PROJECT DELIVERABLE REPORT

Deliverable D3.4: DNA sequences uploaded in BOLD and GENBANK

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	**D3.4: DNA sequences uploaded in GENBANK and BOLD**			
Work Package	WP3 – Development and enhancement of tools and methods for FF prevention			
Date of Delivery	**Contractual**	M52	**Actual**	M48
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	6/3/2023	RMCA, P. Deschepper/S. Vanbergen/A. Kaeyenbergh	Initial draft

0.20	16/6/2023	RMCA, M. De Meyer	Revised draft
0.30	20/6/2023	RMCA , P. Deschepper	Revised draft
0.40	26/6/2023	RMCA , S. Vanbergen	Edited M&M section, counts of individuals in table, comments
0.50	26/6/2023	RMCA, M. De Meyer	Second revised draft
0.60	03/07/2023	RMCA, S. Vanbergen	Cleaned up comments, accepted changes, added average genome wide depths per individual per species
0.70	10/07/2023	UTH, N. Papadopoulos	Revision
0.80	01/08/2023	RMCA, S. Vanbergen, M. De Meyer	Third revised draft, including comments and suggestions by N. Papadopoulos
0.90	01/08/2023	UTH, N. Papadopoulos	Minor edits on the third revision
1.10	10/08/2023	CIRAD, H. Delatte	Revision
1.20	10/08/2023	RMCA, M. De Meyer	Final draft
1.30	10/08/2023	EB	Approval of EB
2	20/09/2023	UTH	Submitted

Table of Contents

1 Summary

This deliverable presents the DNA sequences generated and/or made available within the framework of the FF-IPM project. DNA sequences relate to the three target species: *Bactrocera dorsalis, Bactrocera zonata* and *Ceratitis capitata* in addition to related taxa within the genera and subfamily Dacinae of importance to fruit and vegetable production.

The DNA sequences comprise two groups:

- Short DNA sequences, usually referred to as DNA barcodes, which are used as an identification tool for molecular identification of all life stages of fruit flies.
- Extended DNA sequencing reads, based on whole-genome sequencing, for the three target species, which are used to select diagnostic SNPs for tracing origin and reconstructing invasion histories.

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.

Regarding whole genome sequencing, in total 1,188 fruit fly genomes were re-sequenced for the three target species covering the full geographic range for each of them. Raw sequence reads for those data that are published, are publicly available either through the BioProject storage facility or the China National GeneBank DataBase. Data in process for publication preparation are currently stored on a local NAS at RMCA in the form of fastq.gz files and are available upon request. Ultimately, once published, they will also be uploaded on publicly available storage facilities.

2 Introduction

2.1 Purpose and Scope

The family Tephritidae has more than 5000 species distributed globally (White & Elson-Harris, 1992; Bragard et al. 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.,* 2007, 2016). The larvae feed in fruit and is the life stage detected during inspection of fruit for import or export, but are difficult to identify (Pieterse et al., 2016).

The correct identification of insect species is essential to comply with international biosecurity measures, since not all species are of quarantine importance in all countries (Boykin et al., 2012). Morphological identification is feasible for adult species if they are in intact condition. However, morphological identification of immature stages is less straightforward. Limited keys for third instar larvae (including the one developed under deliverable 3.3.) are available but the lack of a wide range of distinct character states makes identification difficult for the non-specialist. Other immature stages are virtually impossible to identify (i.e. eggs) or are not available (first and second instars, puparia). Genetic data are becoming increasingly important for both fundamental and applied research but also for routine activities such as identification. Such data, in the form of specific sequences of parts of the genome, are a suitable alternative to morphological identification. This is often referred to as DNA barcoding (Hebert et al., 2003). DNA barcoding is a molecular technique, which uses a particular region in the DNA-sequence as a "barcode" enabling taxonomic identification of a specimen, often up to species level. Assuming that the interspecific variation exceeds the intraspecific variation, the mitochondrial cytochrome c oxidase subunit 1 (COI) gene (648 bp) is a frequently used marker for animal species (Hebert et al., 2003). This technique is especially useful for specimens, which cannot be morphologically identified, for example damaged specimens, cryptic species or newly discovered species. After DNA is extracted from the animal tissue, a particular region of the genome (e.g. the COI gene) will be amplified with the use of markers using the Polymerase Chain Reaction (PCR) technique. After visualisation on an agarose gel, the samples are purified and finally sequenced. To identify the species, the obtained barcodes can then be compared to available sequences in an online digital library.

