GENOME NOTE

A new genome sequence resource for five invasive fruit flies of agricultural concern: *Ceratitis capitata*, *C. quilicii*, *C. rosa*, *Zeugodacus cucurbitae* and *Bactrocera zonata* (Diptera, Tephritidae)

[version 1; peer review: 2 approved]

Pablo Deschepper[1], Sam Vanbergen[1], Lore Esselens[1], John S. Terblanche[2], Minette Karsten[2], Maxi Snyman[2], Domingos Cugala[3,4], Laura Canhanga[3,4], Luis Bota[5], Maulid Mwatawala[6], Majubwa Ramadhani[6], Abdul Kudra[6], Jenipher Tairo[6], Jacqueline Bakengesa[7], Pia Addison[2], Aruna Manrakhan[2,8], Corentin Gledel[9], Hélène Delatte[9], Marc De Meyer[1], Massimiliano Virgilio[1]

[1]Biology department, invertebrates section, Royal Museum for Central Africa, Tervuren, Belgium
[2]Stellenbosch University Department of Conservation Ecology and Entomology, maiteland, South Africa
[3]University of Eduardo Mondlane College of Agriculture and Forestry, Maputo, Maputo City, Mozambique
[4]Centre of Excellence in Agri-Food Systems and Nutrition, University of Eduardo Mondlane, Maputo, Mozambique
[5]Provincial Directorate of Agriculture and Food Security, National Fruit Fly Laboratory, Chimoio, Manica, Mozambique
[6]Department of Crop Science and Horticulture, Sokoine University of Agriculture, Morogoro, Morogoro Region, Tanzania
[7]Department of Biology, The University of Dodoma, Dodoma, Dodoma Region, Tanzania
[8]Citrus Research International Pty Ltd, Nelspruit, Mpumalanga, South Africa
[9]CIRAD, UMR PVBMT, Saint-Pierre, La Réunion, 97410, France

v1 **First published:** 06 Dec 2024, **13**:1492
https://doi.org/10.12688/f1000research.157946.1
Latest published: 06 Dec 2024, **13**:1492
https://doi.org/10.12688/f1000research.157946.1

Abstract

Here, we present novel high quality genome assemblies for five invasive tephritid species of agricultural concern: *Ceratitis capitata*, *C. quilicii*, *C. rosa*, *Zeugodacus cucurbitae* and *Bactrocera zonata* (read depths between 65 and 78x). Three assemblies (*C. capitata*, *C. quilicii* and *Z. cucurbitae*) were scaffolded with chromosome conformation data and annotated using RNAseq reads. For some species this is the first reference genome available (*B. zonata*, *C. quilicii* and *C. rosa*), for others we have published improved annotated genomes (*C. capitata* and *Z. cucurbitae*). Together, the new references provide an important resource to advance research on genetic techniques for population control, develop rapid species identification methods, and explore eco-evolutionary studies.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 06 Dec 2024	✓ view	✓ view

1. **Craig Wilding**, Liverpool John Moores University, Liverpool, UK

2. **Lucio Navarro Escalante**, The University of Texas at Austin, Austin, USA

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
genome assembly, invasive species, fruit fly, tephritidae, pest

This article is included in the Agriculture, Food and Nutrition gateway.

Corresponding author: Pablo Deschepper (pablo.deschepper@africamuseum.be)

Author roles: Deschepper P: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Vanbergen S**: Data Curation, Methodology, Writing – Review & Editing; **Esselens L** : Data Curation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Terblanche JS**: Writing – Review & Editing; **Karsten M**: Resources, Writing – Review & Editing; **Snyman M**: Resources, Writing – Review & Editing; **Cugala D**: Resources, Writing – Review & Editing; **Canhanga L**: Resources, Writing – Review & Editing; **Bota L**: Resources, Writing – Review & Editing; **Mwatawala M**: Resources, Writing – Review & Editing; **Ramadhani M**: Resources, Writing – Review & Editing; **Kudra A**: Resources, Writing – Review & Editing; **Tairo J**: Resources, Writing – Review & Editing; **Bakengesa J**: Resources, Writing – Review & Editing; **Addison P**: Resources, Writing – Review & Editing; **Manrakhan A**: Resources, Writing – Review & Editing; **Gledel C**: Resources, Writing – Review & Editing; **Delatte H**: Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **De Meyer M**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Review & Editing; **Virgilio M**: Conceptualization, Data Curation, Funding Acquisition, Methodology, Project Administration, Resources, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was co-funded by the projects: DISPEST (Redefining DISpersal potential for adequate fruit fly PEST management), financed through the framework agreement 2019-2023 of the Royal Museum for Central Africa – and the Belgian Directorate-General for Development Cooperation and Humanitarian Aid (DGD) and EU-FFIPM (Fruit Flies In-silico prevention and management, Grant ID: 818184).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Deschepper P, Vanbergen S, Esselens L *et al.* **A new genome sequence resource for five invasive fruit flies of agricultural concern: *Ceratitis capitata*, *C. quilicii*, *C. rosa*, *Zeugodacus cucurbitae* and *Bactrocera zonata* (Diptera, Tephritidae) [version 1; peer review: 2 approved]** F1000Research 2024, **13**:1492 https://doi.org/10.12688/f1000research.157946.1

