Unveiling biogeographic patterns in the worldwide distributed Ceratitis capitata (medfly) using population genomics and microbiome composition

María Belén Arias[1], Katherine Hartle-Mougiou[1], Sergi Taboada[1], Alfried Vogler[2], Ana Riesgo[1], and SAMIA ELFEKIH[3]

[1]Natural History Museum
[2]Natural History Museum
[3]CSIRO

June 9, 2022

Abstract

Invasive species are among the most important, growing threats to food security and agricultural systems. The Mediterranean fruit fly Ceratitis capitata is one of the most damaging representatives of a group of rapidly expanding species in the family Tephritidae due to their wide host range and high invasiveness. Here, we used restriction site-associated DNA sequencing (RADseq) to investigate population genomic structure and phylogeographic history of medflies collected from six sampling sites, including Africa (South Africa), the Mediterranean (Spain, Greece), Latin America (Guatemala, Brazil) and Australia. A total of 1,907 single nucleotide polymorphisms (SNPs) showed two genetic clusters separating native and introduced ranges, consistent with previous findings. In the introduced range, all individuals were assigned to one genetic cluster except for those in Brazil, which showed introgression of a genetic cluster that also appeared exclusively in South Africa and could not be previously identified using microsatellite markers. Moreover, the microbiome variations in medfly populations from selected sampling sites was assessed by amplicon sequencing of the 16S ribosomal RNA (V4 region). No strong patterns of microbiome variation were detected across geographic regions or host plants, except for the differentiation of the Brazilian specimens which showed increased diversity and unique composition of its microbiome compared to other sampling sites. The unique SNP patterns in the Brazilian specimens could point to a direct migration route from Africa with subsequent adaptation of the microbiota to the specific conditions present in Brazil. These findings significantly improve our understanding of the evolutionary history of global medfly invasions and adaptation to newly colonised environments.

Unveiling biogeographic patterns in the worldwide distributed *Ceratitis capitata* (medfly) using populations genomics and microbiome composition

Running title: Population genomics and microbiome of Medfly

María Belén Arias[1,2*], Katherine Hartle-Mougiou[1,3], Sergi Taboada[1,4,5], Alfried P. Vogler[1,3], Ana Riesgo[1,6], Samia Elfekih[7]

[1]Department of Life Sciences, Natural History Museum, London, UK.

[2]School of Life Sciences, University of Essex, Colchester, UK.

[3]Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, UK.

[4]Departamento de Biodiversidad, Ecología y Evolución, Universidad Complutense de Madrid, Madrid, Spain.

[5]Departamento de Ciencias de la Vida, Universidad de Alcalá de Henares, Madrid, Spain.

[6]Department of Biodiversity and Evolutionary Biology, Museum Nacional de Ciencias Naturales, Madrid, Spain.

[7]CSIRO Health & Biosecurity, Black Mountain, Canberra, Australia.

*Corresponding Author: María Belén Arias; e-mail: mbelen.arias@gmail.com, b.arias@nhm.ac.uk

Keywords: Medfly, phylogeography, invasive species, bacterial community, invasion history

ABSTRACT

Invasive species are among the most important, growing threats to food security and agricultural systems. The Mediterranean fruit fly *Ceratitis capitata* is one of the most damaging representatives of a group of rapidly expanding species in the family Tephritidae due to their wide host range and high invasiveness. Here, we used restriction site-associated DNA sequencing (RADseq) to investigate population genomic structure and phylogeographic history of medflies collected from six sampling sites, including Africa (South Africa), the Mediterranean (Spain, Greece), Latin America (Guatemala, Brazil) and Australia. A total of 1,907 single nucleotide polymorphisms (SNPs) showed two genetic clusters separating native and introduced ranges, consistent with previous findings. In the introduced range, all individuals were assigned to one genetic cluster except for those in Brazil, which showed introgression of a genetic cluster that also appeared exclusively in South Africa and could not be previously identified using microsatellite markers. Moreover, the microbiome variations in medfly populations from selected sampling sites was assessed by amplicon sequencing of the 16S ribosomal RNA (V4 region). No strong patterns of microbiome variation were detected across geographic regions or host plants, except for the differentiation of the Brazilian specimens which showed increased diversity and unique composition of its microbiome compared to other sampling sites. The unique SNP patterns in the Brazilian specimens could point to a direct migration route from Africa with subsequent adaptation of the microbiota to the specific conditions present in Brazil. These findings significantly improve our understanding of the evolutionary history of global medfly invasions and adaptation to newly colonised environments.

INTRODUCTION

Drastic environmental changes affecting natural and anthropogenic ecosystems involve massive shifts in species' geographical distributions and cause worldwide invasions facilitated by the ability of certain species to adapt to newly colonised environments (Heino, Virkkala, & Toivonen, 2009). Invasive species have been associated with an estimated 1.3 trillion USD in economic losses and are responsible for diminishing local species richness (Diagne et al., 2021). Therefore, frequent interventions are required to control the damaging effects of invasive species from new colonisation events. Large-scale range expansions and intervention measures for pest control (e.g., pesticides, Sterile Insect Technique (SIT)) have profound but poorly characterised effects on population structure, invasion pathways, and adaptation to local environments. Among agricultural insect pests, the family Tephritidae (fruit flies) harbours several rapidly expanding species of great concern, including the Mediterranean fruit fly *Ceratitis capitata* (Wiedemann, 1824) (commonly known as medfly), a cosmopolitan, polyphagous species with over 250 different hosts of fruit and vegetables (White & Elson-Harris, 1992). In the past two centuries, *C. capitata* has expanded from a presumed origin in Sub-Saharan Africa to the Mediterranean basin and later spread to tropical regions in all continents (Gasperi, Guglielmino, & Milani, 1991; Malacrida et al., 2007; Malacrida et al., 1992). Its range can shift with climate change and commercial trade (Gutierrez & Ponti, 2011; Hill et al., 2016). It is considered one of the most successful invaders worldwide and a significant economic pest to the fruit market, estimated to cost more than 2 billion USD annually in losses (Sciarretta et al., 2018).

Differences in local climate and exposure to various host plants and natural enemies, reinforced by large geographic distances, may have constrained gene flow, and resulted in differentiation of medfly populations. Current knowledge of medfly population genetics supports low levels of intercontinental connectivity between the native and colonised ranges (Elfékih, Makni, & Haymer, 2010; Gasperi et al., 2002; Karsten, Jansen van Vuuren, Addison, Terblanche, & Leung, 2015) but the rate of dispersal among introduced popu-

lations remains unclear, especially in Central and South America (Arias, Elfekih, & Vogler, 2018; Deschepper et al., 2021; Ruiz-Arce et al., 2020), despite its importance for implementing pest control strategies. One of the potential factors influencing the medfly's propensity for dispersal and adaptation to new environments is the interaction between the microbiome and its host. It has been previously reported that specific microbiome variants exist synergistically with insect hosts and might rapidly spread across populations (Aharon et al., 2013). Evidence from *Drosophila melanogaster* Meigen, 1830 suggests that shifts in microbiome composition can alter population dynamics and, consequently, might be considered a relevant driver of ecological and evolutionary processes at the population level (Rudman et al., 2019). The medfly receives its microbiota during oviposition via maternal inheritance (Behar, Jurkevitch, & Yuval, 2008). The composition during larval development is restructured, with microbial community shifts occurring at different stages (Aharon et al., 2013; Malacrino, Campolo, Medina, & Palmeri, 2018). Such knowledge dramatically enhances our understanding of the influence of the microbial profile on insect development. Exploring potential shifts of microbiota in introduced ranges could reveal further information regarding the medfly's colonisation patterns.

The emergence of next-generation sequencing, including reduced representation genome sequencing techniques such as Restriction Associated DNA-tags sequencing (RADseq), has made it possible to study the population diversity of non-model organisms at the genomic level (Andrews et al., 2016, Davey & Blaxter, 2010). For example, RADseq has been used to address biological questions on demography and dispersal of invasive insect pests (Elfékih et al. 2018; McCormack, Hird, Zellmer, Carstens, & Brumfield, 2013; Schmidt et al. 2020), patterns of gene flow, phylogeography, and species delimitation (Eaton et al., 2013; Elfékih et al. 2021; Emerson et al., 2010; Dong et al. 2021), microbial association and local adaptation (Orantes et al. 2018; van Oppen et al. 2018).

Using RADseq data, we present the genetic relationships, gene flow patterns, possible pathways of invasions, and cross-continent colonisation routes of medfly populations collected from six regions distributed worldwide. In addition, we examine genome-wide SNP variation and explore the bacterial microbiome profile associated with each sampling site and its possible correlation with the medfly presumed dispersal routes.

2. MATERIALS AND METHODS

2.1 Sample collection

A total of 92 adults of *C. capitata* were collected in infected orchards from six sites across all biogeographic regions where the species occurs (Afrotropical, Palaearctic, Neotropical, and Australasian) between 2005 and 2012 (Figure 1; Table 1). All flies were preserved in 80% ethanol and kept at -20^{o}C until used for DNA extraction.

