ154145.1, GQ154144.1, GQ154143.1, GQ154142.1, GQ154141.1, GQ154140.1,	RMCA	RMCA (GenBank)

Bactrocera oleae	GQ154139.1, GQ154138.1, FJ009203.1, KJ703712.1, KM023413.1 KM023417.1	Greece (lab strain)	RMCA (GenBank)
Bactrocera cucurbitae (now *Zeugodacus cucurbitae*)	HQ664517.1 to HQ664547.1, GQ154090.1 to GQ154135.1, KM023407.1 to KM023409.1	Hawaii (lab strain)	RMCA (GenBank)
Bactrocera tau (now *Zeugodacus tau*)	GQ154157 to GQ154161, KM023421	China (lab strain)	GQ154161
Ceratitis capitata		CRI, Nelspruit	

2.2 LAMP primer design

Sequences were aligned using MAFFT (Katoh et al., 2002). First, the same species' sequences were aligned (if more than one representative sequence was present for a species, Table 1) and consensus sequences obtained. Consensus sequences for all *Bactrocera* species (Table 1) were then aligned. Using the COI sequence for *B. zonata* as input sequence, the NEB primer design tool (https://lamp.neb.com/#!/) was used with default settings to design LAMP primers. Primer options given by the NEB primer design tool were placed on the MAFFT sequence alignments of all *Bactrocera* species and primer sets were visually inspected for sequence similarity or dissimilarities. Primers were selected if they showed at least one sequence difference per primer between *B. zonata* and another *Bactrocera* species.

2.3 DNA extraction

For optimization of LAMP assays, DNA from all physical specimens (Table 1) were extracted using the Qiagen DNeasy Blood and Tissue Kit (Qiagen) using the optimized protocol for insect tissue (Qiagen DNeasy Blood and Tissue Kit supplementary protocol DY14). Only one insect was used per extraction and samples were not pooled. Whole insects were extracted individually and DNA quality and quantity was assessed using the NanoDrop1000 (ThermoFisher). Samples were diluted with nuclease free water to 100 ng for use in the LAMP assay.

In order to assess more field-friendly DNA extraction methods, three alternative extraction protocols were tested on *B. zonata* and *B. dorsalis* larvae. Heads of larvae were removed and bodies dissected into posterior and anterior halves. Two larval halves (randomly selected posterior and anterior) were used per DNA extraction.

Method 1:

A method by Kitano and Takakura (2020) was adapted wherein larval posterior and anterior halves were added to 0.2 mL microcentrifuge tubes (Eppendorf, Germany) and 30 µL TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) was added (see supplementary material for detailed buffer preparation method). The solution was incubated at 95 °C for 5 min on a thermal cycler (Labnet MultiGene™ OptiMax thermal cycler), and used as-is for LAMP assays as the supernatant would contain the roughly extracted DNA.

Method 2:

Adapted from Blaser et al., (2018), the posterior and anterior larval halves were added to 600 µM potassium hydroxide (KOH) (see supplementary material for detailed buffer preparation method) to a volume of 30 µl in 0.2 mL microcentrifuge tubes (Eppendorf, Germany). The solution was heated to 95 °C for 5 min on a thermal cycler (Labnet MultiGene™ OptiMax thermal cycler), and used as-is for LAMP assays as the supernatant would contain the roughly extracted DNA.

Method 3:

Lysis buffer from the Qiagen DNeasy Blood and Tissue kit (Qiagen) was added to the posterior and anterior larval halves at 200 µl per sample in 1.5 mL microcentrifuge tubes (Eppendorf, Germany) and incubated at 95 °C for 5 min on a heat block (Thermomixer Comfort; Eppendorf AG, Hamburg, Germany). Proteinase K, though used as part of the lysis process for the Qiagen DNeasy Blood and Tissue kit, was omitted from the reaction in an attempt to design a cost effective extraction procedure. The reaction was used as-is for LAMP assays as the supernatant would contain the roughly extracted DNA.

