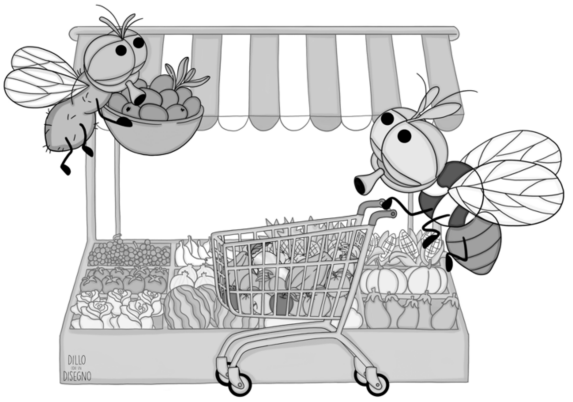
UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Comparative approaches to study the evolution of chemosensory gene families in *Bactrocera*



Federica Valerio

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXIII – A.A. 2017-2020

UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Comparative approaches to study the evolution of chemosensory gene families in *Bactrocera*

Federica Valerio

Supervised by Prof. Lino Ometto

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXIII – A.A. 2017-2020

Table of contents

Abstract

Phytophagous insects rely on chemoreception for various essential activities such as food source localization, mate choice, oviposition site selection and predator avoidance.

Insect chemoreception is mediated by gene families that include odorant binding protein (OBP) genes, chemosensory protein (CSP) genes, olfactory (OR), and gustatory (GR) receptor genes. In several insect species, it has been shown how the evolution of these gene families correlates with adaptation to specific ecological niches. Here we investigate how different degrees of host specialization in frugivorous insects affect the evolution of these genes. As a model group, we chose *Bactrocera*, a genus that comprises over 500 species and is one of the most economically relevant fruit fly genera, with at least 50 species considered important agricultural pests with a worldwide distribution. Species of this genus display an extremely wide range of feeding strategies ranging from extreme polyphagy to strict monophagy.

In this thesis, we first aimed to reconstruct a robust molecular phylogeny of *Bactrocera* species to obtain a useful framework for comparative genomics and comparative biology studies in this genus. For this purpose, we conducted comprehensive multi-locus phylogenetic and dating analyses using genome-wide data of eleven *Bactrocera* species with different feeding strategies. Our phylogenetic analyses clearly reveal that *B. dorsalis* is more closely related to *B. latifrons* than to *B. tryoni*. Our results are in contrast with phylogenies inferred from mitochondrial sequences, which instead supported a closer relationship between *B. dorsalis* and *B. tryoni*, but consistent with yet preliminary (i.e., not statistically supported) recent studies based on nuclear genes. Moreover, our analyses revealed numerous incongruences between gene and species trees, which we interpreted as evidence for incomplete lineage sorting and/or introgression and may explain the discordant results of the mitochondrial phylogenies. Overall, our findings underlined the importance of using genome-wide data to resolve complex phylogenies.

Our second objective was to untangle the evolution of chemoreception in *Bactrocera* species and relate it to the diverse host ranges of these fruit flies. In particular, we explored the evolutionary dynamics of chemosensory gene families in two extremely polyphagous species (*B. dorsalis* and *B. tryoni*), a polyphagous species with a more limited host range (*B. latifrons*), one oligophagous species (*B. cucurbitae*, which feeds mainly on Cucurbitaceae) and one monophagous species (*B. oleae*). We first identified and annotated the repertoire of chemosensory genes in the five species. Then, we analysed the birth and death processes of the four gene families using a comparative approach in a phylogenetic framework. Remarkably, we found that the monophagous *B. oleae* was the species with the lowest number of OBP and OR genes and duplication events,

while the extreme generalist *B. dorsalis* presents the highest number of OBP, OR, and GR genes and several duplications in these gene families. Moreover, the birth-and-death analysis detected that the olive fly was the only species that experienced only losses. Overall, these results were concordant with the idea that specialization in host preference may be accompanied by the loss of genes, and, in the case of the olive fly, this affected mainly the evolution of the OR gene family.

Finally, we analysed the pattern and rate of molecular evolution of these genes, which were then tested for the possible action of positive selection. The results revealed that the monophagous *B. oleae* was the species with the highest number of genes under selection, and only in this species OR and OBP genes are evolving under contrasting selective pressures.

Thus, the evolution of host choice in *Bactrocera* may be the result of a combination of evolutionary processes occurring at both the gene family and the gene sequence levels. Our findings further provide promising candidates for the genetic basis of the adaptation of the olive fly to its plant host.

Abbreviations

aa	amino acid(s)
AIC	Akaike Information Criteria
BD	Birth-and-death
BF	Bayes Factor
BR	Branch-specific
CDS	Coding DNA sequence
CSP	Chemosensory protein
d_N	Nonsynonymous substitution rate
d_S	Synonymous substitution rate
ESS	Effective sample size
FDR	False discovery rate
FR_t	Free Rates
GPCR	G-protein coupled receptor
GR_t	Global Rates
GR	Gustatory receptor
HSD	Honestly significant difference
IR	Ionotropic receptor
nt	nucleotide(s)
ME	Methyl eugenol
Mya	Million years ago
ML	Maximum Likelihood
OBP	Odorant Binding Protein
OR	Odorant Receptor
Orco	Odorant Receptor co-receptor
ORN	Olfactory Receptor Neuron
PCR	Polymerase Chain Reaction
piRNA	PIWI-interacting RNA
qPCR	Quantitative PCR
RBH	Reciprocal-best-hit
RT-PCR	Reverse transcriptase PCR
SD	Standard deviation
TE	Transposable elements

Introduction

In this thesis, we studied the evolution of chemosensory gene families in *Bactrocera* fruit flies using a comparative approach to better understand the genetic basis of their host preference.

To polarize the evolution of these genes and understand their role in the history of the single species, it is of crucial importance to have a robust and reliable phylogeny. For this reason, we also conducted comprehensive multi-locus phylogenetic and dating analyses.

To introduce the experimental chapters, the following literature review explores two broad areas of interest which are essential to understanding the theoretical and experimental background of my thesis: (i) the biology, ecology, and phylogeny of *Bactrocera* fruit flies with a focus on the species considered in this thesis; (ii) the chemoreception mechanisms in insects and, in particular, in *Bactrocera* species.

1.1. Tephritidae family

Tephritidae, commonly known as true fruit flies, are an incredibly diverse group of phytophagous insects and includes more than 5,000 species in over 500 genera [1–3], making it one of the largest families among Diptera.

Several tephritid species cause extensive damage to agriculture [4]. These crop pests are worldwide distributed and are mainly included in four major genera: *Anastrepha* (Schiner), *Ceratitis* (MacLeay), *Rhagoletis* (Loew), and *Bactrocera* (Macquart), one of the most diverse genera and which is the focus of my thesis [5].

Damages are the result of the particular reproductive habits of these flies, whereby females lay eggs within the fruit flesh (i.e., pericarp), and the larvae rapidly develop in the fruit, eating it and inducing bacterial and fungal decay. The consequent economic impacts resulting from such damages are considerable. For example, the olive fruit fly, *Bactrocera oleae* (Rossi), has been estimated to cause an annual economic loss of approximately $800 million in the Mediterranean basin alone [6]. In addition, indirect losses derived from quarantine restrictions imposed by importing countries to block the entry of fruit flies into unaffected areas can cause substantial economic losses [7]. In fact, an additional feature that increases the economic and social impact of these species is their ability to invade new areas and establish viable populations. Therefore, ecological aspects of tephritids such as host plant selection, oviposition, mating behaviour, and invasiveness have been extensively investigated, especially for economically significant pests [8–12].

1.2. *Bactrocera* genus

The genus *Bactrocera* comprises over 500 species [13] and is one of the most economically relevant fruit fly genera, with at least 50 species being considered important pests [1,5]. Members of this genus share key ecological features: high mobility, high fecundity, and relatively long adult life spans [14].

In this section, I will describe the main features of members of this genus, focusing on the five species that will be the subjects of this thesis: *B. cucurbitae*, *B. oleae*, *B. tryoni*, *B. latifrons,* and *B. dorsalis* (**Figure 1**). These species have been chosen because of the availability of genomic and transcriptomic resources and their contrasting feeding preferences.

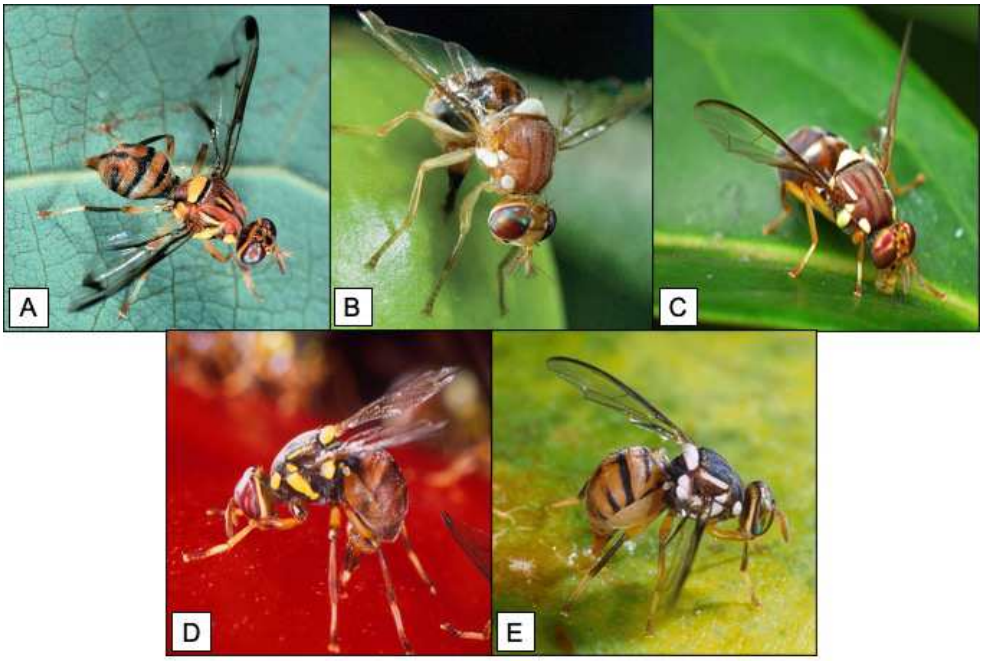


Figure 1 – A) *B. cucurbitae* [228], B) *B. oleae*, C) *B. tryoni* (Photo by James Niland), D) *B. latifrons* [329], E) *B. dorsalis* (Photo by Scott Bauer).

1.2.1. Phylogeny: state of the art

The *Bactrocera* genus, like other genera of the Tephritidae family, has a controversial history of taxonomy and classification. The taxonomic status of this group has been revised many times, and it was established as a genus by Drew only in 1989 [13].

The genus is subdivided into ~ 30 subgenera [15–17], and sometimes within those subgenera species that appeared morphologically highly similar are placed in species complexes (e.g., *B. dorsalis*, *B. tryoni,* and *B. musae* complexes) [13, 18–20]. Changes in the classification involved *Bactrocera* subgenera and the species complexes. For example, several *Bactrocera* species were formerly identified as *Dacus* (e.g., *Dacus dorsalis*, *D. oleae*, *D. latifrons*, *D. tryoni*). Currently, the use of the name *Dacus* was restricted to those members of the tribe Dacini (composed of *Dacus* and *Bactrocera* genera) that have their abdominal terga fused into a single sclerotized plate, while in *Bactrocera* fruit flies the terga are separated [18, 21].

In the last years, the use of molecular data to investigate the phylogeny resulted in relationships that contrast with those obtained using morphological diagnostic data, and it became clear that morphology-based classification had to be revised [21–25].

To date, molecular phylogenies among *Bactrocera* subgenera and within specific subgenus showed different relationships depending on the type of molecular markers (nuclear and/or mitochondrial), on the number of markers, and on the number of taxa analysed [22, 24, 26–29]. In particular, recent phylogenetic studies proposed that the *Bactrocera* (*Zeugodacus*) subgenus had to be elevated to genus level because it is more closely related to the *Dacus* genus than to the *Bactrocera* genus [22, 24, 29, 30]. These analyses supported previous studies on the morphological and ecological differentiation between *Bactrocera* (*Zeugodacus*) and the other *Bactrocera* subgenera [15, 16]. However, because of the ongoing debate on the genus/genera definition, we will use the unambiguous nomenclature and refer to the species used in this study as members of the genus *Bactrocera*.

Besides such nomenclature issues, what is more relevant to our analyses is a correct phylogenetic inference among the studied species. For instance, *B. dorsalis* is usually considered to be more closely related to *B. tryoni* than to *B. latifrons* (e.g., [31, 32]), which may affect the interpretation of shared or diverged characters, as well the implementation of control measures developed for putative “closely related” species.

In our phylogenetic analysis, we included *Bactrocera* species for which genomic and/or transcriptomic resources were available, trying to incorporate species of different subgenera. Our dataset includes members of the subgenus *Afrodacus* (*B. jarvisi*), *Daculus* (*B. oleae*), *Zeugodacus* (*B. cucurbitae*), *Tetradacus* (*B. minax*) and *Bactrocera* (*B. bryoniae*, *B. correcta*, *B. dorsalis*, *B.latifrons*, *B. musae, B. tryoni* and *B. zonata*).

1.2.2. Geographical distribution

The *Bactrocera* genus has a wide ancestral geographical distribution that extends throughout tropical Asia, the South Pacific, and Australia

(**Figure 2**). In fact, several species have expanded their range and have become invasive in Africa, Europe, and the Americas.

Bactrocera oleae is one of the few species that have been found in Africa, Europe, and America. This fruit fly originated in Africa and has been established in the olive-growing regions worldwide [33]. Its current distribution includes South and Central Africa, Mediterranean Europe, Near and Middle East, and it has been recently introduced to California (USA), Hawaii (USA), and Mexico [33–35].

The oriental fruit fly, *B. dorsalis* (Hendel), is native to South-east Asia [36] but, thanks to its invasiveness, it has established in a considerable number of other countries, such as the Hawaiian Islands [37], East Africa [38], West Africa [39] and Italy [40].

Bactrocera cucurbitae, native to India, is abundant throughout Asia, Oceania, and the African continent [16, 41].

Bactrocera latifrons has also an Asian origin and is predominantly distributed in South and South-east Asia. In Africa, it has only been recorded from Tanzania and Kenya [42]. This fruit fly was also introduced to Hawaii [43, 44].

Bactrocera tryoni, also known as the Queensland fruit fly, is an Australian pest, and its native range was considered to be the tropical and subtropical coastal Queensland and northern New South Wales [45]. However, it has been widely established in eastern Australia and some South Pacific islands [46].

Indeed, the fact that all these five species have enlarged their original range, coupled with their feeding habits, is what makes them among the most economically important invasive pest species in the world.

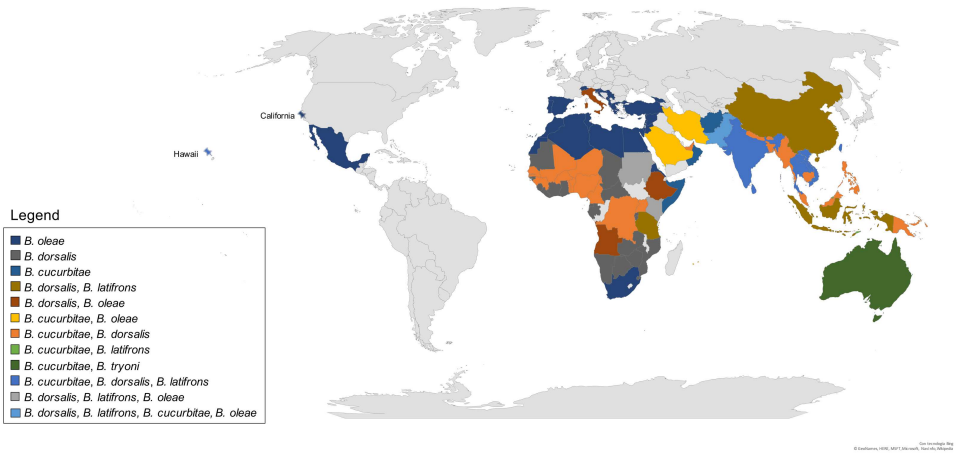


Figure 2 – Geographical distribution of *B. cucurbitae*, *B. oleae*, *B. tryoni*, *B. latifrons* and *B. dorsalis*.

1.2.3. Biology

Bactrocera species have very similar life cycles, differing in few life-history traits, such as the number of eggs produced, the number of generations per year, and, most importantly, the host plant range [47]. As in other insects, life cycle can be divided into five stages: egg, larva, pupa, teneral, and the reproductive adult (**Figure 3**)[48].

Males mate multiple times during their breeding season, while females mate once or a limited number of times in their lives. The gravid female then searches for the fruit where it lays either a single egg or a clutch of eggs under or into the skin of the fruit. The host fruit is generally at the mature or ripening stage, but several exceptions have been reported. Larvae feed inside the fruit, causing direct fruit damage and inducing bacterial decay. Larvae go through three instars before leaving the fruit for pupation, which most frequently occurs in the soil. Teneral adults emerge from the soil after 24–48 hours. Sexual maturation can take an additional 6 to 30 days, depending on the species and environmental conditions [48], as detailed below for the five species considered in this thesis.

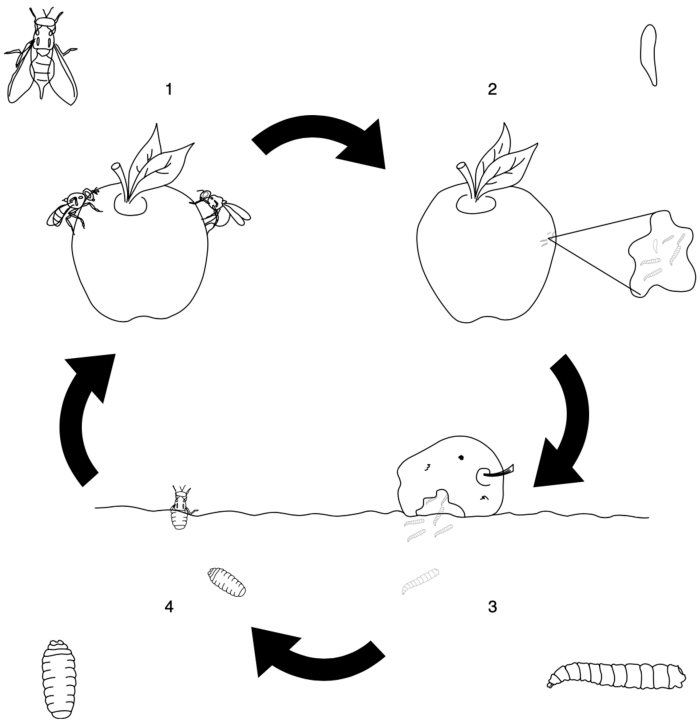


Figure 3 – *Bactrocera* fruit fly life-cycle (from [48]).

The olive fly biology is closely linked to the availability and condition of olives. *Bactrocera oleae* females typically lay one egg under the surface of an olive fruit [49, 50], in whose mesocarp the developing larva then feeds and grows. Adults can feed on several organic sources, including insect honeydews, fruit exudate, plant nectar, and pollens, but, they may also require other types of nutrients, such as bird excrement, bacteria, and yeasts [51–53]. Adult food availability is fundamental for the survival of the populations during periods when olive fruits are unsuitable for oviposition. For this purpose, adult flies can exploit many different plants as adult food sources [54]. Olive fly development in field conditions is strictly dependent on temperature, humidity, microclimate within the olive canopy, and on the availability and quality of the fruit [55, 56].

Besides fruit host preference, there are only few differences in the reproductive biology of the other *Bactrocera* species. For example, *B. cucurbitae* females mainly oviposit in the fruit pulp, but they can also lay eggs in the corolla of flowers and sometimes even into the stem and root tissues [57]. A more glaring difference is found in the number of eggs per clutch that females of different species lay into the hosting fruit. While the olive fly generally lays a single egg per olive, *B. cucurbitae*, *B. tryoni*, *B. latifrons* and *B. dorsalis* may oviposit more eggs per clutch [48].

The number of generations per year also varies among these species. In Southern Europe and Middle East, *B. oleae* is considered to be a “short-day species”, since females can not mature ovaries during late spring and early to mid-summer, probably because of high temperatures and low humidity [50]. As a result, the number of annual generations within the endemic range presents several variations depending on climate: three to five generations have been observed in the Middle East [58–61] and two to five in different parts of Europe [62–65].

Bactrocera cucurbitae is active throughout the year, with a number of generations ranging from 8 to 10 per year [5, 66, 67]. Similarly, *B. dorsalis* is reported to have up to 11 generations per year, but they usually range between 4 and 8, depending on the temperature [68]. Temperature-based development models for *B. tryoni* predicted this number to range from only one generation per year in temperate Tasmania up to 12–15 generations in far northern subtropical Queensland [69]. For *B. latifrons*, the number of generations at 26.6 °C was reported to range between 7 and 8 [70].

1.2.4. Host range

Phytophagous insects are generally grouped according to their host range into three broad classes: i) monophagous, or specialist, insects feed on only one plant species; ii) oligophagous insects on a limited range of related plant species; and iii) polyphagous, or generalist, insects feed on a broad range of unrelated plant species [71, 72]. The definition of these terms is not always in agreement. Monophagy can be intended as feeding on a single plant species [71], as well as feeding on different plant species

within a genus [73] and that share precise fruit features. In this thesis, monophagy will follow the second definition.

Several advantages and disadvantages have been proposed associated with the different host ranges. According to Bernays and Graham, the principal advantage of polyphagy over monophagy is that polyphagous insects are not dependent upon a single plant species for their survival [74]. On the other hand, they are thought to be inefficient because they need to exploit multiple resources [75, 76]. According to other theories, specialization is considered to lower interspecific competition with generalist species, or conversely, to increase the ability to compete against generalists for a specific source [77, 78]. This is supported by the observation that most phytophagous insects feed on only one or a small number of related plants [71, 79].

Studies on host preference have classified host plants for about 200 *Bactrocera* species [19–21]. This genus appears to be composed of a high number of generalist species, 11 of which, including *B. dorsalis* (Hendel), *B. correcta* (Bezzi) and *B. tryoni* (Froggatt) [19–21] show an extreme polyphagy [82], with hosts in more than 20 plant families. Most of the species utilize instead as hosts only members of a plant family and are thus classified as oligophagous species [83]. Finally, only a few species are classified as monophagous species. However, few resources are currently available on the host use and oviposition behaviour of these fruit flies [84].

Below we report host preference for the five species studied in this thesis. The olive fruit fly is a specialist that feeds only on fruits of the genus *Olea*, including *Olea europaea* ssp. *europaea* (cultivated and wild), *O. europaea* ssp. *verrucosa*, and *O. europaea* ssp. *cuspidata* [6]. This preference is expected to have led to specific adaptations to locate and metabolize olives.

Bactrocera cucurbitae, also known as the melon fly, is an oligophagous species that feeds on species of the Cucurbitaceae family but was occasionally detected on members of other plant families. However, these infestations were minor [85] and associated with the lack of optimal host plants.

Bactrocera latifrons attacks mostly species of the Solanaceae family, but it has been shown that other families can be rarely infested (e.g., Cucurbitaceae) [36]. Despite the narrow host range of this species, it has been commonly considered as a generalist species [86].

The other two species are, on the other hand, clear examples of extremely polyphagous species: *B. tryoni* feeds on at least 234 species of plants that belong to 49 plant families [80] and *B. dorsalis* has over 300 host species [36, 80, 81, 87].

The precise mechanisms of the host plant localization by the fruit flies remain to be elucidated. It is known that the selection of the oviposition site by phytophagous insects involves multiple sensory modalities,

including visual, volatile and non-volatile cues, and contact chemical stimuli from host and non-host plants. Thus, the evolution of the genes associated with such perceptions is expected to be tightly associated with the host selection and preference. In particular, we hypothesized that genes involved in the detection of chemical signals emitted by host plants are good candidates for identifying the genetic basis of host choice.

1.3. Insect chemoreception

The reception of chemical signals is crucial for the survival of almost all organisms [87–90], and most terrestrial animals have evolved chemosensory systems with remarkable sensitivities and discriminatory power to detect and process chemical compounds in the environment [92].

In particular, insects rely on chemoreception to perceive chemical signals, localize food sources, oviposition sites, predators, and identify mating partners. Insect chemoreception starts with the primary detection of a chemical molecule by the sensillum, a cuticular structure housing a receptor neuron [93–96]. The main steps include: (1) the uptake of the chemical signal from the external environment by the sensillum, (2) the transport through the sensillum lymph towards the dendritic membrane of the receptor neuron, and (3) the interaction with the chemoreceptor that leads to the activation of the chemosensory neuron.

A pivotal role in these processes is played by proteins encoded by five multigene families (**Figure 4**) (reviewed in [97, 98]), two of which code for receptors characterized by seven transmembrane domains, namely the gustatory receptors (GRs) [99–101] and the odorant receptors (ORs) [100, 102–105]. A third family encodes for ionotropic receptors (IRs), a class within the highly conserved ionotropic glutamate receptor family of ligand-gated ion channels [106–108]. These three families of transmembrane receptors can perceive and differentiate a wide range of semiochemicals [102, 109–111]. However, chemical compounds are generally hydrophobic and therefore need to be solubilized and transported from the external environment to the membrane of chemosensory neurons. Two families of soluble proteins have been proposed to be involved in this process: odorant binding proteins (OBPs) [112, 113] and chemosensory proteins (CSPs) [114, 115].

In this thesis, we refer to these five gene families (ORs, GRs, IRs, CSPs, and OBPs) collectively as chemosensory gene families.

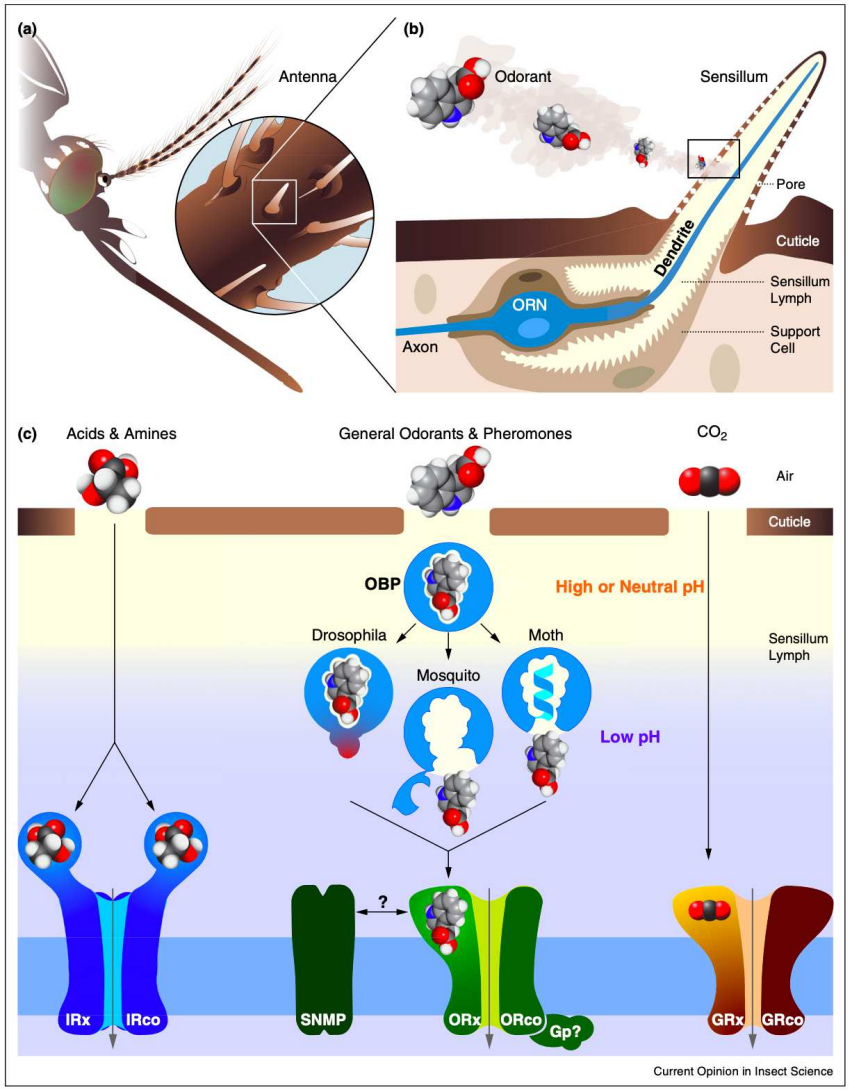


Figure 4 – Schematic representation of insect chemosensory organs and molecular models in signal transduction (from [330]). (A) Insect chemosensory appendages such as the antennae are covered by sensilla. (B) An olfactory sensillum housing an Olfactory Receptor Neuron (ORN, in blue); odorants penetrate the olfactory sensillum through cuticular pores and are detected by the ORN dendrite. (C) Different chemoreceptor families are activated by different classes of odorant molecules via distinct mechanisms. See main text for details.

1.3.1. Odorant binding proteins

Odorant binding proteins are present in both insect sensillum lymph and in the vertebrate olfactory mucus [112, 113]. Although they seem to play a similar role in olfaction, insect, and vertebrate OBPs are not homologous and represent an example of convergent evolution [116, 117]. In the last years, insect odorant binding proteins have drawn significant attention because they are highly expressed in the antennae [118], confirming that they may play a crucial role in odour perception, they have an extreme sequence divergence between and within species [119, 120] and they are coded by many genes in some insect species (e.g., 52 different OBPs are encoded in *Drosophila melanogaster*) [121, 122].

Insect OBPs are small globular proteins (about 135–220 amino acids long) that generally bind, solubilize, and shuttle the odorants that reach the sensillum pore across the aqueous lymph to odorant receptors in the dendrites [123–127].

Despite the high degree of sequence divergence, insect OBPs present a specific structure that distinguishes them from similar proteins. In particular, the typical OBP consists of six α-helical domains characterized by a pattern of six positionally conserved cysteines joined into three interlocked disulphide bridges **(Figure 5)** [126, 128–130]. The OBP family comprises members with a smaller (C-minus) or higher number (C-plus) of cysteines, as well as atypical OBPs containing additional domains [131–134].

These proteins are synthesized by non-neuronal auxiliary cells, trichogen and tormogen cells, and subsequently secreted into the sensillar lymph [112, 135, 136].

Different types of OBPs may be expressed within the same species in different olfactory sensilla [115, 137], but they are not restricted to the olfactory system. Indeed, some are expressed in the tissues involved in taste perception [138, 139] or in other larval chemosensory organs [138, 140]. In fact, while typically OBPs act as solubilizers and carriers of hydrophobic compounds, in particular odorants and pheromones, other roles have also been proposed. For instance, they can be involved as scavengers to remove the pheromones as well as foreign semiochemicals in order to maintain receptor activity [116, 141]. Another possible role is to function as a filter to decrease odorant concentration that, if too high, can lead to long-term receptor desensitization [116]. In addition, OBPs can be active participants together with the pheromone in a specific ternary association in OR activation [142], although their precise role remains unclear. Finally, as mentioned above, the location of OBPs in different insect body parts also suggests different roles from olfaction [143, 144].

In the last years, much information has been obtained on OBP binding affinity and molecular docking to chemical compounds as well as several three-dimensional (3D) crystal structures. Recent studies have also

reported that different OBP types selectively bind defined, partly overlapping spectra of compounds [145].

1.3.2. Chemosensory proteins

Chemosensory proteins (CSPs) are another family of small soluble proteins found in insect chemosensory organs, namely olfactory and gustatory tissues, from which they take the general name [97, 114, 115, 146].

CSPs are smaller (about 100–120 residues) and do not share sequence similarity with OBPs. They have a conserved motif characterized by five α-helical domains formed by four cysteines joined by two disulphide bonds between neighbouring residues **(Figure 5)** [146–148].

To date, two observations support a role of some CSPs in transporting semiochemicals from the environment to the membrane of sensory neurons: the presence of CSPs in the lymph of several chemosensilla and the affinity to pheromones and odorants [146, 149–152]. However, there is no clear evidence of whether they are essential in insect chemosensation.

Other CSPs functions have been prosed, including pheromone delivery, development, solubilization of nutrients, vision, immune response, and insecticide resistance [153–160].

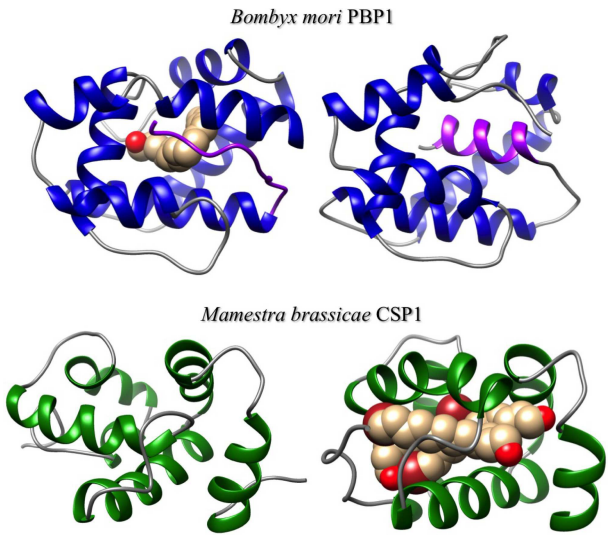


Figure 5 – Crystal structures of the first solved OBP and CSP proteins. Ribbon view representation of the structure of *Bombyx mori* (Lepidoptra) pheromone binding protein (PBP1) [126] and *Mamestra brassicae* (Lepidoptera) chemosensory protein 1 (CSP1) [148] (from [158]).

1.3.3. Odorant receptors

Insect ORs are seven transmembrane domain proteins, with lengths ranging from 350 to 500 amino acids.

They do not appear homologous to their vertebrate counterparts. In fact, they have opposite membrane topology of the vertebrate G-protein coupled receptors (GPCRs), with their N-termini internal to the cell and external C-termini [161, 162].

Insect ORs function as ligand-gated ion channels and are heteromeric complexes formed by two subunits, one of which, the so called OR co-receptor (Orco), is ubiquitarian and highly conserved, while the second is highly variable and divergent and has been proposed to confer odour binding specificity [161–167].

In *Drosophila*, there are two olfactory organs, the antenna and the maxillary palp. Both organs are covered by sensilla that contain the dendrites of up to four olfactory receptor neurons (**Figure 6**). The majority of ORNs of basiconic and trichoid sensilla express a single member of the OR family that provides the odorant response profile of the neuron [168, 169]. They also express Orco that is essential for targeting the heterodimer to the dendritic membrane [163].

OR ligands have been described for most of the *D. melanogaster* and *Anopheles gambiae* ORs, thanks to the development of heterologous *in vitro* and *in vivo* expression systems. More specifically, cultured cell lines and *Xenopus* oocytes are used for *in vitro* analyses, while the “empty neuron system” of *Drosophila* is used for the *in vivo* expression analyses [169–174]. In these deorphanisation studies, panels of odorants are tested against OR/Orco complexes in the *Xenopus* oocytes and two-electrode voltage-clamp method [175, 176] or in neuron mutants (empty) where the endogenous receptor gene Or22a is deleted and replaced by selected odorant receptors using the GAL4/UAS system, thus allowing the measurement of their response profiles [169, 170]. These approaches have revealed a specific difference in odour space covered by *D. melanogaster*, predominantly tuned to esters, compared to *An. gambiae*, which is mainly focused on aromatics. It represents a striking example of how these chemoreceptor genes have evolved to detect odorants important for the ecology of these species. For example, females of *An. gambiae*, which require a blood meal for oviposition, use aromatic compounds found in sweat to locate humans [172].

1.3.4. Gustatory receptors

Insect gustatory receptors are proteins of about 350–500 amino residues in length and contain seven transmembrane domains.

GRs are members of a large GPCR family distantly related to the insect olfactory receptors and unrelated to vertebrate taste receptors. As ORs, they have an inverted membrane topology, i.e., with the C-terminus at the

extracellular surface [161, 177]. However, insect GR topology is not as clearly established, and much is left to be discovered.

GRs are mainly expressed on the membrane of gustatory neurons (GRNs), which are housed in gustatory sensilla located in different insect body parts (e.g., mouthparts, legs, pharynx, and wings) [89, 178]. However, GRs are also expressed in neurons of other body tissues such as the digestive tract, reproductive organs, epidermal cells on the abdomen, in the brain, and in the antenna, either in specific sensilla or olfactory neurons (**Figure 6**) [179–188].