2.2 DNA extraction, RADseq library preparation and sequencing

Genomic DNA was extracted from each whole specimen using a DNeasy Blood & Tissue Spin Column Kit (Qiagen, Venlo, Netherlands), including the RNase treatment suggested by the manufacturer. The extraction was performed in whole specimens previously washed with sterilised water three times to remove any external contaminant in a room never exposed before to medflies. The DNA was quantified using Qubit 2.0 dsDNA HS Kit (Thermo Fisher Scientific, Waltham, US-MA) and sent to GenePool laboratories (Edinburgh, Scotland) for RAD-tag library preparation and sequencing. Briefly, DNA samples were digested using *SbfI* high-fidelity restriction enzyme (NEB Inc., Ipswich, US-MA). P1 adapters, each with a unique eight bp molecular identifying sequence, were ligated using T4 DNA Ligase (NEB) to allow 16 individuals per lane multiplexing. Fragments were pooled and size selected (300-700 bp) prior to purification with Qiagen columns. Fragment ends were repaired using the Quick Blunting Kit (NEB), and P2 adapters were ligated using T4 DNA Ligase. Minelute columns (Qiagen) were used to purify DNA. PCR enrichment of the libraries was performed using Phusion Flash High-Fidelity PCR Master Mix (NEB) and size selected using gel electrophoresis. The sequencing of RAD tags was done on Illumina GAIIx (Illumina Inc., San Diego, US-CA) following standard protocols, and data is available via CSIRO Data Collection (Elfékih, Arias, & Vogler, 2020)

2.3 RAD data processing, SNP calling and filtering

Reads from the six libraries were demultiplexed using *process_radtags* (Stacks v1.13) (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011). Quality assessment and filtering were performed using FASTQC (Andrews, 2012) and low-quality terminal sites were removed using *Fastx_trimmer* to produce reads of exactly 87 bp. A total number of 55,043,269 reads were aligned using the *bwa mem* algorithms (Li & Durbin, 2009) to the updated reference genome of *C. capitata* (NW_019376232.1) (Papanicolaou et al., 2016). Finally, the bam files were sorted and indexed using Samtools (Li et al., 2009).

The STACKS program v2.2 (Rochette, Rivera-Colon, & Catchen, 2019) was used for SNP detection and genotype calling under the Marukilow model using default parameters, except for the minimum mapping quality (*–min-mapq* 20), which resulted in 35,345 loci from 42,8333,799 remaining reads for a mean coverage of 78.2x. The *populations* program of the STACKS v2.2 package was applied to retain only loci with SNPs present in at least 80% of the individuals in a specific population ($r = 0.8$) and the option *order_export* was activated because the reads were mapped to a reference genome. The dataset was filtered by selecting the most informative SNP per locus based on the options *write-single-snps* , to reduce the probability of linkage disequilibrium among loci in the further analyses.

Deviation from Hardy-Weinberg Equilibrium (HWE) for each locus was examined using *populations,* and those loci that deviated from HWE at a significance level of 0.005 were removed. Furthermore, we identified outlier SNPs using ARLEQUIN v.3.5.2.1 (Excoffier & Lischer, 2010) and BAYESCAN v2.1 (Foll & Gaggiotti, 2008). For ARLEQUIN, we used the "no hierarchical island model", 100,000 simulations and 1,000 demes. The *p-* values from this analysis were adjusted for each locus using the *FDR* method to calculate the *q-*values in R (Team, 2020). As for BAYESCAN we ran the analysis with 10,000 output iterations and 100 prior odds. The final putative neutral SNPs dataset consisted of 1,907 SNPs in 92 individuals (Table 1).

2.4 Demographic events and population assignment

Genetic diversity indices for each population, including observed and expected heterozygosity, private alleles, nucleotide diversity and inbreeding coefficient, were calculated using *populations* in STACKS v2.2 (Rochette et al., 2019). In addition, Tajima's D was calculated using VCFtools (Danecek et al., 2011) to assess recent historical demographic events. Finally, population differentiation was analysed using pairwise F_{st}, and the statistical significance was calculated under 20,000 permutations in ARLEQUIN v.3.5.2.1.

We assessed the population structure and differentiation using three clustering approaches. Firstly, we used the discriminant analysis of principal components (DAPC) (Jombart, Devillard, & Balloux, 2010) as implemented in the *adegenet* package (Jombart, 2008) in R (Team, 2020). The cross-validation *xvaldapc* function was used to determine the optimal number of principal components (PCs) using a training set of 0.6 with 1000 replicates (*n.rep = 1000*). The *find.cluster* function was used to infer the most likely number of clusters with the Akaike Information Criterion (AIC) and K-means of 100 (Jombart et al., 2010). The optimal number of genetic clusters was identified as the lowest AIC value. Secondly, we ran STRUCTURE v2.3.4. (Pritchard, Stephens, & Donnelly, 2000) with 200,000 MCMC iterations following a burn-in of 100,000 iterations, using the admixture ancestry model and no sampling location as a prior, setting a putative K from 1 to 6 in a total of 15 independent iterations in each run (Falush, Stephens, & Pritchard, 2003). The web version of STRUCTURE HARVESTER (Earl & vonHoldt, 2011) (http://taylor0.biology.ucla.edu/structureHarvester/) was used to infer the most likely K value based on the Evanno method (Evanno, Regnaut, & Goudet, 2005) and the highest posterior mean log-likelihood (mean LnP(K)). CLUMPP (Jakobsson & Rosenberg, 2007) and DISTRUCT v1.1 (Rosenberg, 2003) were used to summarise the similarity between replicates and visualise the STRUCTURE plots, respectively. Finally, the *fineRADstructure* package (Malinsky, Trucchi, Lawson, & Falush, 2018) was run to understand the recent shared ancestry at high resolution derived from haplotype linkage information among all the samples. The database used contained 13,149 SNPs obtained after running *populations* on the set of 1,907 neutral SNPs deselecting the *-write-single-snps* option; this allowed for the inclusion of linked SNPs in the different RAD-tags. This dataset was reordered first according to linkage disequilibrium using the script *sampleLD.R, fineRADstructure* was run using the default settings. The results were graphically interpreted using the *Finestructure* package (Lawson, Hellenthal, Myers, & Falush, 2012) and the *fineRADstructurePlot.R* script (Malinsky et al., 2018).

2.5 Demographic history scenarios testing and species tree

To investigate which specific demographic history hypotheses were observed to be most compatible with genetic structure, we applied a Bayesian approximate computation approach with supervised machine learning as implemented in DIYABC Random Forest (DIYABC-RF) v 1.0 (Collin et al., 2021). Since the method is computational demanding, the scenarios were kept simple and easily distinguishable from each other (Cabrera & Palsboll, 2017). To save calculation time, we grouped individuals into four groups (1) South Africa, (2) Brazil, (3) Spain and Guatemala (Spain-Guatemala), and (4) Greece and Australia (Greece-Australia) according to genetic groups defined by DAPC and STRUCTURE analyses (see Results, Fig. 3). The modelling strategy was focused on testing different phylogeographic scenarios reflecting hypotheses of colonisation routes from South Africa to Brazil or to the cluster Spain-Guatemala based on the results in Fig. 6 and the historical records (Fig. 1). Main phylogeographic hypotheses are explained graphically in Fig. S1A-B, Supporting information. These scenarios were analysed individually, after selecting the best hypothetical evolutionary scenario, we performed a second analysis only for the three scenarios with the highest classification votes (Fig. S1B, Supporting information).

In accordance with DIYABC's requirements, those SNPs missing in all individuals in one population need to be excluded, which reduced the dataset to 312 SNPs across the 92 individuals. Uniform priors were used as described in Karsten et al. (2015). Effective population sizes (N_e) were allowed to range between 10 to 100,000; the priors for generation times between divergence events were set from 10 to 10,000 and admixture rate ranged from 0.001 to 0.999. A total of 20,000 data sets were simulated for each scenario and all simulated data sets were used in each Random Forest training set. We used five noise variables and ran 2,000 random trees to select the most likely phylogeographic scenario.

The species tree was inferred using SNAPP (Bryant, Bouckaert, Felsenstein, Rosenberg, & RoyChoudhury, 2012), a coalescent-based method analysis that uses SNP markers directly to compute the likelihood of the species tree under a finite-sites mutation model over all possible gene trees. SNAPP is implemented in BEAST v2.6.2 (Bouckaert et al., 2014) and estimates the probability of allele frequency change across ancestor and descendent nodes, giving as results a posterior distribution for the species tree (Bryant et al., 2012).

The putative neutral dataset was additionally filtered for non-polymorphic loci following the model assumptions. Running SNAPP comes at a high computational cost; therefore, we pruned our dataset to include five randomly selected individuals from each population (N=30). The mutation rates (u and v) were calculated and ran for at least one million generations per replicate with sampling every 1,000 generations. We conducted three independent runs and evaluated convergence in Tracer v1.7 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018). The following analyses were carried out with the runs whose parameters presented ESS > 200. We removed 10% of trees as burn-in and merged tree and log files from the different runs using LOGCOMBINER v 2.4.1. Then, TREEANNOTATOR v2.6.2 was used to obtain maximum credibility trees and the trees that were contained in the 95% highest posterior density (HDP). To visualise the posterior distribution of trees we used DENSITREE v2.2.7 (Bouckaert, 2010). This process was repeated further by randomly resampling different individuals than those used for the first run (with the exception of sampling sites from Australia and Guatemala with < 10 individuals) to determine the subsampling effect on the analysis.

2.6 Microbiome analysis

Sixty-three different individuals collected from the populations used in the RADseq analyses were used to study their microbial structure and community composition. Individuals from all populations were used, except for Greece and Guatemala. In addition, we included new samples from Colombia (CO) and Israel (IS) in this analysis, resulting in a total of six sampling sites collected in nine different host plants (Figure 1 and details in Supplementary Table 1). DNA was extracted following the protocol mentioned above, and the V4 region of the 16S rRNA gene was amplified in duplicates using the primers 515F-806R (Caporaso et al., 2011) in one step PCR (95°C 5min, 25 cycles x [95°C 20s, 55°C 20s, 72°C 30s], 72°C 5min). The resulting amplicons

were pooled per sample and cleaned (AMPure XP magnetic beads, Beckman Coulter, Indianapolis, USA) and PCR indexing was carried out using a Nextera XT DNA Library Preparation Kit (Illumina Inc., San Diego, US-CA). The paired-end sequencing was performed at the Molecular Core Labs (Sequencing Facility) of The Natural History Museum on a complete flow cell of the Illumina MiSeq platform using v3 chemistry (2 x 300 bp).