2.4 LAMP assay optimisation

LAMP primer specificity was first assessed using DNA extracted from all *Bactrocera* species using the Qiagen DNeasy Blood and Tissue kit (Qiagen) (Table 1). Assay success was evaluated by running products on a 1.5% Agarose gel (see supplementary material for detailed preparation method) at 90V for 1hr and visualising the resultant products under UV light. A negative control was run in every instance as well as a 100 bp ladder and samples were stained using SmartGlow LD (Accuris). The LAMP assay was set up to a total volume of 25.0 µL using the Warm Start NEB LAMP kit (New England Biolabs, MA, US E1700). Briefly, a 10x Primer Master Mix was prepared from 100µM stock as follows: 2.4µL FIP, 2.4µL BIP, 0.3µL F3, 0.3µL B3, 0.6µL LF and 0.6µL LB added to 8.4µL water. This constituted enough for five reactions. For the LAMP assay, 12.5 µL Warm Start Master mix, 2.5µL 10x Primer Master Mix, 0.5µL fluorescent dye, 8.5µL water and 1.0 µL template DNA were added to 0.2 mL microcentrifuge tubes. Samples were incubated at 65 °C for 30 min and the reaction was terminated at 81 °C for 5 min on a thermal cycler (Labnet MultiGene™ OptiMax thermal cycler). After amplification, 10 µL products were loaded onto the agarose gel for electrophoresis. Primer specificity would be confirmed if a ladder-like product was observed at around 500 bp, only for the *B. zonata* sample, and all other samples resembled the negative control.

In order to optimize the protocol to be more field-friendly, a more rapid assay confirmation method than the Agarose gel was assessed. A LAMP assay was conducted as described above and instead of loading the resultant product on an agarose gel, 10 µL of the product was mixed with 0.5 µL Ethidium Bromide (Promega) in the 0.2 mL microcentrifuge tube that it was amplified in. The tubes were directly visualised under UV light. Primer specificity would be confirmed if fluorescence was observed only for the *B. zonata* sample and all other samples resembled the negative control.

3. RESULTS

3.1 LAMP primer design

LAMP-specific primers for the *B. zonata* COI gene were designed and assessed visually for specificity by placing primers on MAFFT sequence alignments of all *Bactorcera* consensus sequences. Figure 1 shows such an example MAFFT alignment for *B. dorsalis*, *B. latifrons*, *B. oleae* and *B. zonata* to showcase the sequence differences within primer binding sites. Sufficient sequence differences between *B. zonata* and other *Bactrocera* species were needed to ensure primer specificity. Final primer sequences are provided in Table 2.

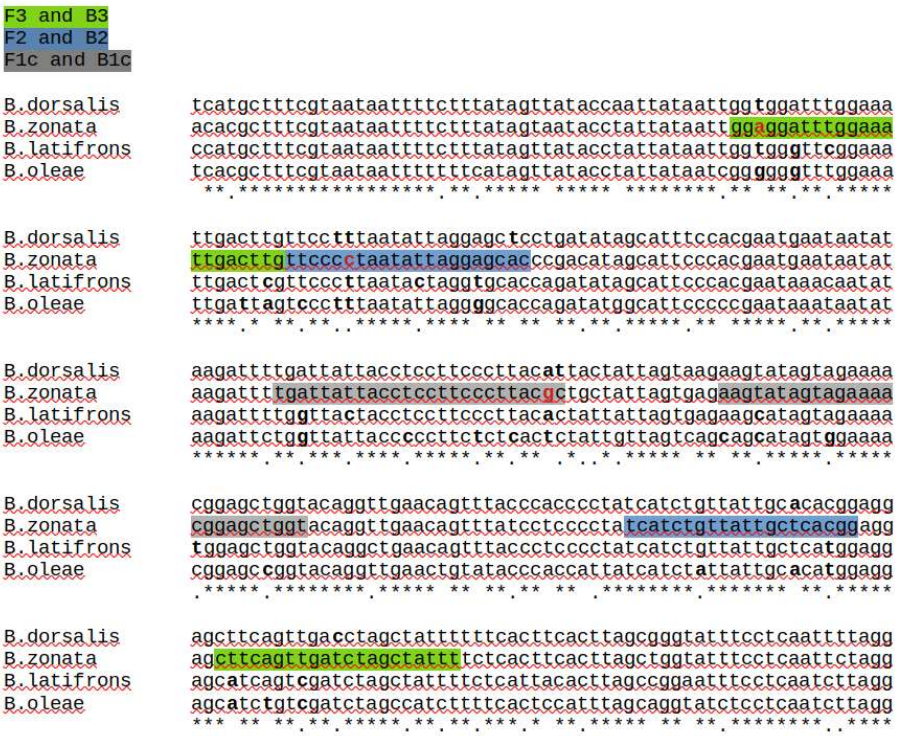


Figure 1: MAFFT sequence alignment of *B. dorsalis*, *B. zonata*, *B. latifrons* and *B. oleae* depicting primer locations of F3 and B3 primers (in green), F2 and B2 primers (in blue) and F1c and B1c primers (in grey). Bold letters highlight sequence differences between *B. zonata* and another *Bactrocera* species while red, bold letters highlight sequence differences present in *B. zonata* and the three other *Bactrocera* species shown here. (All *Bactrocera* species used in the study were aligned and visually inspected for primer specificity, though only four prominent species are shown here, for brevity).