First published: 06 Dec 2024, **13**:1492 https://doi.org/10.12688/f1000research.157946.1

Introduction

A significant number of phytophagous insects within the dipteran family of the Tephritidae (the "true" fruit flies) are considered as serious pests for fruits and vegetables worldwide (White & Elson-Harris 1992). Globalization has led to a surge in intercontinental trade and movement, and has increased the number of incursions of harmful non-native fruit fly species (Bragard *et al.* 2020). Many countries have put costly and elaborate phytosanitary measures in place to prevent entry and establishment of harmful fruit fly species (Bragard *et al.* 2020; Papadopoulos *et al.* 2023a, 2023b). Making resources available that could provide researchers with a better tool for studying fruit fly pests is becoming increasingly important. Agricultural areas with a suitable climate for fruit fly pests are rapidly increasing around the globe (Sultana *et al.* 2020), changing patterns of distribution of fruit fly pests (Ni *et al.* 2011). This leads to more fruit fly incursions and first detections of new fruit fly species in several countries in recent years, e.g. *B. dorsalis* in France, Italy and Belgium; *B. zonata* in France (EPPO alert list, https://www.eppo.int/ACTIVITIES/plant_quarantine/alert_list).

Here, we present high quality reference genome assemblies for five tephritids (*Ceratitis capitata* (Wiedemann), *C. quilicii* (De Meyer, Mwatawala & Virgilio), *C. rosa* (Karsch), *Zeugodacus cucurbitae* (Coquillett), *Bactrocera zonata*) of agricultural importance (Figure 1a). For three (*C. quilicii, C. rosa*, *B. zonata*) of the five species, a genome assembly is completely lacking in public databases and could thus provide a major step forward in accumulating knowledge on those species. Genome assemblies are a valuable resource for both fundamental and applied research and can facilitate the development of new and sustainable pest management methods. The highly contiguous and complete genomes presented here will increase the chances of researchers to find specific genes of interest and investigate changes in genomic architecture. The new assemblies will enable researchers to tackle questions regarding climate adaptation, host and range expansion and niche shifts (Papanicolaou *et al.* 2016).

Results and discussion

PacBio CSS reads covered the genome between 65 and 78 times assuming a genome size of 0.5 Gb (Table 1) for the five fruit fly species shown in Figure 1a. A BUSCO search for genome completeness for all five novel assemblies against the Diptera database delivered a decent genome completeness between 94.6% (*B. zonata*) and 98.8% (*C. capitata*) using the duplicate purged PacBio assemblies (Figure 1d). Total assembly lengths ranged from 410 Mb (*Z. cucurbitae*) to 889 Mb (*C. quilicii*) with L50 values ranging from three (*B. zonata*) to 63 (*C. quilicii)* (Table 1). BlobToolKit results for identifying contaminants are shown in Figure. S1-S5 (Refer extended data) accessible at https://zenodo.org/records/14186560). Physical pairing between chromatin regions is shown in Figure 1b for *C. capitata*, *C. quilicii* and *Z. cucurbitae.*