Sequence data were processed in Mothur v1.41.3 by adapting the MiSeq SOP protocol (Kozich, Westcott, Baxter, Highlander, & Schloss, 2013; Schloss et al., 2009). Following quality filtering, we merged forward and reverse reads. Primers were removed, ambiguous bases were discarded, and sequences with over 15 homopolymers were excluded from further analysis. Unique sequences were aligned against the Silva Seed v132 reference database (Quast et al., 2013), which was reduced and customized to the V4 region to avoid errors and improve alignment quality. The maximum length of the sequences was set to 275 bp, and those that did not align well were removed. Duplicate sequences were merged, and denoising of unique sequences was performed with the command*pre.cluster* with an allowance of 1 difference per 100 bp (Callahan et al., 2016) to infer Amplicon Sequence Variants (ASVs). Singletons that did not cluster were removed at this stage with the command *split.abund.* Chimaeras were removed using UCHIME and the Silva Gold database (Edgar, Haas, Clemente, Quince, & Knight, 2011). ASVs were taxonomically classified using the Silva NR v132 reference database with the command *classify.seqs* , and unwanted lineages, such as Archaea and Eukaryota, were removed with the command*remove.lineage* .

ASVs were summarised at the phylum, family, and genus taxonomic levels for further analysis. All statistical analyses were computed in R (Team, 2020) using the packages *phyloseq, microbiome, ape* and *vegan* . The number of ASVs shared in groups between different countries were visualized with Venn diagrams produced through the Van de Peer lab website (http://bioinformatics.psb.ugent.be/webtools/Venn/). Alpha diversity (within samples and communities) was calculated with the species richness (total number of ASVs) estimator, and the Shannon (Shannon, 1948) and Inverse Simpson's (Simpson, 1949) indexes. In addition, the Bray-Curtis distance metric was used to quantify the dissimilarity of community composition between samples in the different sampling localities, and permutational multivariate analysis of variance (PERMANOVA) was used on the distance matrix to test for statistical differences between groups with the function *adonis()* from the vegan package. Group homogeneity of dispersion was evaluated with the function *betadisper()* from the vegan package (Anderson et al., 2006). Pairwise comparisons were tested with the *pairwise.adonis*wrapper function by Martinez Arbizu (2020), and p-values were adjusted using Bonferroni corrections. We also performed analysis of similarities (ANOSIM) tests, with the Bray-Curtis distance metric, to evaluate whether any differences in microbiome composition between groups were attributed to geographical distance or medfly diet. ANOSIM results in a ratio of between to within groups variation, with an R value closer to 1 suggesting dissimilarity between groups and values closer to 0 suggesting even distribution between and within groups (Clarke, 1993). For these tests we used the *anosim()* function from the vegan package.

Differential abundance analyses were performed on the most abundant ASVs to assess differences in the microbiome composition across sample sites with the package *edgeR* (Robinson, McCarthy, & Smyth, 2010) implemented in R (Team, 2020). First, ASVs with relative abundance [?] 0.01% were chosen, and data normalisation was implemented with function *calcNormFactors()* . Then, an exact binomial test was performed to generalise over dispersed counts with the function *exactTest()* (Xia, Sun, & Chen, 2018). Next, the top differentially abundant ASVs were tabulated with function *topTags()* and the false discovery rate (FDR) was estimated by adjusting the *p-values* with the Benjamini-Hochberg correction method (Benjamini & Hochberg, 1995); ASVs with an FDR < 0.01 were considered significant and were chosen to visualize differential abundance in the microbiome composition with the package *ggplot2* (Wickham, 2016; Xia et al., 2018).

3. RESULTS

3.1 Genetic diversity indices

Overall, genetic diversity as estimated based on the number of private alleles, nucleotide diversity (π), and

expected heterozygosity (H_e) was higher for South Africa (native range) than for the other five populations from the introduced range (i.e., SP, GR, GU, BR, AU) (Table 1). The smallest number of private alleles and lowest genetic diversity were identified in Guatemala and Australia. Inbreeding coefficients (F_{is}) were very similar and around zero across all sampling sites indicating random mating for the different sites (Table 1). Regarding Tajima's D , the South African population had a negative value, showing a possible selective sweep or recent population expansion. The remaining populations presented positive Tajima's D values, indicating balancing selection or sudden population contractions (Table 1).

3.2 Population assignment and structure

Pairwise F_{st} comparisons between regional samples showed low to moderate values ranging from 0.05 to 0.21; all comparisons were significant (Fig. S2, Supporting information). The highest F_{st} value was recovered between Guatemala and Australia, and lowest F_{st} values were obtained comparing Spain and Greece (0.05) and their respective comparisons with Australia (0.08 and 0.09). South African specimens showed moderate F_{st} values in pairwise comparisons with all sites, although slightly higher with the Spanish and Greek (0.18 in both cases) than with Neotropical and Australasian sampling sites.

DAPC detected an optimal genetic cluster of two ($K = 2$) followed by three ($K = 3$) based on the AIC value (Fig. S3A, Supporting information). DAPC results for $K= 2$ identified two clusters (South Africa and Brazil) versus the rest, i.e. Spain, Greece, Guatemala and Australia (Fig. S4A, Supporting information). The DAPC results for $K = 3$, with 16 principal components explaining 46.7% of the variance, showed two separated clusters (South Africa and Brazil) and the rest (Fig. 2A). Group assignment analysis (*compoplot* function) supported these clusters with a high probability of assignation (Fig. S4B, Supporting information). Unexpectedly, the assignment plot placed one individual from SA (SA_8) (Fig. S4B-C, supporting information) into the cluster composed of all other populations, indicating the ancestral presence of the colonising alleles in low frequency or potential back colonisation from the colonised areas. Moreover, STRUCTURE analyses identified K =2 as the most probable number of genetic clusters followed by K =3, with some differences in the groupings as compared to the DAPC analyses (Fig. S3C, Supporting information). For K =2, one cluster grouped all sampling sites except South Africa, which showed some degree of introgression (<20%) from another genetic cluster (dark blue in Fig. 2 top). When considering K =3, Brazilian individuals showed a small proportion of cluster assignation to a cluster also present in the South African population (light blue in Fig. 2B bottom). Further DAPC and STRUCTURE analyses were conducted, excluding samples from South Africa (Fig. 3A-B) to explore the introduced range in more detail. DAPC detected an optimal genetic cluster of three ($K = 3$) (Fig. S3B, Supporting information) with Brazilian samples as the most divergent population and some differentiation between Spain-Guatemala versus Greece-Australia (Fig. 3A). STRUCTURE analyses for K =2 detected a genetic break between Brazil and the rest of populations; for$K$ =3, apart from identifying Brazil as belonging to a different genetic cluster, all individuals from Guatemala were assigned to a distinct genetic cluster, and individuals from Greece and Australia were mostly assigned to a separate cluster (with the exception of two individuals from Greece), while most individuals from Spain had mixed association with the former two clusters in roughly similar proportions (Fig. 3B).

The co-ancestry matrix generated by *fineRADstructure* (Fig. 4) showed that all individuals of the respective locations clustered together, with the exception of European samples from Spain and Greece, which appeared mixed in two different groups. The most distinct group of samples was South Africa, showing the highest values of co-ancestry, followed by Brazil. Both also showed higher ancestry relationships compared to the remaining sampling locations, i.e. they presented fewer allele differences among them compared to the remaining sampling sites. All other samples formed a cluster that could be subdivided into two subclusters: subcluster A, including Australia and a mix of samples from Spain and Greece; and subcluster B, including all samples from Guatemala and the remaining Spanish and Greek samples. The samples from Europe (Spain and Greece) showed strong evidence of heterogeneous gene flow (indistinguishable genetic ancestry).

3.3 Origin of medfly infestation based on demographic history and phylogenetic analysis

In the DIYABC-RF analysis that tested six hypothetical evolutionary scenarios (Fig. 5), the classification

votes from Scenario 1 to Scenario 6 were: 599, 411, 123, 467, 267 and 133, respectively (i.e. the number of times a scenario is selected in a random forest). Based on the classification votes and posterior probabilities, the best fit was Scenario 1, with a posterior probability of 0.596 and global and local error rates of 0.499 and 0.404, respectively (Fig. 5). The projections of data set from the training set on the linear discriminate analyses (LDA) indicated low power to discriminate the tested Scenarios 1, 2 and 4 because the observed data set was located within the cloud of their simulated data (Fig. S5-C. Supplementary information). To improve prediction quality and power of differentiation, we ran a new analysis only for Scenarios 2, 4 and 5 (selected based on the previous test results). The classification votes were 778, 1025 and 197 respectively. The best fit scenario was Scenario 4, with a posterior probability of 0.697 and global and local errors of 0.246 and 0.303, respectively (Fig. 5). The projection of data sets from the training set in this second test was located within the cloud of Scenario 2, indicating substantial power to discriminate among the tested scenarios (Fig. S5-C. Supplementary information). Overall, the demographic colonisation scenarios suggested a long and interconnected history of invasions of*C. capitata* in the studied sites. Both best fit scenarios predicted Brazil divergence from the ancestral South African population. However, Scenario 1 predicted direct colonisation from Brazil to the other sampling sites, while Scenario 4 predicted that the Spain-Guatemala group originated from the admixture between lineages from South Africa and Brazil. According to these results, Brazil specimens were established by direct colonisation from South Africa and likely admixture events leading to the establishment of the remaining lineages (i.e. Spain-Guatemala and Greece-Australia).