Table 2: Primer sequences of LAMP primers designed for specific amplification of *Bactrocera zonata* DNA.

Primer	Sequence (5' - 3')
F3	GGA GGA TTT GGA AAT TGA CTT
B3	AAA TAG CTA GAT CAA CTG AAG C
FIP	GCG TAA GGG AAG GAG GTA ATA ATC AGT TCC CCT AAT ATT AGG AGC AC
BIP	AAG TAT AGT AGA AAA CGG AGC TGG TCC GTG AGC AAT AAC AGA TGA
LF	ATT CAT TCG TGG GAA TGC TAT GTC G
LB	AGG TTG AAC AGT TTA TCC TCC CCT A

3.2 LAMP assay

Primer specificity was experimentally confirmed by performing the LAMP assay with the selected primers (Table 2) using DNA extracted with the Qiagen DNeasy Blood and Tissue kit (Qiagen) from all *Bactrocera* specimens in this study (Table 1). Negative controls for the assay was run in every instance where all components of the assay were added except for DNA. The specificity was visually confirmed on agarose gel electrophoresis where the ladder-like amplification of *B. zonata* only can be seen around 500 bp, with all other samples resembling the negative control (Figure 2).

3.3 Field-friendly optimization

The assessment of a more field-friendly DNA extraction method proved most successful using Method 1 (Kitano and Takakura 2020). The addition of 30 µL TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) and subsequent incubation at 95 °C for 5 min on a thermal cycler (Labnet MultiGene™ OptiMax thermal cycler) yielded DNA of sufficient quality and quantity to be amplified in the LAMP assay (Figure 3).

Visualising LAMP assay products by the addition of Ethidium Bromide (Promega) to the 0.2 mL vials in which amplification took place, proved to be successful given that a negative control is also visualised in order to calibrate accordingly the fluorescence emitted by the positive sample. Figure 3 shows the successful LAMP assay performed using the field-friendly DNA extraction method with in-tube visualization by Ethidium Bromide staining, yielding fluorescent product only for *B. zonata*.