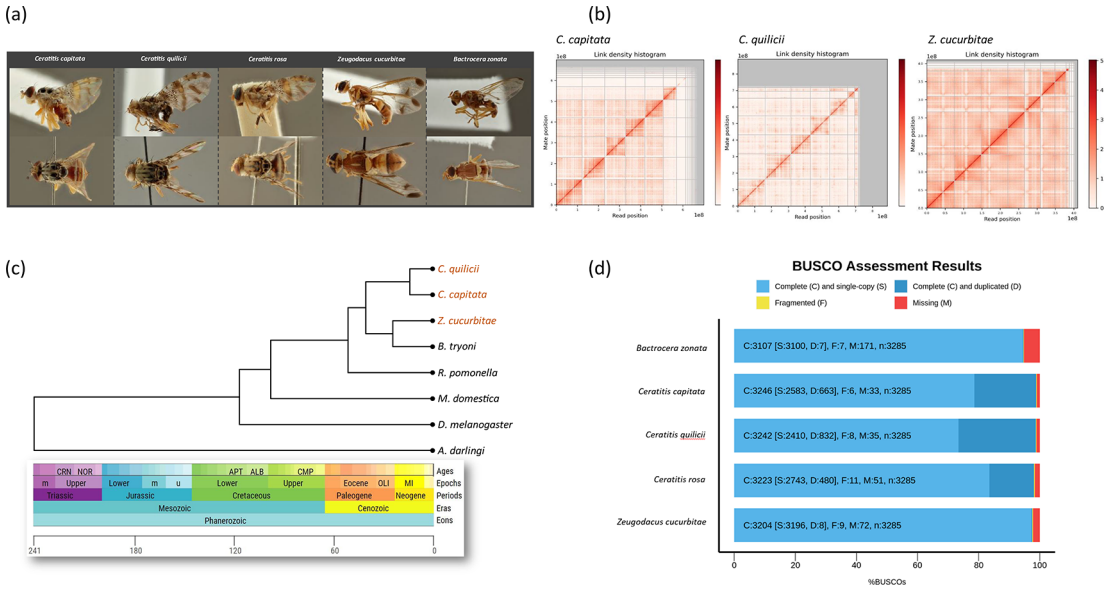

Figure 1. Linkage and BUSCO completeness across the genomes presented here and phylogenetic analysis. (a) Photographs of the five fruit fly pest species from a dorsal and lateral view © RMCA (Royal Museum for Central Africa. (b) Hi-C (Dovetail™ Omni-C™) contact map for three tephritid species showing which reads are in close proximity of each other, revealing the linear representation of the scaffolds/chromosomes within the genome. (c) Phylogenetic tree of the three tephritid fruit flies with annotation and five other diptera species. (d) BUSCO completeness results for each of the assembled tephritid genomes.

Table 1. Comparison of assembly statistics.

	PacBio CSS read data			genome length and contiguity				
Species	**Number of Reads**	**Bp (Gb)**	**Coverage**	**Total Length (bp)**	**N50**	**L50**	**N90**	**L90**
Ceratitis capitata	2,563,883	38.8	78x	699,814,289	8,193,440	25	817,442	121
Ceratitis quilicii	2,498,505	38.4	77x	889,108,370	4,088,374	63	278,990	378
Ceratitis rosa	2,451,133	36.2	72x	650,940,389	17,660,867	12	966,435	75
Zeugodacus cucurbitae	2,332,131	32.4	65x	410,169,932	14,989,679	8	4,412,789	28
Bactrocera zonata	2,522,876	32.4	65x	524,894,629	99,542,525	3	22,789,729	6

The annotated genomes comprise 32,449; 38,590 and 31,422 genes in total for *C. capitata*, *C. quilicii* and *Z. cucurbitae* respectively with a total coding region length (bp) of 39,037,294; 46,768,995 and 41,286,253. The average gene length (bp) is 1,203.04; 1,211.95 and 1,313.93 for *C. capitata*, *C. quilicii* and *Z. cucurbitae* respectively. The most recent *C. capitata* assembly available on NCBI (GCA_905071925.1, published in November 2020) contains 14,054 genes and thus, this novel assembly improves the degree of annotation of the *C. capitata* genome significantly. The same can be observed in *Z. cucurbitae*, where the most recent NCBI reference assembly (GCF_028554725.1) only comprises 17,225 genes. In *Ceratitis* sp. however, a substantial proportion of BUSCO's are duplicated, which suggest the presence of redundant sequences resulting from partial misassemblies. Our recommendation is therefore to be cautious when comparing *Ceratitis* sp. assemblies with other assemblies.

A total of 19,480 gene orthogroups could be found using OrthoFinder and a total of 32,051; 37,950 and 31,009 genes could be attributed to an orthogroup for *C. capitata*, *C. quilicii* and *Z. cucurbitae* respectively. Using these orthogroups as evidence we estimated that the Tephritidae-Drosophilidae split took place around 120 MYA (Figure 1c), which is in line with the estimations of Russo *et al.* (2013) who constructed a drosophilid time tree with two tephritid species as outgroup (*C. capitata* and *B. oleae*) and estimated the split at around 110 MYA.