SNAPP recovered a total of 15 consensus trees topologies. The consensus tree 1 covered 37.18% of the total cumulative trees (Fig. 6) increasing to 67.04% when the consensus trees 2 and 3 were included. The consensus tree topologies were consistent across the independent runs in which different individuals were sampled from each location (Fig. 6; Fig. S6, supporting information), indicating that subsampling did not significantly impact the topology of the SNAPP trees. The species tree revealed three highly supported lineages (PP=1) corresponding to South Africa, Brazil and a third lineage comprised of all other regions, whereas two nodes consistently showed moderate support corresponding to the divergence between Greece (PP=0.81) and Guatemala and Australia (PP=0.84) lineages. These results are consistent with the genetic clusters found in the DAPC and Structure analyses (Fig. 2 and Fig. 3). Effective population size represented in the branch thickness of the consensus tree inferred by theta-estimates showed the highest value in South Africa, followed by Spain, Brazil, and Greece with intermediate values (Fig. 6).

3.4 Microbial taxonomic composition and diversity measures

Amplicon sequencing of associated microbial communities on the Illumina MiSeq produced a total of 2,653,026 raw reads, and after quality filtering, dereplication and removal of non-bacterial sequences (Fig. S7, supporting information), a total of 6,249 different ASVs were identified. The dominant phyla detected across all samples were Proteobacteria (60.4%) (Fig. 7). In all samples from Brazil and some specimens from South Africa, Spain, and Australia, the phylum Firmicutes was also abundant (18.7%). In contrast, Bacteroidetes (7.8%) and Actinobacteria (5.5%) were observed only in some individuals across the localities. Noteworthy, specimens from Brazil exhibited a higher relative abundance of Firmicutes and Bacteroidetes than all other locations (Fig. 7). At the family level, *Enterobacteriaceae* (27.4%), *Anaplasmataceae* (4.4%), unclassified Clostridiales (3.1%) and *Acetobacteraceae* (1.8%) were predominant across all populations.

The individuals collected in South Africa and Brazil presented the lowest and highest number, respectively, of unique bacterial ASVs (581 vs. 2,526) (Fig. S8A-C, Supporting information). South African individuals shared a higher number of ASVs with the Palearctic and Australasian locations (73 and 98 ASVs, respectively) than the Neotropics (42 ASVs) (Fig. S8A-C, Supporting information). Individuals from South Africa and Brazil were collected from the same host plant, guava, but surprisingly the specimens collected from Brazil shared more ASVs with its neighbouring country Colombia (89 ASVs), despite being collected from different host plants, than with their putative native region in South Africa (52 ASVs) (Fig. S8B, Supporting information). This suggests that the bacterial composition remained similar to the native source in some localities, despite the medflies being collected from different host plants and geographically distant areas (see

Fig. 1 and Supplementary Table 1).

Based on Shannon and Inverse Simpson diversity indexes, samples collected in South Africa appeared to have the lowest microbial diversity. At the same time, flies collected from different host plants and geographically distant sampling sites such as Israel, Spain and Australia showed the lowest disparity in diversity indexes, suggesting a similarity in microbial diversity. Conversely, the Neotropical region tended to have a more diverse microbiome, and samples from Brazil presented the highest diversity indexes (Fig. S9, Supporting information).

3.5 Microbial composition turnover among regions and host plants

Microbiome composition comparisons between the putative native South Africa and introduced range of medfly were only statistically significant for the pairwise comparisons with Spain ($F = 6.4$, $R2 = 0.30$, $p = 0.015$), Colombia ($F = 7.4$, $R2 = 0.26$, $p = 0.015$) and Brazil ($F = 7.8$, $R2 = 0.41$, $p = 0.015$). In the introduced range, the microbiome structure of Spain was significantly different to that of Israel ($F = 3.6$, $R2 = 0.15$, $p = 0.015$) despite being collected partly on the same host plant (Fig. 1) and to Brazil ($F = 6.3$, $R2 = 0.28$, $p = 0.015$). Conversely, the microbial communities from Spain, Australia and Colombia showed no significant differences despite their geographic distances and being collected from different host plants. In the neotropical region, the microbiome of medflies from Brazil and Colombia showed significant differences ($F = 7.0$, $R2 = 0.24$, $p = 0.015$), although collected from different host plants (Fig. 1). ANOSIM analysis at the global scale showed significant differences in samples collected from different locations ($R = 0.429$, $p = 0.001$) and different host plants ($R = 0.196$, $p = 0.001$), demonstrating the greater importance of geographical distance for microbiome composition compared to diet.

Exclusively for the sampling sites whose microbiome composition was significantly different to South Africa in the pairwise comparisons, we performed differential abundance analyses on the most abundant ASVs to identify the differences at the genus taxonomic level. The comparisons between South Africa and Spain did not detect any significant changes with the threshold implemented (FDR < 0.01). Indeed, the genus *Klebsiella* was the only altered in high abundance between South African and Colombian samples, while 48 bacterial ASVs were evidently different in the analysis between Brazil and South Africa sampling sites (Fig. 8). Worth mentioning is that samples from both localities were collected in guava and presented a similar relative abundance in the genus *Flavobacterium.* The genus *Gluconobacter* and unclassified Halomonadaceae were detected only at the low relative abundance in all medfly samples collected in South Africa. In contrast, a high abundance of genus *Acinetobacter* , unclassified Burkholderiaceae, unclassified Clostridiales, *Dysgonomonas* and *Escherichia-Shigella,* was harboured exclusively across all the medfly samples collected in Brazil (Fig. 8).

4. DISCUSSION

4.1 Population genomics of *Ceratitis capitata*

The present study is the first using reduced-representation sequencing genome in combination with a microbial characterisation in the medfly *C. capitata* , using samples from key geographic locations to investigate the species' population history and microbiome on a global scale. We find strong evidence for two genetic clusters corresponding to the South African individuals and the other localities in the introduced range, in agreement with virtually all previous studies using allozymes (Gasperi et al., 2002; Gasperi et al., 1991; Kourti, 2004; Malacrida et al., 1992), mitochondrial DNA markers (Arias et al., 2018; Elfékih et al., 2010; Elfékih, Makni, & Haymer, 2013; Karsten, van Vuuren, Barnaud, & Terblanche, 2013; Ruiz-Arce et al., 2020), and microsatellites (Bonizzoni et al., 2004; M. Bonizzoni et al., 2001; Deschepper et al., 2021; Karsten et al., 2015; Nikolouli et al., 2020). In addition, when we analysed the five sampling sites from the introduced range separately (i.e., removing the South African samples), populations from Brazil represented a unique genetic cluster that had not been recognised in previous studies. The Brazil cluster was also characterised by a distinct microbiome and the highest overall bacterial diversity.

Karsten et al. (2013, 2015) observed high genetic diversity in South African medflies due to a large number

of alleles present at low frequency, including many private alleles, which led them to suggest that this population was ancestral and has maintained a large population size over time. Other studies have shown that populations derived from the African lineage exhibited a gradual decrease in genetic variation (Malacrida et al., 2007; Deschepper et al., 2021 and references therein), first to the Mediterranean basin populations and a second towards American populations, thus dividing the colonisation process of the medfly in three main categories: Ancestral populations (Sub-Sahara and Africa), ancient populations (Mediterranean basin) and recent populations (America) (Gasperi et al., 2002; Malacrida et al., 1998). Our results showed a different pattern of genetic variation: across the introduced range, most of the sampled locations belong to the same big genetic cluster, indicating gene flow among these locations, except for Brazil.

Gasperi et al. (2002), using allozymes, found similar levels of genetic variability in South American populations (i.e., Argentina, Brazil and Peru) to African ancestral populations, and they stated that these populations did not have enough time to reach equilibrium and further differentiation. Furthermore, Nikolouli et al. (2020) described a discrete genetic cluster in some South American populations (Argentina, Brazil and Bolivia) using microsatellites; however, they stressed that this genetic cluster was not clearly distinct from other medfly populations worldwide. Our study identified high genetic diversity and a genetically distinct cluster of medflies collected in Brazil. These findings suggest that some South American populations might be derived from different genetic sources.

The combination of population structure and ABC analyses with supervised machine learning allowed us to reconstruct the most probable evolutionary scenario of *C. capitata.* It is important to note that only a portion of the total geographical distribution of medfly is covered in this study. Nevertheless, our limited data set was able to support the initial divergence from South African ancient populations that gave rise to populations in Brazil at a different time than those in the rest of the world. These findings are partially congruent with historical records of medfly distributions (Malacrida et al., 1998), which medfly colonisation route may have occurred through the transatlantic trade of enslaved people, as has been described in *D. melanogaster* , which is also a descendant from the Afrotropical region (David & Capy, 1988). Furthermore, the medfly museum collections at the Natural History Museum of London provided historical records to support this suggested new colonisation route. We found that in the collection, the oldest record was dated in 1904, with specimens collected in the tropical Saint Helena Island (Fig. 1). This island, a UK overseas territory located in the South Atlantic Ocean, midway between Africa and South America, was an important port during the crown colony and an obligate stop for the Trans-Atlantic trade in the colonial period. The human population in Saint Helena has genomic traces linking them with Central-West African populations, moved during the slavery years (Sandoval-Velasco et al., 2019). This colonisation route for the medfly, different from the one connecting the Mediterranean area and South America, had been mentioned by Gasperi et al. (2002) and Ruiz-Arce et al. (2020) but has never been probed at the genetic level. Our results modify the previous belief that only Mediterranean basin medfly populations had contributed to the colonisation of South America, as described in previous publications (Deschepper et al., 2021; Malacrida et al., 1998; Malacrida et al., 2007), and points out new potential ancestry sources for the genetic units in the South American populations that need further investigation.