In the past few years, many attempts have been made to enlighten the molecular functions of GR proteins, but currently, only a small number have been deorphanized to allow a proper functional characterization. From these experiments, it appears that the GR family includes several subfamilies that mediate the perception of different ligands: carbon dioxide [187, 189], several sugars [186, 190–192], various bitter compounds [193] and cuticular hydrocarbons [194]. Moreover, GRs appear to be involved not only in the detection of chemical compounds but also in thermosensation [195].

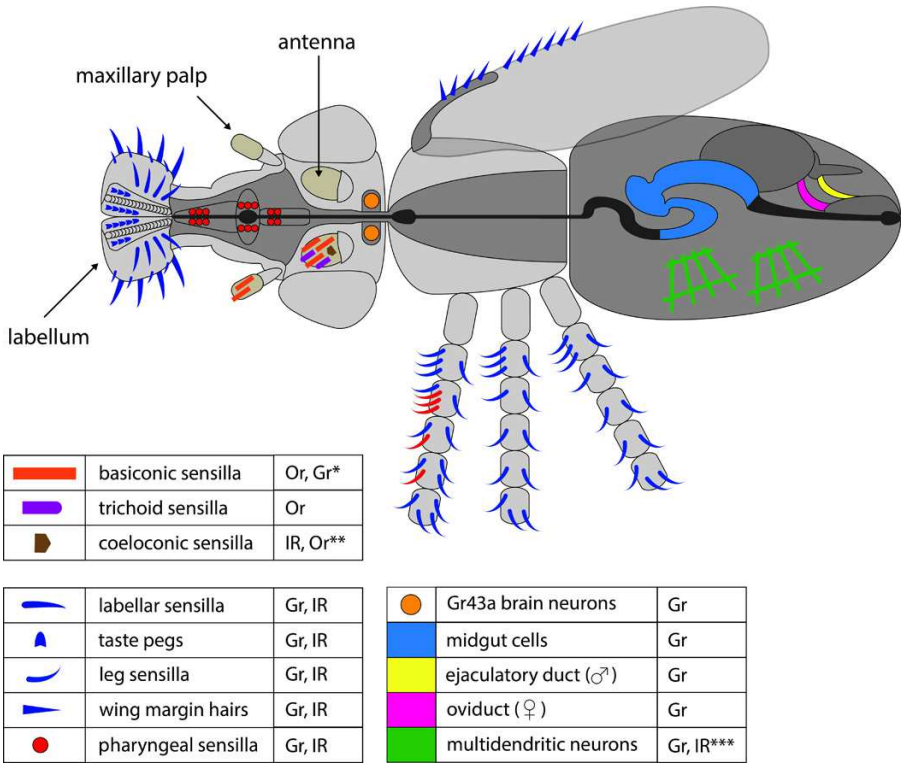


Figure 6 – Expression of the chemoreceptor families in *D. melanogaster* (from [180]).

1.3.5. The evolutionary history of chemosensory gene families

Recent studies have uncovered the evolutionary history of the chemosensory gene families. GRs have been found to be the most ancient of all the eumetazoan chemoreceptors, as revealed by the inferred presence of GR or GR-Like genes in the common ancestor between arthropods and the phylum Placozoa [196–198]. On the other hand, CSPs appear to be an arthropod-specific gene family that emerged after the divergence between the phyla Arthropoda and Onychophora [119, 157, 197]. Finally, ORs and OBPs are present only in the Hexapoda while lacking in all other arthropod taxa (**Figure 7**) [100, 119, 157, 197, 199].

Regarding sequence similarity, GR sequences have an extremely low sequence identity between insect species, and even between paralogues and orthologues within the same species. Moreover, there are no conserved domains among insect GR protein sequences [100, 198]: for example, in *D. melanogaster,* most GRs share less than 8% of their amino acidic sequence [100].

OBPs also appear to be incredibly divergent in their sequences, and the amino acidic identity among paralogs within a species and among species members may be even lower than 10% [157].

CSPs are more evolutionarily conserved than OBPs, with around 50% of sequence identity between orthologues from phylogenetically distant species [200].

Contrary to Orco, which is almost invariant in sequence even across distant insect species, the other OR subunits are highly divergent both in terms of primary amino acid sequence and in their number within the gene family. Indeed, ORs share on average 20% amino acidic sequence between OR members, either within or across species [201].

Beside sequence divergence, chemosensory gene families also differ in the number of genes they contain across species. For instance, OR family size is highly variable, ranging from four members in the damselfly to more than 350 in some ants [202, 203].

A similar trend in gene family size is observed for OBP, which ranges from 13 genes in the silkworm *Bombyx mori* to 66 in the malaria mosquito *Anopheles gambiae* [47, 53, 54, 125–127]. These data leave an open question about the functional importance of these proteins for the specific recognition of odorants. They also may indicate that insects may have evolved different strategies to detect general odour or sex pheromones, namely by divergent selection in the sequence or in the number of OBPs and ORs genes.

Accordingly, GRs also present a variable family size that goes from 10 members in *Apis mellifera* to 220 GR proteins in *Tribolium castaneum* [204, 205].

In contrast, the number of CSP genes tends to be low within arthropod genomes relative to other chemosensory gene families, ranging from one

gene in the tick *Ixodes scapularis* (Ixodida, Arachnida) to 19 genes in *Tribolium castaneum* (Coleoptera, Insecta) [97, 197, 206].

Multiple studies report evolutionary patterns that suggest that chemosensory genes are evolving according to the “birth-and-death” (BD) model [97, 122, 197, 207, 208]. Under this model, new genes are created by gene duplication, and while some of these genes are maintained in the genome for a long time, others become non-functional through the accumulation of deleterious mutations or are lost [209]. In agreement with this model, several studies on chemoreceptors reported lineage-specific expansions or contractions across arthropods [122, 197] as well as a high level of divergence in family composition even between closely related species [97]. These findings suggest that gene family dynamics may be involved in the adaptation to specific ecological niches. In particular, the ability to detect particular odours or tastes can be associated not only to the properties of a given chemosensory gene (i.e., its sequence and thus protein structure and function), but also to the presence of paralogous genes coding for proteins with distinct functions.

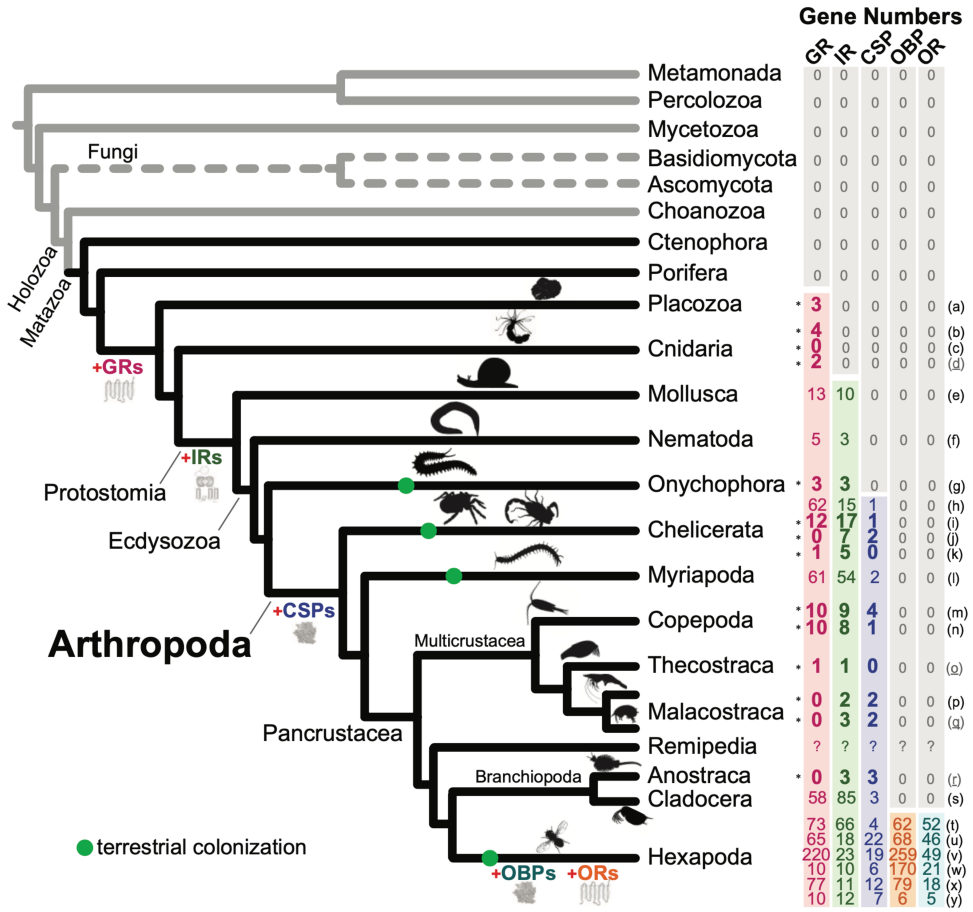


Figure 7 – Evolution of chemosensory gene families in the Arthropoda (from [197]). Numbers of gustatory receptors (GR), ionotropic receptors (IR), chemosensory proteins (CSP), olfactory receptors (OR), and odorant binding proteins (OBP) genes for representative species are reported in the right columns. The appearance of the different gene families is indicated in the phylogeny.

1.3.6. Chemoreception in *Bactrocera*

The molecular components involved in chemoreception have been less studied in the *Bactrocera* genus compared to the model *Drosophila* genus.

Some preliminary and incomplete studies aimed to identify genes involved in chemosensory perception in some *Bactrocera* species, including OBPs, CSPs, ORs, and GRs. Putative OBP genes have been identified using the Expressed Sequence Tag approach in *B. dorsalis* [210, 211]. In these studies, it was also investigated the transcriptional profile of OBP genes. Other putative chemosensory proteins, mainly OBPs, have

been later found by RNA-seq in *B. dorsalis*, *B. cucurbitae*, *B. minax,* and *B*. *oleae* [212–217].

The proteomics profiling of the antennae was analysed in *B. dorsalis* to identify differentially expressed proteins, including OBPs [218].

The OBP role in odour detection has been investigated by functional analyses. In these studies, it was evaluated the gene expression profiles of different tissues, developmental stage in response to sexual maturation, pre-, and post-mating. RNA interference, electrophysiological and behavioural experiments were then applied to determine the involvement of specific OBP targets in odour detection [218–220].

For example, in the oriental fruit fly, five OBPs and one CSP highly expressed in the antennae were tested to determine whether there was a correlation between chemosensory proteins specifically or predominantly expressed in olfactory tissues and behavioural changes [221]. In this work, the authors performed ligand-binding assays using 13 attractant chemical compounds. They showed that BdorOBP83a-1 and BdorOBP83a-2 had the highest affinity to all tested semiochemicals, while the CSP did not exhibit binding ability. Interestingly, the knockdown of BdorOBP83a-2 resulted in a decrease of both the neuronal responses to the tested chemicals and of the associated behavioural responses [221].

Regarding OR, in a recent study where ten *B. dorsalis* ORs were co-expressed with the co-receptor BdorOrco in *Xenopus laevis* oocytes it has been demonstrated that some of the tested ORs respond to plant volatiles (i.e., 1-octen-3-ol, geranyl acetate, farnesenes, and linalyl acetate) [222] or to methyl eugenol (ME) [223].

The function of Orco was also investigated in *B. oleae* [224]. In this work, the authors performed transient knockdown via RNA interference gene silencing in adult olive flies to determine the role of the receptor in mating and oviposition behaviour. They demonstrated that the knockdown of Orco expression in both males and females reduced mating ability. In addition, they showed that Orco silencing caused complete inhibition of oviposition [224].

While these studies provide useful information on the molecular and biological role of (a limited set of) chemosensory proteins in *Bactrocera*, further investigations are required to understand their involvement in host selection.

Aim of the research

As discussed in the introduction chapter, fruit flies of the genus *Bactrocera* are among the most economically relevant insects. This genus comprises species exhibiting very diverse host preferences, ranging from strict monophagy to extreme polyphagy. In several insect species, it has been proposed that changes in chemosensory gene families may be linked to the host adaptation of closely related species [225, 226].

On this basis, we outlined two main objectives.

- We aimed to provide a robust multi-locus phylogeny to facilitate and empower comparative studies between *Bactrocera* species. In particular, we wanted to infer the phylogenetic history and estimate the divergence time of the different *Bactrocera* species using a phylogenomic dataset since the available phylogenesis presented discordant topologies. Comprehensive and robust phylogenetic inferences are crucial to obtain reliable results on the evolutionary history of chemosensory genes.

- Our second aim was to elucidate the genomic basis and the evolution of chemoreception in *Bactrocera* species and link them to differences in their host selection biology. In particular, we focused on five *Bactrocera* species, which are important agricultural pests and are characterized by different host range and for which genomic and transcriptomic data are available. We therefore aimed at first identifying and annotating all chemosensory orthologous genes across these species. Next, we wanted to study the dynamics of the chemosensory gene families and correlate that to the host preference. Finally, we were interested in detecting genes under positive selection, and that could be important for the adaptation to the species-specific feeding preferences. To address this aim, i) we annotated the chemosensory gene repertoire; ii) we studied the evolution of these gene families within a phylogenetic framework and iii) we characterized the genomic events, the genetic changes, and the selective forces that occurred during the evolutionary history of the genus.

Material and Methods

2.1. Multi-locus phylogeny of *Bactrocera* genus

2.1.1. Datasets

We analysed datasets of coding sequences (CDS) and of RNA (transcript) sequences for *Ceratitis capitata* [227], *B. cucurbitae* [228], *B. dorsalis* [229], *B. latifrons* (NCBI accession: MIMC00000000.1), *B. minax* [210], and *B. oleae* [230]. For the six remaining *Bactrocera* species (*B. bryoniae*, *B. correcta*, *B. jarvisi*, *B. musae*, *B. tryoni*, and *B. zonata*), we downloaded the available RNA-Seq raw reads (see Supplemental **Table S1** for SRA accession numbers) and assembled the corresponding transcriptomes using default parameters with Trinity v. 2.7.0 [231].

2.1.2. Assessment of gene data set completeness

We evaluated the quality of the genomic and transcriptomic data and of the annotated gene set (CDS) by assessing their completeness with the BUSCO tool suite v4.1.2 [232]. Because we searched for orthologs in all datasets (CDS, transcriptome, genome) and wanted to assess the probability of false negatives (i.e., absence of an ortholog in a given species), we analysed each dataset independently but also in combined datasets which included all available sequences for each species.

For our analyses, we used the BUSCO set diptera_odb10 (2020-10-16) and the third-party components tBLASTn v2.10.1 [233], Augustus v3.3.3 [227], and HMMER3 v3.3.1 (available at http://hmmer.org/).

2.1.3. Orthologous gene set identification

Orthologs across the eleven *Bactrocera* species and the outgroup *C. capitata* were identified using a reciprocal-best-hit (RBH) approach using pairwise BLASTn searches [233] between each *Bactrocera* dataset and the *C. capitata* CDS sequences.

Putative 1:1 orthologs were first aligned using MAFFT [235] and any incomplete codon (based on the *C. capitata* sequence) was removed. We then re-aligned the ortholog sets using the PRANK algorithm [236] implemented in the tool TranslatorX [237], which aligns protein-coding sequences based on their corresponding amino acid translations. We minimized bias in our datasets by 1) removing alignments containing sequences with internal stop codons and 2) using a custom perl script to remove problematic and ambiguous alignment regions, with an approach similar to that proposed by Han and colleagues [238] (see also [239]).

Using this pipeline, we ultimately identified 110 orthologous gene sets across all the twelve species.

2.1.4. Phylogenetic analyses

We analysed phylogenetic relationships among species using either a maximum likelihood (ML) or a Bayesian approach. We run ML analyses on both the concatenated aminoacidic alignment (63,297 amino acids, aa), using the PROTGAMMAGTR model, and on the concatenated codon alignment (189,891 nucleotides, nt) partitioned in first, second and third (1+2+3) position, using the GAMMAGTR model implemented in RAxML (Randomized Axelerated Maximum Likelihood; [240]). In a third case, we also run a ML analysis based only on the 4-fold degenerate sites (24,885 nt) using the GAMMAGTR model. In all cases, node support was calculated by the rapid bootstrap feature of RAxML (100 replicates). We also estimated bootstrap supports using the coalescent-aware analysis of ASTRAL [241], which was based on all single ML gene trees obtained by RAxML using the same models of the concatenated analyses for either the nucleotide or the aminoacidic sequences. Bootstrap values were estimated by performing either 100 multi-locus bootstrap replicates or gene+site resampling (using the -g option).

The same three datasets (concatenated amino acids, separate codon positions, and 4-fold degenerate sites) were used to run Bayesian analyses in Beast v.2.5.1 [242]. The aminoacidic dataset was analysed with a Blosum62+G4 substitution model. The 4-fold degenerate sites dataset and the codon dataset were analysed with a GTR+G4 substitution model. The codon dataset was split into three partitions, corresponding to the codon positions, setting linked trees across them. To investigate the discordance between the results obtained using the 4-fold degenerate sites dataset and the other datasets (see Results section), we run an additional Bayesian analysis on the 4-fold degenerate sites dataset with PhyloBayes [243]. We run two independent MCMC chains using the CAT model with gamma distribution and checked for convergence using the associated tracecomp and bpcomp commands. We let both chains run until parameters were stabilized with *maxdiff* =0, *reldiff* < 0.25 (with the exception of stat, with *reldiff* < 0.4) and the effective sample size was *effsize* > 170 (exception were nmode, statalpha, kappa and allocent, with *effsize* between 17 and 85). The summary statistics reached stability at the end of the runs and were periodically visualized with the script graphphylo (https://github.com/wrf/graphphylo).

Bayesian analyses were also run using StarBeast2 [244], as implemented in Beast package and according to the tutorial provided by Taming the Beast [245]. StarBeast2 employs a multispecies coalescent method to estimate species trees from multiple sequence alignments (i.e., one for each of the 110 orthologous gene sets). For this analysis we used

either the nucleotide or the amino acidic alignments, linking the site models across the gene sets.

All Bayesian analyses had chains run for 10^8 generations, sampling trees and parameters every 1,000 generations and inspecting convergence and likelihood plateauing in Tracer. Posterior consensus trees were generated after discarding the first 10% of generations as burn-in. For the StarBeast analysis, single gene trees were loaded and visualized by DensiTree.

2.1.5. Dating analysis

Because of the numerous incongruences between the species tree obtained by the multi-locus analyses and the single gene trees (see Results section), which could bias the dating analysis, we produced a conservative dataset by: i) limiting the species sample to 10 representative species (*C. capitata*, *B. bryoniae*, *B. cucurbitae*, *B. dorsalis*, *B. jarvisi*, *B. latifrons*, *B. minax*, *B. musae*, *B. oleae*, *B. tryoni*) and ii) considering only those 37 genes that produced a ML tree supporting the consensus species tree with minimum ML bootstrap values of 50 at each node.

Divergence times were then estimated by Beast 2.5.1 using the 4-fold degenerate sites of the concatenated dataset (11,768 nt). This dataset allowed us to use an instantaneous (neutral) mutation rate as prior. Since mutation rate in Tephritidae is not known yet, we assumed it to be similar to that of *Drosophila* (another Diptera) and used the estimate of 0.0346 (SD = 0.00281) substitutions per base pair per million years provided by [246]. Because in *Bactrocera* we assumed eight generations per year (in nature they range from 3-5 of *B. oleae* and sub-tropical *B. dorsalis* populations, to >12 for the tropical species; [247–250]) and to account for uncertainty, we finally set as prior a normally distributed mean of 0.028 (SD = 0.03). In a second approach, we set a mutation rate lognormal distributed with 'mean in real space' M = 0.028 and S = 0.82 (to produce the same 95% quantile – 0.077 – as the normal distribution). For both approaches we used a strict or a LOGN relaxed clock and either a Yule or a Birth-Death model, for a total of eight different combinations.

In all cases, we employed a GTR+G replacement model and a root prior uniformly distributed between 6 and 65 million years ago (Mya), which correspond to the age of a *Ceratitis* fly fossil [251] and of the Schizophora radiation [252, 253]. Because the Bayesian phylogenetic analysis on the concatenated 4-fold degenerate sites resulted in a topology incongruent to the one supported by all other ML and Bayesian analyses (see Results), the species tree was fixed according to the latter consensus topology.

All analyses were run twice, with chains run for $5x10^7$ generations, sampling trees and parameters every 1,000 generations and inspecting convergence and likelihood plateauing in Tracer. Both chains resulted well mixed, with average effective sample size (ESS) values across posterior

values being well above 200. The posterior consensus trees were generated after discarding the first 20% of generations as burn-in.

We finally identified which model was most appropriate by performing a Bayesian model selection based on the marginal likelihood values with the nested sampling approach implemented in the NS package [254]. Following the recommendations provided by the dedicated Taming the Beast tutorial [245], sub-chain length was set at 50,000, which corresponds to the length of the MCMC run (i.e., $5x10^7$) divided by the smallest ESS value observed across the eight model runs (i.e., ~1,000), and the number of particles was set at 10. A model was considered favoured over another model if the difference between the two marginal likelihoods (i.e., the Bayes Factor (BF) in log space) was more than twice the sum of the corresponding standard deviations (SD).

2.2. Evolution of host selection in *Bactrocera* fruit flies

2.2.1. Identification and annotation of chemosensory genes

The complete repertoire of *C. capitata* OBP, OR and GR amino acid sequences [227] was used as query sequences against the genomes of *B. cucurbitae*, *B. oleae*, *B. dorsalis*, *B. latifrons* and *B. tryoni*, using tBLASTn [233] with an e-value cut-off of 10^{-5}. For CSPs, the amino acid sequences identified in *B. dorsalis* transcriptome were used as a query instead (**Figure 8**) [214].

Sequences that passed the cut-off value were extracted and exons were mapped onto the protein query to manually reconstruct the orthologous coding sequences using UGENE software [255]. The first hit found in tBLASTn analysis was labelled as the putative orthologue, whereas other possible hits were considered putative paralogues. Introns were identified following the GT-AG rule [256] and subsequently removed, and the resulting combined exons were checked for the presence of an in-frame coding sequence. To minimize false negatives and verify putatively incomplete sequences, we checked the chemosensory proteins against all publicly available genomic and transcriptomic data, and we performed a reciprocal BLAST search between the best hits of the initial BLAST against the chemosensory repertoire of *C. capitata* [227] and *B. dorsalis* transcriptome (**Figure 8**) and our genomic and transcriptomic datasets [214].

Orthologous gene groups were inferred using two different strategies. In the first, we used the OMA standalone tool [257] on either the DNA coding sequences or the corresponding translated amino acid sequences. In the second approach, we evaluated the distribution of chemosensory genes on a six-species gene maximum-likelihood phylogeny (described below) using a threshold rule of bootstrap > 70 (**Figure 9, 10, 11, 12**). Combining the two methods, we could assess the orthology of genes previously labelled as paralogs and define lineage-specific duplications.

Genes were named following the *C. capitata* nomenclature and the species' names were represented by a four-letter prefix (one letter for the genus and three letters for the species), e.g., *BdorOr1* corresponds to the *B. dorsalis* ortholog of the *Or1* gene annotated in *C. capitata*. Orthologues were identified by a consecutive number preceded by a point, for example, *BcucObp44.1* and *BcucObp44.2* are the two *B. cucurbitae* paralogues of the *Obp44* of *C. capitata*. Paralogs whose orthologs could not be clearly determined in the other species were named with consecutive letters: for example, the *Or60* has two copies in *B. dorsalis,* which were named *BdorOr60a* and *BdorOr60b*.

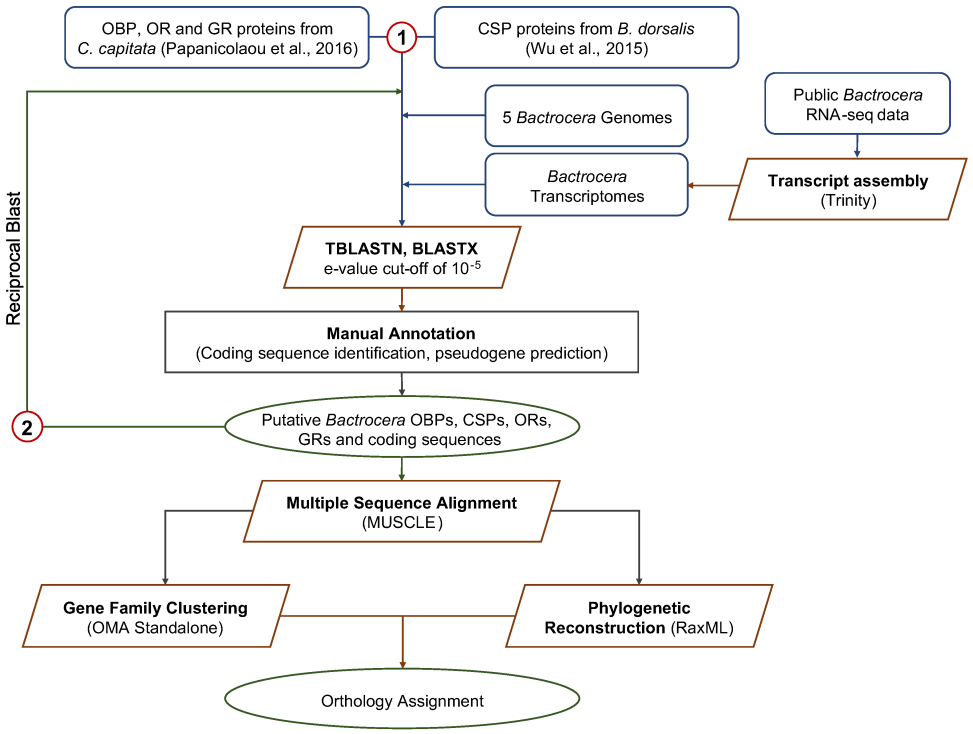


Figure 8 – Flowchart of the gene annotation procedure. 1) tBLASTn analyses using as queries OBPs, ORs and GRs from *C. capitata* ([227]) and CSPs from *B. dorsalis* ([214]) and as subjects *Bactrocera* genomes and transcriptomes. 2) Reciprocal tBLASTn and BLASTX using as queries the best hits from 1) against the chemosensory repertoire of *C. capitata* [227] and *B. dorsalis* transcriptome [214].

2.2.2. Gene Trees

The six-species gene phylogenies were constructed for each of the four gene families (OBP, CSP, OR, and GR) to facilitate the annotation procedure and to gain insights into the chemosensory gene family evolution (**Figure 9, 10, 11, 12**). For each gene family, we produced an alignment of all orthologue sequences using MUSCLE [258] as implemented in the tool TranslatorX [237]. The resulting alignments were manually adjusted to prevent possible ambiguous signals in the phylogeny. Phylogenetic trees were inferred using Maximum Likelihood method with RAxML v8.2.10 (estimating supporting bootstrap values by 100 pseudo-replicates) [240] on both the nucleotide alignment, using the GTRGAMMA model, and on the translated aminoacidic alignment, using the PROTGAMMA+LG+F model.

Phylogenetic trees were visualized and edited using the tool iTOL [259].

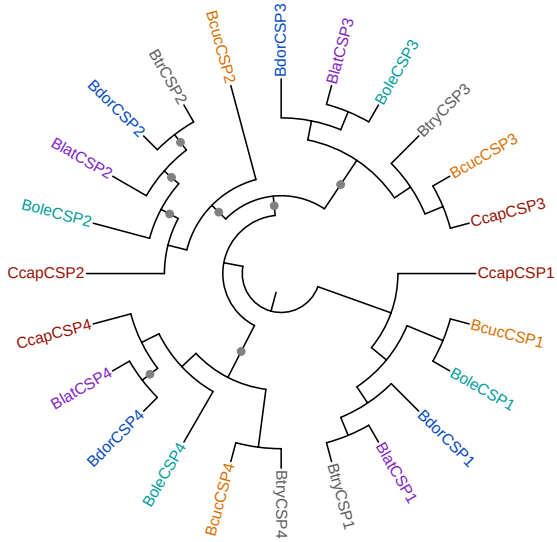


Figure 9 – Maximum likelihood phylogenetic tree of the CSP family in *Bactrocera*, based on the amino acidic alignment. Bootstrap support > 70 is indicated by grey dots.

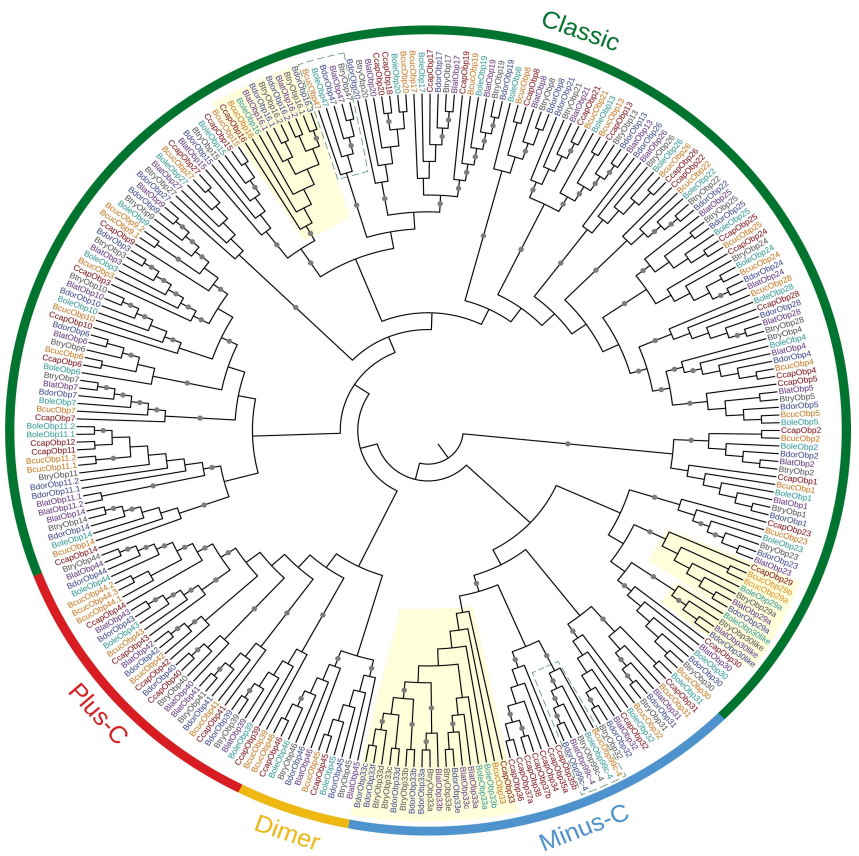


Figure 10 – Maximum likelihood phylogenetic tree of the OBP family in *Bactrocera*, based on the amino acidic alignment. Bootstrap support > 70 is indicated by grey dots. Clades highlighted in yellow were subject to *Bactrocera*-specific gene expansion, while clades highlighted in light blue include genes absent in the outgroup *C. capitata*. The coloured outer circle identifies the OBP subfamilies: Classic, Minus-C, Plus-C, and Dimer.

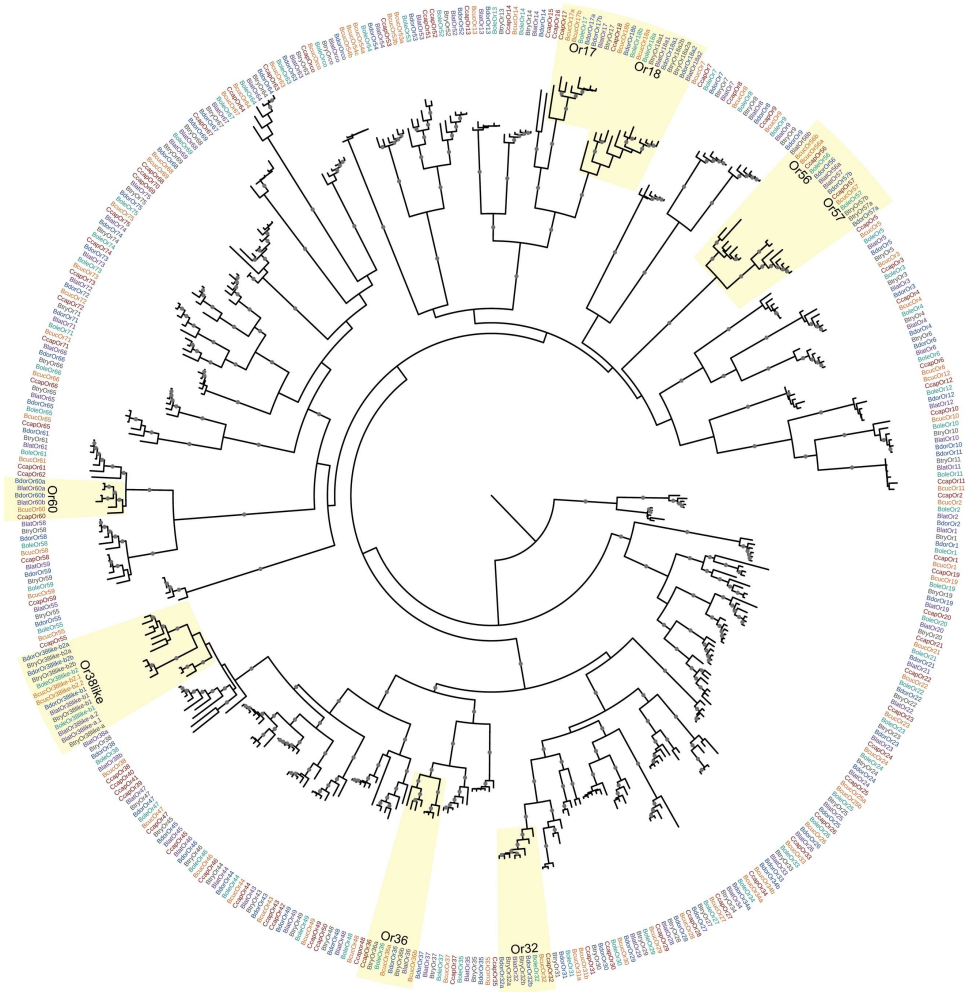


Figure 11 – Maximum likelihood phylogenetic tree of the OR family in *Bactrocera*, based on the amino acidic alignment. Bootstrap support > 70 is indicated by grey dots. Clades highlighted in yellow underwent gene expansion in *Bactrocera*.

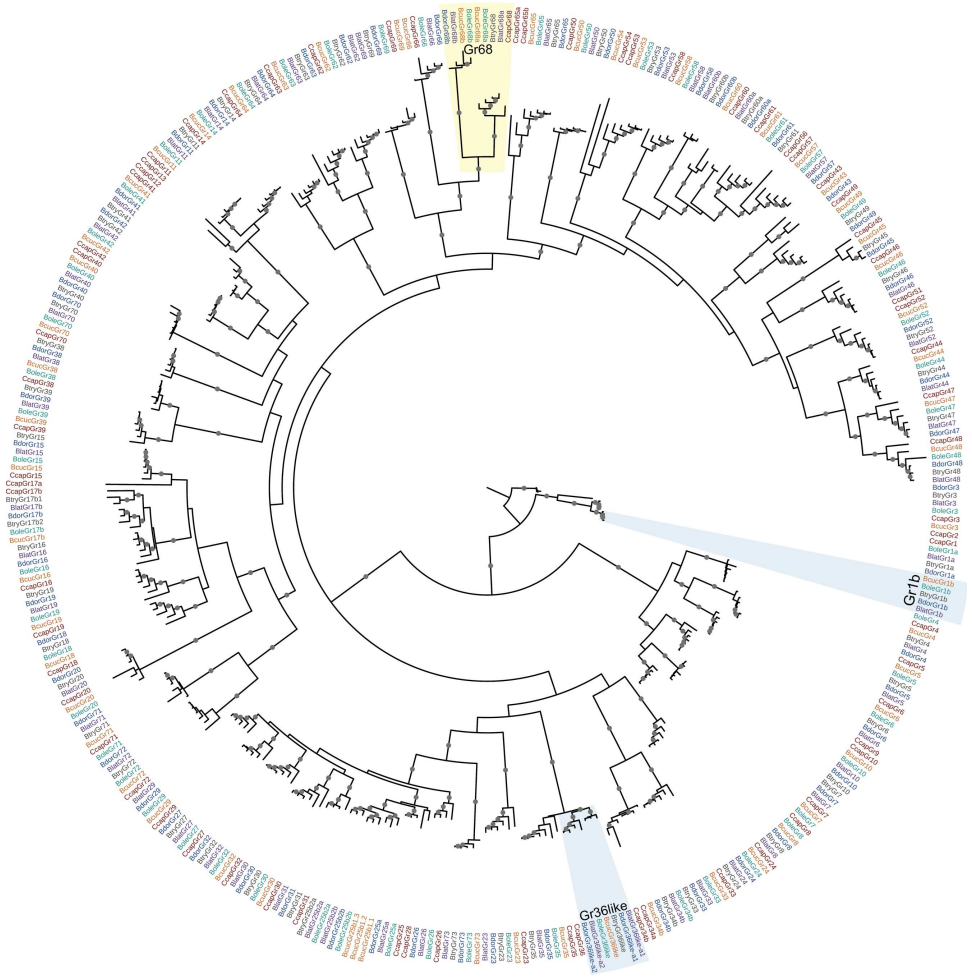


Figure 12 – Maximum likelihood phylogenetic tree of the GR family in *Bactrocera*, based on the amino acidic alignment. Bootstrap support > 70 is indicated by grey dots. Clades highlighted in yellow underwent gene expansion in *Bactrocera*, while clades highlighted in light blue include genes present absent in the outgroup *C. capitata*.