We believe that our contribution will substantially impact tephritid genome research and provides new opportunities for comparative genomics with a focus on characterizing genes related to invasiveness.

Methods

De novo genome assembly

An inbred lab colony of each of the following tephritid species was established in an artificial setting and larvae were collected for subsequent sequencing: *Ceratitis capitata*, *C. quilicii*, *C. rosa*, *Zeugodacus cucurbitae* and *Bactrocera zonata.* Inbred specimens of *C. quilicii*, *C. capitata* and *C. rosa* were produced at Citrus Research International in Mbombela and were originally sourced from wild flies collected in Ermelo (-26.516021, 29.996168), Burgershall (-25.112083, 31.087778) and Mbombela (-25.452258, 30.970778), Mpumalanga Province, South Africa respectively in 2020 (*C. rosa*) and 2021 (*C. capitata* and *C. quilicii*). Species identity was confirmed by Marc De Meyer (*C. quilicii*) and Aruna Manrakhan (*C. capitata* and *C. rosa*). Inbred lines for *Z. cucurbitae* and *B. zonata* were already present at the facilities of CIRAD, Réunion for more than 150 generations and could thus be used for our purposes. Pupae of all species supplied for sequencing originate from a parent x F1 backcross to increase homozygosity. The sequencing and assembly process can be described by three consecutive steps: generation of PacBio CCS reads and primary assembly with Hifiasm, generation of Hi-C (specifically, Dovetail™ Omni-C™ reads) coupled with secondary assembly using HiRise and lastly, generation of an RNAseq library for ab initio genome annotation. Only the assemblies of *C. capitata*, *C. quilicii* and *Z. cucurbitae* comprised the HiRise scaffolding and annotation steps.

De novo PacBio assembly and filtering

A *de novo* assembly was constructed using ±38.8 Gb of PacBio CCS reads resulting in a coverage of around 70x of the tephritid genome (Table 1). The obtained PacBio reads were used as input to Hifiasm v0.15.4-r347 (Cheng *et al.* 2021) with default parameters. Blast results of the Hifiasm output assembly against the nucleotide BLAST database (https://blast.ncbi.nlm.nih.gov/) were used as input for blobtools v1.1.1 (Laetsch and Blaxter 2017) and scaffolds identified as possible contamination were removed from the assembly. Finally, purge_dups3 v1.2.5 (Guan *et al.* 2020) was used to purge haplotigs and contig overlaps. The final assembly was checked for its completeness using BUSCO using the diptera_odb10 dataset (Manni *et al.* 2021).

Chromosome conformation capture and HiRise scaffolding

To construct a Dovetail™ Omni-C™ library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DNAse I, chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The library was sequenced on an Illumina HiSeqX platform to produce approximately 30x sequence coverage.

The input de novo assembly and Dovetail™ Omni-C™ library reads (MQ > 50) were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Putnam *et al.* 2016). Dovetail™ Omni-C™ library sequences were aligned to the draft input assembly using bwa (https://github.com/lh3/bwa). The separations of Dovetail™ Omni-C™ read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold.

Ab initio genome annotation

Firstly, repeat families in the three tephritid genome assemblies (*C. capitata*, *C. quilicii* and *Z. cucurbitae*) were identified *de novo* and classified using the software package RepeatModeler2 (Flynn *et al.* 2020, the original version of RepeatModeler is free and available at https://github.com/Dfam-consortium/RepeatModeler/blob/master/RepeatModeler). The custom repeat library obtained from RepeatModeler2 was used to discover, identify and mask the repeats in the assembly using RepeatMasker (Version 4.1.0, available at https://github.com/rmhubley/RepeatMasker). Secondly, coding sequences from *Bactrocera dorsalis*, *Ceratitis capitata* and *Drosophila melanogaster* available on GenBank were used to train the *ab initio* model in AUGUSTUS (version 2.5.5) by performing six rounds of optimization. Likewise, the same coding sequences were used to train an independent *ab initio* gene model using SNAP (Korf 2004). Furthermore, RNAseq reads were mapped onto the genome using the STAR aligner software (Dobin *et al.* 2013). MAKER (Campbell *et al.* 2014), SNAP and AUGUSTUS (with intron-exon boundary hints provided from RNAseq) were then used to predict genes in the repeat-masked reference genome. To help guide the prediction process, SwissProt peptide sequences from the UniProt database (https://www.uniprot.org/) were downloaded and used in conjunction with the protein sequences from the aforementioned species to generate peptide evidence in the Maker pipeline (Campbell *et al.* 2014). Only genes that were predicted by both SNAP and AUGUSTUS were retained in the final gene sets. To help assess the quality of the gene prediction, AED scores were generated for each of the predicted genes as part of the MAKER pipeline. Genes were further characterised for their putative function by performing a BLAST (Ye *et al.* 2006) search of the peptide sequences against the UniProt database. tRNA were predicted using the software tRNAscan-SE (Lowe & Chan 2016, available at: https://lowelab.ucsc.edu/tRNAscan-SE/).