4.2 Microbiome of *Ceratitis capitata*

Despite a wide range of host plants and large distances among collection sites, many microbial taxa were identified in both the native and introduced ranges of the medfly. The microbiome composition was highly similar across the biological replicates and the sampling sites, except for the specimens collected in Brazil. The microbiome of all samples was dominated by the phylum Proteobacteria followed by a few other predominant phyla: Firmicutes, Bacteroidetes, and Actinobacteria, which is consistent with the significant phyla previously observed in association with tephritids (Deutscher, Chapman, Shuttleworth, Riegler, & Reynolds, 2019; Morrow, Frommer, Royer, Shearman, & Riegler, 2015; Nikolouli et al., 2020). It is important to note that in this study, we used individual mature adults collected in wild conditions, while previous studies were performed on pooled dissected guts of larvae (De Cock et al., 2019; De Cock et al., 2020), adult guts (Nikolouli et al., 2020) or adults collected from infected fruits that emerged in laboratory conditions

(Malacrino et al., 2018). Despite the technical differences, these studies have defined Proteobacteria as the predominant phylum, albeit at a higher abundance ([?] 80%) compared to our study (60.4%). Besides this, the abundance of the phylum Firmicutes in those studies was lower ([?] 10%) compared to our results (18.7%). Differences observed in abundances for these phyla might be explained by differences in the sample source and/or life stage of the medflies analysed. At different taxonomic levels, the prevalence of an unclassified Enterobacteriaceae genus (Gammaproteobacteria) was expected because it is known to dominate the first set of maternally inherited microbiota, which is vertically transmitted during oviposition in tephritids (Aharon et al., 2013; Behar, Yuval, & Jurkevitch, 2005; Deutscher et al., 2019).

The microbiome abundance and composition remained similar to the native range across different host plants and some distant localities despite the low number of ASVs detected in South Africa. Furthermore, the microbiome structure in the introduced range presented similar numbers of ASVs among Spain, Israel, Australia and Colombia, with significant differences in their microbiome composition only found for Spain and Israel. These results of the microbiome structure mirror our population genetic analysis, where we found genetic similarities among all the populations in the introduced range, except Brazil. Therefore, we suggest that these microbial communities are stable regardless of the distances, host plants and differences in environmental conditions, also reflecting the interconnectivity between these localities. In this context, similarities in microbiome composition were found between Spain and Australia, which were also highly connected populations in our DAPC (both populations appeared in the same genetic cluster) and population structure analysis. On this basis, the connectivity of the medfly populations might facilitate a bacterial exchange/transmission across them, suggesting the attainment of an essential microbiome set over time that successfully serves the medfly's polyphagous nature (Gruber et al., 2019). This could represent a solid contributing factor to the species' invasive success by ensuring a universal way of feeding on various plants in introduced ranges.

A particular case is the unique microbiome composition in the Brazilian population. A recognised factor that influences microbiome composition is feeding on different plant species (Malacrino et al., 2018). However, in this study, Brazil and the native South Africa were collected from the same host plant, Guava. Consequently, we suggest that the somewhat isolated flies of Brazilian populations acquired new bacteria from the new environment. Also, exclusively in Brazil, we found in high relative abundance *Acinetobacter* (phylum Proteobacteria) that was previously described in larvae of medfly (De Cock et al., 2019; Malacrino et al., 2018). The genus *Acinetobacter* is associated with plant defence suppression mechanisms in polyphagous insects, which is known to help insects to detoxify phenolic glycosides *in vitro*(Mason, Couture, & Raffa, 2014), although some other bacteria, which are difficult to isolate in culture, may contribute to the metabolism of this metabolite.

Unclassified Burkholderiaceae in Brazil were previously described only in medfly adults collected in Italy (Malacrino et al., 2018). The presence of these bacteria has been associated with nitrogen fixation (i.e. diazotroph microbes), which is an essential mechanism for the fly's nutrition, development, and reproduction (Behar et al., 2005; Raza, Yao, Bai, Cai, & Zhang, 2020). Most insect microbiomes are maintained by strict vertical transmission, however, noteworthy is the case of the bean bug *Riptortus pedestris* (Heteroptera: Alydidae), where it is known that some strains of *Burkholderia* are taken at early stages every generation from the environment (Kikuchi, Hosokawa, & Fukatsu, 2007). Members of the genus *Burkholderia* are known as significant soil bacteria, although the details of the acquisition mechanisms of the bacteria in *R. pedestris* remain unknown (Kikuchi et al., 2007; Kikuchi, Meng, & Fukatsu, 2005). However, in agricultural lands with intensive insecticide applications, an acceleration in the microbial degradation of the insecticide has been observed (Arbeli & Fuentes, 2007; Singh, Walker, & Wright, 2005). Subsequently, when *R. pedestris* occurs in fields heavily treated with the insecticide fenitrothion (one of the most popular organophosphates), some *Burkholderia* strains show the ability to degrade the insecticide and demonstrate that *Burkholderia* confers resistance to the insect against the organophosphate, establishing a beneficial symbiont relationship (Kikuchi et al., 2012; Kikuchi & Yumoto, 2013). This finding raises the possibility that the *Burkholderia* observed only in Brazilian medflies may be associated with the novel acquisition of capabilities to hydrolyse and metabolise insecticides, thus enhancing the fitness of the host insect. Although

this hypothesis should be further tested, it might be an example of horizontal transmission of microbiomes from the soil associated with a demographic response to insecticides and calls for specific experiments to understand the emergence of resistance and the bacterial strains involved. Another example to consider as horizontal transmission observed exclusively in Brazil is *Dysgonomonas.* This member of phylum Bacteroidetes is found on the surface of plant roots in the soil (Liu et al., 2018) and has been described in the guts of wild specimens of*Bactrocera dorsalis,* a species closely related to the medfly, suggesting that they might have been recruited from the surrounding environment (Wang, Jin, & Zhang, 2011).

Overall, most of the prevalent genera identified exclusively in Brazil are associated with the soil microbiome, which raises questions about the possible functions of the microbiome in this specific locality. A recent publication found differences in chromosomes and changes in the allele frequency between experimental and natural populations of *D. melanogaster* that were exposed to different microbiome treatments, suggesting that a shift in microbiome composition may be an agent of selection that drives adaptation at population levels (Rudman et al., 2019). Here, we reveal the correlation between the microbiome composition and the genomic structure of the populations and highlight the importance of the host-microbiome interaction for the adaptation of the medfly to different environments. Therefore, further research at a population level will be required to unveil the role of the microbial communities in the medfly.

5. CONCLUSIONS

To our knowledge, this is the first genome-wide study investigating both population genomics and bacterial communities associated with *Ceratitis capitata.* Our data offer an in-depth view of medfly population structure and yields insights into the influence of invasion pathways on the microbial diversity. It revealed genetic structure, with one genetic cluster in South Africa (the native range of medfly), and then two distinct genetic clusters in the introduced range: one associated to the Brazilian individuals and the other, clustering specimens further distance such as Spain, Greece, Guatemala, and Australia. Furthermore, the microbiome surveys highlighted Brazil as the most diverse set of microbiota compared with all other sampling sites.

Both approaches (genomic and microbiome) emphasize the uniqueness of Brazil's population, which opens the possibility for a completely independent colonisation route from Africa to America during the colonial period, using the transatlantic trade routes that passed through St. Helena Island. Museum records and ABC modelling with supervised machine learning supported this alternative colonisation route. In any case, further research is needed to elucidate the possible effect of microbiota in the genomic divergence at a population level and to identify the possible role and transmission pattern of the microbiome in the host fly.

Worldwide medfly management requires huge financial resources. The use of genomic tools provides an opportunity to develop a framework for a survey of medfly invasion pathways and to identify alleles that might be crucial for population differentiation. We suggest that future studies should use SNPs information to create quick identification methods (e.g., genomic tagging) that will help to inform potential novel outbreaks to the quarantine services and allow for better development and implementation of suitable pest management strategies.

ACKNOWLEDGEMENTS

We are grateful to our collaborators for providing the samples used in this study, particularly Nancy Carrejo (Universidad del Valle, Colombia) and Monica Hernandez (Instituto de Ecologia, A.C, Mexico). We also thank Daniel Whitmore for his assistance with the BMNH collection and to Vanina Tonzo for her feedback and support in SNAPP and ABC model analyses. Finally, we acknowledge Dr Cristina Diez-Vives and thank her for her advice on the microbiome pipeline. MBA was supported by the National Agency for Research and Development (ANID), fellowship program Doctorado en el extranjero/2014 and Postdoctorado en el extranjero/2019 - 74200143. SE was supported by an EMBO grant (ASTF-42-2010) and a CSIRO Julius Career Award (R- 91040-11). ST received funding from the grant PID2020-117115GA-100 funded by MCIN/AEI/ 10.13039/50110001103.

AUTHOR CONTRIBUTIONS

MBA, AR, APV and SE designed and conceived the research study. MBA and SE performed experiments. ST conceived bioinformatic pipelines for the analysis of the data. MBA and KHM analysed the data, prepared figures and/or tables and drafted the manuscript. All authors reviewed, edited, and approved the final manuscript.

DATA ACCESSIBILITY

The radseq and microbiome data are available at the CSIRO Data Collection via the following entry link: https://doi.org/10.25919/7v6t-jw65

FIGURE LEGENDS

Figure 1. Worldwide map showing in colours the sampling locations for the specimens used in RADseq analyses. Dates in parenthesis represent the earliest record based on literature and the Natural History Museum of London collection (e.g. St Helena Island). Dark grey countries: Colombia (CO), Israel (IS) and the asterisk (*) indicated medflies collected for microbiome analyses; the numbers indicate the host plant (note that some flies from the same sampling site were collected in different host plants).