The FF-IPM project aimed at providing additional DNA barcodes to develop or expand existing reference libraries. The most widely used reference libraries are BOLD (Barcode of Life Data System, see https://boldsystems.org/) and GENBANK (https://www.ncbi.nlm.nih.gov/genbank/). The latter is not specifically for DNA barcodes but for all DNA sequences. In addition, FF-IPM was contacted by the European & Mediterranean Plant Protection Organization (EPPO) to specifically expand the European database EPPO-Q-BANK. EPPO-Q-bank (see https://qbank.eppo.int/), which is an open access database administered by EPPO. It was initiated in 2010 as a project in the Netherlands and further supported as the EU funded project QBOL. In 2018, when the initial funding ceased, the administration and maintenance was transferred to EPPO with harmonization of the content and structure between 2019-2020. In 2021 EPPO contacted FF-IPM with the request to incorporate DNA barcode data, both already available and newly generated, for non-EU quarantine fruit flies.

In addition, to mere identification, genetic data can also be used for more detailed investigations at population level. Next-generation sequencing (NGS) technology are becoming increasingly affordable and can determine nucleotide sequences of entire genomes or in target regions of interest. NGS data can produce valuable information on individual genotypes and genetic polymorphisms of which information

can be harnessed to answer questions regarding population demography or adaptation. The FF-IPM project has been exploring this at different levels, in particular for exploring seasonal population dynamics of one of the target species, *Ceratitis capitata*. These results are presented in deliverable 2.4. Secondly in developing molecular ID tools to trace origin of intercepted and detected fruit flies. For *Bactrocera dorsalis* this has been published already (see Zhang et al., 2022 for a worldwide analysis and Deschepper et al., 2022 for a detail analysis of the invasion history in the western Indian Ocean islands). Currently similar analyses for *C. capitata* (in addition to the already first analysis conducted using microsatellites; see under section 4.3 (*cf* below) and Deschepper et al., 2021) and *Bactrocera zonata*. The results of these studies will be incorporated in deliverable 3.5 (manual for the multi-step workflow and decision protocol) through the selection of diagnostic SNPs (single nucleotide polymorphisms) that are characteristic of particular geographic regions. As an output, the project aimed to generate genomic data for extensive samples of the three target fruit flies originating from different populations from the entire geographic distribution of each species.

3 Materials and methods

3.1 Genomic data

The price for whole-genome sequencing (WGS) technologies has dropped dramatically during the last decade while there is an increasing number of companies offering WGS- services. To discriminate between companies in terms of pricing and quality of the service, we first performed several pilot tests by sending comparable samples to different sequencing companies. Berry Genomics (https://www.berrygenomics.com/), located in China, was the best candidate and could deliver more output data for an equal price when compared to competitors. Since population genomic studies rely heavily on the number of samples per population used, Berry Genomics was the best choice for our research purposes and the bulk of our DNA extracts were sequenced at their facilities.

Whole-Genome Sequencing (WGS) data for all three focal species (*C. capitata, B. dorsalis* and *B. zonata*) were generated in a similar manner. Acquisition of raw sequence reads can be described as follows. Once fresh, ethanol preserved fruit fly specimens arrived at the molecular lab at the RMCA, the samples were stored in a -20°C freezer until DNA extraction. We used the Qiagen Blood & Tissue kit for extraction of whole genome DNA. The recommended guidelines of the kit manual were followed with exception of two steps. Firstly, we incubated the specimens at a temperature of 56° C overnight in an Eppendorf tube and removed the exoskeleton from the tube the morning after. Every exoskeleton is stored separately in a tube containing a tracking code and stored in a -80°C freezer, specifically reserved for invertebrate specimens. Secondly, only 120 µl of elution buffer was used to attain a sufficient DNA concentration in the extract. After DNA extraction, a qubit fluorometer was used to quantify the amount of DNA in each sample. A Reconcentrator Plus (Eppendorf) was used in cases where DNA concentration was not sufficiently high according to the recommendations of the sequencing company. Finally, DNA extracts were shipped to the facility of Berry Genomics where an output of 6Gb of 150bp paired-end Illumina sequences were produced using the NovaSeq 6000 platform, roughly generating a 5.0, 7.3 and 7.8x depth of coverage of the 0.5 Gb genome for *B. zonata*, *C. capitata* and *B. dorsalis* respectively. In the special case where samples did not reach a total DNA amount as recommended by Berry Genomics, they were sequenced at the facilities of Novogene (https://www.novogene.com/) using a low input library prep protocol.