Phylogenetic tree reconstruction

We inferred orthogroups using OrthoFinder v2.5.5. (Emms & Kelly 2019) for the three fruit fly species with an annotated genome assembly in this study (*C. capitata*, *C. quilicii* and *Z. cucurbitae*). In addition, we downloaded protein sequence data for *Drosophila melanogaster* Meigen (GCA_000001215.4), *Anopheles darlingi* Root (GCA_000211455.3), *Musca domestica* Linnaeus (GCF_030504385.1), *Rhagoletis pomonella* (Walsh) (GCF_013731165.1) and *Bactrocera tryoni* (Froggatt) (GCF_016617805.1). Sequences were aligned using Diamond and gene trees were inferred using fasttree. The STAG algorithm combined with the STRIDE rooting methods, implemented in OrthoFinder, was then used to infer a species tree with realistic branch lengths from the full set of gene trees (Emms & Kelly 2017). A time-calibrated tree was constructed by transforming the species tree rendered by Orthofinder into a ultrametric tree and calibrating it based on the split between *A. darlingi* and the rest of the taxa (240.8 MYA) as inferred from TIMETREE5 (timetree.org).

Author contributions

PD, SV, LE, MDM, MV (RMCA, BE) – Conceptualization, funding acquisition, original draft preparation and data submission.

PA, JT, MK (SU, ZA) – Conceptualization, development of the inbred lines, provision of field samples, review and editing.

AM (CRI, ZA) - Conceptualization, development of the inbred lines, provision of field samples, review and editing.

DC, LC (EMU, MZ), LB (National FF lab, MZ) - Conceptualization, provision of field samples, review and editing.

MM, RM, AK, JT (SUA, TZ), JB (UDOM, TZ) - Conceptualization, review and editing.

HD (CIRAD – La Réunion, FR) - Conceptualization, funding acquisition, development of the inbred lines, provision of field samples, review and editing.

Data availability statement

All five genome assemblies have been deposited on the NCBI data repository.

National Centre for Biotechnology Information. BioProject: Five new genome assemblies of Tephritid pest species. Accession number: PRJDB18489; https://www.ncbi.nlm.nih.gov/bioproject/PRJDB18489/.

GenBank assemblies for the five tephritid species can be consulted using following identifiers:

National Centre for Biotechnology Information. GCA_043005645.1: *Bactrocera zonata*; https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_043005645.1/.

National Centre for Biotechnology Information. GCA_043005455.1: *Ceratitis capitata*; https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_043005455.1/.

National Centre for Biotechnology Information. GCA_043005495.1: *Ceratitis quilicii*; https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_043005495.1/.

National Centre for Biotechnology Information. GCA_043005725.1: *Ceratitis rosa*; https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_043005725.1/.

National Centre for Biotechnology Information. GCA_043005565.1: *Zeugodacus cucurbitae*; https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_043005565.1/

Annotation files for *C. capitata*, *C. quilicii* and *Z. cucurbitae* are stored at zenodo. https://zenodo.org/records/13928607, Genome sequence and .gff annotation of three pest fruit flies (Tephritidae).

zenodo. Genome sequence and .gff annotation of three pest fruit flies (Tephritidae), DOI: https://doi.org/10.5281/zenodo.13928607 (Royal Museum for Central Africa 2024).