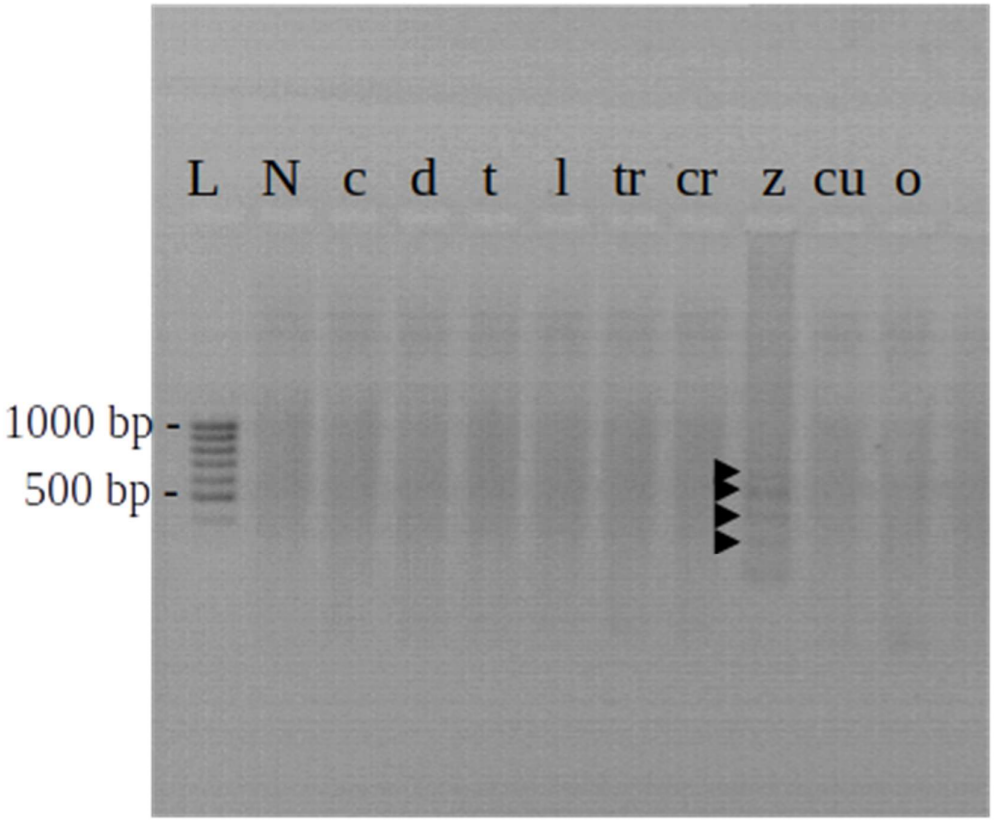


Figure 2: LAMP assay products on 1.5% Agarose gel**.** From left to right lanes contain: L - 100 bp ladder; N - negative control; c - *Ceratitis capitata*; d - *Bactrocera dorsalis*; t - *Bactrocera tau*; l - *Bactrocera latifrons*; tr - *Bactrocera tryoni*; cr - *Bactrocera correcta*; z - *Bactrocera zonata*; cu - *Bactrocera cucurbitae*; o - *Bactrocera oleae*. The ladder-like bands around the 500 bp mark for *B. zonata* (highlighted by black arrows) indicate the successful amplification of the LAMP primers while all other samples resemble the negative control.

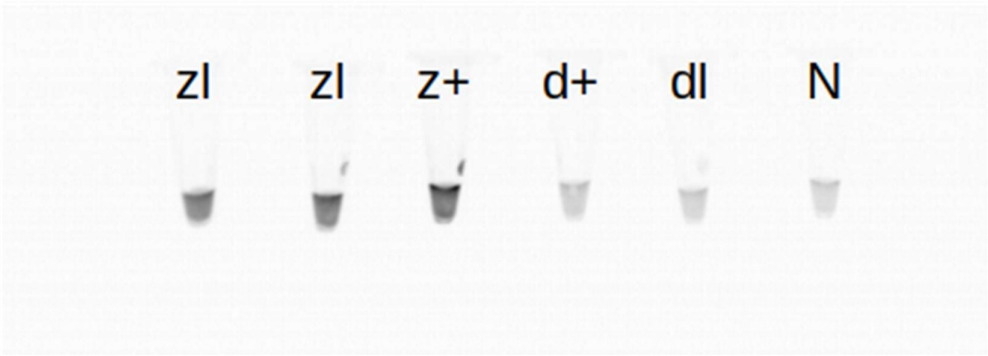


Figure 3: LAMP assay products visualised using the field-friendly in-tube method of adding Ethidium Bromide as fluorescent, intercalating dye in 0.2mL tubes. Samples are extracted using the field-friendly DNA extraction method of incubating larvae in TE buffer and compared to positive controls extracted using the Qiagen DNeasy DNA extraction kit. Samples are as follows from left to right: zl - *Bactrocera zonata* larval segments extracted using field-friendly method; zl - *Bactrocera zonata* larval segments extracted using field-friendly method; z+ - adult *Bactrocera zonata* extracted using the Qiagen DNeasy kit; d+ - adult *Bactrocera dorsalis* extracted using the Qiagen DNeasy kit; dl - *Bactrocera dorsalis* larval segments extracted using field-friendly method; N - negative control. Only samples containing *B. zonata* show fluorescence under UV light.

4. DISCUSSION

The LAMP technique has been applied to other tephritids; however, this study and the CRISPR-based assay of Alon et al., (2023) are the first assays specifically targetting the accurate and rapid identification of *Bactrocera zonata*. The current study, in particular, discerns *B. zonata* from among a series of other agriculturally important *Bactrocera* and *Zeugodacus* species as well as *Ceratitis capitata*. Through a combination of a more rapid DNA extraction and visualization method, the target species can be identified in under 1 hour, providing a field-friendly detection method for *B. zonata*. Figure 4 depicts the workflow for the LAMP assay and indicates the time taken at each step of the workflow. In terms of visual detection, previous studies have used agarose gel electrophoresis (Huang et al., 2009) and speciality fluorescence-based isothermal amplification devices (Dermauw et al., 2022; Blaser et al., 2018; Huang et al., 2009; Blacket et al., 2020). This study provides a methodology for detecting LAMP products in-tube only by adding an intercalating UV fluorescent dye which can be visualized immediately under any UV light source.