Figure 2. Individual genotype assignment for the six populations of medfly distributed worldwide. A) DAPC plot for the clustering obtained in AIC analysis ($K = 3$). B) Individual genotype assignment inferred by STRUCTURE with K =2 and K=3 top and bottom, respectively.

Figure 3. Individual genotype assignment for the locations in the introduced range of medfly. A) DAPC analysis for the clustering obtained in AIC analysis ($K = 3$) of 82 individuals. B) STRUCTURE results based on K =2 and K =3 top and bottom, respectively.

Figure 4. Simple co-ancestry heatmap for medfly populations. Top: the raw data matrix tree (simple hierarchical "tree-like" with posterior population assignment probabilities), each tip corresponds to a sampling site. Left: the location abbreviations. Bottom: the clusters identified in the co-ancestry matrix.

Figure 5. Best fit hypothetical evolutionary scenarios of the invasion routes of *C. capitata.* Left: topology of the best fit scenario selected from six different invasion routes (Scenario 1); South Africa was predicted as the ancestral population followed by colonisation of Brazil, and from there was predicted a divergence to the cluster Spain-Guatemala and Greece-Australia. Right: topology of the best-predicted scenario from three different invasion routes tested (Scenario 4). South Africa was the predicted ancestral population for medfly followed by divergence to Brazil. The clusters Spain-Guatemala and Greece-Australia appeared to be the result of admixture between South Africa and Brazil populations (see Methods and Supplementary Fig.S1 for details). Colours correspond to sampling sites codes in fig 2A (except sampling sites grouped).

Figure 6. Species tree of 30 individuals collected across six medfly sampling sites (SA: South Africa, BR: Brazil, SP: Spain, GU: Guatemala, AU: Australia). A total of 15 consensus tress topologies were obtained, here is represented the consensus tree 1 which covers 37.18% of the total cumulative trees. Branch width is proportional to theta represented in black.

Figure 7. Relative abundance of the different microorganisms by phyla detected across six sampling localities of medfly. Phyla with abundance lower than 1are the abbreviations to each population studied, each bar represented one specimen per sampling site, for more information refer to Table 1 and Supplementary Table 1.

Figure 8. Differential abundance of medfly microbiome at genus taxonomic level in Brazil (BR) and South Africa (SA). On the left, levels represent 48 ASVs with relative abundance above 0.01% and FDR under 0.01. Circles are colour coded according to bacterial phylum and RA stands for relative abundance. The ASV taxonomically assigned to Burkholderiaceae_unclassified is highlighted in blue to indicate its unique presence in all samples from Brazil.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Graphical representation of hypothetical evolutionary scenarios of medfly using DIYABC-RF. The strategy focused on testing hypotheses of colonisation routes from South Africa to Brazil or South Africa to the rest of the sampling sites. A) Analysis 1, in light grey, highlights the best fit Scenario. B) Analysis 2, using only Scenarios 2, 4 and 5 (the highest CV values in Analysis 1). The best fit scenario is highlighted in light blue. CV: corresponded to the classification vote for each scenario obtained in the different analyses.

Figure S2. Pairwise *Fst* values for the six populations of medfly. All calculations were significant, abbreviations are based on where populations are collected. SA, South Africa; SP, Spain; GR, Greece; GU, Guatemala; BR, Brazil; AU, Australia.

Figure S3. AIC statistic plots where lower value indicates optimal clustering. A) Optimal AIC estimation using the dataset with the six populations. B) Optimal AIC estimation using only populations collected in introduced range (i.e. SP, GR, GU, BR, AU). Delta K estimation by Evanno method. C) Best delta K estimation using the dataset with the six populations. D) Best delta K estimation using only populations collected in introduced range (i.e. SP, GR, GU, BR, AU).

Figure S4. Clustering analysis using the 1907 SNPs. A) Group assignment probability plot using K=3 based on DAPC analysis. The colour of cells points out the probability of assigning a given sample (red: high probability) and the blue crosses indicate the true group assignment. The row labels are specimens ID and columns (clusters) correspond to the population groups. B) Hierarchical clustering (Ward clustering), SA: South Africa, BR: Brazil, OP: Other populations. Dark red arrow is highlighting the specimen SA_8 assigned to the major cluster formed by populations collected in the introduced region.

Figure S5. Projection of dataset (observed data) from the training set on Linear Discriminant Analysis plots. C) Analysis 1, six scenarios analysed individually. D) Analysis 2, three scenarios analysed individually.

Figure S6. Tree cloud produced by DENSITREE from SNAPP analysis. *Left:* tree cloud of the 15 consensus trees of Fig. 6.*Right* : tree cloud obtained by independent subsample set from the six populations studied. Maximum-clade-credibility tree estimation is shown in dark blue (most highly supported), red is the next most supported, and green is the least supported. Maximum-clade-credibility tree shown in the black right-angled tree with posterior probabilities at nodes. Branch width is proportional to theta.

Figure S7. Plot of sequencing read depth for microbiota database and the rarefaction curve.

Figure S8. Number of unique and shared ASVs across different biogeographical regions. A) South Africa compared to Palearctic sampling locations. B) South Africa compared to Neotropical sampling locations. C) South Africa compared to Australia (Australasian).

Figure S9. Alpha diversity of medfly microbiome across six sampling locations. Observed (number of ASVs present -species richness), Shannon, Inverse Simpson and Pielou's evenness diversity indexes.

TABLE LEGENDS

Table 1. The collection sites of *C. capitata* used in the population genomic analyses. Sample size (N) and summary of genetic diversity indices, H_o : observed heterozygosity,H_e : expected heterozygosity, π refers to nucleotide diversity and F_{is} to inbreeding coefficient.

Supplementary Table 1. Samples of *C. capitata* collected from six countries in nine different host plants used in the microbiome composition analyses. N corresponds to the number of biological replicates.

REFERENCES

Aharon, Y., Pasternak, Z., Ben Yosef, M., Behar, A., Lauzon, C., Yuval, B., & Jurkevitch, E. (2013). Phylogenetic, metabolic, and taxonomic diversities shape mediterranean fruit fly microbiotas during ontogeny.*Appl Environ Microbiol, 79* (1), 303-313. doi:10.1128/AEM.02761-12

Anderson, M. J., Ellingsen, K. E., & McArdle, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecol Lett* , 9, 683–693. doi:10.1111/j.1461-0248.2006.00926.x

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet, 17* (2), 81-92. doi:10.1038/nrg.2015.28

Andrews, S. (2012). FastQC: a quality control tool for high throughput sequence data. . http://www.bioinformatics.babraham.ac.uk/projects/fastqc: Babraham, UK: Babraham Institute.

Arbeli, Z., & Fuentes, C. L. (2007). Accelerated biodegradation of pesticides: An overview of the phenomenon, its basis and possible solutions; and a discussion on the tropical dimension. *Crop Protection, 26* (12), 1733-1746. doi:10.1016/j.cropro.2007.03.009

Arias, M. B., Elfekih, S., & Vogler, A. P. (2018). Population genetics and migration pathways of the Mediterranean fruit fly Ceratitis capitata inferred with coalescent methods. *PeerJ, 6* , e5340. doi:10.7717/peerj.5340

Behar, A., Jurkevitch, E., & Yuval, B. (2008). Bringing back the fruit into fruit fly-bacteria interactions. *Mol Ecol, 17* (5), 1375-1386. doi:10.1111/j.1365-294X.2008.03674.x

Behar, A., Yuval, B., & Jurkevitch, E. (2005). Enterobacteria-mediated nitrogen fixation in natural populations of the fruit fly Ceratitis capitata. *Mol Ecol, 14* (9), 2637-2643. doi:10.1111/j.1365-294X.2005.02615.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing.*Journal of the Royal Statistical Society. Series B, 57* , 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bonizzoni, M., Guglielmino, C. R., Smallridge, C. J., Gomulski, M., Malacrida, A. R., & Gasperi, G. (2004). On the origins of medfly invasion and expansion in Australia. *Mol Ecol, 13* (12), 3845-3855. doi:10.1111/j.1365-294X.2004.02371.x

Bonizzoni, M., Malacrida, A. R., Guglielmino, C. R., Gomulski, L. M., Gasperi, G., & Zheng, L. (2001). Microsatellite polymorphism in the Mediterranean fruit fly, Ceratitis capitata. *Insect Molecular Biology, 9* (3), 251-261.

Bouckaert, R. (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics, 26* (10), 1372-1373. doi:10.1093/bioinformatics/btq110

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C. H., Xie, D., . . . Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol, 10* (4), e1003537. doi:10.1371/journal.pcbi.1003537

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis.*Mol Biol Evol, 29* (8), 1917-1932. doi:10.1093/molbev/mss086

Cabrera, A. A., & Palsboll, P.J. (2017). Inferring past demographic changes from contemporary genetic data: A simulation-based evaluation of the ABC methods implemented in diyabc. *Mol Ecol Resour, 17* (6), e94-e110. doi: 10.1111/1755-0998.12696.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods, 13* (7), 581-583. doi:10.1038/nmeth.3869

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., . . . Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA, 108 Suppl 1* , 4516-4522. doi:10.1073/pnas.1000080107

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda), 1* (3), 171-182. doi:10.1534/g3.111.000240

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology, 18* , 117-143.

Collin, F. D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., . . . Estoup, A. (2021). Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. *Mol Ecol Resour, 21* (8), 2598-2613. doi:10.1111/1755-0998.13413

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27* (15), 2156-2158. doi:10.1093/bioinformatics/btr330

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in functional genomics, 9*(5-6), 416-423.