2.2.3. Estimation of gene gain, loss and turnover rates

The number of gain and loss events and the turnover rates of the chemosensory gene families were inferred using a gene tree-species tree reconciliation approach as implemented in BadiRate v1.35 [260].

The software takes as input the species phylogenetic tree and the number of genes in the extant species to model changes in the size of the gene family along the phylogeny.

Gene birth and death (BD) turnover rates were estimated comparing different methodologies with an approach similar to that used by Almeida and colleagues [261]. In particular, gene birth (β) rate indicates the birth (duplication) events per gene per million of years, whereas the death (δ) rate refers to deletion and pseudogenization events per gene per million of years. Both are density-dependent rates, as the probability of having a birth or a death event is proportional to the actual family size [260].

We first used the BD-FR_t-CML method, which estimates β and δ rates by counting the gain/loss events along the phylogeny based on the number of family members in the internal nodes inferred by maximum-likelihood and assumes that each branch has its specific turnover rates (FR_t). We also used three additional methods that exploit a full maximum-likelihood framework to estimate the BD rates. Specifically, BD-FR_t-ML estimates BD rates under the BD stochastic model, with each branch having different turnover rates (FR_t). In the BD-GR_t-ML method, the β and δ rates were estimated under the assumption that all branches share the same turnover rates. Lastly, the BD-BR-ML model infers different BD rates in specific lineages. This model is useful to investigate the effect of different ecological features (i.e., host specialization) on gene turnover rates and to evaluate whether different BD rates across the phylogeny can better explain the observed birth and death dynamics. We analysed four different branch models (BR) based on the species group with specific BD rates in which all the internal lineages of the phylogeny shared the same BD rates, while the terminal lineages were assumed to share or have their own turnover rates. In BR1, BR2, and BR3, a different BD rate compared to the internal branches was assumed for the monophagous *B. oleae*, the oligophagous *B. cucurbitae* and the polyphagous B. *tryoni*, *B. latifrons* and *B. dorsalis*, respectively (**Figure 13**). In BR4 the aforementioned monophagous, oligophagous, and polyphagous species had three different BD rates compared to the internal branches.

All models were examined independently for the OBP, OR and GR gene families and for the combined dataset that included all three chemosensory gene families.

The ML-based analyses were performed in 100 independent runs using different random starting values. The Akaike Information Criterion (AIC; [262]) was used to evaluate the goodness of fit of the models.

Based on the BD-FR_t-CML model, we estimated the number of gene family members in the internal nodes of the species tree.

We manually included the information obtained from the BD-FR_t-CML model onto the phylogenetic tree to determine which genes were gained and lost at each node of the *Bactrocera* phylogeny. We further calculated the overall turnover rate on each branch of the phylogeny using the formula:

$$Turnover\ rate = \frac{Nb.of\ gene\ gains + Nb.of\ gene\ losses}{Divergence\ time\ (in\ million\ years)}$$

where divergence time was inferred from the results of the dating analysis.

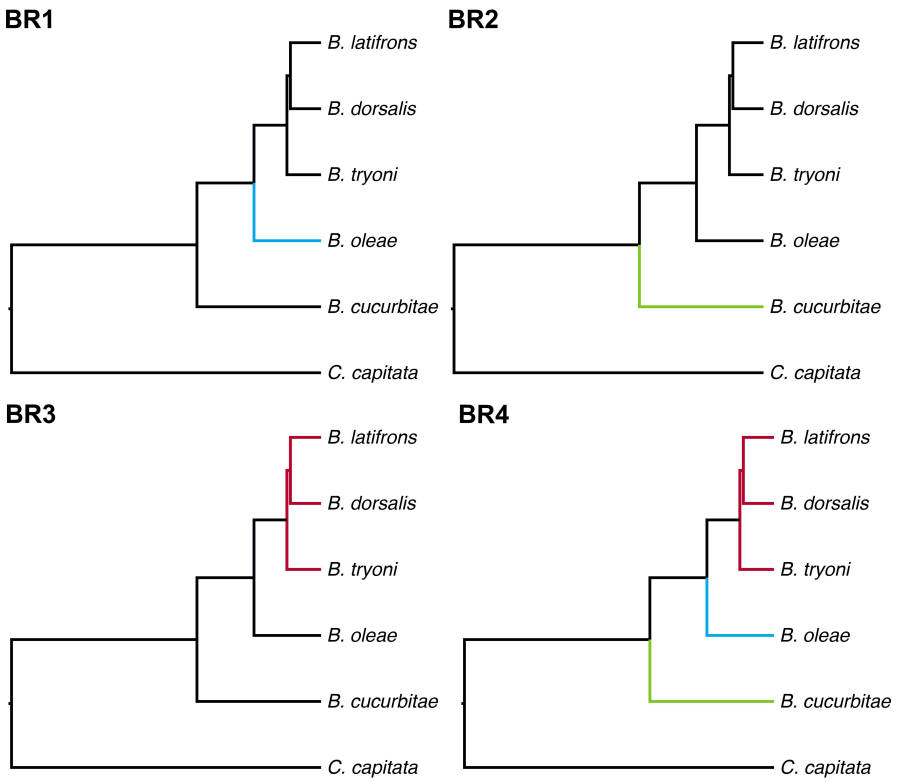


Figure 13 – Branch models used to evaluate differences in BD rates among species groups. Branches indicated in the same colour in the same tree were specified to have the same BD rates and different rates to those in different colours.

2.2.4. Identification and analysis of selective pressures

Rates of molecular evolution were obtained by PAML 4.7 [263] based on the unrooted tree estimated in our phylogenetic analysis, which has topology (((*B.latifrons*, *B. dorsalis*), *B. tryoni*), *B. oleae*, *B. cucurbitae*).

The rate of non-synonymous substitution, d_N (leading to amino acid changes), and synonymous substitution, d_S (which should accumulate

neutrally), have been estimated over all branches of the phylogenetic tree using the ‘‘free- ratio’’ model (M0 [264]; model = 1 and NSsites = 0), which allows $\omega = d_N/d_S$, i.e., the level of selective pressure experienced by a gene, to vary among branches of the tree.

We then used PAML to test different models of substitution rate across coding sites [265, 266], to detect genes that underwent positive selection along with one of the *Bactrocera* lineages. To maximize statistical power, these tests were done only on the orthologous sets containing all species. In the first test, we compared models that assumed one or more substitution rates across the phylogeny. The first of such models is the basic “one-ratio” branch model (M0), which assumes a constant ω across the phylogeny (model = 0 and NSsites = 0). Following the manual recommendations, this model was used to get the branch lengths for each gene tree, which were then copied into the tree structure file to be used with the branch and site substitution models. The likelihood of the M0 model was compared to that of a branch model that assumed two ω values, one for a single *Bactrocera* species (the so-called foreground branch), and a second for the rest of the tree (the background branches; model = 2 and NSsites = 0). Subsequently, we performed a Likelihood ratio test (LRT), whereby the value of twice the difference between the two likelihoods was tested using a χ^2 test with 1 degree of freedom. The presence of positive selection was assessed by the branch-site test. In this test (branch-site model A, test 2 [267]), ω can vary both among sites in the protein and across branches on the tree (model = 2, NSsites = 2). The null model fixed $\omega_2 = 1$ (fix_omega = 1, omega = 1), whereas the positive selection model allowed $\omega_2 > 1$ (fix_omega = 0, omega = 1) in the foreground species. As above, the LRT had 1 degree of freedom. The occurrence of positive selection was also tested by comparing (nearly) neutral models to models that allow for the occurrence of positive selection (site tests). We compared the likelihood of a model where ten site classes have ω values drawn from a β distribution (M7; model = 0 and NSsites = 7) to a model that incorporates an additional class of sites under positive selection (M8; model = 0 and NSsites = 8). In these cases, each comparison was tested using a χ^2 test with 2 degrees of freedom. Tests were performed on the alignments after removing those parts where one or more sequences contained a gap (clean = 1). To account for multiple testing, the false discovery rate (FDR) of each test was estimated using the q-value approach [268] implemented in R [269].

2.2.5. Homology modelling

Computational structural investigations were performed. We initially identified the closest homologs based on sequence identity (using NCBI Blast [233]) and secondary structure matching (using HHPRED [270]). Odorant receptor transmembrane domains were predicted combining the

TMHMM v.2.0 [271] and the TMpred [272] servers. OBPs and CSPs were searched for the occurrence of a signal peptide using SignalP 5.0 [273] and logos were generated using WebLogo [274]. Homology models for each protein of interest were independently generated with Robetta [275], MODELLER [65], and SWISS-MODEL [277]. The seven 3D models produced by the homology-modelling servers were superimposed with PyMOL [278], and the model that best described the superimposition of all the models was selected and the regions of discordance between the models were identified. Model quality was assessed by evaluating average bond lengths, bond angles, clashes, and Ramachandran statistics using Molprobity [279] and the QMEAN server [280]. We also analysed the degree of protein sequence conservation using the ConSurf Server [281], which estimated and mapped divergence across the *Bactrocera* phylogeny onto the 3D protein model. The CASTp server was used to identify all putative binding pockets within the OBP models [282]. Structural figures were generated with PyMOL [278].

Results

3.1. Multi-locus phylogeny of *Bactrocera* genus

3.1.1. Phylogenetic analyses

The ML and Bayesian analyses support a relatively fast and recent radiation of most of the South Eastern *Bactrocera* species considered in this study (**Figure 14**; **Figure S1**). In contrast with most of the phylogenies so far published (e.g. [31, 32]); but see [29]), the results of our analyses indicate that *B. dorsalis* is more closely related to *B. latifrons* than to *B. tryoni*. This relationship is highly supported according to both the ML bootstrap values (≥ 98, **Figure 14A**) and the Bayesian posterior probabilities (PP = 1, **Figure 14B**), no matter whether based on the codon or aminoacidic alignments. The only support for the relationship ((*B. dorsalis*, *B. tryoni*), *B. latifrons*) comes from the Bayesian analysis run in Beast2 and based on the concatenated alignment of the 4-fold degenerate sites (**Figure 15A**). The same dataset analysed with Phylobayes (which also is Bayesian based), however, produced the same topology obtained by the other Bayesian and ML analyses (**Figure 15B**).

Combining single genes analyses into a coalescent framework also support *B. latifrons* as the closest relative of *B. dorsalis*, both when using the coalescent-aware ASTRAL approach [241], which is based on the single ML gene trees, and using StarBeast2 [244], which employs a multispecies coalescent method to estimate species trees from multiple sequence alignments (**Figure 16**; **Figure S3**). It is evident, however, that many genes have phylogenies not consistent with the inferred species phylogeny. For instance, the StarBeast2 approach reveals a high number of genes having an alternative phylogeny within the ((*B. dorsalis*, *B. latifrons*), *B. tryoni*), suggesting incomplete lineage sorting due to fast radiation or (ongoing) hybridization. This uncertainty is also apparent in the ASTRAL results, whereby the support for such clade falls to < 92 when bootstrapping by gene resampling (compare **Figure 17A** with **Figure 17B**; **Figure S2**).

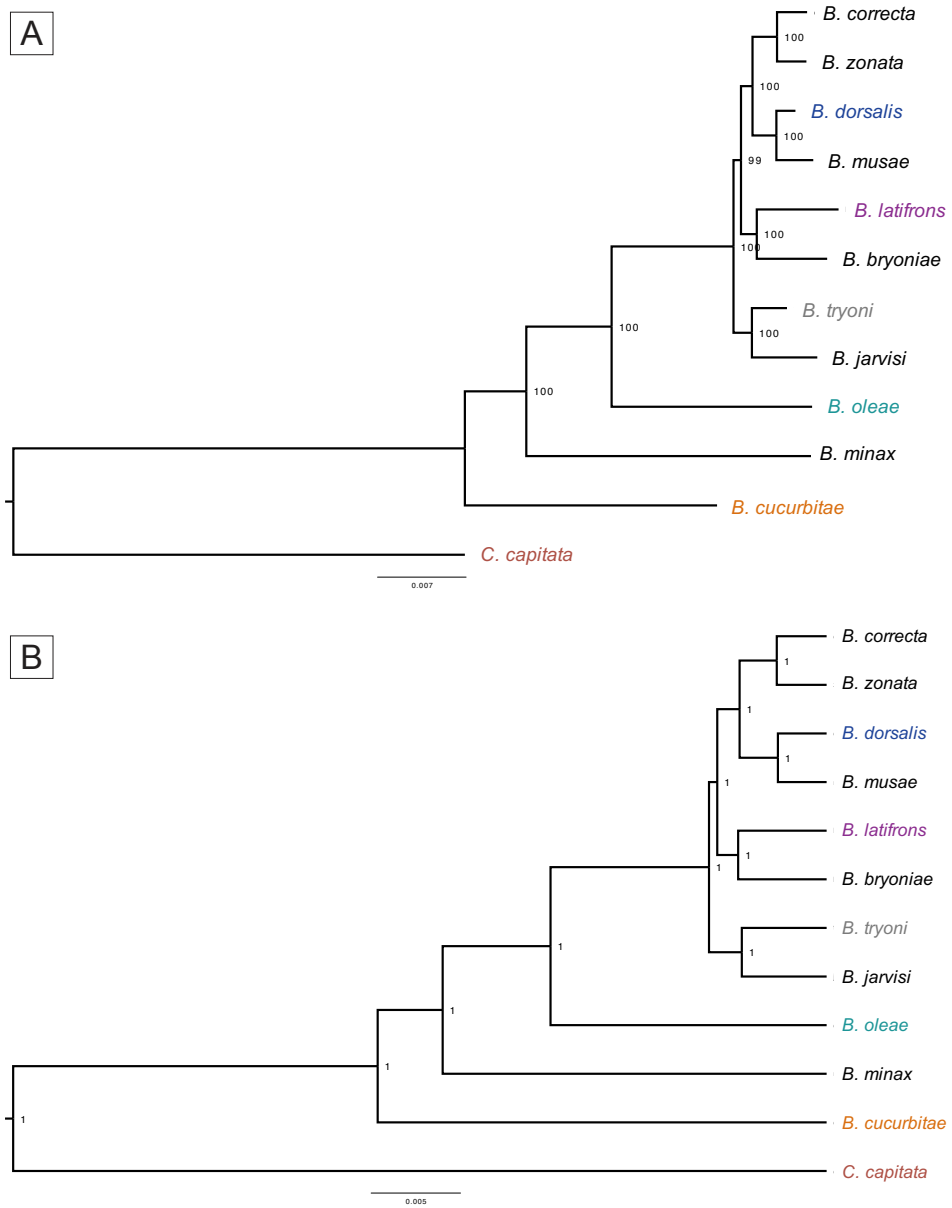


Figure 14 – Phylogenesis of *Bactrocera* inferred from the amino acidic alignments of 110 orthologous nuclear genes. A) Maximum Likelihood phylogenetic analysis (PROTGAMMAGTR model). B) Bayesian phylogenetic analysis (Gamma site model). Both analyses were done using the concatenated amino acidic alignment (63,297 aa) and support a closer relationship of *B. dorsalis* with *B. latifrons* than with *B. tryoni*. Support at nodes is given as bootstrap values for the ML analysis and as posterior probabilities for the Bayesian analysis. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

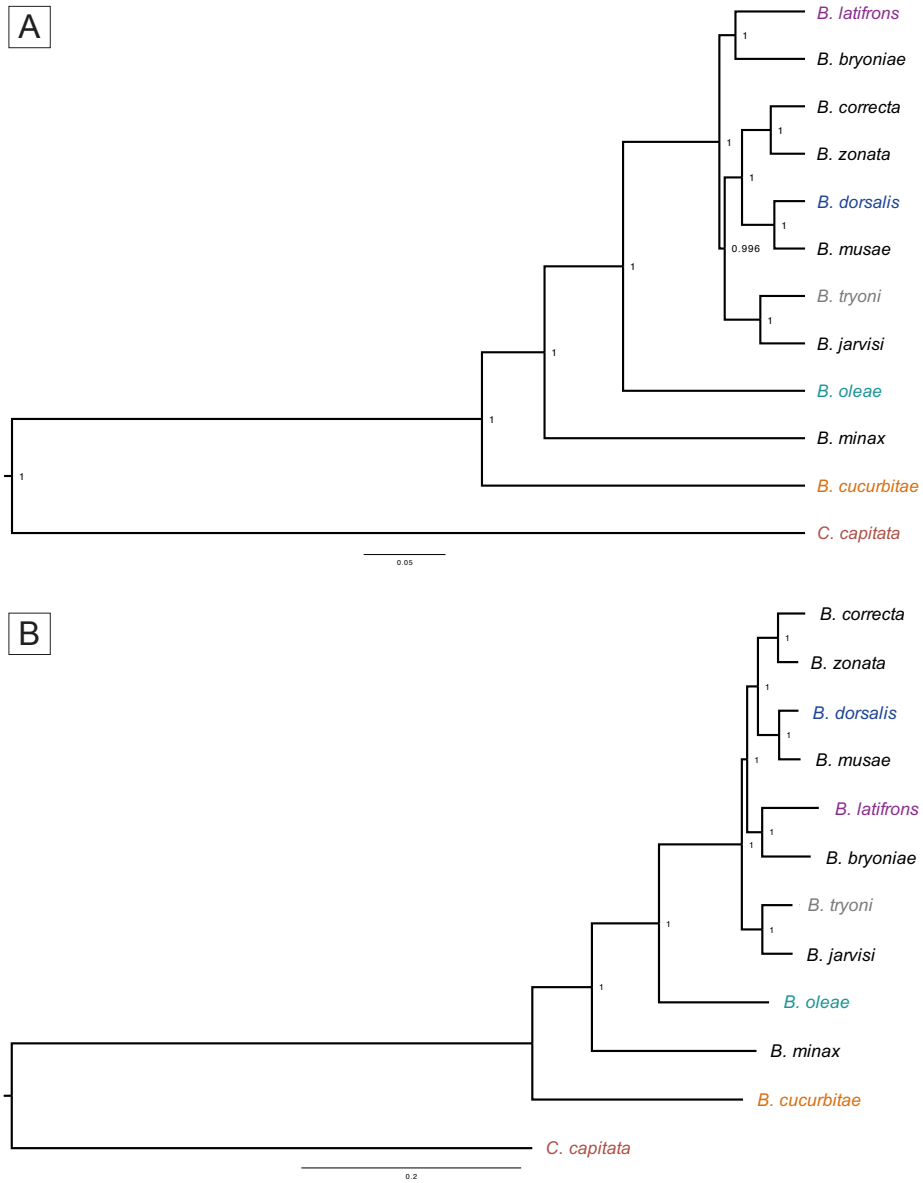


Figure 15 – Phylogenesis of *Bactrocera* inferred from the 4-fold degenerate sites of 110 orthologous nuclear genes. Both trees were obtained by a Bayesian phylogenetic analysis of the concatenated alignment (24,885 bp). Support at nodes is given as posterior probabilities. A) The tree obtained by Beast2 favours a closer relationship of *B. dorsalis* with *B. tryoni* than with *B. latifrons*, while the tree obtained by PhyloBayes (B) favours a closer relationship of *B. dorsalis* with *B. latifrons* than with *B. tryoni*. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

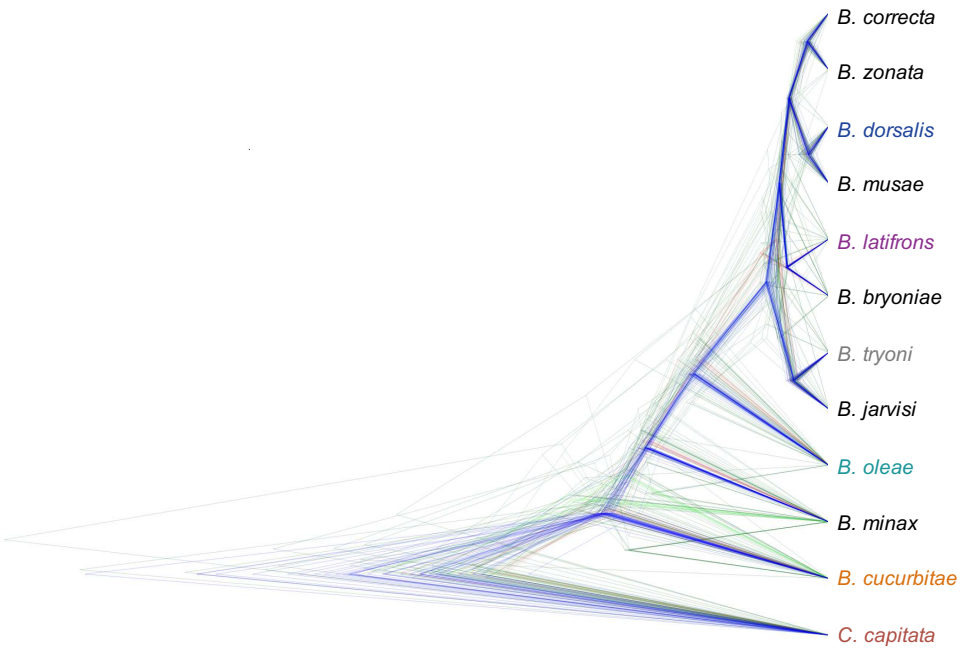


Figure 16 – Multi-locus phylogenesis of *Bactrocera* inferred from 110 orthologous nuclear genes. Bayesian analysis obtained by StarBeast2, which employs a multispecies coalescent method to estimate species trees from multiple sequence alignments (i.e., one for each of the 110 orthologous gene sets). For this analysis we used the amino acidic alignments, linking the site models across the gene sets. Note the numerous discordant gene trees, especially within the *B. dorsalis* - *B. latifrons* - *B. tryoni* clade, compared to the species tree (supported by the gene trees in blue) which supports a closer relationship of *B. dorsalis* with *B. latifrons* than with *B. tryoni*. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

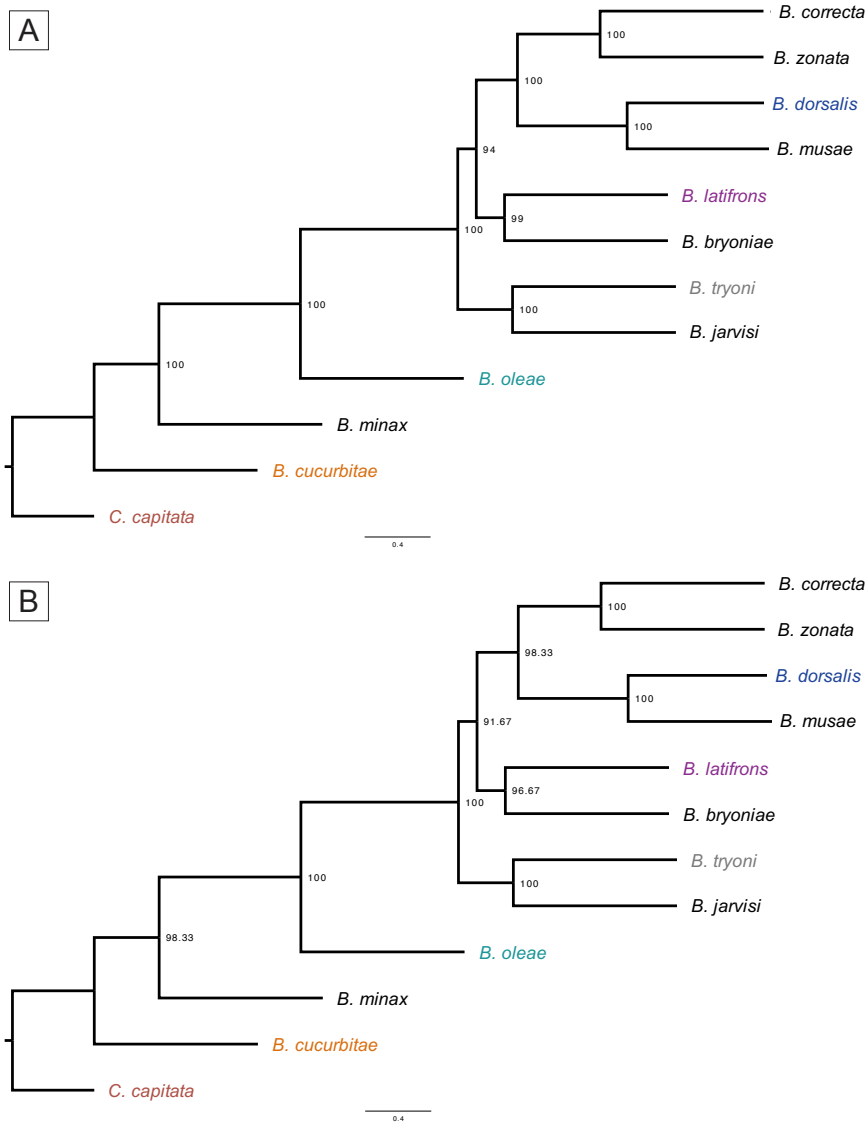


Figure 17 – Multi-locus coalescent-aware phylogenesis of *Bactrocera* inferred from 110 orthologous nuclear genes. Analyses are based on all single ML gene trees obtained using the aminoacidic sequences (PROTGAMMAGTR model). A) Bootstrap values were estimated by performing 100 multi-locus bootstrap replicates; B) Bootstrap values were estimated by performing 100 gene+site resampling. Both analyses support a closer relationship of *B. dorsalis* with *B. latifrons* than with *B. tryoni*. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

3.1.2. Dating analysis

We then estimated divergence times across the phylogeny using the Bayesian approach implemented in Beast2. We run eight different analyses that used different combinations of priors and model settings and performed a model selection to identify the most appropriate for our data (**Figure 18**; **Figure S4–S7**). In particular, the nested sampling approach allowed us to estimate the marginal likelihoods of the different models and make pairwise comparisons using the associated Bayes Factor. All models had marginal likelihoods with a standard deviation ranging from 2.5 to 2.9, which was small enough to assess whether a model was favoured over another one. The model favoured over all the other seven is the one where we set the mutation prior with a lognormal distribution, a strict clock, and a Birth-Death model (**Table 1**). The BF values, even after correcting for uncertainty by subtracting the corresponding standard deviations, are well above two, which provides overwhelming support for that model [283]. The fact that a strict clock is favoured over a relaxed clock is consistent with the low mean value of the coefficient of variation parameter (i.e., the standard deviation of branch rates divided by the mean rate), which equals 0.24. Therefore, unless otherwise stated, in the following we will report the results obtained by this analysis.

The split between the outgroup *C. capitata* and the genus *Bactrocera* dates at around 18.9 million years ago (mya) (**Figure 18**). Consistent with the rapid radiation of the (*B. dorsalis*, *B. latifrons*, *B. tryoni*) clade inferred by the ML and Bayesian analyses, the results of the clock analysis place its origin at ~2.08 mya, with a subsequent very close cladogenesis, at ~1.87 mya, separating *B. dorsalis* and *B. latifrons* (**Figure 18**). The close proximity of these two events and the large overlap of their 95% confidence intervals agrees with a rapid radiation, which could have resulted in frequent incomplete lineage sorting, as revealed by the results of the StarBeast2 analysis (**Figure 18**).

Table 1 – Results of the model selection of the BEAST analyses used to estimate divergence times.

		Mut-normal				**Mut-lognormal**			
		yule_strict	yule_rlxln	bd_strict	bd_rlxln	yule_strict	yule_rlxln	bd_strict	bd_rlxln
Mut-normal	yule_strict	-	62.2 (56.8)	-	54.3 (49.1)	9.2 (3.5)	59.6 (54.3)	-	57.2 (51.9)
	yule_rlxln	-	-	-	-	-	-	-	-
	bd_strict	-	58.6 (53.2)	-	50.7 (45.4)	-	55.9 (50.6)	-	53.5 (48.2)
	bd_rlxln	-	7.9 (2.8)	-	-	-	5.2 (0.2)	-	-
Mut-lognormal	yule_strict	-	53.1 (47.6)	-	45.2 (39.8)	-	50.4 (45)	-	48 (42.6)
	yule_rlxln	-	-	-	-	-	-	-	
	bd_strict	**6.1 (0.7)**	**68.3 (63.1)**	**9.8 (4.4)**	**60.4 (55.3)**	**15.3 (9.7)**	**65.7 (60.5)**	**-**	**63.3 (58.1)**
	bd_rlxln	-	-	-	-	-	-	-	-

The eight models combined a mutation prior with a normal (Mut-normal) or lognormal (Mut-lognormal) distribution, either a Yule or a Birth-Death (*bd*) model, and a *strict* or a LOGN relaxed (*rlxln*) clock. Models were compared in a pairwise fashion, by first estimating their Marginal Likelihoods (mL) and corresponding Standard Deviation (SD) and then calculating the Bayes Factors (BF) as $BF=mL_1-mL_2$, where model 1 and model 2 are those given in the respective row and column. Only BF values that satisfied the condition $BF-(SD_1+SD_2) > 0$, where SD_1+SD_2 are the corresponding SDs estimated for model 1 and 2, are reported. Highlighted on bold are the values for the model favored over all other seven models.

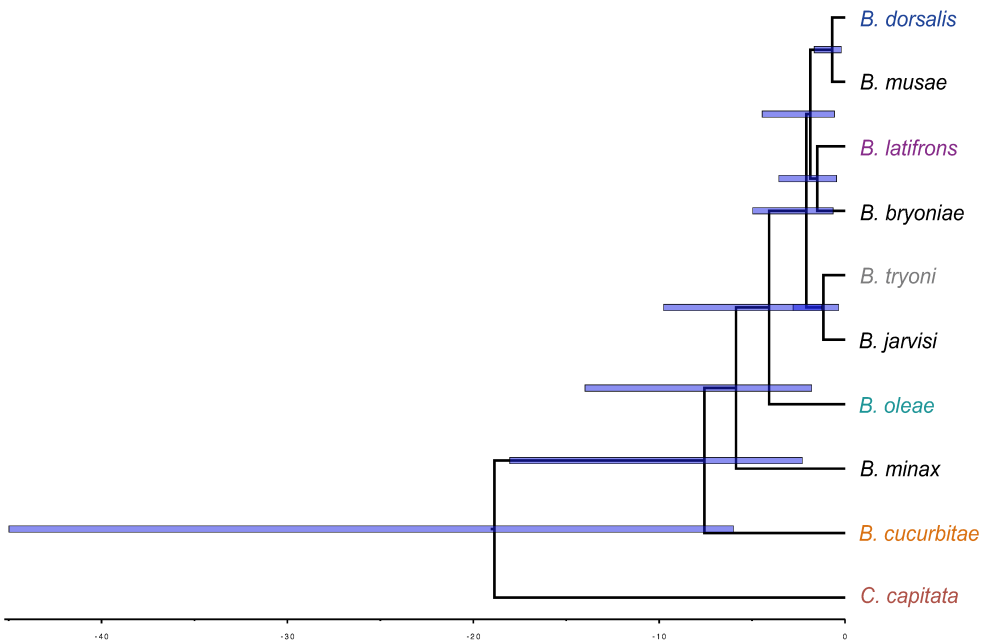


Figure 18 – Molecular time tree of *Bactrocera*. *Bactrocera* originated during the Miocene optimum (around 19 million years ago) and experienced recent fast cladogenetic events around 2 million years ago. The analysis was done setting the mutation rate log-normally distributed as prior, a strict clock and a Birth-Death model. Blue bars at nodes identify the 95% confidence interval of the inferred node age. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

3.2. The evolution of host selection in *Bactrocera* fruit flies

3.2.1. Gene data set completeness

BUSCO analyses revealed high completeness (i.e., > 95% of complete BUSCOs) for most of the datasets (**Figure 19**), with the lowest values found for the *B. tryoni* transcripts (78.8%) and the *B. oleae* genome (76.7%). Since we searched for orthologs in all available datasets, such low completeness values are only partially informative about the possibility of false negatives. In fact, low completeness values were largely compensated by the high values of the other datasets (**Figure 19**): indeed, the percentage of complete BUSCOs in the *B. tryoni* and *B. oleae* combined datasets (respectively 96.6% and 98.1%) are very close to those estimated for the other species (98.3-99.6%). These results suggest a very low probability of false negatives.

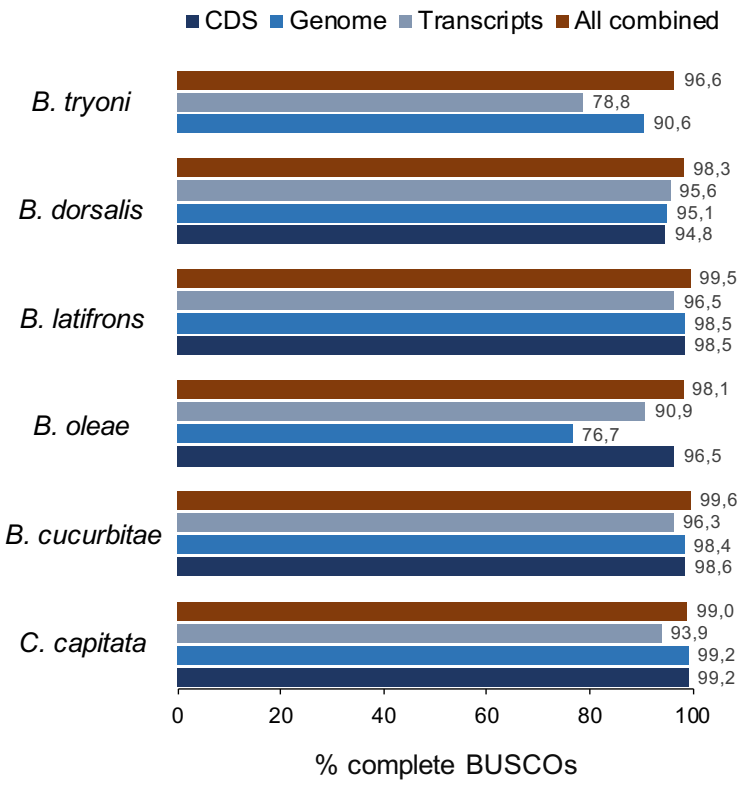


Figure 19 – BUSCO Assessment Results

3.2.2. Chemosensory gene repertoire

We identified and manually annotated the complete repertoire of CSP, OBP, OR, and GR genes present in the genomes of *B. cucurbitae*, *B. oleae*, *B. dorsalis*, *B. latifrons,* and *B. tryoni* (**Table S2**).

A total of 45 OBPs were found in the *B. cucurbitae* genome, 40 in *B. oleae*, 45 in *B. tryoni*, 46 in *B. latifrons,* and 49 in *B. dorsalis* (**Figure 20**). All OBP genes have a similar intron-exon structure among the five species: the number of exons ranges from 1 to 5, with most genes having three exons and only one intronless gene in *B. oleae* (*BoleObp11*), two in *B. cucurbitae* (*BcucObp11* and *21*), *B. latifrons* (*BlatObp11* and *21*), three in *B. tryoni* (*BtryObp11*, *21* and *47*), and four in *B. dorsalis* (*Obp11*, *Obp16*, *Obp21*, *Obp47*) (**Table S3**).