To increase the potential of the read data generated, a good reference genome to align them to is key. Therefore, new reference genomes were generated in parallel of the FF-IPM project for *C. capitata* and *B. zonata* by Royal Museum for Central Africa (RMCA) and the French Agricultural Research Centre for International Development (CIRAD) and for *B. dorsalis* by the Chinese Agriculturel University (CAU). For this undertaking, RMCA relied on the facilities of Dovetail Genomics (https://dovetailgenomics.com/) and CAU partnered with Berry Genomics for sequencing and assembly of the de novo genomes. The different steps in the de novo assembly of *C. capitata* (Annex I and II) and *B. dorsalis* (Annex III) reference genomes can briefly be described as follows. A tissue sample of sufficient quality was submitted to the facilities of Dovetail Genomics. Next, PacBio HiFi long reads were generated, and a primary assembly was mode using the HiFiasm software. To obtain information on chromosome conformation, Dovetail Omni-C (Dovetail Genomics) or Hi-C libraries (Berry Genomics) were made and HiRise was used to obtain a final assembly (*C. capitata* only). Annotation of the assembly was performed by using a combination of RNA-seq data and sequence evidence of related species. All three newly acquired de-novo reference genomes are a significant improvement to the ones currently available on public platforms. The new assemblies have been or are being used as a reference to align read data of *C. capitata* and *B. zonata* produced in the framework of FFIPM and will be made open-access upon publication of our findings of the species' phylogeography.

3.2 DNA barcodes

DNA barcodes were generated in two ways: either developed directly through Sanger Sequencing or extracted from genomic data obtained through Next Generation Sequencing (cf 3.1).

Regarding the former, existing and new sequence data were gathered from the RMCA and the smaller set from Stellenbosch University (SU). These were developed mainly along the lines as described in Van Houdt et al. (2010) and Virgilio et al. (2015): DNA was extracted from both pinned and ethanol preserved specimens using the DNeasy Blood and Tissue Kit (Qiagen) and following the manufacturer's protocol. The standard COX1 barcoding primers, LCO1490 and HCO2198 (Folmer et al. 1994), were adjusted for the Tephritidae based on a comparative analysis of full mitochondrial genome sequences from *C. capitata* (NC_000857), *B. dorsalis* (NC_008748), *Bactrocera oleae* (NC_005333), *Bactrocera papayae* (NC_009770), *Bactrocera philippinensis* (NC_009771), *Bactrocera carambolae* (NC_009772) and *Anopheles gambiae* (NC_002084). The PCR was carried out in 30 lL containing 1X PCR buffer (Qiagen), 1–20 ng template DNA, 2.0 mM MgCl2, 0.2 mM dNTPs, 0.6 units of Taq DNA polymerase (Qiagen) and 0.4 lM of forward and reverse primer. The PCR profile starts with an initial denaturation of 3 min at 94 _C, followed by 40 cycles of 30 s at 94 _C, 30 s at 50 _C and 30 s at 72 _C. The PCR ends with a final step of 7 min at 72 _C. PCR products were purified by using Nucleofast PCR cleanup (Machery Nagel) or with 'GFX PCR DNA and Gel Band Purification kit' (GE Healthcare). PCR products were purified by means of GFX purification columns (GE Healthcare). Cleaned PCR products were sequenced in both directions using the BigDye version 3.1 cycle sequencing kit (Applied Biosystems) and finally sequenced in both directions with an ABI Prism 3100 Genetic Analyzer (Applied Biosystems). Nucleotide sequences were aligned using the muscle routine implemented by SeaView 4. Before analyses, coding regions were translated into amino acids to verify the possible presence of internal stop codons. All sequences were validated against known sequences to rule out contamination.