David, J., & Capy, P. (1988). Genetic variation of *Drosophila melanogaster* natural populations. *Trends in Genetics, 4* (4), 106-111.

De Cock, M., Virgilio, M., Vandamme, P., Augustinos, A., Bourtzis, K., Willems, A., & De Meyer, M. (2019). Impact of Sample Preservation and Manipulation on Insect Gut Microbiome Profiling. A Test Case With Fruit Flies (Diptera, Tephritidae). *Front Microbiol, 10* , 2833. doi:10.3389/fmicb.2019.02833

De Cock, M., Virgilio, M., Vandamme, P., Bourtzis, K., De Meyer, M., & Willems, A. (2020). Comparative Microbiomics of Tephritid Frugivorous Pests (Diptera: Tephritidae) From the Field: A Tale of High Variability Across and Within Species. *Front Microbiol, 11* , 1890. doi:10.3389/fmicb.2020.01890

Deschepper, P., Todd, T. N., 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, 23* (11), 3529-3543. doi:10.1007/s10530-021-02595-4

Deutscher, A. T., Chapman, T. A., Shuttleworth, L. A., Riegler, M., & Reynolds, O. L. (2019). Tephritid-microbial interactions to enhance fruit fly performance in sterile insect technique programs. *BMC Microbiol, 19* (Suppl 1), 287. doi:10.1186/s12866-019-1650-0

Diagne, C., Leroy, B., Vaissiere, A. C., Gozlan, R. E., Roiz, D., Jaric, I., . . . Courchamp, F. (2021). High and rising economic costs of biological invasions worldwide. *Nature, 592* (7855), 571-576. doi:10.1038/s41586-021-03405-6

Dong, X., Yi, W., Zheng, C., Zhu, X., Wang, S., Xue, H., ... & Bu, W. (2022). Species delimitation of rice seed bugs complex: Insights from mitochondrial genomes and ddRAD-seq data. *Zoologica Scripta* , 51 (2), 185-198. doi: 10.1111/zsc.12523

Earl, D. A., & vonHoldt, B. M. (2011). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources, 4* (2), 359-361. doi:10.1007/s12686-011-9548-7

Eaton, D. A., & Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Systematic biology, 62* (5), 689-706.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection.*Bioinformatics, 27* (16), 2194-2200. doi:10.1093/bioinformatics/btr381

Elfékih, S., Arias, M., & Vogler, A. (2020). Population genomics of the Mediterranean fruit fly Ceratitis capitata using RADseq (Data Collection.) (Publication no. https://doi.org/10.25919/5f7ba8554fc6d). CSIRO Data Access Portal, from CSIRO.

Elfékih, S., Makni, M., & Haymer, D. S. (2010). Detection of novel mitochondrial haplotype variants in populations of the Mediterranean fruit fly, Ceratitis capitata, from Tunisia, Israel and Morocco.*Journal of Applied Entomology, 134, 647-651* . doi:10.1111/j.1439-0418.2009.01500.x

Elfékih, S., Etter, P., Tay, W. T., Fumagalli, M., Gordon, K., Johnson, E., & De Barro, P. (2018). Genome-wide analyses of the Bemisia tabaci species complex reveal contrasting patterns of admixture and complex demographic histories. *PLoS One, 13* (1), e0190555.

Elfékih, S., Tay, W. T., Polaszek, A., Gordon, K. H. J., Kunz, D., Macfadyen, S., ... & De Barro, P. J. (2021). On species delimitation, hybridization and population structure of cassava whitefly in Africa. *Scientific reports, 11* (1), 1-11.

Elfékih, S., Makni, M., & Haymer, D. S. (2013). Genetic Diversity of ND5 mitochondrial patterns inCeratitis capitata (Diptera: Tephritidae) populations from Tunisia. *Annales de la Société entomologique de France (N.S.), 46* (3-4), 464-470. doi:10.1080/00379271.2010.10697682

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proc Natl Acad Sci USA, 107* (37), 16196-16200. doi:10.1073/pnas.1006538107

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol, 14* (8), 2611-2620. doi:10.1111/j.1365-294X.2005.02553.x

Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour, 10* (3), 564-567. doi:10.1111/j.1755-0998.2010.02847.x

Falush, D., Stephens, M., & Pritchard, J. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics, 164* (4), 1567-1587.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics, 180* (2), 977-993. doi:10.1534/genetics.108.092221

Gasperi, G., Bonizzoni, M., Gomulski, L.M., Murelli, V., Torti, C., Malacrida, A.R., & Guglielmino, C. R. (2002). Genetic Differentiation, Gene Flow and the Origin of Infestations of the Medfly, Ceratitis Capitata *Genetica, 116* , 125–135. doi:doi.org/10.1023/A:1020971911612

Gasperi, G., Guglielmino, C. R., & Milani, R. (1991). Genetic variability and gene flow in geographical populations of Ceratitis capitata (Wied.) (medfly) *Heredity 67* , 347-356.

Gruber, M. A. M., Quinn, O., Baty, J. W., Dobelmann, J., Haywood, J., Wenseleers, T., & Lester, P. J. (2019). Fitness and microbial networks of the common wasp, Vespula vulgaris (Hymenoptera: Vespidae), in its native and introduced ranges. *Ecological Entomology, 44* (4), 512-523. doi:10.1111/een.12732

Gutierrez, A. P., & Ponti, L. (2011). Assessing the invasive potential of the Mediterranean fruit fly in California and Italy. *Biological Invasions, 13* (12), 2661-2676. doi:10.1007/s10530-011-9937-6

Heino, J., Virkkala, R., & Toivonen, H. (2009). Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions. *Biol Rev Camb Philos Soc, 84*(1), 39-54. doi:10.1111/j.1469-185X.2008.00060.x

Hill, M. P., Bertelsmeier, C., Clusella-Trullas, S., Garnas, J., Robertson, M. P., & Terblanche, J. S. (2016). Predicted decrease in global climate suitability masks regional complexity of invasive fruit fly species response to climate change. *Biological Invasions, 18*(4), 1105-1119. doi:10.1007/s10530-016-1078-5

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics, 23* (14), 1801-1806. doi:10.1093/bioinformatics/btm233

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24* (11), 1403-1405. doi:10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet, 11* , 94. doi:10.1186/1471-2156-11-94

Karsten, M., Jansen van Vuuren, B., Addison, P., Terblanche, J. S., & Leung, B. (2015). Deconstructing intercontinental invasion pathway hypotheses of the Mediterranean fruit fly (Ceratitis capitata) using a Bayesian inference approach: are port interceptions and quarantine protocols successfully preventing new invasions? *Diversity and Distributions, 21* (7), 813-825. doi:10.1111/ddi.12333

Karsten, M., van Vuuren, B. J., Barnaud, A., & Terblanche, J. S. (2013). Population genetics of Ceratitis capitata in South Africa: implications for dispersal and pest management. *PLoS One, 8* (1), e54281. doi:10.1371/journal.pone.0054281

Kikuchi, Y., Hayatsu, M., Hosokawa, T., Nagayama, A., Tago, K., & Fukatsu, T. (2012). Symbiont-mediated insecticide resistance. *Proc Natl Acad Sci USA, 109* (22), 8618-8622. doi:10.1073/pnas.1200231109

Kikuchi, Y., Hosokawa, T., & Fukatsu, T. (2007). Insect-microbe mutualism without vertical transmission: a stinkbug acquires a beneficial gut symbiont from the environment every generation.*Appl Environ Microbiol, 73* (13), 4308-4316. doi:10.1128/AEM.00067-07

Kikuchi, Y., Meng, X. Y., & Fukatsu, T. (2005). Gut symbiotic bacteria of the genus Burkholderia in the broad-headed bugs Riptortus clavatus and Leptocorisa chinensis (Heteroptera: Alydidae). *Appl Environ Microbiol, 71* (7), 4035-4043. doi:10.1128/AEM.71.7.4035-4043.2005

Kikuchi, Y., & Yumoto, I. (2013). Efficient colonization of the bean bug Riptortus pedestris by an environmentally transmitted Burkholderia symbiont. *Appl Environ Microbiol, 79* (6), 2088-2091. doi:10.1128/AEM.03299-12

Kourti, A. (2004). Estimates of gene flow from rare alleles in natural populations of medfly Ceratitis capitata (Diptera: Tephritidae).*Bull Entomol Res, 94* (5), 449-456. doi:10.1079/ber2004324

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol, 79* (17), 5112-5120. doi:10.1128/AEM.01043-13

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet, 8* (1), e1002453. doi:10.1371/journal.pgen.1002453

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25* (14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25* (16), 2078-2079. doi:10.1093/bioinformatics/btp352

Liu, S. H., Chen, Y., Li, W., Tang, G. H., Yang, Y., Jiang, H. B., . . . Wang, J. J. (2018). Diversity of Bacterial Communities in the Intestinal Tracts of Two Geographically Distant Populations of Bactrocera dorsalis (Diptera: Tephritidae). *J Econ Entomol, 111* (6), 2861-2868. doi:10.1093/jee/toy231

Malacrida, A. R., Marinoni, F., Torti, C., Gomulski, L., Sebastiani, F., Bonvicini, C., . . . Guglielmino, C. R. (1998). Genetic aspects of the worldwide colonization process of Ceratitis capitata. *Journal of Heredity* (6), 501-507.

Malacrida, A. R., Gomulski, L. M., Bonizzoni, M., Bertin, S., Gasperi, G., & Guglielmino, C. R. (2007). Globalization and fruitfly invasion and expansion: the medfly paradigm. *Genetica, 131* (1), 1-9. doi:10.1007/s10709-006-9117-2

Malacrida, A. R., Guglielmino, C. R. G., Gasperi, G., Baruffi, L., & Milani, R. (1992). Spatial and Temporal Differentiation in Colonizing Populations of Ceratitis capitata. . *Heredity, 69* , 101-111.