The OBPs identified in *Bactrocera* have been classified into several groups on the basis of distinctive structural features, functional information, and phylogenetic relationships: the Classic, Minus-C, Plus-C, and Dimer [121, 122, 284]. All identified genes have the characteristic hallmarks of the OBP gene family: the presence of a signal peptide, the six-helix pattern, and the highly conserved six cysteine profiles (**Figure 21**).

A total of 4 CSPs were observed in all the *Bactrocera* genomes exhibiting the typical cysteine profile (**Figure 21**).

We found 74 OR genes in *B. cucurbitae*, 63 in *B. oleae*, 65 in *B. tryoni*, 70 in *B. latifrons,* and 75 in *B. dorsalis*. Consistent with OBPs, ORs show an analogous exon-intron structure. The number of exons ranges from one (only in *BcucOr73*) to ten, with most genes having four exons (**Table S4**).

B. cucurbitae shows the highest number of duplications (duplication of *Or17*, *Or18*, *Or25*, *Or31*, *Or34*, *Or36*, *Or53*, *Or56*, *Or68* and triplication of *Or38like* and *Or54*).

Regarding GR genes, we were able to annotate 63 genes in *B. cucurbitae*, 61 in *B. oleae*, 58 in *B. tryoni*, 63 in *B. latifrons* and 66 in *B. dorsalis*. Even in this case, GR genes are characterized by a conserved exon-intron structure, with most genes containing four exons. However, in some cases, they contained much more exons, up to 12 (**Table S5**).

Interestingly, the OR gene family size is positively correlated with both OBP (Spearman's rank correlation coefficient $\rho = 0.8407$, $P = 0.0444$) and GR (Spearman's $\rho = 0.9276$, $P = 0.0167$) family gene sizes in the six analysed genomes (**Figure 22**). *Ceratitis capitata* and *B. dorsalis* are the species with the highest number of genes in the OBP, OR, and GR gene families, while the monophagous *B. oleae* and *B. tryoni* have the lowest number across the three gene families (**Figure 20**).

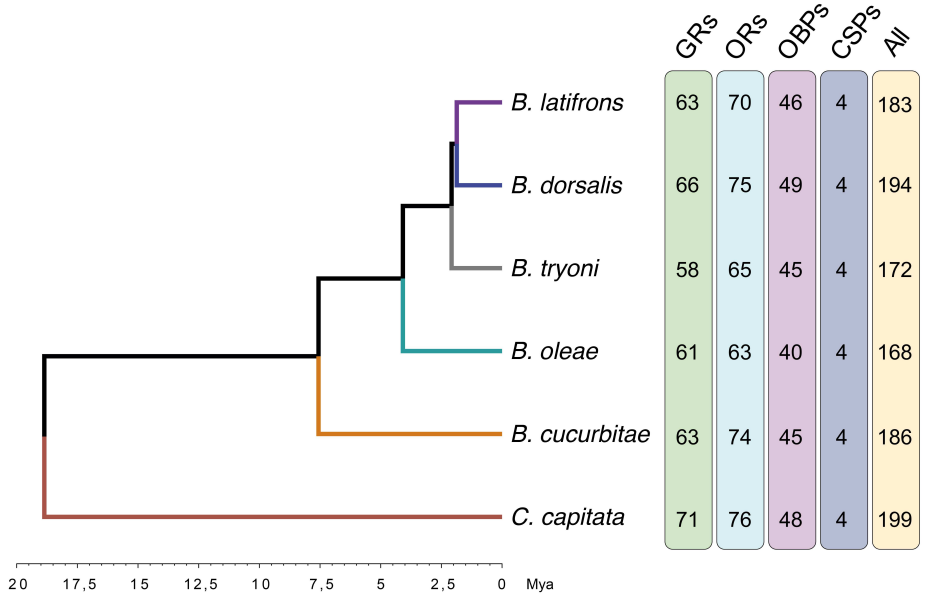


Figure 20 – Phylogenetic relationships across the six investigated species. Numbers in the right boxes represent the number of chemosensory protein encoding sequences per each gene family. Divergence times are given in millions of years (Mya).

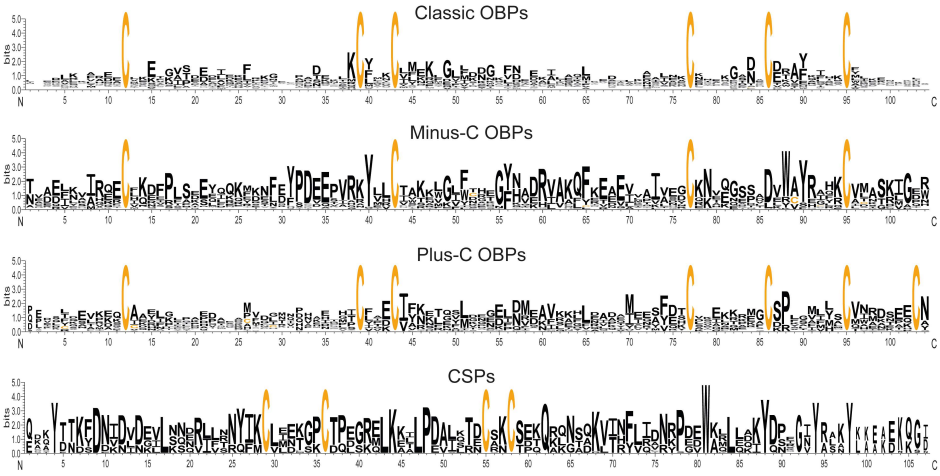


Figure 21 – Conserved OBP and CSP cysteine pattern in *Bactrocera*. Conserved amino acid patterns, based on the multiple sequence alignments are shown using the sequence logo [274]. The height of each amino-acid letter is proportional to its frequency of occurrence at a given position. The highly conserved cysteines are highlighted in orange.

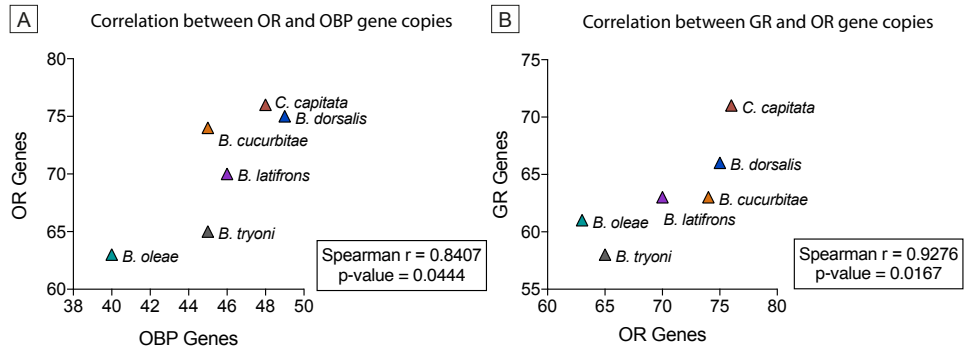


Figure 22 – Correlation between the number of chemosensory gene copies along the *Bactrocera* phylogeny. Scatter plots and Spearman rank correlations between A) the number of total OR and OBP genes and between B) the number of total GR and OR genes. The scatter graphs show a positive correlation between the variables.

We found 22 genes absent in all *Bactrocera* species and, conversely, five absent in *C. capitata* (**Figure 23**). Interestingly, we observed that three genes (*Obp40*, *Or45*, *Gr31*) are present in the four polyphagous species but not in the monophagous and oligophagous fruit flies (**Figure 23**).

We compared the amino acidic identities between proteins encoded by CSP, OBP, OR and GR genes among species (**Figure 24**). We found that divergence between orthologous copies increases with the phylogenetic distance in the majority of CSPs and ORs, while proteins encoded by OBP and GR genes do not always reflect the phylogenetic relationships (e.g., *Obp25*, *Gr42*). OBPs, ORs and GRs showed a higher rate of amino acidic divergence among species compared to CSPs. Indeed, CSPs present an average amino acidic identity of about 93.0%, while in the other protein families, the average ranges from 80.2% to 84.2%.

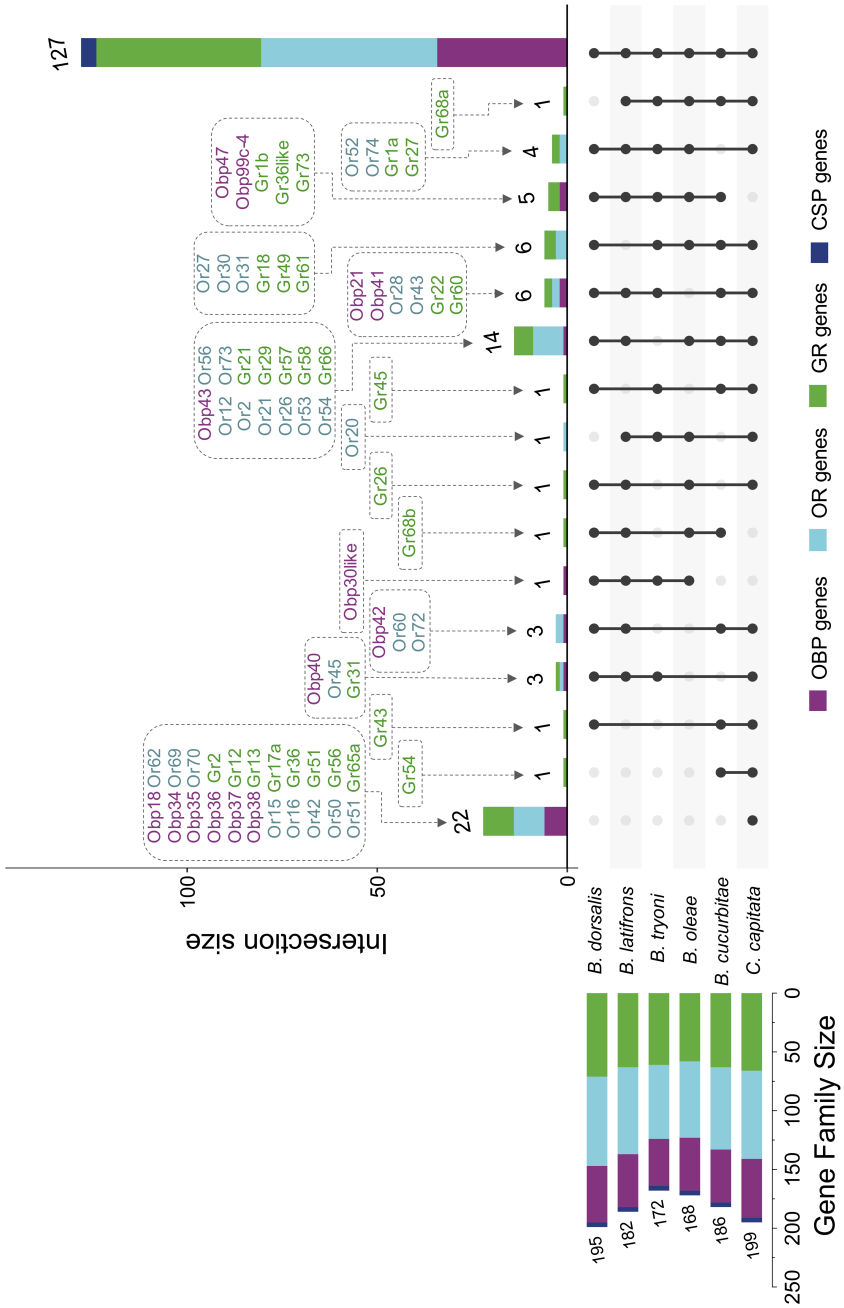


Figure 23 – UpSet plot showing unique and overlapping genes across the six analysed species. The intersection matrix is sorted based on the phylogenetic relationships. Different gene families are indicated in different colours. Connected dots represent intersections of overlapping genes and horizontal bars show the total number of genes identified in each species. The plot was generated using UpSetR package [331] in R [332].

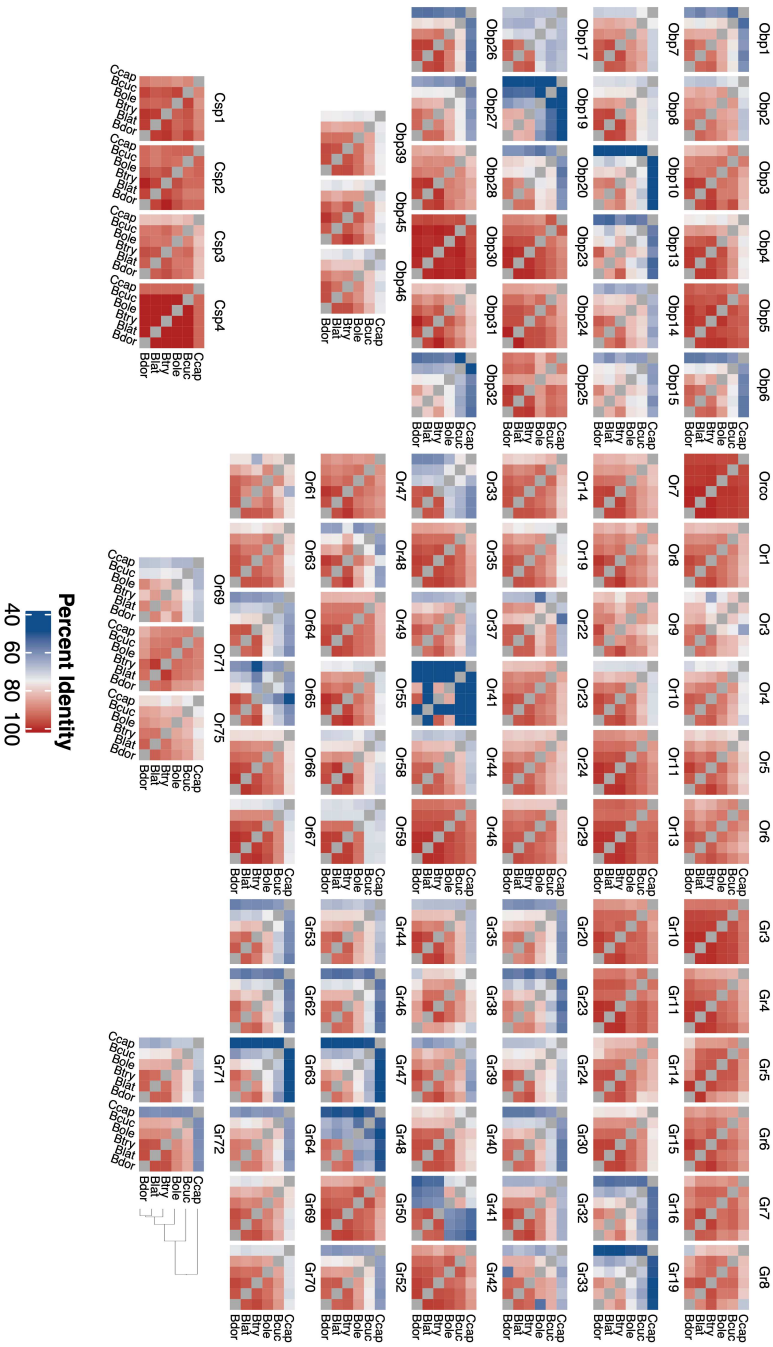


Figure 24 – Heatmaps representing the average amino acidic identities in CSPs, OBPs, ORs and GRs. The percentages of amino acidic identity were based on BLOSUM62 matrix and displayed as colours ranging from blue to red as shown in the key. Heatmaps were generated using ComplexHeatmap [333] package in R [332].

3.2.3. Chemosensory gene family evolution

We analysed the dynamics of chemosensory gene families using Badirate. The preliminary analyses of the birth and death rates estimation for the *Bactrocera* chemosensory gene families revealed differences between the methods used. Thus, we compared several models of rate heterogeneity among lineages and estimated their Akaike Information Criteria to infer whether one model was statistically favoured over the other (a lower AIC value indicates a better fit).

AIC value for the BD-FR_t-ML was much higher than those obtained with the other models, while the BD-GR_t-ML produced AIC values comparable to the three of the BR ones (**Table 2**). AIC values for the BR4 model were not reported because they did not achieve convergence of the likelihood values in any of the gene families, even after 100 independent runs.

For the odorant and gustatory receptor families and all gene families combined, the best-fitting model was the BR3 (**Table 2**), suggesting that for these gene families, the three polyphagous species (*B. dorsalis*, *B. latifrons* and *B. tryoni*) have a different gene turnover rates from those of the other species analysed. In contrast, the best-fitting model for the OBP gene family was the BR1 model (in which the monophagous species *B. oleae* had a different gene turnover rate) (**Table 2**). We note however, that the AIC differences between all the models were small and nonsignificant, thus the best fitting models should be considered only as indicative.

The results obtained with the BD-FR_t-CML displayed that birth (gain) and death (loss) rates were extremely heterogeneous along the *Bactrocera* evolutionary history. OR gene family was the most dynamic, experiencing a total of 29 loss and 20 gain events, followed by GR gene family that experienced three gain and 28 loss events and OBP gene family with eight loss and 17 gain events (**Figure 25**).

We also found that OR death rates were positively correlated with GR death rates (Spearman's ρ = 0.982, P < 0.001), suggesting a shared evolutionary trend in the expansion/contraction of these chemosensory gene families.

The turnover rates of OBPs were, in general, lower than those of GRs and ORs, pointing to the presence of more constraints in the former. The species that showed the highest turnover rate in all the gene families was *B. tryoni* (OR = 5.29, OBP = 2.88, GR = 5.29, **Figure 26**).

In *B. oleae*, there is a remarkable departure in the evolutionary patterns compared with other *Bactrocera* species. In the OR gene family, both the turnover rate and the β rate (1.47 and 0.00, respectively) were among the lowest of the entire phylogeny (**Figure 26**), and the β rates were also 0 for all other chemosensory gene families. In fact, the turnover rates of *B. oleae* are only associated with loss events.

Table 2 – AIC values for the global model (inferred by BD-GR$_t$-ML method), free model (estimated by BD-FR$_t$-ML method), and the different BR models (obtained with BD-BR-ML). The best AIC value for each gene family is highlighted in bold; K indicates the number of parameters.

Model	K	GR	OBP	OR	All Genes
global	2	380.9	224.6	509.9	1112.7
free	20	841.9	604.8	903.2	2252.1
BR1	4	382.8	**223.3**	505.5	1101.5
BR2	4	382.9	223.4	507.0	1107.2
BR3	4	**357.6**	228.3	**479.9**	**1063.6**

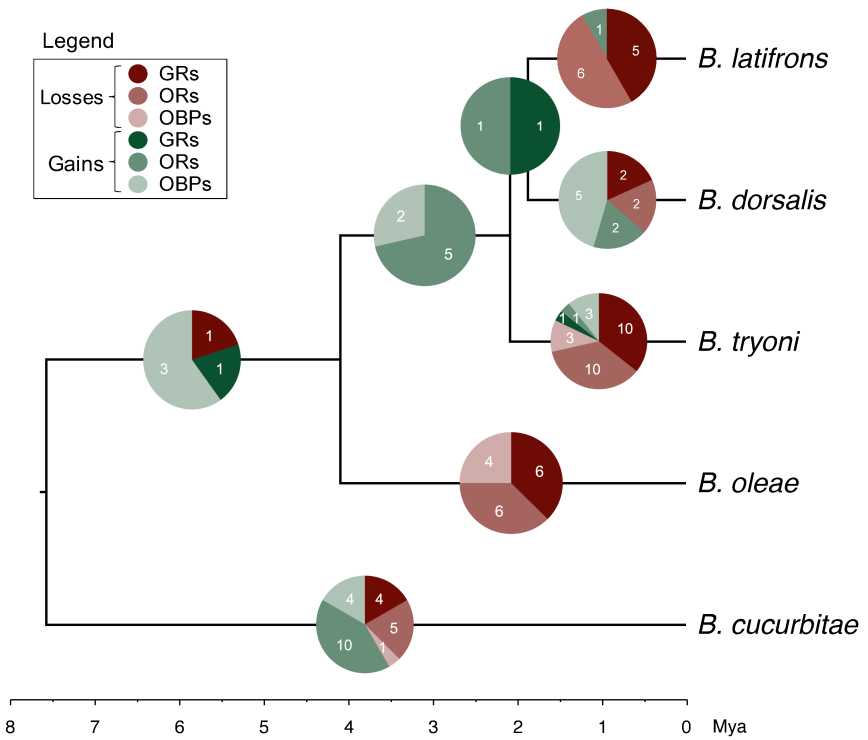


Figure 25 – Evolution of chemosensory receptors on the *Bactrocera* phylogeny. Pie charts have slices proportional to the relative contribution of gene gain (green) and loss (red) events to the total gene families turnover rate. Numbers reported on the pie slices indicate the number of gain and loss events. Divergence times are given in millions of years (Mya).

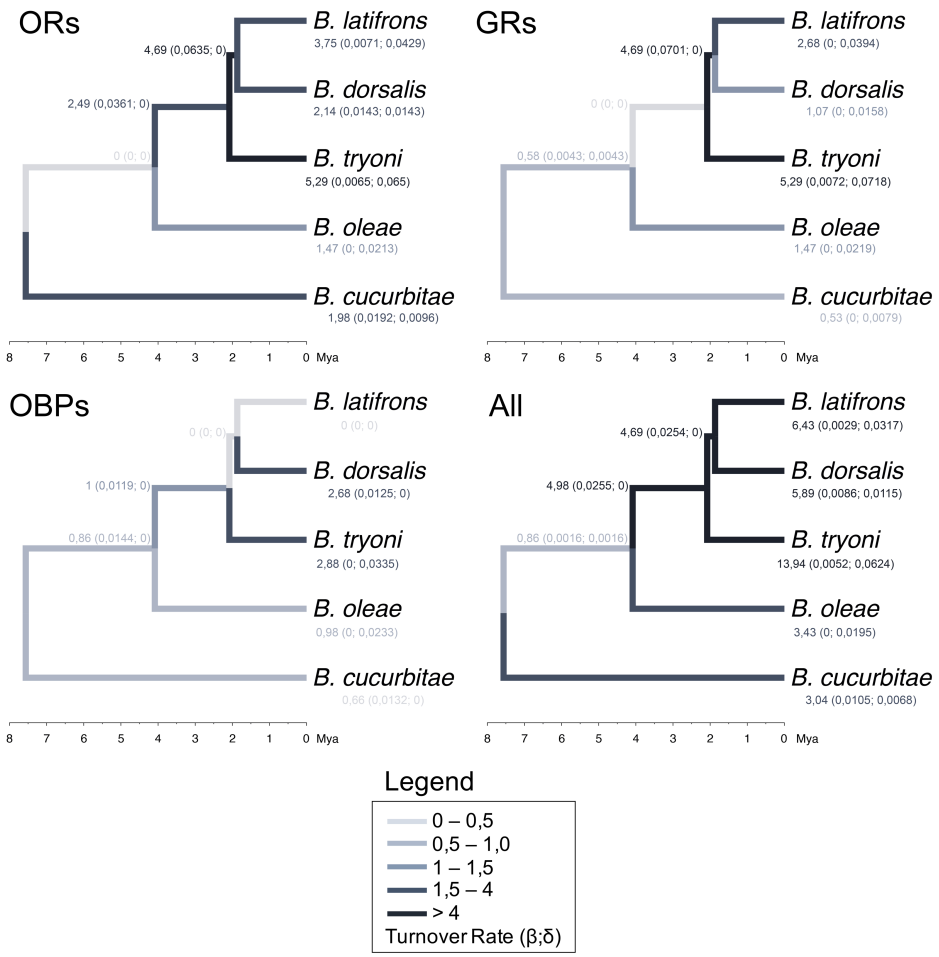


Figure 26 – Evolution of chemosensory receptors on the *Bactrocera* phylogeny. For each branch in the tree, we report the turnover rate, the birth (β) and death (δ) rates obtained by the BadiRate analysis. Divergence times are given in millions of years (Mya).

3.2.4. Signatures of selection

We investigated the pattern of molecular evolution of GR, OR, and OBP genes using codon substitution models with PAML [263].

Branch tests identified four OBPs (FDR < 0.05) that were evolving under positive selection in *B. oleae* (**Table 3**). Among ORs, this test identified six genes in *B. cucurbitae*, two in *B. oleae*, three in *B. dorsalis*, three in *B. latifrons* and one in *B. tryoni* that bear signatures of positive selection pressure (FDR < 0.05, **Table 3).** In the GR gene family, we detected seven GRs evolving under positive selection in *B. oleae*, three in *B. latifrons*, two in *B. dorsalis* and one in *B. cucurbitae* (FDR < 0.05, **Table 3**).

We also applied a branch-site test and obtained evidence for site-specific selection (**Table 3**) in specific branches of the phylogeny. We found one candidate OBP gene in *B. oleae* (*BoleObp25*, FDR < 0.05). Analysing OR genes, we identified two genes in *B. cucurbitae* (*BcucOr11* and *BcucOr38*, FDR < 0.05), one in B. *dorsalis* (*BdorOr71*, FDR < 0.05) and one in *B. tryoni* (*BtryOr75*, FDR < 0.05).

Finally, we investigated whether the three gene families are evolving under similar constraints by analysing their rate of evolution. In particular, we analysed the rate of synonymous and non-synonymous substitution, and of their ratio ω. We note that, when comparing substitution rates across species, we used patristic distances instead of the terminal branch distances (i.e., we summed substitution rates over all branches leading to the species starting from the *Bactrocera* common ancestor). While this has the caveat of introducing partial pseudo-replication (internal branches are summed to each of the descendant branches), and of possibly using less reliable estimates of deep branches and nodes (compared to the external branches), this allowed us to perform direct comparisons between the tips of the phylogeny.

The comparison of ω estimated assuming a constant selective pressure over the entire phylogeny (model M0) revealed that the three gene families are evolving under similar constraints (ANOVA, $F_{(2,96)}=0.326$, $P = 0.722$). We then analysed patterns of molecular evolution when assuming different selective pressure across the phylogeny (free-ratio model).

Gene family and species are both strong determinants of the variance of d_N (ANOVA, $F_{(2,480)} = 4.94$, $P = 0.008$; and $F_{(4,480)} = 5.10$, $P = 0.0005$, respectively), d_S (ANOVA, $F_{(2,480)} = 5.03$, $P = 0.007$; and $F_{(4,480)} = 16.06$, $P < 10^{-10}$, respectively) and ω values (ANOVA, $F_{(2,479)} = 5.58$, $P = 0.004$; and $F_{(4,479)} = 2.74$, $P = 0.028$, respectively). However, pairwise comparisons revealed only few statistically significant differences between the species. The rate d_N was similar across species for all three gene families (Tukey's HSD – honestly significant difference – multiple comparison test, all $P >$ 0.05), whereas d_S was significantly lower in *B. cucurbitae* than in *B. latifrons* for both OBPs ($P = 0.0040$) and for ORs ($P = 0.0002$) (**Figure 27**). Regarding GRs, d_S was significantly lower in *B. cucurbitae* than *B. latifrons*

(P = 0.0002), *B. dorsalis* (P = 0.0017) and *B. tryoni* (P = 0.0295) (**Figure 27**). By inspecting the distribution of the rate values across species, it seems that most of the differences are between the polyphagous species and the two monophagous/oligophagous ones. While it is tempting to associate this pattern to the different ecology of the species, we caution on the possibility that the aforementioned bias (i.e., the use of patristic distances) may be at least partly responsible for the results.

We then analysed patterns within each species. We found that gene family is a determinant of the variance of ω only in *B. oleae* (ANOVA, $F_{(2,96)}$ = 3.38, P = 0.038). In particular, ORs are under more selective constraints than OBPs (mean (SD) ω = 0.227 (0.176) vs. mean (SD) ω = 0.379 (0.305), respectively; Tukey's HSD test P = 0.029) (**Figure 28**), suggesting different selective pressures in these two classes of chemosensory genes.

Table 3 – Number of GR, OBP and OR candidate genes for positive selection according to the branch and the branch-site tests.

		Branch test				**Branch-site test**			
		P value			FDR	*P* value			FDR
		<0.01	<0.005	<0.001	<0.05	<0.01	<0.005	<0.001	<0.05
GRs	*B. cucurbitae*	2	1	1	1	1	0	0	0
	B. oleae	7	7	1	7	1	1	1	1
	B. tryoni	0	0	0	0	0	0	0	0
	B. latifrons	3	3	0	3	3	2	1	1
	B. dorsalis	2	2	1	2	2	2	2	2
ORs	*B. cucurbitae*	9	6	3	6	4	2	2	2
	B. oleae	3	2	2	2	0	0	0	0
	B. tryoni	3	1	1	1	1	1	1	1
	B. latifrons	4	3	1	3	0	0	0	0
	B. dorsalis	3	3	3	3	3	3	1	1
OBPs	*B. cucurbitae*	0	0	0	0	1	1	0	0
	B. oleae	5	4	2	4	1	1	1	1
	B. tryoni	0	0	0	0	0	0	0	0
	B. latifrons	0	0	0	0	0	0	0	0
	B. dorsalis	0	0	0	0	0	0	0	0
All genes	*B. cucurbitae*	11	7	4	7	6	3	2	2
	B. oleae	15	13	5	13	2	2	2	2
	B. tryoni	3	1	1	1	1	1	1	1
	B. latifrons	7	6	1	6	3	2	1	1
	B. dorsalis	5	5	4	5	5	5	3	3

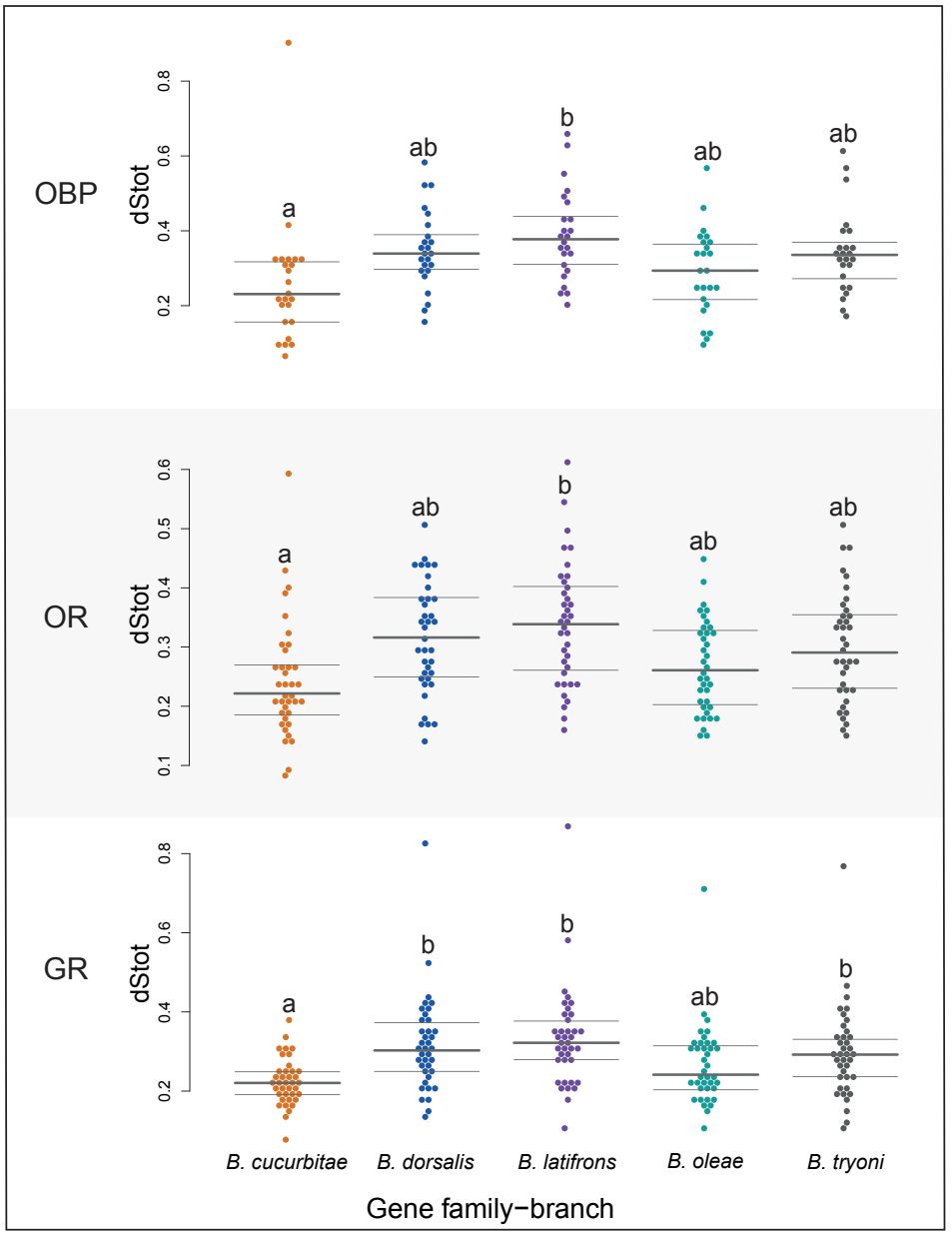


Figure 27 – Boxplots showing synonymous substitution rate for OBPs, ORs and GRs in the five *Bactrocera* species (patristic distances). Different letters identify significant statistical differences at adjusted $P < 0.05$ according to a Tukey's HSD multiple comparison test. Median and quantiles are shown as grey lines for each gene.

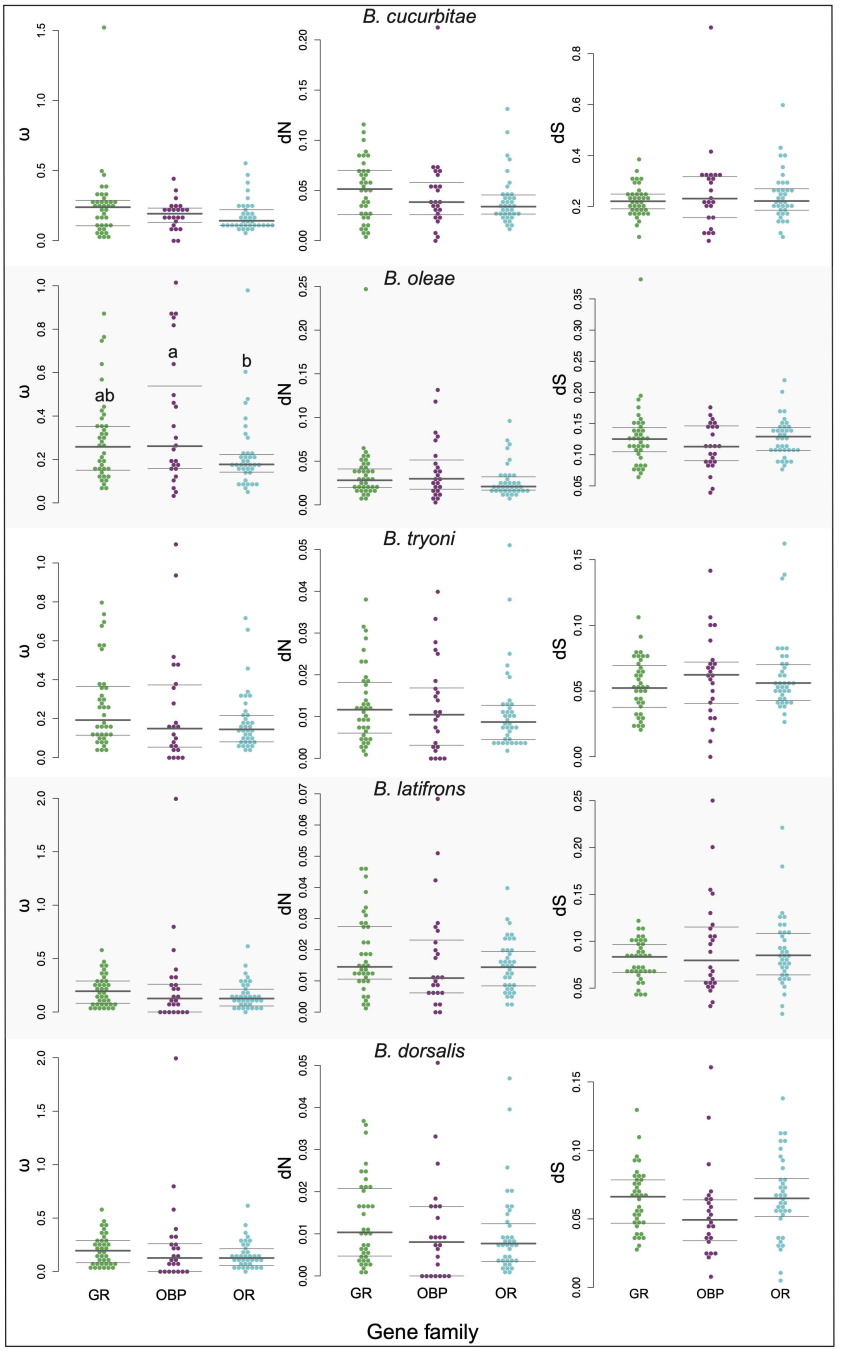


Figure 28 – Tukey boxplots showing ω, d_S and d_N for OBPs, ORs and GRs in the five *Bactrocera* species. Different letters identify significant statistical differences at adjusted $P < 0.05$ according to a Tukey's HSD multiple comparison test. Median and quantiles are shown as grey lines for each gene.