For barcodes extracted from genomic data, the methodology for obtaining genomic data is explained below in section 3.1. After performing Illumina Next Generation Sequencing, 150bp paired-end reads were aligned on the mitochondrial genome of *C. capitata*

(https://www.ncbi.nlm.nih.gov/nucleotide/NC_000857.1) and *B. dorsalis* (https://www.ncbi.nlm.nih.gov/nuccore/NC_008748.1) and variant sites have been called with the Genome Analysis Toolkit (GATK). To extract the COI gene sequence, raw genomic reads were filtered using fastp using the following settings, --qualified_quality_phred 30, --correction, --length_required 100. The filtered reads were then aligned to the mitochondrial reference genome (NCBI reference sequence accession number NC_000857.1) using bwa mem. The resulting bam files were then processed with samtools sort -n, samtools fixmate, samtools sort, and samtools markdup in that order. Mitochondrial consensus sequences were then generated per species using samtools consensus ---mode bayesian ---cutoff 50 ---format FASTA ---min-MQ 30 ---P-het 0. All consensus sequences were then combined in a single fasta file and manually trimmed in MEGA 11 to only contain the COI barcode and a flanking region of 500 bp on both sides. The resulting dataset was then realigned in MEGA 11 using ClustalW, and any leading or trailing gaps were trimmed manually. The individual sequences were then uploaded to BOLD.

Final data were first analysed, cleaned by Eric Pierre and Jean-Claude Streito of INRAE, UMR-Centre de Biologie pour la Gestion des Populations (Montferriez-sur-Lex, France) as part of the collaboration with ANSES, one of the European Reference Laboratories in charge of insects together with AGES of Austria. After this analysis, data were forwarded to EPPO for incorporation and upload on the EPPO Q-Bank.

4 Results

4.1 Genomic data

In the framework of T2.7 and T3.4, whole genome sequencing reads of the three focal tephritid species were generated. More specifically, a total of 1,255 fruit fly genomes were re-sequenced. Table 1 gives a full overview of the number of individuals per species and population that were sequenced. On average, roughly 27,000,000 reads were generated per sample across species, rendering an average genome wide depth of coverage ranging between 5 and 7.8 depending on species. All raw data is stored on a local NAS in the form of fastq.gz files.

To this date, two peer-reviewed publications have been produced using the sequence data generated (Deschepper *et al.* 2022, Zhang *et al.* 2022). Raw sequence reads for both publications have been made available: raw paired-end read data are stored as a BioProject (PRJNA893460) (Deschepper *et al.* 2022) and similarly, raw data for Zhang *et al.* (2022) are stored at the China National GeneBank DataBase (https://db.cngb.org/).

Table 1: Number of genomes re-sequenced per population and species.

B. dorsalis		***C. capitata***		***B. zonata***	
Country/Region	***#sequenced***	***Country***	***#sequenced***	***Country***	***#sequenced***
Anjouan	15	Argentina	6	Egypt	7
Grande Comore	15	Austria	62	India	18
Madagascar	48	Benin	5	Iran	66
Mauritius	15	Bolivia	2	Israel	7
Mayotte	20	Brazil	6	Mauritius	7
Moheli	15	Burundi	5	Oman	6

Réunion	23	Costa Rica	7	Pakistan	7
China	160	Croatia	104	Reunion	2
Laos	10	El Salvador	5	Sudan	7
Vietnam	10	Greece	14	Thailand	7
Myanmar	10	Guatemala	6	***Total***	***134***
Thailand	20	Italy	207		
Philippines	20	Kenya	5		
Malaysia	10	Mozambique	5		
Indonesia	10	Nicaragua	6		
Papua New Guinea	10	Panama	6		
Pakistan	10	Senegal	5		
India	44	South Africa	5		
Sri Lanka	10	Spain	13		
Bangladesh	10	Switzerland	11		
Nepal	10	Tanzania	5		
Sudan	10	Zambia	3		
Senegal	10	***Total***	***493***		
Mali	10				
Ghana	10				
Ivory Coast	10				
Nigeria	3				
Ethiopia	10				
Uganda	10				
Kenya	10				
DRC	10				
Burundi	10				
Malawi	10				
South Africa	10				
Hawaii	10				
Total	**628**				