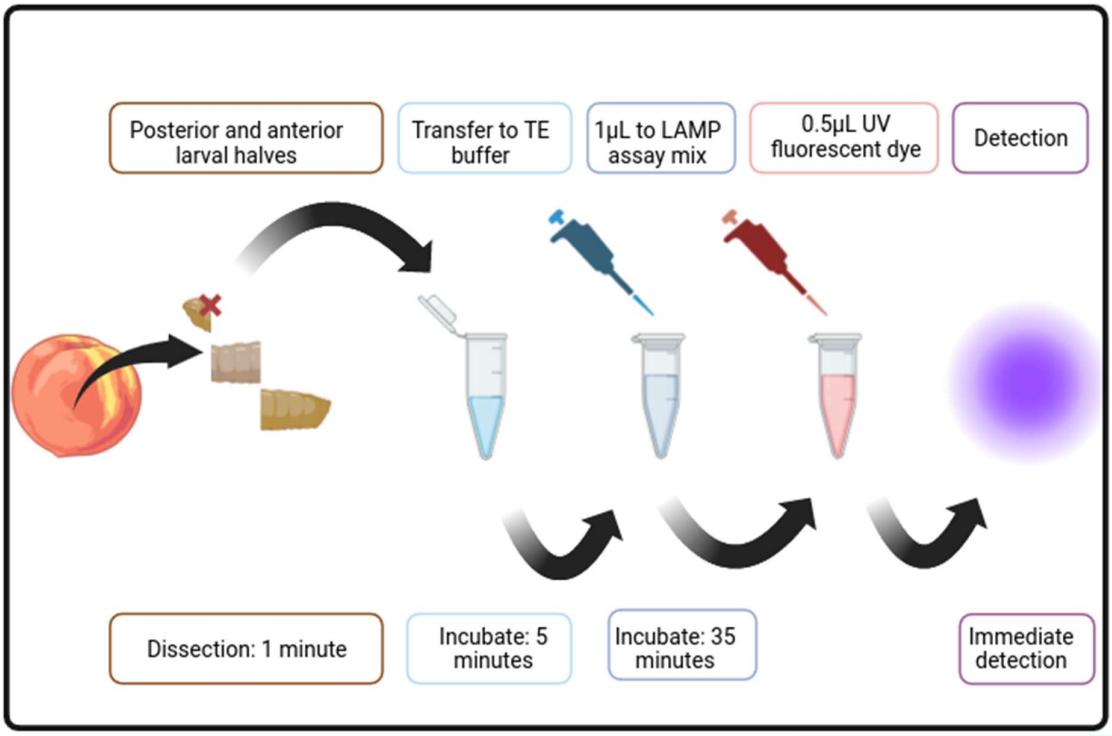


Figure 4: LAMP assay workflow depicting the steps and duration in time for each part of the workflow from dissection of larval samples on the left, through to the DNA extraction in TE buffer and transfer of 1µL crude extract to the LAMP assay mixture followed by addition of 0.5 µL of a UV-fluorescent dye such as Ethidium Bromide and immediate visualization under UV light. The complete work-flow can be accomplished in under 1 hour if sample material is at hand.

Alon et al., (2023) recently presented a protocol for detecting *B. zonata* and *C. capitata* using a CRISPR-Cas12a detection assay. Several notable differences exist between our study and the study conducted by Alon et al., (2023). First, while Alon et al., (2023) solely focused on *C. capitata* and *B. zonata*, our study included a greater diversity of closely related species. Though the assay of Alon et al., (2023) aimed to detect either *B. zonata* or *C. capitata*, the authors did not test the assay for specificity on any other species as our study did. Second, the field-friendly DNA extraction method presented in our study does not require manual grinding of samples or ultra-pure water, unlike the Chelex-based protocol used by Alon et al., (2023). The completion time for the protocol by Alon et al., (2023) is approximately 1.5 hours.

In contrast, our protocol can be completed in under 1 hour and does not require a benchtop centrifuge that might be cumbersome or unavailable to field-based applications. Lastly, the protocol by Alon et al., (2023) utilizes a specialized plate reader for visualization, mentioning that hand-held fluorometers could be used. However, this assessment was not conducted. In contrast, our protocol only requires a UV light source, which is less expensive and more readily available than fluorometers.

LAMP is an isothermal process and does not require temperature ramping up and down as in conventional PCR reactions (Notomi et al., 2000). This shortens the reaction time and simplifies the equipment (i.e. a thermal cycler might not always be necessary). In addition to the low complexity equipment required to perform LAMP assays, the lack of background interference due to the high specificity of LAMP primers to the target sequence makes LAMP particularly suited to field-based applications (Notomi et al., 2000). The advantages of LAMP including its low cost, rapidity, sensitivity and amenability to field applications make it ideally suited to species confirmation of economically important insects such as tephritids.

The advantages also far outweigh the disadvantages, including the slightly complicated (but not insurmountable) primer design and challenges in compiling multiplex reactions and quantifying target DNA. However, care should be taken as cross-contamination by material present in aerosols is still a risk and could lead to false positive results (Soroka et al., 2021) and the difficulty in detecting inhibitors in the reaction could lead to false negatives.