The project contains the following underlying data:

- Zcucurbitae_DDBJ_100Ngaps.gff
- Zcucurbitae_DDBJ_100Ngaps.fa
- Cquilicii_DDBJ_100Ngaps.gff
- Cquilicii_DDBJ_100Ngaps.fa
- Ccapitata_DDBJ_100Ngaps.gff
- Ccapitata_DDBJ_100Ngaps.fa

Extended data

Zenodo: A new genome sequence resource for five invasive fruit flies of agricultural concern: Ceratitis capitata, C. quilicii, C. rosa, Zeugodacus cucurbitae and Bactrocera zonata (Diptera, Tephritidae), DOI: https://doi.org/10.5281/zenodo.14186560 (Deschepper 2024).

The project contains the following extended data:

- FigS1.png
- FigS2.png
- FigS3.png
- FigS4.png
- FigS5.png

References

Bragard CEFSA Panel on Plant Health (PLH): **Pest categorisation of non-EU Tephritidae.** *EFSA J.* 2020.
Publisher Full Text

Campbell MS, Holt C, Moore B, *et al.*: **Genome Annotation and Curation Using MAKER and MAKER-P.** *Curr. Protoc. Bioinformatics.* 2014; **48**: 4.11.1–4.11.39.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm.** *Nat. Methods.* 2021; **18**: 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Deschepper P, *et al.*: A new genome sequence resource for five invasive fruit flies of agricultural concern: Ceratitis capitata, C. quilicii, C. rosa, Zeugodacus cucurbitae, and Bactrocera zonata (Diptera, Tephritidae). [Dataset]. *Zenodo.* 2024.
Publisher Full Text

Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: Ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**: 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

Emms DM, Kelly S: **STRIDE: Species Tree Root Inference from Gene Duplication Events.** *Mol. Biol. Evol.* 2017; **34**: 3267–3278.
PubMed Abstract | Publisher Full Text | Free Full Text

Emms DM, Kelly S: **OrthoFinder: phylogenetic orthology inference for comparative genomics.** *Genome Biol.* 2019; **20**: 238.
PubMed Abstract | Publisher Full Text | Free Full Text

Flynn JM, Hubley R, Goubert C, *et al.*: **RepeatModeler2 for automated genomic discovery of transposable element families.** *Proc. Natl. Acad. Sci.* 2020; **117**: 9451–9457.
PubMed Abstract | Publisher Full Text | Free Full Text

Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics.* 2004; **5**.
PubMed Abstract | Publisher Full Text | Free Full Text

Laetsch DR, Blaxter ML: **BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations].** *F1000Res.* 2017; **6**.
Publisher Full Text

Lowe TM, Chan PP: **tRNAscan-SE On-line: Integrating search and context for analysis of transfer RNA genes.** *Nucleic Acids Res.* 2016; **44**: W54–W57.
PubMed Abstract | Publisher Full Text | Free Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**: 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO: Assessing genomic data quality and beyond.** *Current Protocols.* 2021; **1**: e323.
PubMed Abstract | Publisher Full Text

Ni WL, Li ZH, Chen HJ, *et al.*: **Including climate change in pest risk assessment: the peach fruit fly, *Bactrocera zonata* (Diptera: Tephritidae).** *Bull. Entomol. Res.* 2011; **102**: 173–183.
Publisher Full Text

Papadopoulos NT, Camilleri M, Graziosi I: **Surveillance of non-EU Tephritidae in the EU: a guide for using EFSA pest survey cards.** *EFSA Supporting Publications.* 2023a.
Publisher Full Text

Papadopoulos NT, De Meyer M, Terblanche JS, *et al.*: **Fruit flies: challenges and opportunities to stem the tide of global invasions.** *Annu. Rev. Entomol.* 2023b; **69**: 355–373.
PubMed Abstract | Publisher Full Text

Papanicolaou A, *et al.*: **The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species.** *Genome Biol.* 2016; **17**: 192.
PubMed Abstract | Publisher Full Text | Free Full Text

Putnam NH, O'Connell BL, Stites JC, *et al.*: **Chromosome-scale shotgun assembly using an in vitro method for long-range linkage.** *Genome Res.* 2016; **26**: 342–350.
PubMed Abstract | Publisher Full Text | Free Full Text

Royal Museum for Central Africa, Vanbergen S, Esselens L, *et al.*: **Genome sequence and .gff annotation of three pest fruit flies (Tephritidae).** *Zenodo.* 2024.
Publisher Full Text

Russo CAM, Mello B, Frazão A, *et al.*: **Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae).** *Zool. J. Linn. Soc.* 1 December 2013; **169**(4): 765–775.
Publisher Full Text

Sultana S, Baumgartner JB, Dominiak BC, *et al.*: **Impacts of climate change on high priority fruit fly species in Australia.** *PLoS One.* 2020; **15**: e0213820.
PubMed Abstract | Publisher Full Text | Free Full Text

White IM, Elson-Harris MM: *Fruit flies of economic significance: their identification and bionomics.* Wallingford, UK: CAB International; 1992. 9780851987903.