4.2. DNA barcodes

Two datasets were created. The first dataset contained 164 species and 853 sequences (see Annex IV). After analysis, 17 sequences were removed from the dataset (pseudogenes, contaminations, questionable or unverifiable IDs, missing locality or specimen information) resulting in 163 and 836 species and sequences, respectively. A total of 49 of these sequences are temporarily invisible in BLAST as these are mostly sequences with gaps and might contain pseudogenes. All these are available in BOLD and in EPPO Q-bank.

The second data with newly generated COI sequences from the RMCA (mainly through mining of the genome databases, see above) contained a total of 1,035 sequences: 329 for *C. capitata* and 236 for *B. dorsalis* (see Annex V), 330 for *Ceratitis quilicii* and *Ceratitis rosa* together (see Annex VI) and 15 *Bactrocera* sp. and

125 *Dacus* sp. (see Annex VII). In addition, 46 COI sequences generated with sanger sequencing were received from the University of Stellenbosch (see Annex VIII). All these have been uploaded and are available in BOLD. These data also have been made available to the EPPO Q-Bank for upload selection in March 2023.

Below, we provide an overview of the contents of the Annexes related to barcoding sequences (table 2).

Table 2: Summary of annexes containing sequencing (meta) data. We distinguish data generated by traditional sanger sequencing (Sanger) and whole-genome-sequencing (WGS).

Annex	Species	Method	**Number of samples**	Content
IV	167 species	Sanger (COI)	853	Sequences and metadata
V	*C. capitata, B. dorsalis*	WGS	236	Metadata
VI	*C. quilicii, C. rosa*	WGS	330	Metadata
VII	55 species	WGS	204	Metadata
VIII	13 species	Sanger (COI)	46	Metadata

4.3. Microsatellite data

The recent study on the worldwide phylogeography of *C. capitata* generated a large dataset of microsatellite data (Deschepper *et al.* 2021). The results of this study served as a cornerstone to select samples of *C. capitata* for whole-genome sequencing and provided us with the information to cover all main genetic groups within the distribution of *C. capitata*. All microsatellite genotypes have been made available at: https://doi.org/10.5281/zenodo.3967065.

5 Conclusion

Genetic and genomic data generated and made available by the FF-IPM project contribute to the development of identification tools.

The genomic data in the form of paired-end 150 bp providing a depth of coverage of approximately 5x to 8x, depending on the species and reference genome. This data type allows for in-depth analysis of population structure or genome-wide association studies (GWAS). Sequences are gradually made available on public platforms (NCBI genbank and https://db.cngb.org/) upon publication of associated papers while also being stored on a local NAS system.

The genetic data allow an identification of all life stages (immature as well as adult), and conditions (fragmented as well as intact specimens). Such identification relies mainly on DNA barcoding and in particular on DNA barcodes of the mitochondrial DNA part of COI, considered as the main reference barcode region for insects (Hebert et al., 2003).

In total 1,917 specimens were DNA barcoded representing 173 taxa and made available through BOLD. Also, a substantial contribution was provided to EPPO-Q-Bank which is a reference database more specifically geared towards pest species of relevance to the European Union. With the DNA data received from the RMCA the number available sequences on the EPPO-Q-Bank has increased from 2524 specimens (1198 visible; 1326 invisible) to 3355 specimens (2034 visible; 1321 invisible) and from 80 different species (49 visible; 31 invisible) to a total of 210 species (176 visible; 34 invisible). Additional large datasets of the FF-IPM target species, of interest to chart the intra-specific variation, are submitted to EPPO-Q-Bank and are currently under investigation by the latter for further incorporation if considered relevant.

The above output will contribute significantly to ongoing and planned activities in the EU and beyond.