Most DNA-based species diagnostic tools utilize the Cytochrome Oxidase subunit I (COI) gene region or the 16S ribosomal RNA gene. Though these regions are often sufficient, instances do occur where species cannot be discriminated based on these gene regions alone. As discussed by Alon et al., (2023), the increased availability of genomic data for fruit flies of economic importance will be a significant benefit to the future development of DNA-based assays for species identification allowing researchers to widen their searches for unique DNA regions outside of the traditional Cytochrome Oxidase subunit I (COI) gene region.

Declaration of Competing Interest

The authors have not disclosed any competing interests.

Authorship contribution statement

All authors contributed to the study conception and design. Specimen collection, sample preparation and experimental analysis were conducted by ***Anandi Bierman, Annelie Smit*** *and* ***Minette Karsten*** *and* ***Marc de Meyer*** *(sample collection). Data analysis were performed by* ***Anandi Bierman****.* ***Anandi Bierman****,* ***Minette Karsten****, and* ***John S. Terblanche*** *contributed to the first draft of the manuscript and all authors commented on subsequent versions of the manuscript, revisions and responses to referees. All authors read and approved the final manuscript.*

Data availability

All data are available from the corresponding author upon reasonable request.

Acknowledgements

The authors wish to acknowledge the European Union's Horizon 2020 research and innovation programme under grant agreement No 818184 for funding and support of this project. The authors would also like to acknowledge contributions of sample material from the International Atomic Energy Agency (IAEA, Austria), Royal Museum of Central Africa (RMCA, Tervuren, Belgium) and Stellenbosch University, Conservation Ecology and Entomology Department with particular thanks to Maxi Snyman, Kurt Jordaens, John Midgeley and Marc de Meyer for assisting with obtaining samples.

REFERENCES

Alon, D.M., Partosh, T., Burstein, D. and Pines, G., 2023. Rapid and sensitive on-site genetic diagnostics of pest fruit flies using CRISPR-Cas12a. Pest Manag. Sci. 79, 68-75.

Armstrong, K.F., Cameron, C.M. and Frampton, E.R., 1997. Fruit fly (Diptera: Tephritidae) species identification: a rapid molecular diagnostic technique for quarantine application. Bull. Entomol. Res. 87, 111-118.

Awad, M., Ben Gharsa, H., ElKraly, O.A., Leclerque, A. and Elnagdy, S.M., 2022. COI Haplotyping and Comparative Microbiomics of the Peach Fruit Fly, an Emerging Pest of Egyptian Olive Orchards. Biology 12, 27.

Bacon, S.J., Bacher, S. and Aebi, A., 2012. Gaps in border controls are related to quarantine alien insect invasions in Europe. PLoS One 7, e47689.

Balmès, V. and Mouttet, R., 2017. Development and validation of a simplified morphological identification key for larvae of tephritid species most commonly intercepted at import in Europe. EPPO Bulletin 47, 91-99.

Barr, N.B., Copeland, R.S., De Meyer, M., Masiga, D., Kibogo, H.G., Billah, M.K., Osir, E., Wharton, R.A. and McPheron, B.A., 2006. Molecular diagnostics of economically important *Ceratitis* fruit fly species (Diptera: Tephritidae) in Africa using PCR and RFLP analyses. Bull. Entomol. Res. 96, 505-521.

Barr, N.B., Islam, M.S., De Meyer, M. and McPheron, B.A., 2012. Molecular identification of *Ceratitis capitata* (Diptera: Tephritidae) using DNA sequences of the COI barcode region. Ann. Entomol. Soc. Am. 105, 339-350.

Ben-Yosef, M., Verykouki, E., Altman, Y., Nemni-Lavi, E., Papadopoulos, N.T. and Nestel, D., 2021. Effects of Thermal Acclimation on the Tolerance of *Bactrocera zonata* (Diptera: Tephritidae) to Hydric Stress. Front. Physiol. 12, 686424.

Ben-Yosef, M., Altman, Y., Nemni-Lavi, E., Papadopoulos, N.T. and Nestel, D., 2023. Larval nutritional-stress and tolerance to extreme temperatures in the peach fruit fly, *Bactrocera zonata* (Diptera: Tephritidae). Fly 17, 2157161.

Blacket, M.J., Agarwal, A., Zheng, L., Cunningham, J.P., Britton, D., Schneider, I. and Rodoni, B.C., 2020. A LAMP assay for the detection of *Bactrocera tryoni* Queensland fruit fly (Diptera: Tephritidae). Sci. Rep. 10, 9554.

Blaser, S., Diem, H., von Felten, A., Gueuning, M., Andreou, M., Boonham, N., Tomlinson, J., Müller, P., Utzinger, J., Frey, J.E. and Bühlmann, A., 2018. From laboratory to point of entry: development and implementation of a loop-mediated isothermal amplification (LAMP)-based genetic identification system to prevent introduction of quarantine insect species. Pest Manag. Sci. 74, 1504-1512.