3.2.5. Homology modelling of olfactory proteins

We computed predictions of 3D molecular structures of eight candidate ORs and two OBPs that were identified as candidates for the action of positive selection. These models allowed us to gain insights in the degree of conservation of the different olfactory proteins and to map residues under positive selection (see 3.2.4 paragraph) (**Figure 29**).

Our results revealed that, despite sequence divergence among the different OR and OBP genes, the overall structure of the proteins is well conserved within each gene family (**Figure 29**), in line with conservation of their molecular function.

In ORs, amino acids under positive selection were mainly positioned within the transmembrane portion of the protein, indicating that such residues are unlikely to be involved in protein binding activities. The few sites under positive selection that were predicted to be located in the extra-membrane portion were characterized by low prediction scores, thus their actual role in the protein cannot be assessed (**Figure 29**). In Obp25, among the seven residues we found to be under positive selection in *B. oleae*, the cysteine at position 25 was substituted by a serine (C025S). Interestingly, in our computation homology model, this cysteine appeared to participate in a potential disulphide bond with cysteine at position 56 (**Figure 30**).

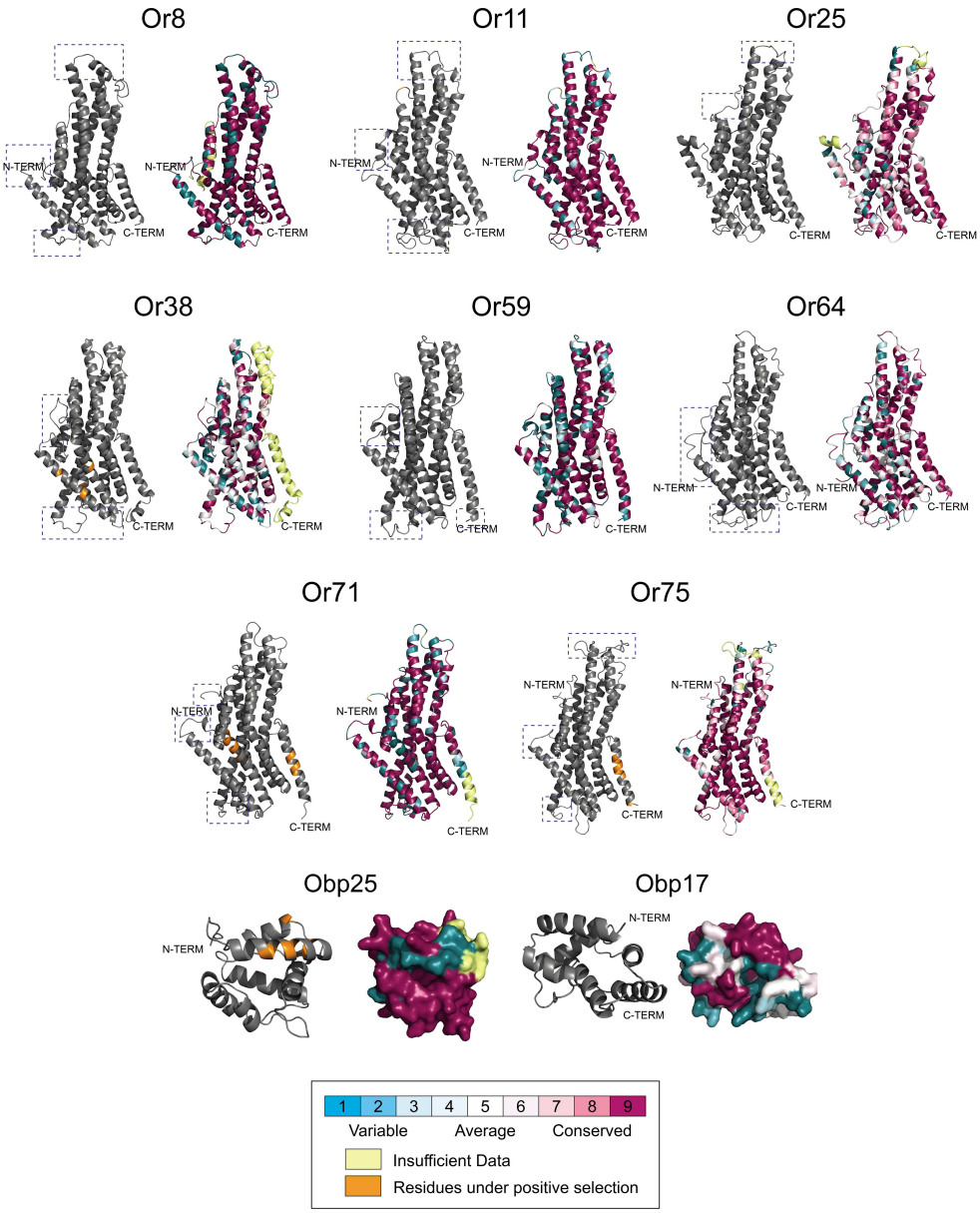


Figure 29 – Computational homology models of the *Bactrocera* OR and OBP proteins. Cartoon representations of eight OR and two OBP homology models in grey. Residues under positive selection are show in orange, regions of discordance between the models were highlighted by blue rectangles. On the side of each model, ConSurf overview of the aminoacid sequence conservation in *Bactrocera* mapped on three-dimensional homology models, coloured by conservation.

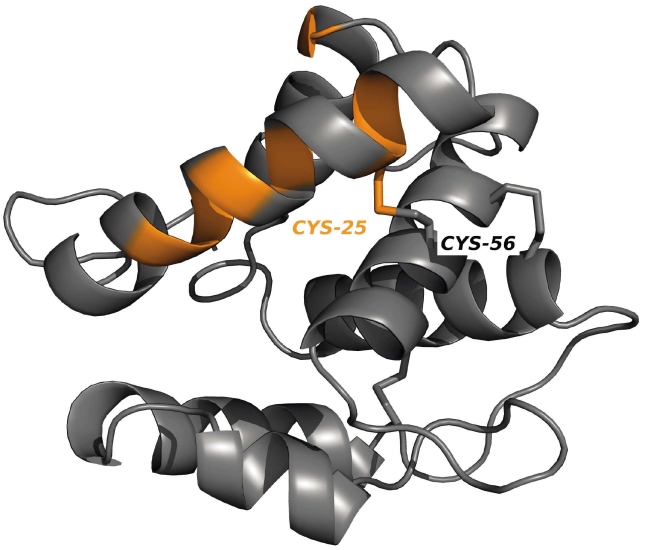


Figure 30 – Cartoon representation of Obp25 based on the 3D model. Predicted disulphide bond between Cys25 and Cys56 shown as sticks.

In addition, we mapped the degree of sequence conservation observed across the five *Bactrocera* species onto our computational models using the ConSurf server (**Figure 29, 31**).

We found that the amino acid conservation across ORs and OBPs was extremely variable. Overall, ORs reveal a high level of divergence between species. In particular, the receptors that appeared to be less conserved were Or38 with 38.2% of residues in the highest conservation grades 7–9, Or59 with 42.0%, Or64 with 42.4% and Or8 with 47.2% (**Figure 31**; **Table S6**).

In OBPs, Obp25 is highly conserved with 85.7% of residues with conservation grade 9, while only 14.3% of residues are highly variable with grade 1 (**Figure 31**; **Table S6**). Instead, the ConSurf profile of Obp17 showed a greater divergence among *Bactrocera* species with 55.2% of residues in the highest conservation grades 7–9, the 21.5% with grades 4–6 and the 23.3% of residues are highly variable with grades 1–3 (**Figure 31**; **Table S6**).

Figure 31 – Residue conservation among ORs and OBPs from the five *Bactrocera* species calculated using ConSurf.

Discussion

4.1. Multi-locus phylogeny of *Bactrocera* genus

Our phylogenetic analyses clearly reveal that *B. dorsalis* is more closely related to *B. latifrons* than to *B. tryoni*. This finding is in contrast with all phylogenies inferred from mitochondrial sequence (e.g. [31, 32]), which supported a closer relationship between *B. dorsalis* and *B. tryoni*.

Our results are instead consistent with what reported by a recent study that examined phylogenetic relationships among 167 Dacini species, including *Bactrocera* [29]. This study, however, was not conclusive in determining the relationships between these three species, since the phylogeny had many unresolved nodes, including those relative to the most common ancestors of *B. dorsalis*, *B. latifrons* and *B. tryoni*. The low power to disentangle such relationships likely derives from the small dataset – seven nuclear genes – used to produce their phylogeny. The same inconclusive result was reported by Choo and colleagues [285], who analysed 116 orthologous genes across 11 *Bactrocera* species. Despite the larger dataset, however, their results are not adequately supported: their Bayesian tree does not provide support values, and the ML analysis has a bootstrap value of 70 for the split between (*B. dorsalis*, *B. latifrons*) and *B. tryoni*, indicating lack of statistical confidence. The discordance between the results points to a complex evolutionary history of the group. In fact, our StarBeast2 results revealed numerous incongruences between gene and species trees (**Figure 16**), which can be interpreted as an evidence for incomplete lineage sorting and/or introgression (e.g., [286]). This would also explain the discordant results of the mitochondrial phylogenies, a finding that is commonly reported in many organisms, including insect species [287–290].

Incomplete lineage sorting is expected for rapid radiations, which is exactly what it is revealed by our molecular clock analyses. The clade that includes *B. dorsalis*, *B. latifrons,* and *B. tryoni* experienced a fast radiation about 2 million years ago (mid-Pliocene), when sea levels rose at peak level [291] and thus increased distances between islands and island groups, possibly facilitating allopatric speciation (the three species have native ranges in south east Asia and Australia). The rapid and successive speciation events were likely associated with incomplete lineage sorting. Moreover, we cannot exclude the possibility that these species experienced, or even still experience, hybridization events, which could then result in widespread introgression events.

Our dating analysis also revealed that the split between *Bactrocera* and *C. capitata* occurred less than 20 million years ago. The number of generations per year of *Bactrocera* is very similar to that estimated for *Drosophila* (around 10), a genus for which the date for the most recent

common ancestor has been estimated at 25–40 mya [246]. We, therefore, think that the relatively young age of the *Bactrocera* genus should advise against proposing splits into sub-clades (e.g., *Dacus*, *Zeugodacus*, etc.) that are not fully supported by significant genetic divergence.

The results of our dating analysis were also discordant with previously published works. In these studies, it has been proposed that *C. capitata* and *Bactrocera* diverged around 24.9 mya (no confidence intervals reported; [292]), 83 mya (95% high probability distribution: 64–103 mya; [293]), 110.9 mya (95% confidence interval: 91.2–131.4 mya; [24]) and 31.21 mya (95% confidence interval: 21.16–41.27 mya; [285]). The difference in the estimates could be due to several factors such as the type of molecular markers (nuclear or mitochondrial), the model choice, the calibration points, the mutation rate, and the number of generations per year. The latter greatly influences the divergence time estimation. For example, if in our analysis we assumed five generations per year (instead of 8) without changing the other parameters, the divergence time would increase from ~20 to ~32 mya. The latter is close to what estimated by Choo and colleagues, who performed phylogenetic and dating analyses using nuclear (116) and mitochondrial (13) genes of a similar dataset of species [285]. The results of their molecular clock analyses were not in agreement with our estimations (~31 vs. ~20 mya), but they used a less powerful approach, whereby they specified as prior only the divergence between *Rhagoletis* (Diptera: Tephritidae) and *Drosophila* previously inferred in [294].

Overall, our results highlight the importance of using genome-wide data to resolve complex phylogenies and provide a useful framework for future comparative genomics and comparative biology studies in *Bactrocera*.

4.2. The evolution of host selection in *Bactrocera* fruit flies

4.2.1. Chemosensory gene repertoire

Phytophagous insects rely on chemosensory perception to recognize chemical stimuli in their environment, such as food resources, mating partners, and oviposition sites.

Therefore, it has been proposed that chemosensory genes are likely to be under selective forces during host specialization.

In this thesis, we analysed the evolutionary dynamics of CSP, OBP, OR, and GR gene families among five *Bactrocera* species, including species with very diverse host ranges: two extreme polyphagous species (*B. dorsalis* and *B. tryoni*), a polyphagous species with a narrower host range (*B. latifrons*), an oligophagous species (*B. cucurbitae*), and a monophagous species (*B. oleae*).

We hypothesized that differences in the evolutionary behaviour of chemosensory gene families might be correlated with the different host ranges of these pest species.

For this purpose, we manually annotated the complete repertoire of chemosensory genes in the genomes of the five *Bactrocera* species. We performed various rounds of iterative searches and combined different resources to minimize false negatives.

Consistent with our hypothesis, we found that the specialist *B. oleae* was the species with the lowest number of OBP and OR genes and duplication events (*Or18* and *Or38like*). This is reasonable if we assume that genes involved in the perception of odorants absent in olives are not under purifying selection in *B. oleae* (unless they have pleiotropic effects on other phenotypic traits) and therefore may be lost (or degenerate) without consequences. In comparison, the extreme generalist *B. dorsalis* presents the highest number of OBP, OR, and GR genes and several duplications in these gene families (**Figure 20**). Again, this is expected for species that need to be efficient in locating a huge variety of fruits and therefore many different chemical signals.

Interestingly, we observed a burst of OR duplications in *B. cucurbitae* (**Figure 11**), which may be associated with its specialization towards Cucurbitaceae hosts. This, in turn, may be associated with species-specific abilities to respond to host volatiles, which could ultimately be used as lures in traps against this pest [295, 296].

We also noted that *B. tryoni* has a relatively low number of OBP, OR, and GR genes compared to the closest relatives (**Figure 20**). While this may indeed be a consequence of its biology, we cannot completely exclude a bias effect due to the lower coverage of the available genomic resources. The results of the BUSCO analyses point to slightly lower gene completeness of the *B. tryoni* datasets compared to the *B. dorsalis* one

(~2% less), however, not enough to explain the >10% difference in chemosensory genes between the two species.

Overall, our annotations displayed fair conservation of the CSP gene family size across the *Bactrocera* species, while the OR, GR, and OBP gene families were highly variable. A similar pattern was observed comparing the amino acidic identities (**Figure 24**). This result is consistent with what reported in *Drosophila* and in other dipteran species [119, 297, 298].

We identified several *Bactrocera*-specific expansions, especially in the olfactory genes (*Obp29*, *Obp33like*, *Or17*, *Or18*, *Or36*, *Or38like*, *Or56,* and *Gr68*) (**Figure 10, 11, 12**). These expansions may have an adaptive role since they may be associated with the higher diversification observed in the genus *Bactrocera* compared to *Ceratitis,* which includes only polyphagous species, or with the higher invasiveness of the *Bactrocera* species [299, 300]. We cannot exclude, however, a contraction of OR genes in *C. capitata*, since without an outgroup we are not able to polarize the gain/loss events between the two genera.

In *Bactrocera*, we also noticed a smaller GR family size compared to the outgroup *C. capitata*. On one hand, we may hypothesize that the contraction of this gene family indicates a less critical role of GRs in the detection of non-volatile compounds in this genus. On the other hand, we cannot exclude that the high sequence divergence of this gene family did not allow us to detect all the putative orthologs. However, we think that this is improbable because of the accurate annotation procedure that involved rounds of within-species BLAST searches. As for the OR genes, another possibility is an increase of GR genes in *C. capitata*. A more comprehensive comparative analysis is required to elucidate the role of GRs in tephritid ecology.

Our findings were also consistent with what was found in other insects that shifted their host preference compared to the closest relatives. For instance, in *Drosophila sechellia* and in Lepidoptera, it has been observed that repertoire sizes of chemosensory gene families are associated with differences in host use [225, 226, 301–304]. In all these studies, it seems that specialization on a novel host plant is correlated with a contraction of the GR family.

4.2.2. Chemosensory gene family evolution

We performed a birth-and-death analysis based on the time tree of the five *Bactrocera* species.

The monophagous *B. oleae* was the only species that experienced only losses (**Figure 25**). As a result, the β rate was zero for all the chemosensory genes and, in the OR gene family, the turnover rate was the lowest of the phylogeny (**Figure 26**). Instead, the polyphagous species exhibit mainly high turnover rates both in the internal branch and in the terminal nodes (**Figure 26**). As expected, the turnover rates were

particularly high in the branch leading to *B. tryoni* for all the gene families and this was mainly due to losses.

As discussed above, these results were concordant with the idea that host adaptation may be accompanied by the loss of genes and, in the case of the olive fly, host specialization affected mainly the evolution of the OR gene family. It is possible that *B. oleae* monophagy resulted in many chemorsensory genes to be not necessary because involved in the detection of odorants not found on olives.

In the OBP gene family, we noticed that *B. latifrons* showed a turnover rate more similar to the values observed in the oligophagous *B. cucurbitae* and monophagous *B. oleae* than to those of the polyphagous species (**Figure 26**). These data are in agreement with the feeding behaviour of this species, which mainly feeds on species of the Solanaceae family and support the hypothesis that *B. latifrons* should be considered an oligophagous species [36, 305].

In general, the OR and GR gene family were the most dynamic along the *Bactrocera* phylogeny. In these two gene families, we observed a positive correlation both in the gene family sizes and in the δ rates, highlighting the shared evolutionary trend in the expansion/contraction of these chemosensory gene families.

4.2.3. Signatures of selection

Our molecular evolution results showed that the olive fly had the highest numbers of chemosensory genes under positive selection (**Table 3**).

Interestingly, among them, we found the *BoleOr64,* whose ortholog in *B. dorsalis* was reported to mediate methyl eugenol perception [223]. ME is a powerful attractant for male fruit flies, including many *Bactrocera* species. It is a compound naturally occurring in many plant species and it is extensively used for the monitoring and trapping of fruit flies populations [306–308]. Notably, *B. oleae* is not attracted to methyl eugenol [309]. The selective pressure that we detected on this gene may indicate that *Or64* in *B. oleae* experienced a functional diversification. However, an accurate experimental characterization is needed to unravel the specific role of this receptor in the olive fly and the ME-responders *Bactrocera*.

In general, these genes that we found under selective pressure may have played an active role in the olive fly specialization, and they may be considered good candidates for functional studies.

We also found that the OR gene family was under more selective constraints than the OBP gene family, suggesting different selective pressures in these two classes of chemorsensory genes. This finding makes sense if we think that OBPs are supposed to be the direct interactors with the odorant molecules and thus are likely to evolve faster than the OR counterparts whenever the olfactory stimuli change, for instance, because of a host plant shift.

Bactrocera cucurbitae had the highest number of OR genes under selection. Interestingly, in recent works, *BcucOr59* (candidate according to the branch-test*)*, *BcucOr11* (branch-site) and *BcucOr38* (branch-site) were found to be expressed in the antenna [310], suggesting an active role of these genes in olfaction. These results, together with the high number of duplicated ORs in this species, may indicate that these events may have an adaptive role in the expansion of *B. cucurbitae* distribution to diverse ecological niches as previously proposed in *D. suzukii* [311]. This hypothesis is partially supported by studies on this species in the Afrotropical region. Vayssières and colleagues noted that the host range of *B. cucurbitae* presented geographical variations. It appeared to be more oligophagous on La Réunion Island while having a broader host range in western Africa with different infestations rates according to the region [41]. However, functional information will be fundamental to elucidate the role of these genes in oviposition preference.

4.2.4. Homology modelling of olfactory proteins

We computationally modelled the candidate ORs and OBPs under selection to investigate the putative role of the sites under selection on the protein function (**Figure 29, 31**).

This analysis allowed us to map the sites under positive selection in the protein's tertiary structure and evaluate whether these sites may be associated with specific domains of the protein. Moreover, we complemented this analysis by estimating the distribution of conservation across the protein to evaluate if sites candidate for positive selection resided in regions that are under specific selective regimes in *Bactrocera*.

This may help understand how these changes would affect their structures and, consequently, their functions. However, we note that we performed only computational inferences based on sequence similarity with proteins for which a structure has been determined, and thus our models correspond to the putative structure of these proteins in *Bactrocera*. Indeed, all models were generated using as reference three-dimensional structures of proteins of other organisms. Thus, we can only make tentative hypotheses on the implications of our findings since experimental validation is necessary to formulate significant conclusions.

In Obp25, we found seven sites under positive selection in *B. oleae,* which may promote functional divergence on the binding specificities. Interestingly, we note that in this species, a cysteine was substituted by a serine (C025S) and, in our computational model, this cysteine appeared to be involved in a potential disulphide bond (**Figure 30**). This change may impair the folding and the stability of the protein since it is well known that the disulphide bonds stabilize the tertiary structure [128, 129, 147] and help define the hydrophobic binding cavity [133, 312].

We also investigated whether sites under positive selection are located in the OBP binding cavity. In the olive fly, the alanine at position 18 is

substituted by an isoleucine (A018I) and this residue is located in the top-one internal binding pocket of the CASTp ranking [282] (**Figure 32**).

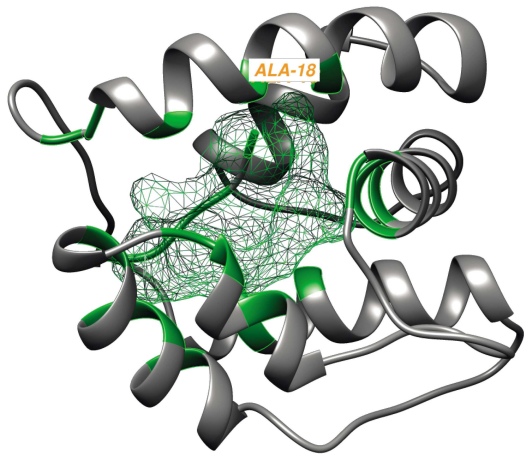


Figure 32 – Cartoon representations of Obp25 based on the 3D model. The odorant binding cavity estimated by CASTp is represented by green mesh as well as the residues that form it. Labelled in orange, the residue under positive selection in *B. oleae*.

This finding suggests that BoleObp25 may recognize and bind a different range of organic molecules and naturally occurring odorants. These changes may also be not directly involved with the binding function but relevant because they may impact the size and shape of the binding cavity. Obp25 ConSurf profile highlighted overall conservation among orthologous, with divergence between species observed mainly in the proximity of the binding cavity (**Figure 29, 31**; **Table S6**). Instead, the Obp17 protein appeared to be quite dissimilar in their sequences among the *Bactrocera* species (**Figure 29, 31**; **Table S6)**.

However, in the absence of functional evidence, it is unclear whether these OBPs are truly involved in olfaction. Further experiments are needed to determine if these proteins are expressed in olfactory tissues and what type of molecule they bind.

Regarding ORs, we noticed that residues under positive selection were mainly located within the predicted transmembrane regions of the proteins suggesting that these residues are not involved in the binding activities of the receptors, although they might affect protein stability (**Figure 29**).

The ConSurf analyses displayed a high level of divergence among the modelled ORs between species, especially for the Or8, the Or38 (which

presents site under positive selection in *B. cucurbitae*), Or59, and Or64 (**Figure 31**; **Table S6**). Even if we could not predict whether or how these differences have implications on the binding activity of these proteins, the fact that these genes are allowed to accumulate more substitutions than the other genes suggests that they are under relaxed selective pressure. Thus, even if such differences are not the result of positive selection, they can provide a source of variation which could eventually be co-opted for other purposes (i.e., other binding capabilities).

In the case of ORs, it is more challenging to speculate on the effect of the sites under positive selection because the only three-dimensional structure currently available is for a tetramer of the Odorant receptor co-receptor [167]. Indeed, very little is known about the role of amino acid changes on these transmembrane proteins.

Our 3D computational models of OBPs and ORs suggest that changes in the primary sequence may promote functional divergence of the binding specificities among species and may enable these proteins to recognize and bind new odorants. The above-mentioned proteins are promising targets for further structural characterization. This holds for OBPs in particular since their heterologous expression, and the subsequent three-dimensional structure determination are more feasible compared to ORs. In fact, to date, the structures of more than 20 OBPs have been solved by X-ray crystallography and/or nuclear magnetic resonance spectroscopy, some also complexed with ligands (reviewed in [313]). OBPs are thus appealing candidates for insect control, since determining their structural features may be the key to the design of more effective and species-specific attractants and repellents. Indeed, they can be exploited as new tools to hamper fruit fly olfaction [120, 314, 315], or to improve the efficiency of trapping systems [316]. On the other hand, it is well known that OBPs are not restricted to the olfactory organs, and there is an ongoing debate about their ligand binding specificity with some authors supporting a broad binding capacity to multiple volatiles [157, 317, 318], and others sustaining the striking specificity of some of them [319–321]. For this reason, we think that to achieve a complete understanding of the biological roles of these proteins, it is important to establish an integrated approach combing both in vitro and in vivo analyses. Such strategy should include molecular and structural biology, behavioural genetics, and electrophysiology.

Conclusions and perspectives

In this thesis, we conducted a comprehensive multi-locus phylogenetic analysis using genome-wide data of eleven *Bactrocera* fruit flies (*B. cucurbitae*, *B. dorsalis*, *B. latifrons*, *B. minax*, *B. oleae*, *B. bryoniae*, *B. correcta*, *B. jarvisi*, *B. musae*, *B. tryoni*, and *B. zonata*), including species with different feeding preferences. We also performed a molecular clock analysis to infer the time of divergence among the species.

Our results revealed that *B. dorsalis* is more closely related to *B. latifrons* than to *B. tryoni*. These findings are in contrast with all phylogenies inferred from mitochondrial sequences [31, 32], which supported a closer relationship between *B. dorsalis* and *B. tryoni*, and consistent with recent studies based on nuclear genes [29, 285], which however did not provide statistical support for the inferred topology. Our analyses revealed numerous incongruences between gene and species trees, which we interpreted as evidence for incomplete lineage sorting and/or introgression and would also explain the discordant results of the mitochondrial phylogenies. Overall, our results underlined the importance of using genome-wide data to resolve complex phylogenies. With this analysis, we provide a useful framework for comparative biology studies. The possibility that hybridization can still occur between closely related species also earns about the possibility that selective events in one species (for instance, resistance to insecticides) may be readily transferred to other species by introgression.

We also examined the role of the chemosensory gene families on host selection in five *Bactrocera* fruit flies with contrasting host preferences.

Interestingly, we noted that the specialist *B. oleae* was the species with the lowest number of OBP and OR genes and duplication events, while the extreme polyphagous *B. dorsalis* showed the highest number of OBP, OR, and GR genes and several duplications in these gene families. This is consistent with a reduction of the sensory spectrum in those species that need only a fine-tuned sensory apparatus to locate a single (or few) host plant. In fact, genes that are involved in the detection of chemicals unrelated to the host fruit are not essential and, therefore, can be lost or degraded without affecting the fitness of the fly. The birth-and-death analysis revealed that the olive fly was the only species that experienced exclusively gene losses. Overall, these results support the idea that host specialization may be followed by the loss of genes, and, in the case of *B. oleae*, this impacted mainly the evolution of the OR gene family. The results obtained testing the rate, and the pattern of molecular evolution revealed that the olive fly was the species with the highest number of genes under selection, and only in this species OR and OBP genes are evolving under contrasting selective pressures. Thus, the evolution of host choice in this genus may be the result of several evolutionary processes occurring at both the gene family and the gene sequence levels.

Our future studies will aim both at the analysis of the expression profile of the chemosensory genes under selection in olive fly and at their functional characterization to determine the exact involvement of these genes in the olfactory and gustatory functions and, ultimately, in its oviposition behaviour. Moreover, untangling the molecular machinery of chemoreception in the olive fly may lay the foundation for future research directed at the development of innovative insect control strategies against this species.

References

1. Norrbom AL, Carroll LE, Thompson FC, White IM, Freidberg A. Systematic database of names. Myia. 1999;:64–251.

2. Pape T, Bickel D, Meier R. Diptera Diversity: Status, Challenges and Tools. BRILL; 2009. doi: 10.1163/ej.9789004148970.I-459.

3. Schutze MK, Bourtzis K, Cameron SL, Clarke AR, De Meyer M, Hee AKW, et al. Integrative taxonomy versus taxonomic authority without peer review: the case of the Oriental fruit fly, *Bactrocera dorsalis* (Tephritidae). Syst Entomol. 2017;42:609–20. doi: 10.1111/syen.12250.

4. Thompson FC. Fruit fly expert identification system and systematic information database: A Resource for Identification and Information on Fruit Flies and Maggots, with Information on Their Classification, Distribution and Documentation. 1999.

5. White IM, Elson-Harris MM. Fruit flies of economic significance: their identification and bionomics. 1992.

6. Daane KM, Johnson MW. Olive Fruit Fly: Managing an Ancient Pest in Modern Times. Annu Rev Entomol. 2010;55:151–69. doi: 10.1146/annurev.ento.54.110807.090553.

7. De Meyer M, Robertson MP, Mansell MW, Ekesi S, Tsuruta K, Mwaiko W, et al. Ecological niche and potential geographic distribution of the invasive fruit fly *Bactrocera invadens* (Diptera, Tephritidae). Bull Entomol Res. 2010;100:35–48. doi: 10.1017/S0007485309006713.

8. Shelly TE. Sexual Selection on Leks: A Fruit Fly Primer. J Insect Sci. 2018;18. doi: 10.1093/jisesa/iey048.

9. Benelli G, Giunti G, Canale A, Messing RH. Lek dynamics and cues evoking mating behavior in tephritid flies infesting soft fruits: implications for behavior-based control tools. Appl Entomol Zool. 2014;49:363–73. doi: 10.1007/s13355-014-0276-9.

10. Prokopy RJ. Evidence for a Marking Pheromone Deterring Repeated Oviposition in Apple Maggot Flies. Environ Entomol. 1972;1:326–32.

11. Papaj DR, Katsoyannos BI, Hendrichs J. Use of fruit wounds in oviposition by Mediterranean fruit flies. Entomol Exp Appl. 1989.

12. Prokopy RJ, Roitberg BD. Foraging Behavior of True Fruit Flies: Concepts of foraging can be used to determine how tephritids search for food, mated and egg-laying sites and to help control these pests. Am Sci. 1984;72:41–9. https://www.jstor.org/stable/27852437?seq=1.

13. Drew RAI. The tropical fruit flies (Diptera: Tephritidae: Dacinae) of the Australasian and Oceanian regions. 1989.

14. Fletcher BS. Life history strategies of tephritid fruit flies. In: Fruit Flies: Their Biology, Natural Enemies and Control. Amsterdam, The Netherlands, Elsevier Science Publ.; 1989. p. 195–208.

15. Drew RAI, Hancock DL. Phylogeny of the tribe Dacini (Dacinae) based on morphological, distributional, and biological data. In: Fruit Flies (Tephritidae): Phylogeny and Evolution of Behavior. 1999.

16. White IM. Taxonomy of the Dacina (Diptera: Tephritidae) of Africa and the Middle East. African Entomol Mem. 2006.

17. Drew RAI. The generic and subgeneric classification of Dacini (Diptera: Tephritidae) from the South Pacific Area. Aust J Entomol. 1972.

18. Munro HK. A taxonomic treatise on the Dacidae (Tephritoidea, Diptera) of Africa. Entomol Mem S Afr Dept Agric. 1984.

19. Drew RAI, Hancock DL. The *Bactrocera dorsalis* complex of fruit flies (Diptera: Tephritidae: Dacinae) in Asia. Bull Entomol Res. 1994;2:1–68. doi: 10.1017/s1367426900000278.

20. White IM. Morphological Features of the Tribe Dacini (Dacinae): Their Significance to Behavior and Classification. 1999;:523–52. doi: 10.1201/9781420074468-29.

21. Drew RAI, Romig MC. Keys to the Tropical Fruit Flies (Tephritidae: Dacinae) of South-East Asia: Indomalaya to North-West Australasia. CAB International; 2016. https://books.google.it/books?id=QYflDQAAQBAJ.

22. Virgilio M, Jordaens K, Verwimp C, White IM, De Meyer M. Higher phylogeny of frugivorous flies (Diptera, Tephritidae, Dacini): Localised partition conflicts and a novel generic classification. Mol Phylogenet Evol. 2015;85:171–9. doi: 10.1016/j.ympev.2015.01.007.

23. De Meyer M, Delatte H, Mwatawala M, Quilici S, Vayssieres J-F, Virgilio M. A review of the current knowledge on *Zeugodacus cucurbitae* (Coquillett) (Diptera, Tephritidae) in Africa, with a list of species included in *Zeugodacus*. Zookeys. 2015;540:539–57. doi: 10.3897/zookeys.540.9672.

24. Krosch MN, Schutze MK, Armstrong KF, Graham GC, Yeates DK, Clarke AR. A molecular phylogeny for the Tribe Dacini (Diptera: Tephritidae): systematic and biogeographic implications. Mol Phylogenet Evol. 2012;64:513–23. doi: 10.1016/j.ympev.2012.05.006.

25. Freidberg A, Kovac D, Shiao SF. A revision of Ichneumonopsis Hardy, 1973 (Diptera: Tephritidae: Dacinae: Gastrozonini), Oriental bamboo-shoot fruit flies. Eur J Taxon. 2017;2017:1–23. doi: 10.5852/ejt.2017.317.

26. Muraji M, Nakahara S. Phylogenetic relationships among fruit flies, *Bactrocera* (Diptera, Tephritidae), based on the mitochondrial rDNA

sequences. Insect Mol Biol. 2001;10:549–59. doi: 10.1046/j.0962-1075.2001.00294.x.

27. Nakahara S, Muraji M. Phylogenetic Analyses of *Bactrocera* Fruit Flies (Diptera: Tephritidae) Based on Nucleotide Sequences of the Mitochondrial COI and COII Genes. 2008.

28. Smith PT, Kambhampati S, Armstrong KA. Phylogenetic relationships among *Bactrocera* species (Diptera: Tephritidae) inferred from mitochondrial DNA sequences. Mol Phylogenet Evol. 2003;26:8–17. doi: 10.1016/S1055-7903(02)00293-2.

29. San Jose M, Doorenweerd C, Leblanc L, Barr N, Geib S, Rubinoff D. Incongruence between molecules and morphology: A seven-gene phylogeny of Dacini fruit flies paves the way for reclassification (Diptera: Tephritidae). Mol Phylogenet Evol. 2018;121:139–49. doi: 10.1016/j.ympev.2017.12.001.

30. Dupuis JR, Bremer FT, Kauwe A, San Jose M, Leblanc L, Rubinoff D, et al. HiMAP: Robust phylogenomics from highly multiplexed amplicon sequencing. Mol Ecol Resour. 2018;18:1000–19. doi: 10.1111/1755-0998.12783.

31. Zhang A Bin, Liu YH, Wu WX, Wang Z Le. Molecular Phylogeny of *Bactrocera* Species (Diptera: Tephritidae: Dacini) Inferred from Mitochondrial Sequences of 16S rDNA and COI Sequences. Florida Entomol. 2010;93:369. doi: 10.1653/024.093.0308.

32. Yong H-S, Song S-L, Lim P-E, Eamsobhana P, Suana IW. Complete Mitochondrial Genome of Three *Bactrocera* Fruit Flies of Subgenus *Bactrocera* (Diptera: Tephritidae) and Their Phylogenetic Implications. PLoS One. 2016;11:e0148201. doi: 10.1371/journal.pone.0148201.

33. Nardi F, Carapelli A, Dallai R, Roderick GK, Frati F. Population structure and colonization history of the olive fly, *Bactrocera oleae* (Diptera, Tephritidae). Mol Ecol. 2005;14:2729–38. doi: 10.1111/j.1365-294X.2005.02610.x.

34. Rice RE, Phillips PA, Stewart-Leslie J, Sibbett GS. Olive fruit fly populations measured in Central and Southern California. Calif Agric. 2003;57:122–7. doi: 10.3733/ca.v057n04p122.