The extended DNA barcode libraries will facilitate identification activities of European reference laboratories as well as NPPOs involved in fruit fly identification. In addition, it will serve bodies and organizations in non-EU countries for similar activities.

The published genomic data can be integrated into datasets for future research on the species of interest. Such data spanning over multiple years can provide valuable insights into fruit fly evolution and population dynamics. Additionally, reference databases are useful in a more direct manner, for instance when being employed to reveal the geographic origin of (novel) populations or intercepted individuals outside of established populations' range (Tietjen et al., 2023).

REFERENCES

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

EFSA Panel on Plant Health (PLH), Bragard C., Dehnen-Schmutz K., Di Serio F., Gonthier P., Jacques M. A., MacLeod A., Miret J. A. J., Justesen A. F., Magnusson C. S., Milonas P., Navas-Cortes J. A., Parnell S., Potting R., Reignault P. L., Thulke H.-H., Van der Werf W., Vicent Civera A., Yuen J., Zappalà L. (2020). Pest categorisation of non-EU Tephritidae. EFSA Journal, 18(1), e05931.

Deschepper, P., T.N. Todd, Virgilio M., De Meyer M., Barr N.B. & Ruiz-Arce R. 2021. Looking at the big picture: worldwide population structure and range expansion of the cosmopolitan pest *Ceratitis capitata* (Diptera, Tephritidae). Biological Invasions doi.org/10.1007/s10530-021-02595-4

Deschepper P., Vanbergen S., Zhang Y., Li Z., Hassani I., Patel N.A., Rasolofoarivao H., Singh S., Wee S.L., De Meyer M.,. Virgilio M. & Delatte H. 2022. *Bactrocera dorsalis* in the Indian Ocean: a tale of two invasions. Evolutionary Applications DOI: 10.1111/eva.13507

Hebert, P., D., Cywinska, A., Ball, S., L., deWaard, J., R. (2003). Biological identification through DNA barcodes. Proceedings of the Royal Society of London *B.* 270: 313-321.

Pieterse, W., Manrakhan, A., Ramukhesa, H.R., Rosenberg, SM.M, Addison, P. The use of shape análisis to differentiate between the mandibles of four economically important tephritid species. Journal of Applied Entomology doi: 10.1111/jen.12368

Tietjen M., Arp P.A.,. Lohmeyer K.H. Development of a diagnostic single nucleotide polymorphism (SNP) panel for identifying geographic origins of Cochliomyia hominivorax, the New World screwworm, Veterinary Parasitology 315, 2023, 109884, https://doi.org/10.1016/j.vetpar.2023.109884.

Van Houdt, J.K.J., F.C. Breman, Virgillio M., De Meyer M. 2010. Covering full DNA barcodes from natural history collections of tephritid fruitflies (Tephritidae, Diptera) using mini barcodes. Molecular Ecology Resources 10: 459-465. (IF 1.25)

Virgilio M., Delatte H., Nzogela Y.B., Simiand C., Quilici S., De Meyer M. , Mwatawala M. 2015. Population structure and cryptic genetic variation in the mango fruit fly, *Ceratitis cosyra* (Diptera , Tephritidae). In: De Meyer, M., A. Clarke, T. Vera & J. Hendrichs (Eds). Resolution of Cryptic Species Complexes of Tephritid Pests to Enhance SIT Application and Facilitate International Trade. ZooKeys 540: 525-538. (IF 0.864)