CABI, 2017. *Bactrocera zonata* (peach fruit fly) In: Invasive Species Compendium datasheet 17694. Wallingford, UK: CAB International. Available: http://www.cabi.org/isc. 2018.

De Meyer, M., Mohamed, S., White, I.M., 2007. Invasive fruit fly pests in Africa. Available: http://www.africamuseum.be/fruitfly/AfroAsia.htm.

Dermauw, W., Van Moerkercke, Y., Ebrahimi, N., Casteels, H., Bonte, J. and Witters, J., 2022. A loop-mediated isothermal amplification (LAMP) assay for rapid identification of *Ceratitis capitata* and related species. CRIS. 2, 100029.

EFSA Panel on Plant Health (PLH), Bragard C, Dehnen-Schmutz K, Di Serio F, Gonthier P, Jacques M.A., Jaques Miret JA, Justesen AF, Magnusson CS, Milonas P, Navas-Cortes JA, Parnell S, Potting R, Reignault PL, Thulke HH, Van der Werf W, Civera AV, Yuen J, Zappalà L, Bali EM, Papadopoulos N, Papanastassiou S, Czwienczek E, MacLeod A. 2020. Pest categorisation of non-EU Tephritidae. EFSA Journal 18, e05931.

EPPO, 2010. *Bactrocera zonata*: Procedure for official control. OEPP/ EPPO Bulletin, 40, 390–395.

EPPO, 2002. EPPO Workshop on *Bactrocera zonata*, Paris, UNESCO, 2002-03-05

EPPO, 2011. *Ceratitis capitata*. EPPO Bulletin, 41, 340–346.

Essl, F., Dullinger, S., Rabitsch, W., Hulme, P.E., Hülber, K., Jarošík, V., Kleinbauer, I., Krausmann, F., Kühn, I., Nentwig, W. and Vilà, M., 2011. Socioeconomic legacy yields an invasion debt. PNAS. 108, 203-207.

Fukuta, S., Ohishi, K., Yoshida, K., Mizukami, Y., Ishida, A. and Kanbe, M., 2004. Development of immunocapture reverse transcription loop-mediated isothermal amplification for the detection of tomato spotted wilt virus from chrysanthemum. J. Virol. Methods. 121, 49-55.

Gazit, Y., Akiva, R., 2017. Toxicity of malathion and spinosad to *Bactrocera zonata* and *Ceratitis capitata* (Diptera: Tephritidae). Fla. Entomol. 100, 385-389.

Horton, D.R., Lewis, T.M. and Dobbs, T.T., 2013. Interceptions of Anthocoridae, Lasiochilidae, and Lyctocoridae at the Miami plant inspection station (Hemiptera: Heteroptera). Fla. Entomol. 482-497.

Huang, C.G., Hsu, J.C., Haymer, D.S., Lin, G.C., Wu, W.J. 2009. Rapid identification of the Mediterranean fruit fly (Diptera: Tephritidae) by loop-mediated isothermal amplification. J. Econ. Entomol. 102, 1239-46.

Itakura, S., Masuta, T., Tanaka, H. and Enoki, A., 2006. Identification of two subterranean termite species (Isoptera: Rhinotermitidae) using cellulase genes. J. Econ. Entomol. 99, 123-128.

Iwamoto, T., Sonobe, T. and Hayashi, K., 2003. Loop-mediated isothermal amplification for direct detection of Mycobacterium tuberculosis complex, *M. avium*, and *M. intracellulare* in sputum samples. J. Clin. Microbiol. 41, 2616-2622.

Kakouli-Duarte, T., Casey, D.G. and Burnell, A.M., 2001. Development of a diagnostic DNA probe for the fruit flies *Ceratitis capitata* and *Ceratitis rosa* (Diptera: Tephritidae) using amplified fragment-length polymorphism. J. Econ. Entomol. 94, 989-997.

Katoh, K., Misawa, K., Kuma, K.I. and Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059-3066.

Kitano, D. and Takakura, K.I., 2020. Simple and on-site DNA purification for LAMP reaction applicable to non-adult tephritid fruit fly (Diptera: Tephiritidae). J. Appl. Entomol. 144, 824-829.

Levine, J.M. and D'Antonio, C.M., 2003. Forecasting biological invasions with increasing international trade. Conserv. Biol. 17, 322-326.

Nagamine, K., Hase, T. and Notomi, T.J.M.C.P., 2002. Accelerated reaction by loop-mediated isothermal amplification using loop primers. Moll. Cell. Probes. 16, 223-229.

Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N. and Hase, T., 2000. Loop-mediated isothermal amplification of DNA. Nucleic Acids Res. 28, e63-e63.

OEPP/EPPO, 2005. *Bactrocera zonata.* Data sheets on quarantine pests. European and Mediterranean Plant Protection Organization. EPPO Bulletin, *35* 371-373.

Papadopoulos, N.T., Plant, R.E. and Carey, J.R., 2013. From trickle to flood: the large-scale, cryptic invasion of California by tropical fruit flies. Proc. Royal Soc. B. 280, 20131466.

Parida, M.M., 2008. Rapid and real-time detection technologies for emerging viruses of biomedical importance. J. Biosci. 33, 617-628.

Poon, L.L., Wong, B.W., Ma, E.H., Chan, K.H., Chow, L.M., Abeyewickreme, W., Tangpukdee, N., Yuen, K.Y., Guan, Y., Looareesuwan, S. and Peiris, J.M., 2006. Sensitive and inexpensive molecular test for falciparum malaria: detecting Plasmodium falciparum DNA directly from heat-treated blood by loop-mediated isothermal amplification. Clin. Chem. 52, 303-306.

Qureshi, Z.A., Ashraf, M., Bughio, A.R. and Siddiqui, Q.H., 1975. Population fluctuation and dispersal studies of the fruit fly, Dacus zonatus Saunders. In *International Atomic Energy Agency; Food and Agriculture Organization: Sterility principle for insect control 1974. Proceedings of the symposium on the sterility principle for insect control jointly organized by the IAEA and the FAO of the United Nations and held in Innsbruck, 22-26 July 1974.* (pp. 201-206). International Atomic Energy Agency.

Sabahi, S., Fekrat, L., Zakiaghl, M. and Moravej, G.H., 2018. Loop-mediated isothermal amplification combined with PCR for rapid identification of the Ethiopian fruit fly (Diptera: Tephritidae). Neotrop. Entomol. 47, 96-105.

Soroka, M., Wasowicz, B. and Rymaszewska, A., 2021. Loop-mediated isothermal amplification (LAMP): The better sibling of PCR? Cells 10, 1931.

Stonehouse, J.M., Mumford, J.D. and Mustafa, G., 1998. Economic losses to tephritid fruit flies (Diptera: Tephritidae) in Pakistan. Crop Prot. 17, 159-164.

Szyniszewska, A.M., Leppla, N.C., Huang, Z. and Tatem, A.J., 2016. Analysis of seasonal risk for importation of the Mediterranean fruit fly, *Ceratitis capitata* (Diptera: Tephritidae), via air passenger traffic arriving in Florida and California. J. Econ. Entomol. 109, 2317-2328.

Thai, H.T.C., Le, M.Q., Vuong, C.D., Parida, M., Minekawa, H., Notomi, T., Hasebe, F. and Morita, K., 2004. Development and evaluation of a novel loop-mediated isothermal amplification method for rapid detection of severe acute respiratory syndrome coronavirus. J. Clin. Microbiol. 42, 1956-1961.

Toriniwa, H. and Komiya, T., 2006. Rapid detection and quantification of Japanese encephalitis virus by real-time reverse transcription loop-mediated isothermal amplification. Microbiol. Immunol. 50, 379-387.

Van Houdt, J.K.J., Breman, F.C., Virgilio, M. and De Meyer, M., 2010. Recovering full DNA barcodes from natural history collections of Tephritid fruitflies (Tephritidae, Diptera) using mini barcodes. Mol. Ecol. Resour. 10, 459-465.

Virgilio, M., Manrakhan, A., Delatte, H., Daneel, J.H., Mwatawala, M.W., Meganck, K., Barr, N.B. and De Meyer, M., 2017. The complex case of *Ceratitis cosyra* (Diptera: Tephritidae) and relatives. A DNA barcoding perspective. J. Appl. Entomol. 141, 788-797.

Virgilio, Massimiliano, J-H. Daneel, Aruna Manrakhan, Hélène Delatte, Kenny Meganck, and Marc De Meyer. 2019. An integrated diagnostic setup for the morphological and molecular identification of the *Ceratitis* FAR complex (*C. anonae*, *C. fasciventris*, *C. rosa*, *C. quilicii*, Diptera, Tephritidae). Bull. Entomol. Res. 109, 376-382.

Zhang, Y., De Meyer, M., Virgilio, M., Feng, S., Badji, K., Li, Z., 2021. Phylogenomic resolution of the Ceratitis FARQ complex (Diptera: Tephritidae). Mol. Phylogenet. Evol. 161,107160.