35. Augustinos AA, Stratikopoulos EE, Zacharopoulou A, Mathiopoulos KD. Polymorphic microsatellite markers in the olive fly, *Bactrocera oleae*. Molecular Ecology Notes. 2002. doi: 10.1046/j.1471-8286.2002.00222.x.

36. Allwood AJ, Chinajariyawong A, Kritsaneepaiboon S, Drew RAI, Hamacek EL, Hancock DL, et al. Host plant records for fruit flies (Diptera: Tephritidae) in Southeast Asia. Raffles Bull Zool. 1999;47 7 SUPPL.:1–92.

37. Barr NB, Ledezma LA, Leblanc L, San Jose M, Rubinoff D, Geib SM, et

al. Genetic Diversity of *Bactrocera dorsalis* (Diptera: Tephritidae) on the Hawaiian Islands: Implications for an Introduction Pathway Into California. J Econ Entomol. 2014;107:1946–58. doi: 10.1603/EC13482.

38. Lux SA, Copeland RS, White IM, Manrakhan A, Billah MK. A New Invasive Fruit Fly Species from the *Bactrocera dorsalis* (Hendel) Group Detected in East Africa. Int J Trop Insect Sci. 2003;23:355–61. doi: 10.1017/S174275840001242X.

39. Vayssières JF, Goergen G, Lokossou O, Dossa P, Akponon C. A new *Bactrocera* species in Benin among mango fruit fly (Diptera: Tephritidae) species. Fruits. 2005;60:371–7.

40. Nugnes F, Russo E, Viggiani G, Bernardo U. First Record of an Invasive Fruit Fly Belonging to *Bactrocera dorsalis* Complex (Diptera: Tephritidae) in Europe. Insects. 2018;9:182. doi: 10.3390/insects9040182.

41. Vayssières J-F, Rey J-Y, Traoré L. Distribution and host plants of *Bactrocera cucurbitae* in West and Central Africa. Fruits. 2007;62:391–6. doi: 10.1051/fruits:2007037.

42. Mwatawala M, Makundi R, Maerere, Amon P and De Meyer M. Occurrence of the Solanum fruit fly *Bactrocera latifrons* (Hendel) (Diptera: Tephritidae) in Tanzania. J Afrotrop Zool. 2010;:83–9.

43. Vargas RI, Nishida T. Survey for *Dacus latifrons* (Diptera: Tephritidae). J Econ Entomol. 1985.

44. Liquido NJ, Harris EJ, Dekker LA. Ecology of *Bactrocera latifrons* (Diptera: Tephritidae) populations: Host plants, natural enemies, distribution, and abundance. Ann Entomol Soc Am. 1994;87:71–84.

45. Gilchrist AS, Dominiak B, Gillespie PS, Sved JA. Variation in population structure across the ecological range of the Queensland fruit fly, *Bactrocera tryoni*. Aust J Zool. 2006;54:87. doi: 10.1071/ZO05020.

46. Drew RAI, Hooper GHS, Bateman MA. Economic fruit flies of the South Pacific Region. Econ fruit flies South Pacific Reg. 1978.

47. Bateman MA. The Ecology of Fruit Flies. Annu Rev Entomol. 1972.

48. Clarke A. Biology and management of *Bactrocera* and related fruit flies. Wallingford: CABI; 2019. doi: 10.1079/9781789241822.0000.

49. Ant T, Koukidou M, Rempoulakis P, Gong H-F, Economopoulos A, Vontas J, et al. Control of the olive fruit fly using genetics-enhanced sterile insect technique. BMC Biol. 2012;10:51. doi: 10.1186/1741-7007-10-51.

50. Tzanakakis M. Seasonal development and dormancy of insects and mites feeding on olive: a review. Netherlands J Zool. 2003;52:87–224. doi: 10.1163/156854203764817670.

51. Tsiropoulos GJ. Reproduction and Survival of the Adult *Dacus oleae* Feeding on Pollens and Honeydews. Environ Entomol. 1977;6:390–2. doi: 10.1093/ee/6.3.390.

52. Tsiropoulos GJ. The importance of dietary amino acids on the reproduction and longevity of adult *Dacus oleae* (Gmelin) (Diptera Tephritidae). Arch Int Physiol Biochim. 1983.

53. Tsiropoulos GJ. Amino-acid synthesis in adult *Dacus oleae* (Gmelin) (Diptera Tephritidae) determined with [U-^{14}C] glucose. Arch Physiol Biochem. 1984.

54. Athar M. Infestation of olive fruit fly, *Bactrocera oleae*, in California and taxonomy of its host trees. Agric Conspec Sci. 2005.

55. Joy Burrack H, Zalom FG. Olive Fruit Fly (Diptera: Tephritidae) Ovipositional Preference and Larval Performance in Several Commercially Important Olive Varieties in California. J Econ Entomol. 2008;101:750–8. doi: 10.1093/jee/101.3.750.

56. Kounatidis I, Papadopoulos NT, Mavragani-Tsipidou P, Cohen Y, Tertivanidis K, Nomikou M, et al. Effect of elevation on spatio-temporal patterns of olive fly (*Bactrocera oleae*) populations in northern Greece. J Appl Entomol. 2008.

57. Back EA, Pembërton CE. Life history of the melon fly. J Agric Res. 1914;3:269–74.

58. Al-Zaghal K. Studies on the pupation of the olive fruit fly *Dacus oleae* Gmel. (Diptera, Tephritidae) in Jordan. J Appl Entomol. 1987.

59. Donia AR, El-Sawaf SK, Abou-Ghadir MF. Number of generations and seasonal abundance of the olive fruit fly, *Dacus oleae* (Gmel.) and the susceptibility of different olive varieties to infestation. Diptera: Trypetidae. Soc Entomol Egypte Bull. 1972. https://agris.fao.org/agris-search/search.do?recordID=US201303263089.

60. Mustafa TM, Al-Zaghal K. Frequency of *Dacus oleae* (Gmelin) immature stages and their parasites in seven olive varieties in Jordan. Int J Trop Insect Sci. 1987.

61. Sharaf NS. Life history of the olive fruit fly, *Dacus oleae* Gmel. (Diptera: Tephritidae), and its damage to olive fruits in Tripolitania. Zeitschrift für Angew Entomol. 1980.

62. Economopoulos AP, Haniotakis GE, Michelakis S, Tsiropoulos GJ, Zervas GA, Tsitsipis JA, et al. Population studies on the olive fruit fly, *Dacus oleae* (Gmel.) (Dipt., Tephritidae) in Western Crete. Zeitschrift für Angew Entomol. 1982;93:463–76. doi: 10.1111/j.1439-0418.1982.tb03621.x.

63. Kapatos ET FB. The Phenology of the Olive Fly, *Dacus oleae* (Gmel.) (Diptera, Tephritidae), in Corfu. Zeitschrift für Angew Entomol. 1984.

64. Michelakis SE. The olive fly (*Dacus oleae* Gmel.) in Crete, Greece. Acta Hortic. 1990;:371–4.

65. Neuenschwander P, Michelakis S. Determination of the lower thermal thresholds and day-degree requirements for eggs and larvae of *Dacus oleae* (Gmel.) (Diptera: Tephritidae) under field conditions in Crete, Greece. Mitteilungen der Schweizerischen Entomol Gesellschaft. 1979.

66. Dhillon MK, Singh R, Naresh JS, Sharma HC. The melon fruit fly, *Bactrocera cucurbitae*: A review of its biology and management. Journal of Insect Science. 2005;5. doi: 10.1093/jis/5.1.40.

67. Weems H V, Heppner JB, Fasulo TR. Melon Fly, *Bactrocera cucurbitae* (Coquillett) (Diptera: Tephritidae). In: SpringerReference. 2012. doi: 10.1007/springerreference_89671.

68. Zhi P. Development of resistance to trichlorophon, alphamethrin,and abamectin in laboratory populations of the oriental fruit fly, *Bactrocera dorsalis* (Hendel)(Diptera:Tephritidae). 2008.

69. Sutherst RW, Yonow T. The geographical distribution of the Queensland fruit fly, *Bactrocera* (*Dacus*) *tryoni*, in relation to climate. Aust J Agric Res. 1998;49:935. doi: 10.1071/A97152.

70. Vargas RI, Nishida T. Life History and Demographic Parameters of *Dacus latifrons* (Diptera: Tephritidae)1. J Econ Entomol. 1985.

71. Bernays EA, Chapman RE. Host-Plant Selection by Phytophagous Insects. 1994.

72. Prokopy RJ, Owens ED. Visual detection of plants by herbivorous insects. Annu Rev Entomol Vol 28. 1983;28:337–64. doi: 10.1146/annurev.en.28.010183.002005.

73. Stark RW. Generalized ecology and life cycles of bark beetles. In: Bark Beetles in North American Conifers: A System for the Study of Evolutionary Biology. 1982. p. 21–45.

74. Bernays E, Graham M. On the Evolution of Host Specificity in Phytophagous Arthropods. Ecology. 1988;69:886–92. doi: 10.2307/1941237.

75. Michaud JP. Conditions for the Evolution of Polyphagy in Herbivorous Insects. Oikos. 1990;57:278. doi: 10.2307/3565951.

76. Walter GH. Insect Pest Management and Ecological Research. Cambridge University Press; 2003. doi: 10.1017/CBO9780511525612.

77. Bernays EA. Host specificity in phytophagous insects: selection pressure

from generalist predators. Entomol Exp Appl. 1988.

78. Loxdale HD, Lushai G, Harvey JA. The evolutionary improbability of 'generalism' in nature, with special reference to insects. Biol J Linn Soc. 2011;103:1–18. doi: 10.1111/j.1095-8312.2011.01627.x.

79. Jaenike J. Host Specialization in Phytophagous Insects. Annu Rev Ecol Syst. 1990;21:243–73. doi: 10.1146/annurev.es.21.110190.001331.

80. Hancock, D. L., Hamacek, E. L., Lloyd, A. C., & Elson-Harris MM. The distribution and host plants of fruit flies (Diptera: Tephritidae) in Australia. 2000.

81. Leblanc L, Vueti ET, Allwood A. Host Plant Records for Fruit Flies (Diptera: Tephritidae: Dacini) in the Pacific Islands: 2. Infestation Statistics on Economic Hosts. 2013.

82. Normark BB, Johnson NA. Niche explosion. Genetica. 2011;139:551–64. doi: 10.1007/s10709-010-9513-5.

83. Drew RAI. Biogeography and speciation in the Dacini (Diptera: Tephritidae: Dacinae). Bish Museum Bull Entomol. 2004;12:165–78. http://hdl.handle.net/10072/5230.

84. De Meyer M. Phylogeny of the genus *Ceratitis* (Dacinae: Ceratitidini). In: Aluja M, Norrbom A, editors. Fruit Flies (Tephritidae): Phylogeny and Evolution of Behavior. CRC Press; 1999. p. 409–28. doi: 10.1201/9781420074468.

85. Mwatawala M, Maerere AP, Makundi R, De Meyer M. Incidence and host range of the melon fruit fly *Bactrocera cucurbitae* (Coquillett) (Diptera: Tephritidae) in Central Tanzania. Int J Pest Manag. 2010;56:265–73. doi: 10.1080/09670871003596792.

86. Clarke AR, Allwood A, Chinajariyawong A, Drew RAI, Hengsawad C, Jirasurat M, et al. Seasonal abundance and host use patterns of seven *Bactrocera* macquart species (Diptera: Tephritidae) in Thailand and Peninsular Malaysia. Raffles Bull Zool. 2001.

87. Leblanc L, Vueti E, Drew R, Allwood A. Host Plant Records for Fruit Flies (Diptera: Tephritidae: Dacini) in the Pacific Islands. Proc Hawaiian Entomol Soc. 2012;44:11–53.

88. Bargmann CI. Chemosensation in *C. elegans*. WormBook : the online review of *C. elegans* biology. 2006;doi/10.189:1–29.

89. Vosshall LB, Stocker RF. Molecular Architecture of Smell and Taste in *Drosophila*. Annu Rev Neurosci. 2007;30:505–33. doi: 10.1146/annurev.neuro.30.051606.094306.

90. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet.

2008;9:951–63. doi: 10.1038/nrg2480.

91. Kaupp UB. Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci. 2010;11:188–200. doi: 10.1038/nrn2789.

92. Touhara K, Vosshall LB. Sensing Odorants and Pheromones with Chemosensory Receptors. Annu Rev Physiol. 2009;71:307–32. doi: 10.1146/annurev.physiol.010908.163209.

93. Stengl M, Ziegelberger G, Boekhoff I, Krieger J. Perireceptor Events and Transduction Mechanisms in Insect Olfaction. In: Insect Olfaction. 1999.

94. Hildebrand JG, Shepherd GM. Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. Annual Review of Neuroscience. 1997.

95. Shanbhag SR, Müller B, Steinbrecht RA. Atlas of olfactory organs of *Drosophila melanogaster* 1. Types, external organization, innervation and distribution of olfactory sensilla. Int J Insect Morphol Embryol. 1999.

96. Pelosi P. Perireceptor events in olfaction. Journal of Neurobiology. 1996.

97. Sánchez-Gracia A, Vieira FG, Rozas J. Molecular evolution of the major chemosensory gene families in insects. Heredity (Edinb). 2009;103:208–16. doi: 10.1038/hdy.2009.55.

98. Robertson HM. Molecular Evolution of the Major Arthropod Chemoreceptor Gene Families. Annu Rev Entomol. 2019;64:227–42. doi: 10.1146/annurev-ento-020117-043322.

99. Clyne PJ. Candidate Taste Receptors in *Drosophila*. Science. 2000;287:1830–4. doi: 10.1126/science.287.5459.1830.

100. Robertson HM, Warr CG, Carlson JR. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 2003;100:14537–42. doi: 10.1073/pnas.2335847100.

101. Scott K, Brady R, Cravchik A, Morozov P, Rzhetsky A, Zuker C, et al. A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila*. Cell. 2001;104:661–73. doi: 10.1016/S0092-8674(01)00263-X.

102. Clyne PJ, Warr CG, Freeman MR, Lessing D, Kim J, Carlson JR. A novel family of divergent seven-transmembrane proteins: Candidate odorant receptors in *Drosophila*. Neuron. 1999;22:327–38. doi: 10.1016/S0896-6273(00)81093-4.

103. Gao Q, Chess A. Identification of candidate *Drosophila* olfactory receptors from genomic DNA sequence. Genomics. 1999;60:31–9. doi: 10.1006/geno.1999.5894.

104. Fox AN, Pitts RJ, Robertson HM, Carlson JR, Zwiebel LJ. Candidate odorant receptors from the malaria vector mosquito *Anopheles gambiae* and evidence of down-regulation in response to blood feeding. Proc Natl Acad Sci U S A. 2001;98:14693–7. doi: 10.1073/pnas.261432998.

105. Engsontia P, Sanderson AP, Cobb M, Walden KKO, Robertson HM, Brown S. The red flour beetle's large nose: An expanded odorant receptor gene family in *Tribolium castaneum*. Insect Biochem Mol Biol. 2008;38:387–97. doi: 10.1016/j.ibmb.2007.10.005.

106. Abuin L, Bargeton B, Ulbrich MH, Isacoff EY, Kellenberger S, Benton R. Functional architecture of olfactory ionotropic glutamate receptors. Neuron. 2011;69:44–60. doi: 10.1016/j.neuron.2010.11.042.

107. Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. Cell. 2009;136:149–62. doi: 10.1016/j.cell.2008.12.001.

108. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, et al. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 2010;6:e1001064. doi: 10.1371/journal.pgen.1001064.

109. Tegoni M, Pelosi P, Vincent F, Spinelli S, Campanacci V, Grolli S, et al. Mammalian odorant binding proteins. Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology. 2000;1482:229–40. doi: 10.1016/S0167-4838(00)00167-9.

110. Leal WS. Odorant reception in insects: Roles of receptors, binding proteins, and degrading enzymes. Annu Rev Entomol. 2013;58 September 2012:373–91. doi: 10.1146/annurev-ento-120811-153635.

111. Wicher D. Olfactory signaling in insects. In: Progress in Molecular Biology and Translational Science. Elsevier B.V.; 2015. p. 37–54.

112. Vogt RG, Riddiford LM. Pheromone binding and inactivation by moth antennae. Nature. 1981;293:161–3.

113. Pelosi P. Odorant-binding proteins. Crit Rev Biochem Mol Biol. 1994.

114. McKenna MP, Hekmat-Scafe DS, Gaines P, Carlson JR. Putative *Drosophila* pheromone-binding proteins expressed in a subregion of the olfactory system. J Biol Chem. 1994;269:16340–7. https://www.jbc.org/content/269/23/16340.

115. Pikielny CW, Hasan G, Rouyer F, Rosbash M. Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs. Neuron. 1994.

116. Pelosi P, Maida R. Odorant-binding proteins in vertebrates and insects: similarities and possible common function. Chem Senses. 1990;15:205–

15. doi: 10.1093/chemse/15.2.205.

117. Tegoni M, Pelosi P, Vincent F, Spinelli S, Campanacci V, Grolli S, et al. Mammalian odorant binding proteins. Biochim Biophys Acta - Protein Struct Mol Enzymol. 2000;1482:229–40. doi: 10.1016/S0167-4838(00)00167-9.

118. Menuz K, Larter NK, Park J, Carlson JR. An RNA-Seq Screen of the *Drosophila* Antenna Identifies a Transporter Necessary for Ammonia Detection. PLoS Genet. 2014;10:e1004810. doi: 10.1371/journal.pgen.1004810.

119. Vieira FG, Rozas J. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol. 2011;3:476–90. doi: 10.1093/gbe/evr033.

120. Pelosi P, Mastrogiacomo R, Iovinella I, Tuccori E, Persaud KC. Structure and biotechnological applications of odorant-binding proteins. Appl Microbiol Biotechnol. 2014;98:61–70. doi: 10.1007/s00253-013-5383-y.

121. Hekmat-Scafe DS. Genome-Wide Analysis of the Odorant-Binding Protein Gene Family in *Drosophila melanogaster*. Genome Res. 2002;12:1357–69. doi: 10.1101/gr.239402.

122. Vieira FG, Sánchez-Gracia A, Rozas J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. Genome Biol. 2007;8:R235. doi: 10.1186/gb-2007-8-11-r235.

123. Xu P, Atkinson R, Jones DNM, Smith DP. *Drosophila* OBP LUSH Is Required for Activity of Pheromone-Sensitive Neurons. Neuron. 2005;45:193–200. doi: 10.1016/j.neuron.2004.12.031.

124. Biessmann H, Andronopoulou E, Biessmann MR, Douris V, Dimitratos SD, Eliopoulos E, et al. The *Anopheles gambiae* Odorant Binding Protein 1 (AgamOBP1) Mediates Indole Recognition in the Antennae of Female Mosquitoes. PLoS One. 2010;5:e9471. doi: 10.1371/journal.pone.0009471.

125. Pelletier J, Guidolin A, Syed Z, Cornel AJ, Leal WS. Knockdown of a Mosquito Odorant-binding Protein Involved in the Sensitive Detection of Oviposition Attractants. J Chem Ecol. 2010;36:245–8. doi: 10.1007/s10886-010-9762-x.

126. Sandler BH, Nikonova L, Leal WS, Clardy J. Sexual attraction in the silkworm moth: Structure of the pheromone-binding-protein-bombykol complex. Chem Biol. 2000;7:143–51. doi: 10.1016/S1074-5521(00)00078-8.

127. Wojtasek H, Leal WS. Conformational Change in the Pheromone-

binding Protein from *Bombyx mori* Induced by pH and by Interaction with Membranes. J Biol Chem. 1999;274:30950–6. doi: 10.1074/jbc.274.43.30950.

128. Scaloni A, Monti M, Angeli S, Pelosi P. Structural analysis and disulfide-bridge pairing of two odorant-binding proteins from *Bombyx mori*. Biochem Biophys Res Commun. 1999;266:386–91. doi: 10.1006/bbrc.1999.1791.

129. Leal WS, Nikonova L, Peng G. Disulfide structure of the pheromone binding protein from the silkworm moth, *Bombyx mori*. FEBS Lett. 1999;464:85–90. doi: 10.1016/S0014-5793(99)01683-X.

130. Briand L, Nespoulous C, Huet J-C, Takahashi M, Pernollet J-C. Ligand binding and physico-chemical properties of ASP2, a recombinant odorant-binding protein from honeybee (*Apis mellifera* L.). Eur J Biochem. 2001;268:752–60. doi: 10.1046/j.1432-1327.2001.01927.x.

131. Xu PX, Zwiebel LJ, Smith DP. Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, *Anopheles gambiae*. Insect Mol Biol. 2003;12:549–60. doi: 10.1046/j.1365-2583.2003.00440.x.

132. Zhou J-J, Huang W, Zhang G-A, Pickett JA, Field LM. "Plus-C" odorant-binding protein genes in two *Drosophila* species and the malaria mosquito *Anopheles gambiae*. Gene. 2004;327:117–29. doi: 10.1016/j.gene.2003.11.007.

133. Lagarde A, Spinelli S, Qiao H, Tegoni M, Pelosi P, Cambillau C. Crystal structure of a novel type of odorant-binding protein from *Anopheles gambiae*, belonging to the C-plus class. Biochem J. 2011;437:423–30. doi: 10.1042/BJ20110522.

134. Spinelli S, Lagarde A, Iovinella I, Legrand P, Tegoni M, Pelosi P, et al. Crystal structure of *Apis mellifera* OBP14, a C-minus odorant-binding protein, and its complexes with odorant molecules. Insect Biochem Mol Biol. 2012;42:41–50. doi: 10.1016/j.ibmb.2011.10.005.

135. Klein U. Sensillum-lymph proteins from antennal olfactory hairs of the moth *Antheraea polyphemus* (Saturniidae). Insect Biochem. 1987;17:1193–204. doi: 10.1016/0020-1790(87)90093-X.

136. Jin X, Brandazza A, Navarrini A, Ban L, Zhang S, Steinbrecht RA, et al. Expression and immunolocalisation of odorant-binding and chemosensory proteins in locusts. C Cell Mol Life Sci. 2005;62:1156–66. doi: 10.1007/s00018-005-5014-6.

137. Schultze A, Pregitzer P, Walter MF, Woods DF, Marinotti O, Breer H, et al. The Co-Expression Pattern of Odorant Binding Proteins and Olfactory Receptors Identify Distinct Trichoid Sensilla on the Antenna of the Malaria Mosquito *Anopheles gambiae*. PLoS One. 2013;8:e69412. doi:

10.1371/journal.pone.0069412.

138. Galindo K, Smith DP. A large family of divergent *Drosophila* odorant-binding proteins expressed in gustatory and olfactory sensilla. Genetics. 2001;159:1059–72.

139. Jeong YT, Shim J, Oh SR, Yoon HI, Kim CH, Moon SJ, et al. An Odorant-Binding Protein Required for Suppression of Sweet Taste by Bitter Chemicals. Neuron. 2013;79:725–37. doi: 10.1016/j.neuron.2013.06.025.

140. Park S-K, Shanbhag SR, Wang Q, Hasan G, Steinbrecht RA, Pikielny CW. Expression patterns of two putative odorant-binding proteins in the olfactory organs of *Drosophila melanogaster* have different implications for their functions. Cell Tissue Res. 2000;300:181–92. doi: 10.1007/s004410000187.

141. Kaissling K. Chemo-Electrical Transduction in Insect Olfactory Receptors. Annu Rev Neurosci. 1986;9:121–45. doi: 10.1146/annurev.ne.09.030186.001005.

142. Pophof B. Pheromone-binding Proteins Contribute to the Activation of Olfactory Receptor Neurons in the Silkmoths *Antheraea polyphemus* and *Bombyx mori*. Chem Senses. 2004;29:117–25. doi: 10.1093/chemse/bjh012.

143. Arya GH, Weber AL, Wang P, Magwire MM, Negron YLS, Mackay TFC, et al. Natural Variation, Functional Pleiotropy and Transcriptional Contexts of Odorant Binding Protein Genes in *Drosophila melanogaster*. Genetics. 2010;186:1475–85. doi: 10.1534/genetics.110.123166.

144. Findlay GD, Yi X, MacCoss MJ, Swanson WJ. Proteomics Reveals Novel *Drosophila* Seminal Fluid Proteins Transferred at Mating. PLoS Biol. 2008;6:e178. doi: 10.1371/journal.pbio.0060178.

145. Qiao H, He X, Schymura D, Ban L, Field L, Dani FR, et al. Cooperative interactions between odorant-binding proteins of *Anopheles gambiae*. Cell Mol Life Sci. 2011;68:1799–813. doi: 10.1007/s00018-010-0539-8.

146. Angeli S, Ceron F, Scaloni A, Monti M, Monteforti G, Minnocci A, et al. Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. Eur J Biochem. 1999;262:745–54.

147. Tegoni M, Campanacci V, Cambillau C. Structural aspects of sexual attraction and chemical communication in insects. Trends Biochem Sci. 2004;29:257–64. doi: 10.1016/j.tibs.2004.03.003.

148. Campanacci V, Lartigue A, Hallberg BM, Jones TA, Giudici-Orticoni M-T, Tegoni M, et al. Moth chemosensory protein exhibits drastic conformational changes and cooperativity on ligand binding. Proc Natl Acad Sci. 2003;100:5069–74. doi: 10.1073/pnas.0836654100.

149. Monteforti G, Angeli S, Petacchi R, Minnocci A. Ultrastructural characterization of antennal sensilla and immunocytochemical localization of a chemosensory protein in *Carausius morosus* Brünner (Phasmida: Phasmatidae). Arthropod Struct Dev. 2002;30:195–205. doi: 10.1016/S1467-8039(01)00036-6.

150. Briand L, Swasdipan N, Nespoulous C, Bézirard V, Blon F, Huet J-C, et al. Characterization of a chemosensory protein (ASP3c) from honeybee (*Apis mellifera* L.) as a brood pheromone carrier. Eur J Biochem. 2002;269:4586–96. doi: 10.1046/j.1432-1033.2002.03156.x.

151. González D, Zhao Q, McMahan C, Velasquez D, Haskins WE, Sponsel V, et al. The major antennal chemosensory protein of red imported fire ant workers. Insect Mol Biol. 2009;18:395–404. doi: 10.1111/j.1365-2583.2009.00883.x.

152. Ozaki M. Ant Nestmate and Non-Nestmate Discrimination by a Chemosensory Sensillum. Science. 2005;309:311–4. doi: 10.1126/science.1105244.

153. Baer B, Zareie R, Paynter E, Poland V, Millar AH. Seminal fluid proteins differ in abundance between genetic lineages of honeybees. J Proteomics. 2012;75:5646–53. doi: 10.1016/j.jprot.2012.08.002.

154. Xuan N, Guo X, Xie H-Y, Lou Q-N, Lu X-B, Liu G-X, et al. Increased expression of CSP and CYP genes in adult silkworm females exposed to avermectins. Insect Sci. 2015;22:203–19. doi: 10.1111/1744-7917.12116.

155. Liu G, Ma H, Xie H, Xuan N, Guo X, Fan Z, et al. Biotype Characterization, Developmental Profiling, Insecticide Response and Binding Property of *Bemisia tabaci* Chemosensory Proteins: Role of CSP in Insect Defense. PLoS One. 2016;11:e0154706. doi: 10.1371/journal.pone.0154706.

156. Zhu J, Iovinella I, Dani FR, Liu Y-L, Huang L-Q, Liu Y, et al. Conserved chemosensory proteins in the proboscis and eyes of Lepidoptera. Int J Biol Sci. 2016;12:1394–404. doi: 10.7150/ijbs.16517.

157. Pelosi P, Iovinella I, Felicioli A, Dani FR. Soluble proteins of chemical communication: an overview across arthropods. Front Physiol. 2014;5:1–13. doi: 10.3389/fphys.2014.00320.

158. Pelosi P, Iovinella I, Zhu J, Wang G, Dani FR. Beyond chemoreception: diverse tasks of soluble olfactory proteins in insects. Biol Rev. 2018;93:184–200. doi: 10.1111/brv.12339.

159. Sabatier L, Jouanguy E, Dostert C, Zachary D, Dimarcq J-L, Bulet P, et al. Pherokine-2 and -3. Two *Drosophila* molecules related to pheromone/odor-binding proteins induced by viral and bacterial infections. Eur J Biochem. 2003;270:3398–407. doi: 10.1046/j.1432-

1033.2003.03725.x.

160. Forêt S, Wanner KW, Maleszka R. Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. Insect Biochem Mol Biol. 2007;37:19–28. doi: 10.1016/j.ibmb.2006.09.009.

161. Benton R, Sachse S, Michnick SW, Vosshall LB. Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors in vivo. PLoS Biol. 2006;4:e20. doi: 10.1371/journal.pbio.0040020.

162. Smart R, Kiely A, Beale M, Vargas E, Carraher C, Kralicek A V., et al. *Drosophila* odorant receptors are novel seven transmembrane domain proteins that can signal independently of heterotrimeric G proteins. Insect Biochem Mol Biol. 2008;38:770–80. doi: 10.1016/j.ibmb.2008.05.002.

163. Larsson MC, Domingos AI, Jones WD, Chiappe ME, Amrein H, Vosshall LB. Or83b Encodes a Broadly Expressed Odorant Receptor Essential for *Drosophila* Olfaction. Neuron. 2004;43:703–14. doi: 10.1016/j.neuron.2004.08.019.

164. Neuhaus EM, Gisselmann G, Zhang W, Dooley R, Störtkuhl K, Hatt H. Odorant receptor heterodimerization in the olfactory system of *Drosophila melanogaster*. Nat Neurosci. 2005;8:15–7. doi: 10.1038/nn1371.

165. Sato K, Pellegrino M, Nakagawa T, Nakagawa T, Vosshall LB, Touhara K. Insect olfactory receptors are heteromeric ligand-gated ion channels. Nature. 2008;452:1002–6. doi: 10.1038/nature06850.

166. Wicher D, Schäfer R, Bauernfeind R, Stensmyr MC, Heller R, Heinemann SH, et al. *Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. Nature. 2008;452:1007–11. doi: 10.1038/nature06861.

167. Butterwick JA, del Mármol J, Kim KH, Kahlson MA, Rogow JA, Walz T, et al. Cryo-EM structure of the insect olfactory receptor Orco. Nature. 2018;560:447–52. doi: 10.1038/s41586-018-0420-8.

168. Couto A, Alenius M, Dickson BJ. Molecular, Anatomical, and Functional Organization of the *Drosophila* Olfactory System. Curr Biol. 2005;15:1535–47. doi: 10.1016/j.cub.2005.07.034.

169. Hallem EA, Ho MG, Carlson JR. The molecular basis of odor coding in the *Drosophila* antenna. Cell. 2004;117:965–79. doi: 10.1016/j.cell.2004.05.012.

170. Hallem EA, Carlson JR. Coding of Odors by a Receptor Repertoire. Cell. 2006;125:143–60. doi: 10.1016/j.cell.2006.01.050.

171. Wang G, Carey AF, Carlson JR, Zwiebel LJ. Molecular basis of odor coding in the malaria vector mosquito *Anopheles gambiae*. Proc Natl Acad

Sci. 2010;107:4418–23. doi: 10.1073/pnas.0913392107.

172. Carey AF, Wang G, Su C-Y, Zwiebel LJ, Carlson JR. Odorant reception in the malaria mosquito *Anopheles gambiae*. Nature. 2010;464:66–71. doi: 10.1038/nature08834.

173. Kreher SA, Kwon JY, Carlson JR. The Molecular Basis of Odor Coding in the *Drosophila* Larva. Neuron. 2005;46:445–56. doi: 10.1016/j.neuron.2005.04.007.

174. Mathew D, Martelli C, Kelley-Swift E, Brusalis C, Gershow M, Samuel ADT, et al. Functional diversity among sensory receptors in a *Drosophila* olfactory circuit. Proc Natl Acad Sci. 2013;110:E2134–43. doi: 10.1073/pnas.1306976110.

175. Xia Y, Wang G, Buscariollo D, Pitts RJ, Wenger H, Zwiebel LJ. The molecular and cellular basis of olfactory-driven behavior in *Anopheles gambiae* larvae. Proc Natl Acad Sci U S A. 2008;105:6433–8. doi: 10.1073/pnas.0801007105.

176. Lu T, Qiu YT, Wang G, Kwon JY, Rutzler M, Kwon H-W, et al. Odor Coding in the Maxillary Palp of the Malaria Vector Mosquito *Anopheles gambiae*. Curr Biol. 2007;17:1533–44. doi: 10.1016/j.cub.2007.07.062.

177. Zhang H-J, Anderson AR, Trowell SC, Luo A-R, Xiang Z-H, Xia Q-Y. Topological and Functional Characterization of an Insect Gustatory Receptor. PLoS One. 2011;6:e24111. doi: 10.1371/journal.pone.0024111.

178. Liman ER, Zhang Y V., Montell C. Peripheral Coding of Taste. Neuron. 2014;81:984–1000. doi: 10.1016/j.neuron.2014.02.022.

179. Freeman EG, Dahanukar A. Molecular neurobiology of *Drosophila* taste. Curr Opin Neurobiol. 2015;34:140–8. doi: 10.1016/j.conb.2015.06.001.

180. Joseph RM, Carlson JR. *Drosophila* Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain. Trends Genet. 2015;31:683–95. doi: 10.1016/j.tig.2015.09.005.

181. French A, Moutaz AA, Mitra A, Yanagawa A, Sellier M-J, Marion-Poll F. *Drosophila* Bitter Taste(s). Front Integr Neurosci. 2015;9 November:1–13. doi: 10.3389/fnint.2015.00058.

182. Park J-H, Kwon JY. A systematic analysis of *Drosophila* gustatory receptor gene expression in abdominal neurons which project to the central nervous system. Mol Cells. 2011;32:375–81. doi: 10.1007/s10059-011-0128-1.

183. Park J-H, Kwon JY. Heterogeneous Expression of *Drosophila* Gustatory Receptors in Enteroendocrine Cells. PLoS One. 2011;6. doi: 10.1371/journal.pone.0029022.

184. Miyamoto T, Slone J, Song X, Amrein H. A Fructose Receptor Functions as a Nutrient Sensor in the *Drosophila* Brain. Cell. 2012;151:1113–25. doi: 10.1016/j.cell.2012.10.024.

185. Miyamoto T, Amrein H. Diverse roles for the *Drosophila* fructose sensor Gr43a. Fly (Austin). 2014;8:19–25. doi: 10.4161/fly.27241.

186. Fujii S, Yavuz A, Slone J, Jagge C, Song X, Amrein H. *Drosophila* Sugar Receptors in Sweet Taste Perception, Olfaction, and Internal Nutrient Sensing. Curr Biol. 2015;25:621–7. doi: 10.1016/j.cub.2014.12.058.

187. Jones WD, Cayirlioglu P, Grunwald Kadow I, Vosshall LB. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. Nature. 2007;445:86–90. doi: 10.1038/nature05466.

188. Yao CA, Carlson JR. Role of G-Proteins in Odor-Sensing and CO2-Sensing Neurons in *Drosophila*. J Neurosci. 2010;30:4562–72. doi: 10.1523/JNEUROSCI.6357-09.2010.

189. Kwon JY, Dahanukar A, Weiss LA, Carlson JR. The molecular basis of CO_2 reception in *Drosophila*. Proc Natl Acad Sci. 2007;104:3574–8. doi: 10.1073/pnas.0700079104.

190. Dahanukar A, Lei Y-T, Kwon JY, Carlson JR. Two Gr Genes Underlie Sugar Reception in *Drosophila*. Neuron. 2007;56:503–16. doi: 10.1016/j.neuron.2007.10.024.

191. Jiao Y, Moon SJ, Montell C. A *Drosophila* gustatory receptor required for the responses to sucrose, glucose, and maltose identified by mRNA tagging. Proc Natl Acad Sci. 2007;104:14110–5. doi: 10.1073/pnas.0702421104.

192. Jiao Y, Moon SJ, Wang X, Ren Q, Montell C. Gr64f Is Required in Combination with Other Gustatory Receptors for Sugar Detection in *Drosophila*. Curr Biol. 2008;18:1797–801. doi: 10.1016/j.cub.2008.10.009.

193. Delventhal R, Carlson JR. Bitter taste receptors confer diverse functions to neurons. Elife. 2016;5. doi: 10.7554/eLife.11181.