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

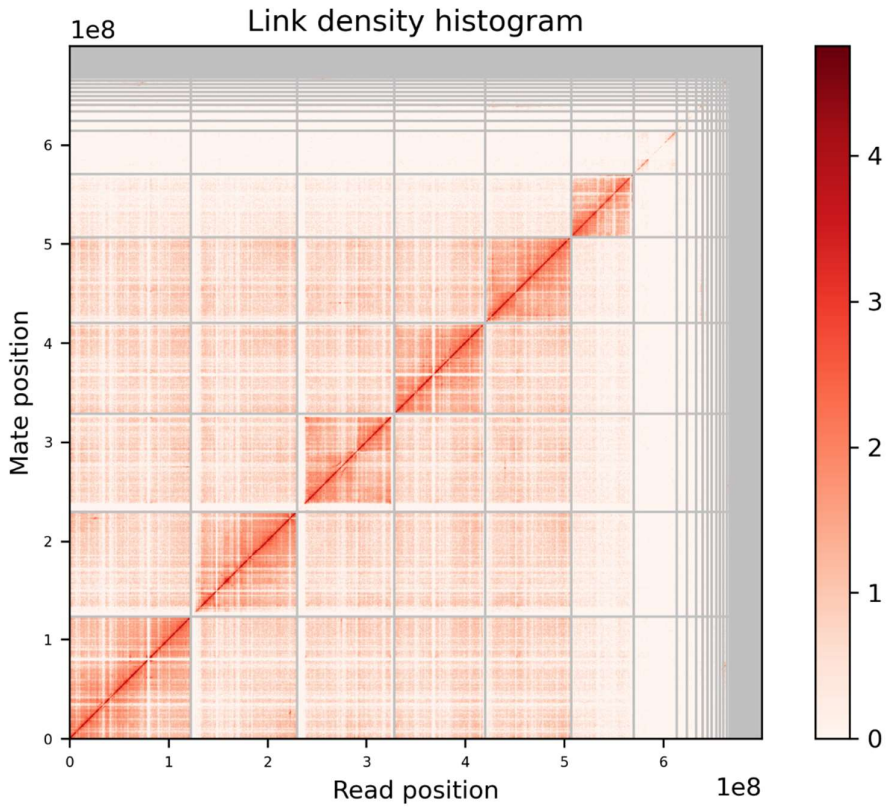
Zhang, Y., Liu S., De Meyer M., Liao Z., Zhao Y., Virgilio M., Feng S., Qin Y., Singh S., Wee S.L., Jiang F., Guo S., Li H., Deschepper P., Vanbergen S., Delatte H., van Sauers-Muller A., Syamsudin T.S., Kawi A.P., Kasina M., Badji K., Said F., Liu L.,. Zhao Z, Li Z.. 2022. Genomes of the cosmopolitan fruit pest *Bactrocera dorsalis* (Diptera: Tephritidae) reveal its global invasion history and thermal adaptation. Journal of Advanced Research doi.org/10.1016/jare.2022.12.012

6 List of Annexes

Annex I: *Ceratitis capitata* genome assembly statistics

Assembly	Total Length (bp)	N50	L50	N90	L90
Input Assembly	699,814,289	8,193,440	25	817,442	121
Dovetail HiRise Assembly	699,845,986	91,678,140	4	9,478,669	9

Annex II: *Ceratitis capitata* genome assembly contact map



Annex III: *Bactrocera zonata* genome assembly statistics

Assembly	Total Length (bp)	N50	L50	N90	L90
Hifiasm Assembly	767,067,092	84,542,130	4	404,732	88
Primary Filtered Assembly	524,894,629	99,542,525	3	22,789,729	6

Annex IV – Sanger sequencing derived DNA sequences metadata and sequences General dataset 1

https://fruitflies-ipm.eu/wp-content/uploads/2023/06/Annex_IV_DNA_sequences_General_dataset_1.xlsx

Annex V – WGS derived DNA sequences metadata Dataset *Ceratitis capitata* _ *Bactrocera dorsalis*

https://fruitflies-ipm.eu/wp-content/uploads/2023/06/Annex_V_DNA_sequences_Dataset_Ceratitis_capitata_Bactrocera_dorsalis.xls

Annex VI – WGS derived DNA sequences metadata Dataset *Ceratitis quilicii* _ *Ceratitis rosa*

https://fruitflies-ipm.eu/wp-content/uploads/2023/06/Annex_VI_DNA_sequences_Dataset_Ceratitis_quilicii-_Ceratitis-rosa.xls

Annex VII – WGS derived DNA sequences metadata General dataset 2

https://fruitflies-ipm.eu/wp-content/uploads/2023/06/Annex_VII_DNA_sequences_General_dataset_2.xls

Annex VIII – Sanger sequencing derived DNA sequences General dataset 3

https://fruitflies-ipm.eu/wp-content/uploads/2023/06/Annex_VIII_DNA_sequences_General_dataset_3.xls