Ye J, McGinnis S, Madden TL: **BLAST: Improvements for better sequence analysis.** *Nucleic Acids Res.* 2006; **34**: W6–W9.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 28 February 2025

https://doi.org/10.5256/f1000research.173473.r365152

✔ **Lucio Navarro Escalante**
The University of Texas at Austin, Austin, USA

Report:
The manuscript reports the sequencing and assembly for the genomes of five tephritid fly species of agricultural importance (Ceratitis capitata, .C. quilicii, C. rosa, Bactrocera zonata and Zeugodacus cucurbitae). PacBio sequences were used to assembly contig level high quality draft genomes for the five species with completeness ranged between 95%-99%. Three of them (C. capitata, C. quilicii and Z. cucurbitae) were further scaffolded using Hi-C chromosome conformation data. Additionally, for three of these flies (C. quilicii, C. rosa and B. zonata) this data represents the first reported assembled genome. RNAseq data was also produced and used to predict gene contents in scaffolded genomes (C. capitata, C. quilicii and Z. cucurbitae) by performing a combination of ab-initio and evidence-based methods. Finally, a phylogenetic tree was built with the annotated genomes using orthologous protein sequences as evidence.

The significance for the sequenced species is clearly defined in the manuscript. The described methods offer enough details and are appropriate for the goals, however part of the datasets does not seem to be fully available. Specifically, I have not been able to find the original raw genome sequences (PacBio and Hi-C Illumina libraries), nor the manuscript indicating where this data is stored. Authors should provide SRA accession numbers for this raw data.

Minor comments:

- Are there any references about the number of chromosomes in these particular fly species? If so, that should be reported and could be used to discuss any correlation with the number of chromatin regions detected in the Hi-C interaction matrix. In fact, the authors do not provide any detail or discussion about these observations in the Hi-C analysis.
- The total percentage of repetitive sequences for each assembled genome should be also reported.
- Authors should simplify users' access to data by providing FASTA files containing the predicted protein and CDS sequences for the annotated genomes in this study.
- Include more details about the methods used for DNA extraction, at least indicate what type of method was used.

- For the ab initio genome annotation methods, specify the meaning of 'AED'.
- In supplementary figures S1-S5, each image should clearly indicate to what fly species they correspond. Additionally, images with better resolution should be provided.
- Check for missing or spare parenthesis across the manuscript.

Are the rationale for sequencing the genome and the species significance clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Insect genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 26 February 2025

https://doi.org/10.5256/f1000research.173473.r365150

Craig Wilding
Liverpool John Moores University, Liverpool, England, UK

Deschepper et al. describe the production of high-quality chromosomal assemblies for each of five invasive fruit flies. This is an excellent body of work comprising not just PacBio sequencing, but Hi-C data.
The methods used are both clearly described and appropriate, and produced good to excellent (dependent upon species) results in terms of genome contiguity and annotation completeness. There are differences in quality of assemblies, as adjudged by BUSCO data where two species assemblies have substantial apparent duplicated genes (which the authors discuss) but, nevertheless, these are good quality genomes which will be of value to those studying these species from a control perspective, as well as in understanding their relationships.

Comments:

- The statement in the results that there is "a decent genome completeness" is vague and non-quantitative. I would avoid vague terms like decent.
- The use of BlobTools is mentioned in the methods but there is then no comment on the output from this in the results and no blobplot provided. What did it show? (Perhaps some of these flies have, for instance, *Wolbachia*?)
- Why are Hi-C maps provided for just 3 of the 5 genomes?
- I am unclear on the point of Fig 1c given its limited number of species. What is the message from this?
- Aside from the differences in gene content for the new versus existing *C. capitata* and *Z. cucurbitae* assemblies, how different are these in other parameters e.g. gene identity of coding genes? Will there be any future attempt to utilise these assemblies in combination to examine variability such as SNPs and CNVs?

Are the rationale for sequencing the genome and the species significance clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary genetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.