194. Bray S, Amrein H. A putative *Drosophila* pheromone receptor expressed in male-specific taste neurons is required for efficient courtship. Neuron. 2003;39:1019–29. doi: 10.1016/S0896-6273(03)00542-7.

195. Ni L, Bronk P, Chang EC, Lowell AM, Flam JO, Panzano VC, et al. A gustatory receptor paralogue controls rapid warmth avoidance in *Drosophila*. Nature. 2013;500:580–4. doi: 10.1038/nature12390.

196. Robertson HM. The Insect Chemoreceptor Superfamily Is Ancient in Animals. Chem Senses. 2015;40:609–14. doi: 10.1093/chemse/bjv046.

197. Eyun S, Soh HY, Posavi M, Munro JB, Hughes DST, Murali SC, et al. Evolutionary History of Chemosensory-Related Gene Families across the

Arthropoda. Mol Biol Evol. 2017;34:1838–62. doi: 10.1093/molbev/msx147.

198. Saina M, Busengdal H, Sinigaglia C, Petrone L, Oliveri P, Rentzsch F, et al. A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. Nat Commun. 2015;6:6243. doi: 10.1038/ncomms7243.

199. Brand P, Robertson HM, Lin W, Pothula R, Klingeman WE, Jurat-Fuentes JL, et al. The origin of the odorant receptor gene family in insects. Elife. 2018;7. doi: 10.7554/eLife.38340.

200. Pelosi P, Zhou JJ, Ban LP, Calvello M. Soluble proteins in insect chemical communication. Cellular and Molecular Life Sciences. 2006;63:1658–76. doi: 10.1007/s00018-005-5607-0.

201. Vosshall LB. Olfaction in *Drosophila*. Curr Opin Neurobiol. 2000;10:498–503. doi: 10.1016/S0959-4388(00)00111-2.

202. McKenzie SK, Fetter-Pruneda I, Ruta V, Kronauer DJC. Transcriptomics and neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in chemical communication. Proc Natl Acad Sci. 2016;113:14091–6. doi: 10.1073/pnas.1610800113.

203. Ioannidis P, Simao FA, Waterhouse RM, Manni M, Seppey M, Robertson HM, et al. Genomic features of the damselfly *Calopteryx splendens* representing a sister clade to most insect orders. Genome Biol Evol. 2017;:evx006. doi: 10.1093/gbe/evx006.

204. Robertson HM, Wanner KW. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: Expansion of the odorant, but not gustatory, receptor family. Genome Res. 2006;16:1395–403. doi: 10.1101/gr.5057506.

205. Richards S, Gibbs RA, Weinstock GM, Brown S, Denell R, Beeman RW, et al. The genome of the model beetle and pest *Tribolium castaneum*. Nature. 2008;452:949–55. doi: 10.1038/nature06784.

206. Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, et al. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat Commun. 2016;7:10507. doi: 10.1038/ncomms10507.

207. Sánchez-Gracia A, Vieira FG, Almeida FC, Rozas J. Comparative Genomics of the Major Chemosensory Gene Families in Arthropods. In: Encyclopedia of Life Sciences. Chichester, UK: John Wiley & Sons, Ltd; 2011. doi: 10.1002/9780470015902.a0022848.

208. Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 1997;94:7799–806. doi: 10.1073/pnas.94.15.7799.

209. Nei M, Rooney AP. Concerted and Birth-and-Death Evolution of Multigene Families. Annu Rev Genet. 2005;39:121–52. doi: 10.1146/annurev.genet.39.073003.112240.

210. Zheng W, Peng T, He W, Zhang H. High-throughput sequencing to reveal genes involved in reproduction and development in *Bactrocera dorsalis* (Diptera: Tephritidae). PLoS One. 2012;7. doi: 10.1371/journal.pone.0036463.

211. Zheng W, Peng W, Zhu C, Zhang Q, Saccone G, Zhang H. Identification and expression profile analysis of odorant binding proteins in the oriental fruit fly *Bactrocera dorsalis*. Int J Mol Sci. 2013;14:14936–49. doi: 10.3390/ijms140714936.

212. Wang J, Xiong K-C, Liu Y-H. De novo Transcriptome Analysis of Chinese Citrus Fly, *Bactrocera minax* (Diptera: Tephritidae), by High-Throughput Illumina Sequencing. PLoS One. 2016;11:e0157656. doi: 10.1371/journal.pone.0157656.

213. Elfekih S, Chen CY, Hsu JC, Belcaid M, Haymer D. Identification and preliminary characterization of chemosensory perception-associated proteins in the melon fly *Bactrocera cucurbitae* using RNA-seq. Sci Rep. 2016;6 June 2015:1–10. doi: 10.1038/srep19112.

214. Wu Z, Zhang H, Wang Z, Bin S, He H, Lin J. Discovery of Chemosensory Genes in the Oriental Fruit Fly, *Bactrocera dorsalis*. PLoS One. 2015;10. doi: 10.1371/journal.pone.0129794.

215. Xu P, Wang Y, Akami M, Niu C-Y. Identification of olfactory genes and functional analysis of BminCSP and BminOBP21 in *Bactrocera minax*. PLoS One. 2019;14:e0222193. doi: 10.1371/journal.pone.0222193.

216. Sagri E, Koskinioti P, Gregoriou ME, Tsoumani KT, Bassiakos YC, Mathiopoulos KD. Housekeeping in Tephritid insects: The best gene choice for expression analyses in the medfly and the olive fly. Sci Rep. 2017;7 April:1–9. doi: 10.1038/srep45634.

217. Xu L, Tang K, Chen X, Tao Y, Jiang H, Wang J. Comparative transcriptomic analysis reveals female-biased olfactory genes potentially involved in plant volatile-mediated oviposition behavior of *Bactrocera dorsalis*. BMC Genomics. 2021;22:25. doi: 10.1186/s12864-020-07325-z.

218. Liu H, Zhao X-F, Fu L, Han Y-Y, Chen J, Lu Y-Y. BdorOBP2 plays an indispensable role in the perception of methyl eugenol by mature males of *Bactrocera dorsalis* (Hendel). Sci Rep. 2017;7:15894. doi: 10.1038/s41598-017-15893-6.

219. Liu Z, Liang X-F, Xu L, Keesey IW, Lei Z-R, Smagghe G, et al. An Antennae-Specific Odorant-Binding Protein Is Involved in *Bactrocera dorsalis* Olfaction. Front Ecol Evol. 2020;8:63. doi:

10.3389/fevo.2020.00063.

220. Zhang J, Luo D, Wu P, Li H, Zhang H, Zheng W. Identification and expression profiles of novel odorant binding proteins and functional analysis of OBP99a in *Bactrocera dorsalis*. Arch Insect Biochem Physiol. 2018;98. doi: 10.1002/arch.21452.

221. Wu Z, Lin J, Zhang H, Zeng X. BdorOBP83a-2 Mediates Responses of the Oriental Fruit Fly to Semiochemicals. Front Physiol. 2016;7 Oct:1–15. doi: 10.3389/fphys.2016.00452.

222. Miyazaki H, Otake J, Mitsuno H, Ozaki K, Kanzaki R, Chui-Ting Chieng A, et al. Functional characterization of olfactory receptors in the Oriental fruit fly *Bactrocera dorsalis* that respond to plant volatiles. Insect Biochem Mol Biol. 2018;101 May:32–46. doi: 10.1016/j.ibmb.2018.07.002.

223. Liu H, Chen Z-S, Zhang D-J, Lu Y-Y. BdorOR88a Modulates the Responsiveness to Methyl Eugenol in Mature Males of *Bactrocera dorsalis* (Hendel). Front Physiol. 2018;9 JUL:1–17. doi: 10.3389/fphys.2018.00987.

224. Tsoumani KT, Belavilas-Trovas A, Gregoriou M-E, Mathiopoulos KD. Anosmic flies: what Orco silencing does to olive fruit flies. BMC Genet. 2020;21:140. doi: 10.1186/s12863-020-00937-0.

225. McBride CS, Arguello JR. Five *Drosophila* Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily. Genetics. 2007;177:1395–416. doi: 10.1534/genetics.107.078683.

226. McBride CS. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proc Natl Acad Sci. 2007;104:4996–5001. doi: 10.1073/pnas.0608424104.

227. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, 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. doi: 10.1186/s13059-016-1049-2.

228. Sim SB, Geib SM. A Chromosome-Scale Assembly of the *Bactrocera cucurbitae* Genome Provides Insight to the Genetic Basis of white pupae. G3 Genes, Genomes, Genet. 2017;7:1927–40. doi: 10.1534/g3.117.040170.

229. Geib SM, Calla B, Hall B, Hou S, Manoukis NC. Characterizing the developmental transcriptome of the oriental fruit fly, *Bactrocera dorsalis* (Diptera: Tephritidae) through comparative genomic analysis with *Drosophila melanogaster* utilizing modENCODE datasets. BMC Genomics. 2014;15. doi: 10.1186/1471-2164-15-942.

230. Bayega A, Djambazian H, Tsoumani KT, Gregoriou ME, Sagri E, Drosopoulou E, et al. De novo assembly of the olive fruit fly (*Bactrocera oleae*) genome with linked-reads and long-read technologies minimizes gaps and provides exceptional y chromosome assembly. BMC Genomics. 2020;21. doi: 10.1186/s12864-020-6672-3.

231. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52. doi: 10.1038/nbt.1883.

232. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. In: Methods in Molecular Biology. Humana Press Inc.; 2019. p. 227–45. doi: 10.1007/978-1-4939-9173-0_14.

233. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:421. doi: 10.1186/1471-2105-10-421.

234. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44. doi: 10.1093/bioinformatics/btn013.

235. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. doi: 10.1093/molbev/mst010.

236. Löytynoja A, Goldman N. A model of evolution and structure for multiple sequence alignment. Philos Trans R Soc B Biol Sci. 2008;363:3913–9.

237. Abascal F, Zardoya R, Telford MJ. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 2010;38 SUPPL. 2:W7-13. doi: 10.1093/nar/gkq291.

238. Han M V., Demuth JP, McGrath CL, Casola C, Hahn MW. Adaptive evolution of young gene duplicates in mammals. Genome Res. 2009;19:859–67. doi: 10.1101/gr.085951.108.

239. Ramasamy S, Ometto L, Crava CM, Revadi S, Kaur R, Horner DS, et al. The evolution of olfactory gene families in *Drosophila* and the genomic basis of chemical-ecological adaptation in *Drosophila suzukii*. Genome Biol Evol. 2016;8:2297–311. doi: 10.1093/gbe/evw160.

240. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3. doi: 10.1093/bioinformatics/btu033.

241. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19:153. doi: 10.1186/s12859-018-2129-y.

242. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15:e1006650. doi: 10.1371/journal.pcbi.1006650.

243. Lartillot N, Philippe H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Mol Biol Evol. 2004;21:1095–109. doi: 10.1093/molbev/msh112.

244. Ogilvie HA, Bouckaert RR, Drummond AJ. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol. 2017;34:2101–14. doi: 10.1093/molbev/msx126.

245. Barido-Sottani J, Bošková V, Plessis L Du, Kühnert D, Magnus C, Mitov V, et al. Taming the BEAST - A Community Teaching Material Resource for BEAST 2. Syst Biol. 2018;67:170–4. doi: 10.1093/sysbio/syx060.

246. Obbard DJ, MacLennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. Mol Biol Evol. 2012;29:3459–73. doi: 10.1093/molbev/mss150.

247. Vargas RI, Walsh WA, Kanehisa D, Jang EB, Armstrong JW. Demography of four Hawaiian fruit flies (Diptera: Tephritidae) reared at five constant temperatures. Ann Entomol Soc Am. 1997;90:162–8. doi: 10.1093/aesa/90.2.162.

248. Stephens AEA, Kriticos DJ, Leriche A. The current and future potential geographical distribution of the oriental fruit fly, *Bactrocera dorsalis* (Diptera: Tephritidae). Bull Entomol Res. 2007;97:369–78. doi: 10.1017/S0007485307005044.

249. Theron CD, Manrakhan A, Weldon CW. Host use of the oriental fruit fly, *Bactrocera dorsalis* (Hendel) (Diptera: Tephritidae), in South Africa. J Appl Entomol. 2017;141:810–6. doi: 10.1111/jen.12400.

250. Li X, Yang H, Wang T, Wang J, Wei H. Life history and adult dynamics of *Bactrocera dorsalis* in the citrus orchard of Nanchang, a subtropical area from China: implications for a control timeline. ScienceAsia. 2019;45:212–20. doi: 10.2306/scienceasia1513-1874.2019.45.212.

251. Norrbom A. New genera of Tephritidae (Diptera) from Brazil and Dominican Amber, with phylogenetic analysis of the tribe Ortalotrypetini. Insecta mundi. 1994;8:1–15. https://digitalcommons.unl.edu/insectamundi/289.

252. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, et al. Episodic radiations in the fly tree of life. Proc Natl Acad Sci U S A. 2011;108:5690–5. doi: 10.1073/pnas.1012675108.

253. Junqueira ACM, Azeredo-Espin AML, Paulo DF, Marinho MAT, Tomsho

LP, Drautz-Moses DI, et al. Large-scale mitogenomics enables insights into Schizophora (Diptera) radiation and population diversity. Sci Rep. 2016;6. doi: 10.1038/srep21762.

254. Russel PM, Brewer BJ, Klaere S, Bouckaert RR. Model Selection and Parameter Inference in Phylogenetics Using Nested Sampling. Syst Biol. 2019;68:219–33. doi: 10.1093/sysbio/syy050.

255. Okonechnikov K, Golosova O, Fursov M, UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28:1166–7. doi: 10.1093/bioinformatics/bts091.

256. Rédei GP, editor. GT – AG RULE (Chambon's rule) BT - Encyclopedia of Genetics, Genomics, Proteomics and Informatics. Dordrecht: Springer Netherlands; 2008. p. 828. doi: 10.1007/978-1-4020-6754-9_7198.

257. Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Warwick Vesztrocy A, Dalquen DA, et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 2019;29:1152–63. doi: 10.1101/gr.243212.118.

258. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7. doi: 10.1093/nar/gkh340.

259. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23:127–8. doi: 10.1093/bioinformatics/btl529.

260. Librado P, Vieira FG, Rozas J. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics. 2012;28:279–81. doi: 10.1093/bioinformatics/btr623.

261. Almeida FC, Sánchez-Gracia A, Campos JL, Rozas J. Family Size Evolution in *Drosophila* Chemosensory Gene Families: A Comparative Analysis with a Critical Appraisal of Methods. Genome Biol Evol. 2014;6:1669–82. doi: 10.1093/gbe/evu130.

262. Akaike H. Information theory and the maximum likelihood principle. In: 2nd International Symposium on Information Theory. 1973.

263. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24:1586–91. doi: 10.1093/molbev/msm088.

264. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 1998;15:568–73. doi: 10.1093/oxfordjournals.molbev.a025957.

265. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000;17:32–43. doi: 10.1093/oxfordjournals.molbev.a026236.

266. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000;155:431–49. http://www.ncbi.nlm.nih.gov/pubmed/10790415.

267. Yang Z. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol Biol Evol. 2005;22:1107–18. doi: 10.1093/molbev/msi097.

268. Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B (Statistical Methodol. 2002;64:479–98. doi: 10.1111/1467-9868.00346.

269. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. https://www.r-project.org/.

270. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018;430:2237–43. doi: https://doi.org/10.1016/j.jmb.2017.12.007.

271. Sonnhammer ELL, Von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. 1998. www.aaai.org.

272. Hofmann K, Stoffel W. TMbase: A Database of Membrane Spanning Protein Segments. Biol Chem. 1993.

273. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37:420–3. doi: 10.1038/s41587-019-0036-z.

274. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004.

275. Song Y, DiMaio F, Wang RY-R, Kim D, Miles C, Brunette T, et al. High-resolution comparative modeling with RosettaCM. Structure. 2013;21:1735–42. doi: 10.1016/j.str.2013.08.005.

276. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinforma. 2016;54:5.6.1-5.6.37. doi: 10.1002/cpbi.3.

277. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46:W296–303. doi: 10.1093/nar/gky427.

278. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version~1.8. 2015.

279. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 2018;27:293–315. doi: 10.1002/pro.3330.

280. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009;37 suppl_2:W510–4. doi: 10.1093/nar/gkp322.

281. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016;44:W344–50. doi: 10.1093/nar/gkw408.

282. Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. 2018;46:W363–7. doi: 10.1093/nar/gky473.

283. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995.

284. Valenzuela JG, Pham VM, Garfield MK, Francischetti IMB, Ribeiro JMC. Toward a description of the sialome of the adult female mosquito *Aedes aegypti*. Insect Biochem Mol Biol. 2002;32:1101–22. doi: 10.1016/S0965-1748(02)00047-4.

285. Choo A, Nguyen TNM, Ward CM, Chen IY, Sved J, Shearman D, et al. Identification of Y-chromosome scaffolds of the Queensland fruit fly reveals a duplicated gyf gene paralogue common to many *Bactrocera* pest species. Insect Mol Biol. 2019;28:873–86. doi: 10.1111/imb.12602.

286. Pollard DA, Iyer VN, Moses AM, Eisen MB. Widespread Discordance of Gene Trees with Species Tree in *Drosophila*: Evidence for Incomplete Lineage Sorting. PLoS Genet. 2006;2:e173. doi: 10.1371/journal.pgen.0020173.

287. DeSalle R, Giddings L V. Discordance of nuclear and mitochondrial DNA phylogenies in Hawaiian *Drosophila*. Proc Natl Acad Sci U S A. 1986.

288. Beltrán M, Jiggins CD, Bull V, Linares M, Mallet J, McMillan WO, et al. Phylogenetic Discordance at the Species Boundary: Comparative Gene Genealogies Among Rapidly Radiating *Heliconius* Butterflies. Mol Biol Evol. 2002;19:2176–90. doi: 10.1093/oxfordjournals.molbev.a004042.

289. Putnam AS, Scriber JM, Andolfatto P. Discordant divergence times among Z-chromosome regions between two ecologically distinct swallowtail butterfly species. Evolution (N Y). 2007;61:912–27. doi: 10.1111/j.1558-5646.2007.00076.x.

290. Toews DPL, Brelsford A. The biogeography of mitochondrial and nuclear discordance in animals. Mol Ecol. 2012;21:3907–30. doi: 10.1111/j.1365-294X.2012.05664.x.

291. Zhong G, Geng J, Wong HK, Ma Z, Wu N. A semi-quantitative method for the reconstruction of eustatic sea level history from seismic profiles and its application to the southern South China Sea. Earth Planet Sci Lett. 2004;223:443–59.

292. Yaakop S, Ibrahim NJ, Shariff S, Md. Zain BM. Molecular clock analysis on five *Bactrocera* species flies (Diptera: Tephritidae) based on combination of COI and NADH sequences. Orient Insects. 2015;49:150–64. doi: 10.1080/00305316.2015.1081421.

293. Nardi F, Carapelli A, Boore JL, Roderick GK, Dallai R, Frati F. Domestication of olive fly through a multi-regional host shift to cultivated olives: Comparative dating using complete mitochondrial genomes. Mol Phylogenet Evol. 2010;57:678–86.

294. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–7. doi: 10.1126/science.1257570.

295. Vargas RI, Piñero JC, Miller NW. Effect of Physiological State on Female Melon Fly (Diptera: Tephritidae) Attraction to Host and Food Odor in the Field. J Econ Entomol. 2018;111:1318–22. doi: 10.1093/jee/toy092.

296. Piñero JC, Jácome I, Vargas R, Prokopy RJ. Response of female melon fly, *Bactrocera cucurbitae*, to host-associated visual and olfactory stimuli. Entomol Exp Appl. 2006;121:261–9. doi: 10.1111/j.1570-8703.2006.00485.x.

297. Liu R, He X, Lehane S, Lehane M, Hertz-Fowler C, Berriman M, et al. Expression of chemosensory proteins in the tsetse fly *Glossina morsitans morsitans* is related to female host-seeking behaviour. Insect Mol Biol. 2012;21:41–8. doi: 10.1111/j.1365-2583.2011.01114.x.

298. Iovinella I, Bozza F, Caputo B, della Torre A, Pelosi P. Ligand-binding study of *Anopheles gambiae* chemosensory proteins. Chem Senses. 2013.

299. Ekesi S, Mohamed SA, De Meyer M. Fruit Fly Research and Development in Africa - Towards a Sustainable Management Strategy to Improve Horticulture. Springer International Publishing; 2016. https://books.google.it/books?id=YAykDQAAQBAJ.

300. Rwomushana I, Ekesi S, Ogol CKPO, Gordon I. Mechanisms contributing to the competitive success of the invasive fruit fly *Bactrocera invadens* over the indigenous mango fruit fly, *Ceratitis cosyra*: The role of temperature and resource pre-emption. Entomol Exp Appl. 2009.

301. Pearce SL, Clarke DF, East PD, Elfekih S, Gordon KHJ, Jermiin LS, et al. Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and

invasive *Helicoverpa* pest species. BMC Biol. 2017;15:1–30. doi: 10.1186/s12915-017-0402-6.

302. Cheng T, Wu J, Wu Y, Chilukuri R V., Huang L, Yamamoto K, et al. Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. Nat Ecol Evol. 2017;1:1747–56. doi: 10.1038/s41559-017-0314-4.

303. Xu W, Papanicolaou A, Zhang H-J, Anderson A. Expansion of a bitter taste receptor family in a polyphagous insect herbivore. Sci Rep. 2016;6:23666. doi: 10.1038/srep23666.

304. Gouin A, Bretaudeau A, Nam K, Gimenez S, Aury J-M, Duvic B, et al. Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda*, Noctuidae) with different host-plant ranges. Sci Rep. 2017;7:11816. doi: 10.1038/s41598-017-10461-4.

305. Clarke AR. Why so many polyphagous fruit flies (Diptera: Tephritidae)? A further contribution to the 'generalism' debate. Biol J Linn Soc. 2016. doi: 10.1111/bij.12880.

306. Tan KH, Nishida R. Methyl Eugenol: Its Occurrence, Distribution, and Role in Nature, Especially in Relation to Insect Behavior and Pollination. J Insect Sci. 2012;12:1–60. doi: 10.1673/031.012.5601.

307. Vargas RI, Mau RFL, Stark JD, Piñero JC, Leblanc L, Souder SK. Evaluation of Methyl Eugenol and Cue-Lure Traps With Solid Lure and Insecticide Dispensers for Fruit Fly Monitoring and Male Annihilation in the Hawaii Areawide Pest Management Program. J Econ Entomol. 2010;103:409–15. doi: 10.1603/EC09299.

308. Vargas RI, Shelly TE, Leblanc L, Piñero JC. Recent Advances in Methyl Eugenol and Cue-Lure Technologies for Fruit Fly Detection, Monitoring, and Control in Hawaii. In: Vitamins and Hormones. 2010. p. 575–95. doi: 10.1016/S0083-6729(10)83023-7.

309. IAEA (International Atomic Energy Agency). Trapping guidelines for area-wide fruit fly programmes. IAEA. 2003;:47. http://www-naweb.iaea.org/nafa/ipc/public/trapping-web.pdf.

310. Wu Z, Cui Y, Ma J, Qu M, Lin J. Analyses of chemosensory genes provide insight into the evolution of behavioral differences to phytochemicals in *Bactrocera* species. Mol Phylogenet Evol. 2020;151 May:106858. doi: 10.1016/j.ympev.2020.106858.

311. Crava CM, Ramasamy S, Ometto L, Anfora G, Rota-Stabelli O. Evolutionary Insights into Taste Perception of the Invasive Pest *Drosophila suzukii*. G3 Genes, Genomes, Genet. 2016;6:4185–96. doi: 10.1534/g3.116.036467.

312. Leite NR, Krogh R, Xu W, Ishida Y, Iulek J, Leal WS, et al. Structure of

an Odorant-Binding Protein from the Mosquito *Aedes aegypti* Suggests a Binding Pocket Covered by a pH-Sensitive "Lid." PLoS One. 2009;4:e8006. doi: 10.1371/journal.pone.0008006.

313. Brito NF, Moreira MF, Melo ACA. A look inside odorant-binding proteins in insect chemoreception. Journal of Insect Physiology. 2016. doi: 10.1016/j.jinsphys.2016.09.008.

314. Carey AF, Carlson JR. Insect olfaction from model systems to disease control. Proc Natl Acad Sci U S A. 2011;108:12987–95. doi: 10.1073/pnas.1103472108.

315. Suckling DM, Dymock JJ, Park KC, Wakelin RH, Jamieson LE. Communication Disruption of Guava Moth (*Coscinoptycha improbana*) Using a Pheromone Analog Based on Chain Length. J Chem Ecol. 2013;39:1161–8. doi: 10.1007/s10886-013-0339-3.

316. Tan KH, Nishida R, Jang E, Shelly T. Pheromones, Male Lures, and Trapping of Tephritid Fruit Flies. In: Trapping and the Detection, Control, and Regulation of Tephritid Fruit Flies: Lures, Area-wide Programs, and Trade Implications. 2014. p. 15–74.

317. Campanacci V, Krieger J, Bette S, Sturgis JN, Lartigue A, Cambillau C, et al. Revisiting the Specificity of *Mamestra brassicae* and *Antheraea polyphemus* Pheromone-binding Proteins with a Fluorescence Binding Assay. J Biol Chem. 2001;276:20078–84. doi: 10.1074/jbc.M100713200.

318. Zhou J-J, Zhang G-A, Huang W, Birkett MA, Field LM, Pickett JA, et al. Revisiting the odorant-binding protein LUSH of *Drosophila melanogaster* : evidence for odour recognition and discrimination. FEBS Lett. 2004;558:23–6. doi: 10.1016/S0014-5793(03)01521-7.

319. Qiao H, Tuccori E, He X, Gazzano A, Field L, Zhou JJ, et al. Discrimination of alarm pheromone (E)-β-farnesene by aphid odorant-binding proteins. Insect Biochem Mol Biol. 2009;39:414–9. doi: 10.1016/j.ibmb.2009.03.004.

320. Damberger FF, Michel E, Ishida Y, Leal WS, Wüthrich K. Pheromone discrimination by a pH-tuned polymorphism of the *Bombyx mori* pheromone-binding protein. Proc Natl Acad Sci U S A. 2013;110:18680–5. doi: 10.1073/pnas.1317706110.

321. Li N, Sun X, Wang M-Q. Expression pattern and ligand-binding properties of odorant-binding protein 13 from *Monochamus alternatus* hope. J Appl Entomol. 2017;141:751–7. doi: 10.1111/jen.12396.

322. Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci. 2015;112:E5907–15. doi: 10.1073/pnas.1516410112.

323. Palatini U, Miesen P, Carballar-Lejarazu R, Ometto L, Rizzo E, Tu Z, et al. Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. BMC Genomics. 2017;18:512. doi: 10.1186/s12864-017-3903-3.

324. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303. doi: 10.1101/gr.107524.110.

325. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *Drosophila*. Cell. 2007;128:1089–103. doi: 10.1016/j.cell.2007.01.043.

326. Miesen P, Joosten J, van Rij RP. PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector Mosquitoes. PLoS Pathog. 2016;12:e1006017. doi: 10.1371/journal.ppat.1006017.

327. Akbari OS, Antoshechkin I, Amrhein H, Williams B, Diloreto R, Sandler J, et al. The Developmental Transcriptome of the Mosquito *Aedes aegypti*, an Invasive Species and Major Arbovirus Vector. G3 Genes, Genomes, Genet. 2013;3:1493–509. doi: 10.1534/g3.113.006742.

328. Miesen P, Girardi E, van Rij RP. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in *Aedes aegypti* mosquito cells. Nucleic Acids Res. 2015;43:6545–56. doi: 10.1093/nar/gkv590.

329. Mcquate GT, Liquido NJ. Annotated World Bibliography of Host Fruits of *Bactrocera latifrons* (Hendel) (Diptera: Tephritidae). Insecta mundi. 2013;2013:1–61.

330. Suh E, Bohbot JD, Zwiebel LJ. Peripheral olfactory signaling in insects. Curr Opin Insect Sci. 2014;6:86–92. doi: 10.1016/j.cois.2014.10.006.

331. Conway JR, Lex A, Gehlenborg N. UpSetR: An R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33:2938–40. doi: 10.1093/bioinformatics/btx364.

332. RStudio | Open source & professional software for data science teams - RStudio. https://rstudio.com/.

333. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847–9. doi: 10.1093/bioinformatics/btw313.

334. Pischedda E, Scolari F, Valerio F, Carballar-Lejarazú R, Catapano PL, Waterhouse RM, et al. Insights into an unexplored component of the mosquito repeatome: Distribution and variability of viral sequences integrated into the genome of the arboviral vector *Aedes albopictus*. Front

Genet. 2019;10 FEB. doi: 10.3389/fgene.2019.00093.

335. Marconcini M, Hernandez L, Iovino G, Houé V, Valerio F, Palatini U, et al. Polymorphism analyses and protein modelling inform on functional specialization of Piwi clade genes in the arboviral vector *Aedes albopictus*. PLoS Negl Trop Dis. 2019;13:e0007919. doi: 10.1371/journal.pntd.0007919.

Appendix A - Supporting Materials

Multi-locus phylogeny of *Bactrocera* genus

Table S1 – SRA accession numbers

Species	SRA databases (https://www.ncbi.nlm.nih.gov/sra)
Bactrocera bryoniae	SRX2791703
Bactrocera correcta	SRX2013592, SRX2013591, SRX2013590, SRX2013589, SRX2372821, SRX2372818, SRX2372817
Bactrocera jarvisi	SRX2791705, SRX697442, SRX697441, SRX697440, SRX697437, SRX697435, SRX697434, SRX697431, SRX697428
Bactrocera musae	SRX2791704
Bactrocera tryoni	
Bactrocera zonata	SRX2016849, SRX2016848, SRX2016847, SRX2016846

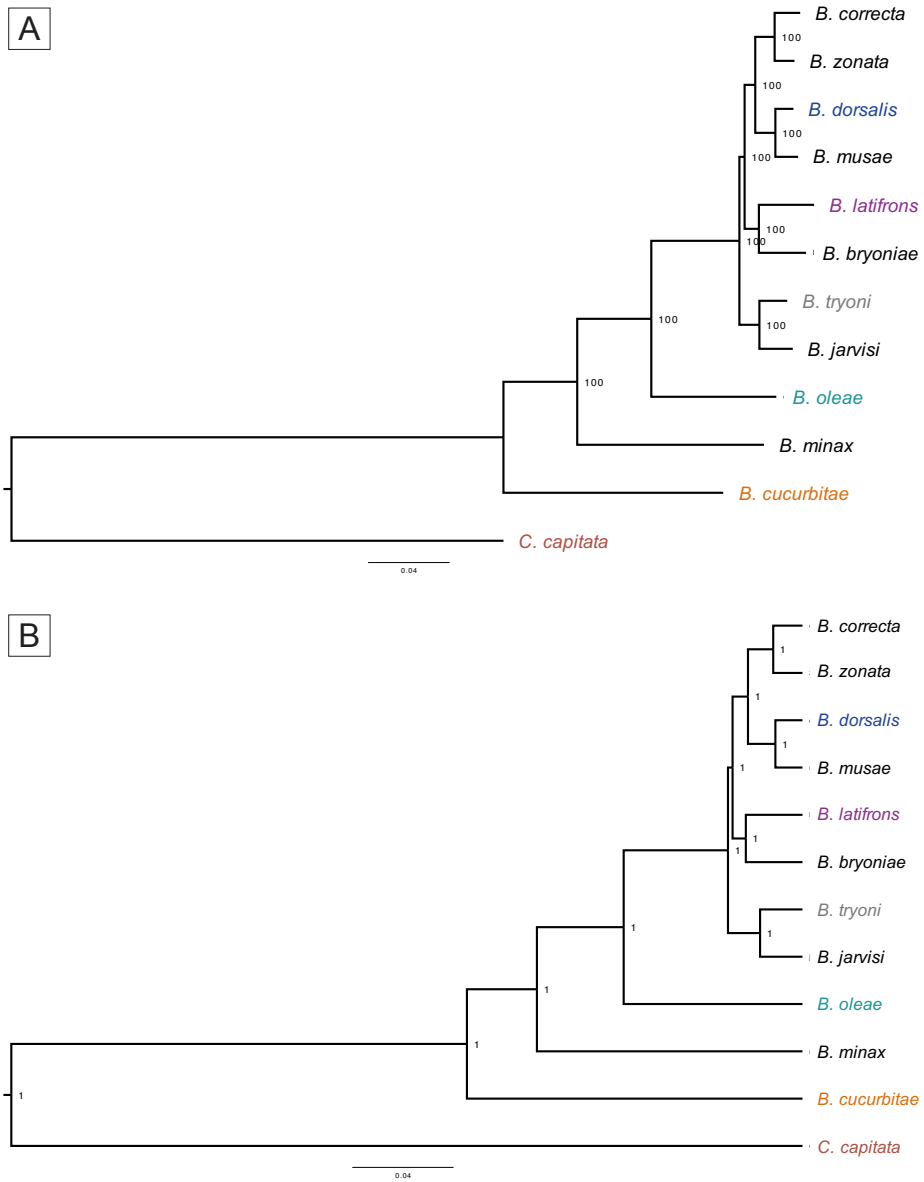


Figure S1 – Phylogenesis of *Bactrocera* inferred from the nucleotide alignments of 110 orthologous nuclear genes. A) Maximum Likelihood phylogenetic analysis (GAMMAGTR model). B) Bayesian phylogenetic analysis (Gamma site model). Both analyses were done using the concatenated codon alignment (189,891 nt) and support a closer relationship of *B. dorsalis* with *B. latifrons* than with *B. tryoni*. Support at nodes is given as bootstrap values for the ML analysis and posterior probabilities for the Bayesian analysis. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

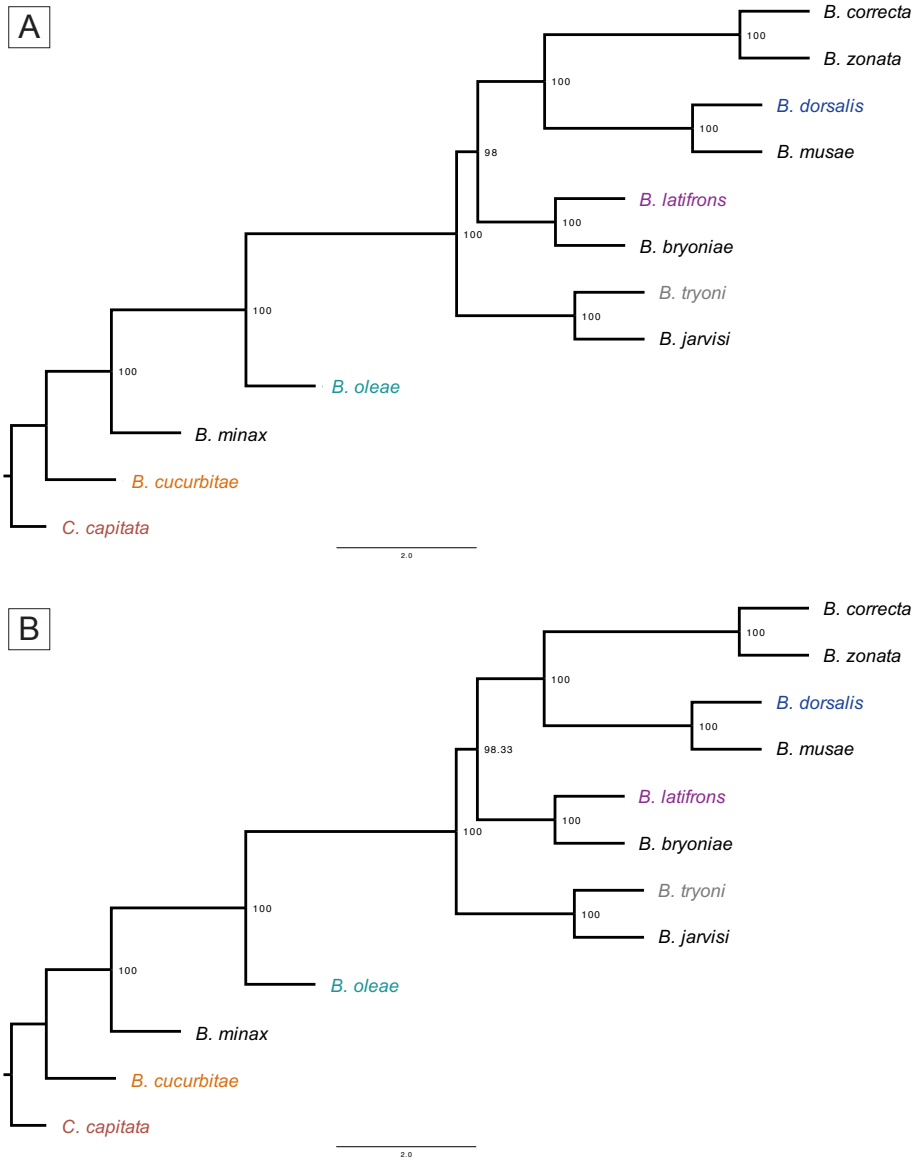


Figure S2 – Multi-locus coalescent-aware phylogenesis of *Bactrocera* inferred from 110 orthologous nuclear genes. Analyses are based on all single ML gene trees obtained using the nucleotide sequences (GAMMAGTR model). A) Bootstrap values were estimated by performing 100 multi-locus bootstrap replicates; B) Bootstrap values were estimated by performing 100 gene+site resampling. Both analyses support a closer relationship of *B. dorsalis* with *B. latifrons* than with *B. tryoni*. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