Malacrino, A., Campolo, O., Medina, R. F., & Palmeri, V. (2018). Instar- and host-associated differentiation of bacterial communities in the Mediterranean fruit fly Ceratitis capitata. *PLoS One, 13* (3), e0194131. doi:10.1371/journal.pone.0194131

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADpainter and fineRADstructure: Population Inference from RADseq Data.*Mol Biol Evol, 35* (5), 1284-1290. doi:10.1093/molbev/msy023

Martinez Arbizu, P. (2020). pairwiseAdonis: Pairwise multilevel comparison using adonis. R package version 0.4

Mason, C. J., Couture, J. J., & Raffa, K. F. (2014). Plant-associated bacteria degrade defense chemicals and reduce their adverse effects on an insect defoliator. *Oecologia, 175* (3), 901-910. doi:10.1007/s00442-014-2950-6

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol, 66* (2), 526-538. doi:10.1016/j.ympev.2011.12.007

Morrow, J. L., Frommer, M., Royer, J. E., Shearman, D. C., & Riegler, M. (2015). Wolbachia pseudogenes and low prevalence infections in tropical but not temperate Australian tephritid fruit flies: manifestations of lateral gene transfer and endosymbiont spillover?*BMC Evol Biol, 15* , 202. doi:10.1186/s12862-015-0474-2

Nikolouli, K., Augustinos, A. A., Stathopoulou, P., Asimakis, E., Mintzas, A., Bourtzis, K., & Tsiamis, G. (2020). Genetic structure and symbiotic profile of worldwide natural populations of the Mediterranean fruit fly, Ceratitis capitata. *BMC Genet, 21* (Suppl 2), 128. doi:10.1186/s12863-020-00946-z

Orantes, L. C., Monroy, C., Dorn, P. L., Stevens, L., Rizzo, D. M., Morrissey, L., ... & Helms Cahan, S. (2018). Uncovering vector, parasite, blood meal and microbiome patterns from mixed-DNA specimens of the Chagas disease vector Triatoma dimidiata. *PLoS neglected tropical diseases* , *12* (10), e0006730.

Papanicolaou, A., Schetelig, M. F., Arensburger, P., Atkinson, P. W., Benoit, J. B., Bourtzis, K., . . . Handler, A. M. (2016). 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, 17* (1), 192. doi:10.1186/s13059-016-1049-2

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics, 155* (2), 945-959.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res, 41* (Database issue), D590-596. doi:10.1093/nar/gks1219

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol, 67* (5), 901-904. doi:10.1093/sysbio/syy032

Raza, M. F., Yao, Z., Bai, S., Cai, Z., & Zhang, H. (2020). Tephritidae fruit fly gut microbiome diversity, function and potential for applications. *Bull Entomol Res, 110* (4), 423-437. doi:10.1017/S0007485319000853

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26* (1), 139-140. doi:10.1093/bioinformatics/btp616

Rochette, N. C., Rivera-Colon, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol Ecol, 28* (21), 4737-4754. doi:10.1111/mec.15253

Rosenberg, N. A. (2003). distruct: a program for the graphical display of population structure. *Molecular Ecology Notes, 4* (1), 137-138. doi:10.1046/j.1471-8286.2003.00566.x

Rudman, S. M., Greenblum, S., Hughes, R. C., Rajpurohit, S., Kiratli, O., Lowder, D. B., . . . Schmidt, P. (2019). Microbiome composition shapes rapid genomic adaptation of Drosophila melanogaster. *Proc Natl Acad Sci USA, 116* (40), 20025-20032. doi:10.1073/pnas.1907787116

Ruiz-Arce, R., Todd, T. N., Deleon, R., Barr, N. B., Virgilio, M., De Meyer, M., & McPheron, B. A. (2020). Worldwide Phylogeography of Ceratitis capitata (Diptera: Tephritidae) Using Mitochondrial DNA.*J Econ Entomol* , *113* (3), 1455-1470. doi:10.1093/jee/toaa024

Sandoval-Velasco, M., Jagadeesan, A., Ávila-Arcos, M. C., Gopalakrishnan, S., Ramos-Madrigal, J., Moreno-Mayar, V., . . . Schroeder, H. (2019). The genetic origins of Saint Helena's liberated Africans. *bioRxiv* 787515; doi: doi.org/10.1101/787515

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol, 75* (23), 7537-7541. doi:10.1128/AEM.01541-09

Schmidt, T. L., Chung, J., Honnen, A. C., Weeks, A. R., & Hoffmann, A. A. (2020). Population genomics of two invasive mosquitoes (Aedes aegypti and Aedes albopictus) from the Indo-Pacific. *PLoS neglected tropical diseases, 14* (7), e0008463.

Sciarretta, A., Tabilio, M. A., Lampazzi, E., Ceccaroli, C., Colacci, M., & Trematerra, P. (2018). Analysis of the Mediterranean fruit fly [*Ceratitis capitata* (Wiedemann)] spatio-temporal distribution in relation to sex and female mating status for precision IPM.*PLoS One, 4* , e0195097.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27* , 623-656. doi:10.1002/j.1538-7305.1948.tb00917.x

Simpson, E. H. (1949). Measurement of Diversity. *Nature, 163* , 688. doi:doi.org/10.1038/163688a0

Singh, B. K., Walker, A., & Wright, D. J. (2005). Cross-enhancement of accelerated biodegradation of organophosphorus compounds in soils: Dependence on structural similarity of compounds. *Soil Biology and Biochemistry, 37* (9), 1675-1682. doi:10.1016/j.soilbio.2005.01.030

Team, R. C. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

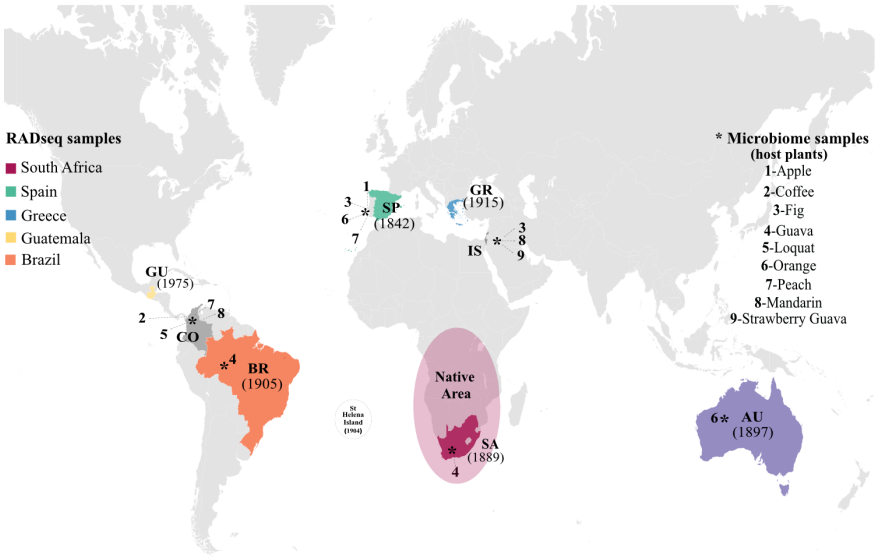
van Oppen, M. J., Bongaerts, P., Frade, P., Peplow, L. M., Boyd, S. E., Nim, H. T., & Bay, L. K. (2018). Adaptation to reef habitats through selection on the coral animal and its associated microbiome.*Molecular ecology* , *27* (14), 2956-2971.

Wang, H., Jin, L., & Zhang, H. (2011). Comparison of the diversity of the bacterial communities in the intestinal tract of adult Bactrocera dorsalis from three different populations. *J Appl Microbiol, 110*(6), 1390-1401. doi:10.1111/j.1365-2672.2011.05001.x

White, I. M., & Elson-Harris, M. M. (1992). *Fruit flies of economic significance: their identification and bionomics.* (Vol. xii). UK: CAB International Wallingford UK.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* : Springer-Verlag New York.

Xia, Y., Sun, J., & Chen, D. (2018). *Statistical Analysis of Microbiome Data with R* : ICSA Book Series in Statistics.



Biogeographic region	Country	Province	Area	N	Private allele	H_o	H_e	π	F_{is}	Tajima's D
Afrotropical	South Africa (SA)	Stellenbosh	Hex River Valley	10	907	0.1262 ± 0.003	0.1218 ± 0.003	0.1291 ± 0.003	0.0059 ± 0.031	-0.4549
Palaearctic	Spain (SP)	Valencia	Moncada	34	221	0.0451 ± 0.002	0.0520 ± 0.003	0.0529 ± 0.003	0.0265 ± 0.063	0.0757
	Greece (GR)	Thessaloniki	Thessaloniki	18	108	0.0421 ± 0.003	0.0445 ± 0.003	0.0460 ± 0.003	0.0127 ± 0.027	0.1456
Neotropical	Guatemala (GU)	Guatemala	Santa Barbara Mazatenanga	8	94	0.0358 ± 0.003	0.0306 ± 0.002	0.0330 ± 0.002	-0.0059 ± 0.026	0.0724
	Brazil (BR)	Bahia	Salvador	17	105	0.0551 ± 0.003	0.0538 ± 0.003	0.0560 ± 0.003	0.0027 ± 0.035	0.3613
Australasian	Australia (AU)	Western Australia	Perth	5	28	0.0314 ± 0.003	0.0327 ± 0.002	0.0373 ± 0.003	0.0117 ± 0.016	0.3309