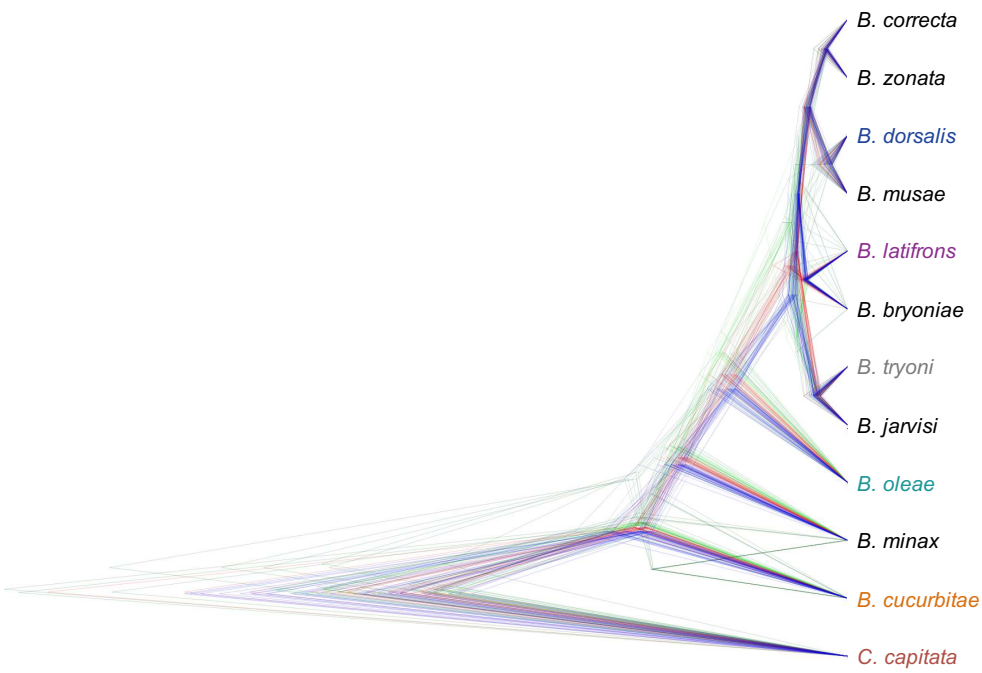


Figure S3 – Multi-locus phylogenesis of *Bactrocera* inferred from 110 orthologous nuclear genes. Bayesian analysis obtained by StarBeast2, which employs a multispecies coalescent method to estimate species trees from multiple sequence alignments (i.e., one for each of the 110 orthologous gene sets). For this analysis we used the nucleotide alignments, linking the site models across the gene sets. Note the numerous discordant gene trees, especially between the *B. dorsalis* - *B. latifrons* - *B. tryoni* clade, compared to the species tree (supported by the trees in blue, which supports a closer relationship of *B. dorsalis* with *B. latifrons* than with *B. tryoni*). The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

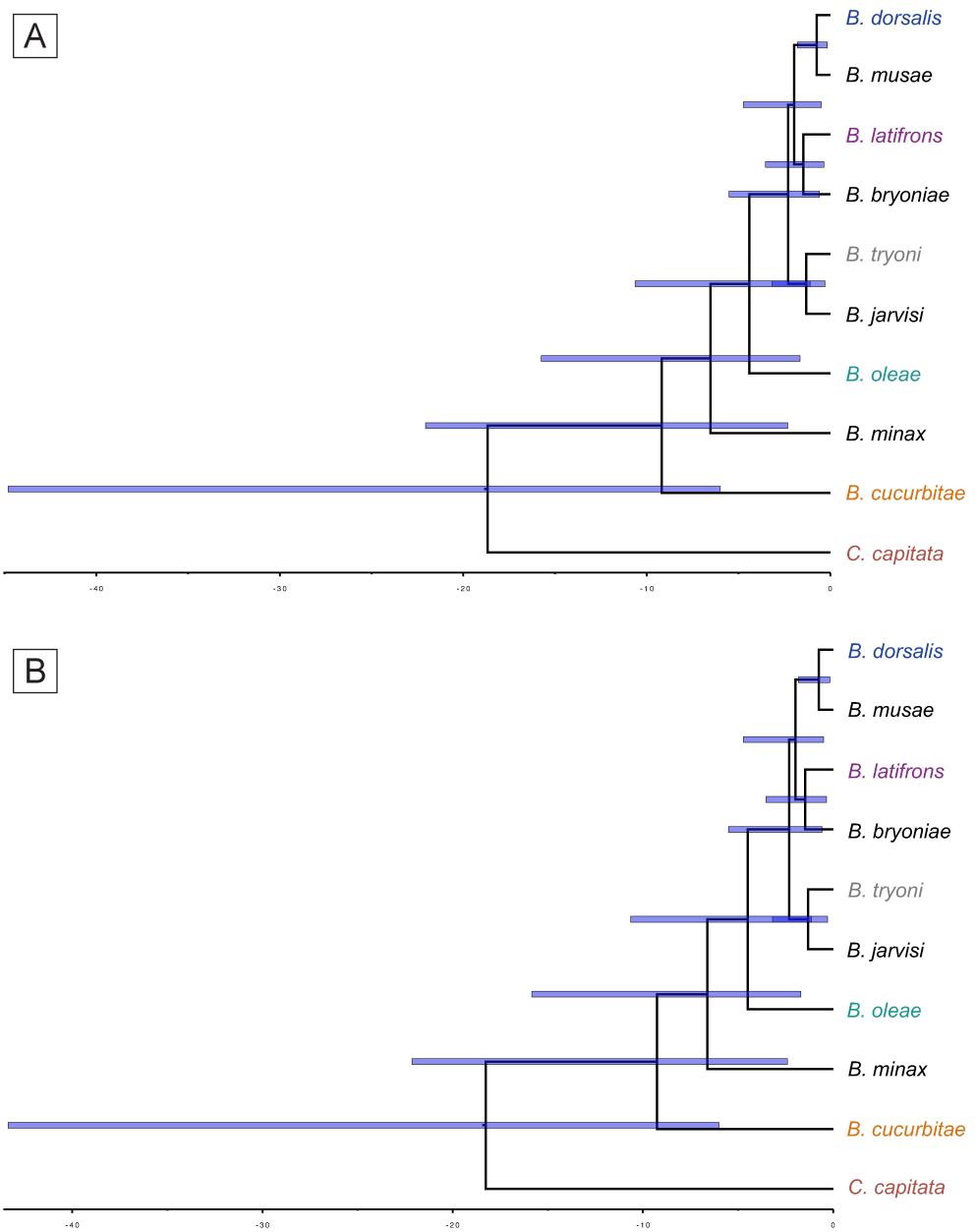


Figure S4 – Molecular time tree of *Bactrocera*. The analysis was done setting A) the mutation rate log-normally distributed as prior, a LOGN relaxed clock and a Birth-Death model, B) the mutation rate log-normally distributed as prior, a LOGN relaxed clock and a Yule model. Blue bars at nodes identify the 95% confidence interval of the inferred node age. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

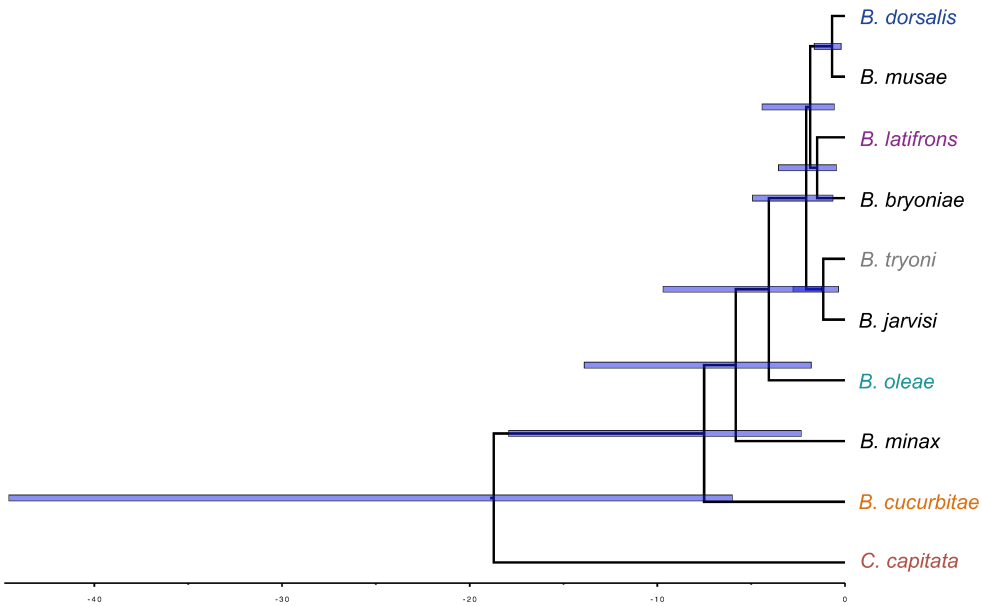


Figure S5 – Molecular time tree of *Bactrocera*. The analysis was done setting the mutation rate log-normally distributed as prior, a strict relaxed clock and a Yule model. Blue bars at nodes identify the 95% confidence interval of the inferred node age. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

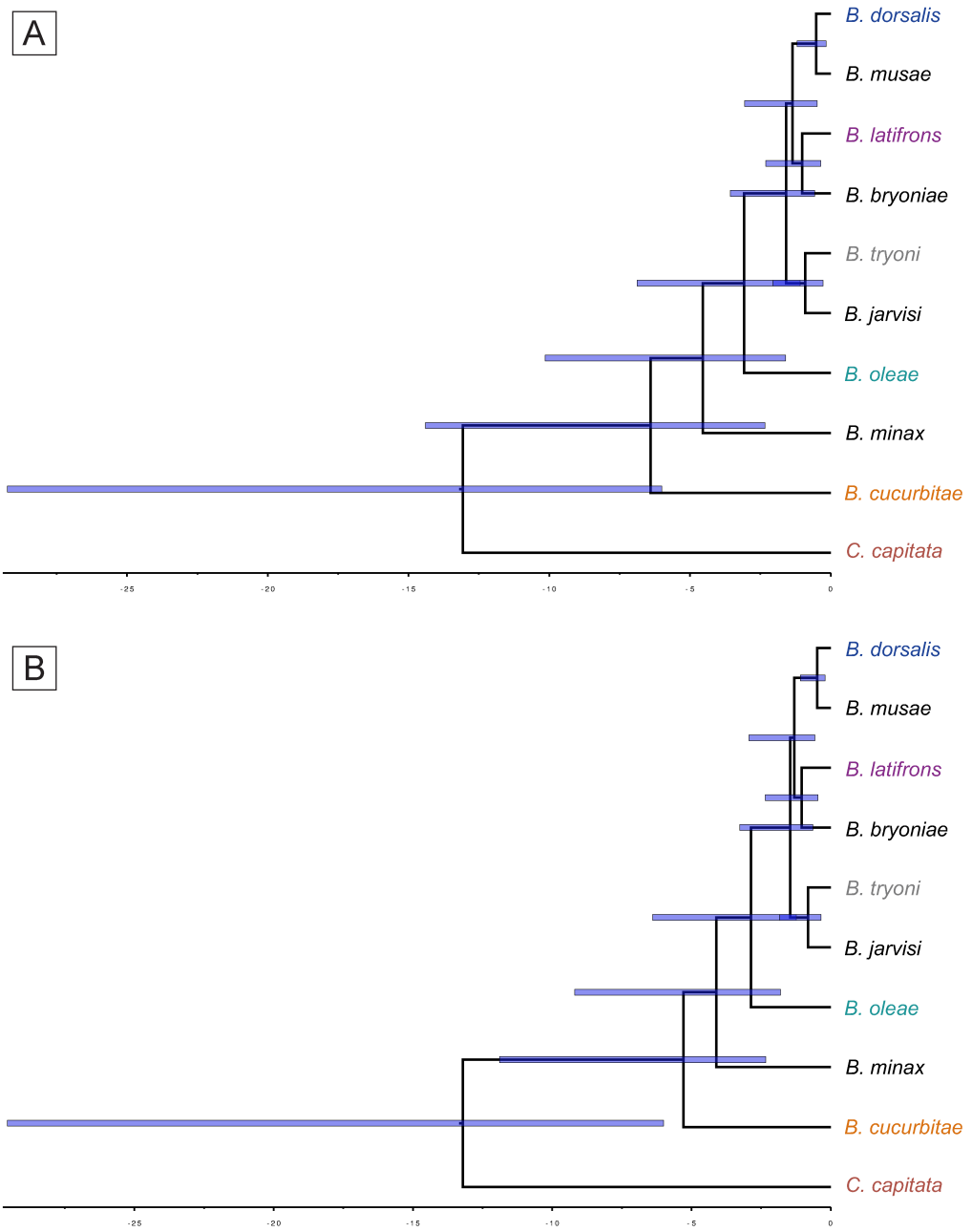


Figure S6 – Molecular time tree of *Bactrocera*. The analysis was done setting A) the mutation rate normal distribution as prior, a LOGN relaxed clock and a Birth-Death model, B) the mutation rate log-normal distribution as prior, a strict clock and a Birth-Death model. Blue bars at nodes identify the 95% confidence interval of the inferred node age. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

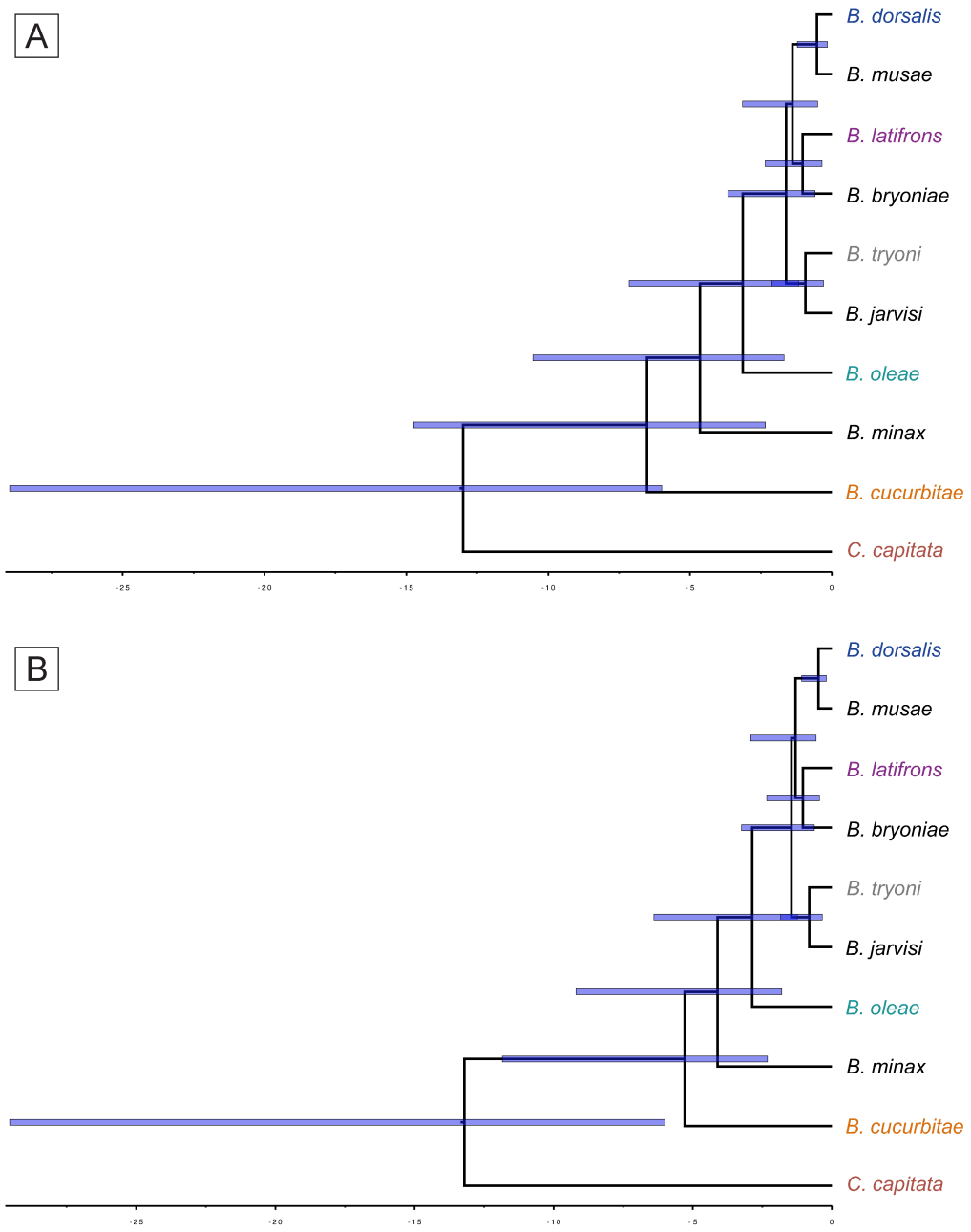


Figure S7 – Molecular time tree of *Bactrocera*. The analysis was done setting A) the mutation rate normal distribution as prior, a LOGN relaxed clock and a Yule model, B) the mutation rate log-normal distribution as prior, a strict clock and a Yule model. Blue bars at nodes identify the 95% confidence interval of the inferred node age. The five *Bactrocera* species that are the focus of the chemosensory evolution analyses and the outgroup *C. capitata* are highlighted in colour.

The evolution of host selection in *Bactrocera* fruit flies

Table S2 – Chemosensory gene repertoire in *Bactrocera* species. "Absents" indicates genes which are not present in the analysed genomes, "Singletons" contains single copy genes, "Duplications" includes multiple copy genes (total number of copies in brackets).

		Species					
		B. dorsalis	*B. latifrons*	*B. tryoni*	*B. oleae*	*B. cucurbitae*	*C. capitata*
GR Genes	Tot. Genes	66	63	58	61	63	71
	Absents	9	13	16	13	11	3
	Singletons	60	56	54	58	60	67
	Duplications	3(6)	3(7)	2(4)	1(3)	1(3)	2(4)
OR Genes	Tot. Genes	75	70	65	63	74	76
	Absents	9	11	18	13	12	0
	Singletons	56	57	49	58	50	72
	Duplications	8(19)	5(13)	6(16)	2(5)	11(24)	1(4)
OBP Genes	Tot. Genes	49	46	45	40	45	48
	Absents	6	6	8	10	8	3
	Singletons	39	39	38	36	36	42
	Duplications	3(10)	3(7)	2(7)	2(4)	4(9)	3(6)
CSP Genes	Tot. Genes	4	4	4	4	4	4
	Absents	0	0	0	0	0	0
	Singletons	4	4	4	4	4	4
	Duplications	0	0	0	0	0	0

Table S3 – Number of introns in full-length OBP genes. The asterisk symbol (*) indicates genes in which it was not possible to identify the total number of introns (i.e., because of partial gene sequence), "X" indicates genes that were absent in that species.

Genes	*B. cucurbitae*	*B. oleae*	*B. tryoni*	*B. latifrons*	*B. dorsalis*
Obp1	4	3	3	3	4
Obp2	3	3	3	3	3
Obp3	2	2	2	2	2
Obp4	3	3	3	3	*
Obp5	2	2	2	2	2
Obp6	4	4	4	4	4
Obp7	3	3	3	3	3
Obp8	4	4	*	4	4
Obp9.1	4	4	*	4	4
Obp9.2	4	X	X	X	X
Obp10	4	4	*	4	4
Obp11.1	0	0	0	0	0
Obp11.2	*	0	X	0	0
Obp13	1	1	1	1	1
Obp14	1	1	1	1	1
Obp15	1	1	1	1	1
Obp16.1	1	2	*	1	0
Obp16.2	X	X	1	1	0
Obp16.3	X	X	X	X	0
Obp17	1	1	*	1	1
Obp19	1	1	1	1	1
Obp20	1	1	1	1	1
Obp21	0	X	0	0	0
Obp22	1	1	1	1	1
Obp23	1	1	1	1	1
Obp24	3	3	3	3	3
Obp25	3	1	3	3	3
Obp26	3	3	3	3	3
Obp27	1	1	*	1	*
Obp28	4	4	4	4	4
Obp29	1	1	1	1	1
Obp30	1	1	1	1	1
Obp30like	X	1	1	1	1
Obp31	1	1	1	1	*
Obp32	1	1	1	1	*
Obp33	1	1	1	1	1
Obp39	2	2	2	*	2
Obp40	X	X	4	4	4
Obp41	2	X	2	3	2
Obp42	3	X	X	3	3
Obp43	4	4	X	4	*
Obp44.1	1	3	*	3	*
Obp44.2	1	X	X	X	X
Obp44.3	3	X	X	X	X
Obp45	2	2	2	2	2
Obp46	2	2	2	1	2
Obp47	1	1	0	1	0
Obp99c-4	1	1	1	1	1

Table S4 – Number of introns in full-length OR genes. The asterisk symbol (*) indicates genes in which it was not possible to identify the total number of introns (i.e., because of partial gene sequence), “X” indicates genes that were absent in that species.

Genes	*B. cucurbitae*	*B. oleae*	*B. tryoni*	*B. latifrons*	*B. dorsalis*
Or1	5	5	5	5	*
Or2	2	5	X	5	5
Or3	5	5	*	5	5
Or4	*	4	4	4	*
Or5	4	4	4	4	4
Or6	5	5	*	5	5
Or10	4	4	4	4	4
Or11	6	6	6	7	6
Or12	*	6	X	*	6
Or13	6	6	6	6	*
Or14	6	6	6	6	*
Or17a	*	6	*	6	6
Or17b	6	X	X	*	6
Or18a	6	6	6	6	6
Or18b	6	6	X	X	*
Or19	2	2	2	2	3
Or20	X	3	3	3	X
Or21	4	3	X	3	3
Or22	3	*	3	3	3
Or23	3	3	3	3	3
Or24	3	3	3	3	3
Or25a	3	3	3	3	*
Or25b	3	X	X	X	X
Or26	2	2	X	2	2
Or27	1	1	1	X	1
Or28	1	X	1	1	*
Or29	1	1	1	1	1
Or30	1	1	1	X	1
Or31a	1	1	1	X	1
Or31b	1	X	X	X	X
Or32(a)	1	1	1	1	1
Or32b	X	X	1	X	1
Or33	1	1	1	1	1
Or34a	1	1	*	1	1
Or34b	1	X	X	X	1
Or35	2	2	2	2	2
Or36a	2	2	2	2	2
Or36b	*	X	2	2	X
Or37	2	2	2	2	2
Or38	2	*	2	2	2
Or43	2	X	2	2	2
Or44	2	2	2	2	*
Or45	X	X	2	2	2
Or46	2	2	2	2	2
Or47	2	2	2	2	2
Or48	3	3	2	3	3
Or49	3	3	3	3	3
Or52	X	4	4	4	4
Or53a	5	5	X	*	*
Or53b	5	X	X	X	X
Or54a	4	4	*	4	4
Or54b	4	X	X	X	X
Or54c	4	X	X	X	X
Or55	5	5	5	5	5
Or56a	9	*	X	9	9
Or56b	9	X	X	9	X
Or57(a)	9	9	9	9	*
Or57b	X	X	9	X	9
Or58	3	3	3	3	3
Or59	3	3	3	3	3
Or60(a)	3	X	X	3	3
Or60b	X	X	X	3	3
Or61	3	3	*	*	3
Or63	3	3	3	3	3
Or64	3	3	3	3	*
Or65	4	4	*	4	4
Or66	4	4	3	4	4
Or67	3	3	3	3	3
Or68	3	X	3	3	3
Or69	3	3	3	3	3
Or7	5	5	5	5	5
Or71	3	3	3	3	2
Or72	3	X	X	*	3
Or73	0	3	X	3	3
Or74	X	3	3	3	3
Or75	3	*	2	3	3
Or8	4	4	4	*	4
Or9	4	4	*	4	1
Orco	6	6	6	6	*

Table S5 – Number of introns in full-length GR genes. The asterisk symbol (*) indicates genes in which it was not possible to identify the total number of introns (i.e., because of partial gene sequence), “X” indicates genes that were absent in that species.

Genes	*B. cucurbitae*	*B. oleae*	*B. tryoni*	*B. latifrons*	*B. dorsalis*
Gr1a	X	3	3	3	3
Gr1b	3	3	3	3	*
Gr3	6	4	6	6	6
Gr4	8	9	8	8	8
Gr5	8	*	8	7	7
Gr6	8	8	8	8	8
Gr7	9	9	9	9	9
Gr8	2	2	2	2	*
Gr9	7	7	7	7	7
Gr10	5	5	5	5	7
Gr11	11	11	11	11	11
Gr12	X	6	X	X	X
Gr14	11	11	11	11	11
Gr15	3	3	3	3	3
Gr16	1	1	3	2	3
Gr17b	3	3	3	4	3
Gr18	*	4	3	X	3
Gr19	3	*	3	3	3
Gr20	3	3	3	3	3
Gr21	2	2	X	2	2
Gr22	2	X	*	2	2
Gr23	2	2	2	*	2
Gr24	4	4	4	4	4
Gr26	X	2	X	2	2
Gr27	X	2	2	2	2
Gr29	*	2	X	2	2
Gr30	*	2	2	2	2
Gr31	X	X	2	2	2
Gr32	2	2	2	2	2
Gr33	4	1	4	*	4
Gr34b	4	4	4	4	4
Gr35	3	3	3	3	3
Gr38	5	5	5	5	4
Gr39	5	4	5	6	5
Gr40	3	3	4	4	3
Gr41	3	3	*	3	3
Gr42	3	3	*	3	3
Gr43	1	X	X	X	1
Gr44	2	1	1	1	1
Gr45	1	X	1	X	1
Gr46	*	1	1	1	1
Gr47	1	1	1	1	1
Gr48	1	3	1	1	1
Gr49	1	1	1	X	1
Gr50	2	1	1	1	1
Gr52	1	1	1	1	1
Gr53	1	1	1	1	1
Gr57	1	1	X	1	1
Gr58	1	1	X	1	1
Gr60a	*	X	1	1	1
Gr60b	X	X	1	1	1
Gr61	1	1	1	X	1
Gr62	2	2	2	2	*
Gr63	1	2	2	2	*
Gr64	3	3	3	3	3
Gr65b	*	1	1	1	1
Gr66	1	1	X	1	1
Gr68a	1	1	1	1	X
Gr68b	1	1	X	1	1
Gr69	2	2	2	1	2
Gr70	2	3	3	3	3
Gr71	2	2	2	2	2
Gr72	2	2	2	2	2
Gr73	2	2	1	1	*

Table S6 – Percentages of variable (1–3), intermediately variable (4–6) and conserved (7–9 grade) residues of olfactory proteins calculated by the ConSurf server [281].

	Consurf Conservation Grades		
	1–3	4–6	7–9
Obp17	23.3%	21.6%	55.2%
Obp25	14.3%	0.0%	85.7%
Or59	20.4%	37.6%	42.0%
Or38	18.6%	43.2%	38.2%
Or8	16.4%	36.5%	47.2%
Or64	14.0%	43.6%	42.4%
Or71	13.4%	38.2%	48.4%
Or11	10.2%	40.4%	49.4%
Or25	7.9%	45.7%	46.4%
Or75	4.4%	45.7%	49.8%

Side projects

During my PhD, I was involved in two other projects besides my main project.

In this section, I indicated the experiments that I conducted using the first-person pronoun (I), together with my colleagues using the plural form (we), and I adopted the third singular personal pronoun referring to the work done entirely by colleagues of mine.

1. NIRVS landscape in *Ae. albopictus*

I worked with Elisa Pischedda (PhD candidate) in her project on viral integrations in *Aedes albopictus*. I performed the Southern blotting experiments and contributed with the bioinformatic analysis (for more details, see page 118).

Ae. albopictus is an invasive mosquito and a competent vector for many arboviruses such as Chikungunya (*Alphavirus*), Dengue, and Zika (*Flavivirus*) viruses. Unexpectedly, its genome sequencing (AaloF1, [322]) harboured an unusually high number of integrated sequences with similarities to non-retroviral RNA viruses of the *Flavivirus* and *Rhabdovirus* genera. In particular, 32 Non-retroviral Integrated RNA Virus Sequences (NIRVS) with similarities to *Flaviviruses* (F-NIRVS) and 40 NIRVS similar to *Rhabdoviruses* (R-NIRVS) were reported [323].

In her project, one of the questions she aimed to address was whether NIRVS identified in the reference genome were conserved within mosquitoes of the Foshan reference strain.

For this purpose, we analysed the genome sequencing of 16 mosquitoes of the Foshan strain (i.e., single-sequenced mosquitoes) [323], and we compared their NIRVS pattern with the list of viral integrations previously characterized in the reference genome [323]. NIRVS presence in the samples was established using the DepthOfCoverage function of the GATK tool [324]. This function allowed the detection of the coverage of the region of interest and to filter out reads with Phred mapping quality lower than 20. NIRVS with a minimum of five reads of depth of coverage and 30 consecutive nucleotides with that depth of coverage were considered as present in the sample. Our criterion was more stringent than the one previously adopted by [323]. The ratio between the number of R-NIRVS present in a sample and the total R-NIRVS of Foshan (40) was used to estimate R-NIRVS prevalence. The same calculation was done for F-NIRVS. With this approach, we found that eleven NIRVS (i.e., AlbFlavi19, AlbFlavi31, AlbFlavi32, AlbFlavi33, AlbFlavi38, AlbFlavi39, AlbFlavi40, AlbRha43, AlbRha79, AlbRha80, AlbRha95) were absent in all the 16 samples. A total of 20 NIRVS were detected in the samples, with statistical enrichment for NIRVS with similarities to *Rhabdovirus* (R-NIRVS) (Hypergeometric test, $P = 0.022$)

and NIRVS mapping in gene exons (Hypergeometric test, *P* = 0.006) (**Figure 1A**). Conversely, F-NIRVS were variably distributed across the samples (**Figure 1A**). Of note is AlbFlavi4, a sequence of 512 bp with similarity to the capsid gene of *Aedes flavivirus* [323] and which was annotated within the second exon of the putative gene AALF0033. In the single-sequenced mosquitoes, variants were identified for this putative gene, only one of which includes AlbFlavi4. To confirm the presence of this sequence in Foshan mosquito, I performed a southern-blotting experiment on a pool of 20 Foshan mosquitoes using a probe for AlbFlavi4 (**Figure 1B**).

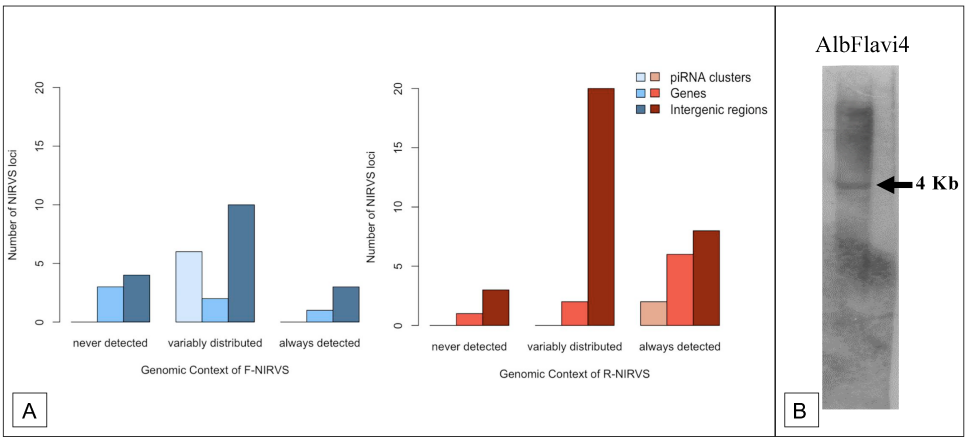


Figure 1 – (A) Number of F-NIRVS and R-NIRVS mapping within genes, piRNA clusters or intergenic regions, classified on the basis of read-coverage across the single-sequenced mosquitoes. F-NIRVS displayed in blue, R-NIRVS in red. (B) Southern-blotting result for AlbFlavi4 (adapted from [334]).

Our results showed that the landscape of viral integrations is variable among the single-sequenced mosquitoes. We found a core set of NIRVS, which is enriched for integrations with similarity to *Rhabdoviruses* and NIRVS mapping in CDS, this result was significant because it demonstrated that viral integrations are a dynamic component of the mosquito repeatome.

2. Identification of *Piwi* genes in *Ae. albopictus*

I worked with Dr. Michele Marconcini in his project on the characterization of the genes of the piRNA pathway in *Ae. albopictus.* I contributed to the Northern blotting experiments (for more details, see page 118).

The PIWI-interacting RNA (piRNA) pathway is the most recently identified of the small RNA-based pathways within RNA interference. It has been mostly studied in *D. melanogaster* where the three proteins of the Piwi subclade, namely Argonaute-3, PIWI, and Aubergine, interact with piRNAs to silence transposable elements (TE), preserving genome integrity in gonadal tissues [325]. Recent literature supports the involvement of the piRNA pathway in antiviral immunity in *Aedes* spp. mosquitoes, differently than in *D. melanogaster* [326]. It has been reported that in *Ae aegypti*, the *Piwi* subclade has undergone expansion with seven genes (i.e., *Ago3*, *Piwi2*, *Piwi3*, *Piwi4*, *Piwi5*, *Piwi6,* and *Piwi7*). They displayed differences in their tissue and developmental expression profile and were associated with either TE-derived or viral piRNAs [327, 328]. However, the genes and the function of this pathway in *Ae. albopictus* are still enigmatic.

As a consequence, one of the aims of his project was to characterize the genes in *Ae. albopictus*.

He bioinformatically identified and molecularly validated seven Piwi genes in the genome of *Ae. albopictus*: *Ago3*, *Piw1/3*, *Piwi2*, *Piwi4*, *Piwi5*, *Piwi6,* and *Piwi7*. He determined their copy number by real-time PCR.

He also molecularly analysed their transcription by RT-PCR and their expression throughout the developmental stages and the adult life of the mosquito by RNA-seq analyses. He found a single transcript sequence for *Ago3*, *Piwi1/3*, *Piwi2*, *Piwi4,* and *Piwi6* that corresponded to the predictions on the identified DNA sequences. Sequencing results of the transcript from Piwi5 showed a sequence 27 bp shorter than predicted on the reference genome, due to a 45bp gap followed by a 18bp insertion, 110 and 333 bases after the ATG starting codon, respectively. Also, the transcript sequence of *Piwi7* appeared shorter than the predicted. We investigated the presence of *Piwi5* and *Piwi7* transcripts by Northern-blot and our results for *Piwi5* indicate the presence of a transcript of 3 kb, while we were not able to detect *Piwi7* transcript (**Figure 2**).

Combing these approaches, he was able to determine that the genome of *Ae. albopictus* harbours one copy of *Ago3* and six *Piwi* genes (i.e., *Piwi1/3*, *Piwi2*, *Piwi4*, *Piwi5*, *Piwi6,* and *Piwi7*). He also investigated their transcriptional profiles and confirmed their expression throughout the developmental stages and the adult life of the mosquito, both in ovaries and somatic tissues. Interestingly, *Piwi7* transcript expression decreased following early embryogenesis, to the point that we could detect it neither in RNA-seq analyses nor in Northern-blot experiments (**Figure 2**).

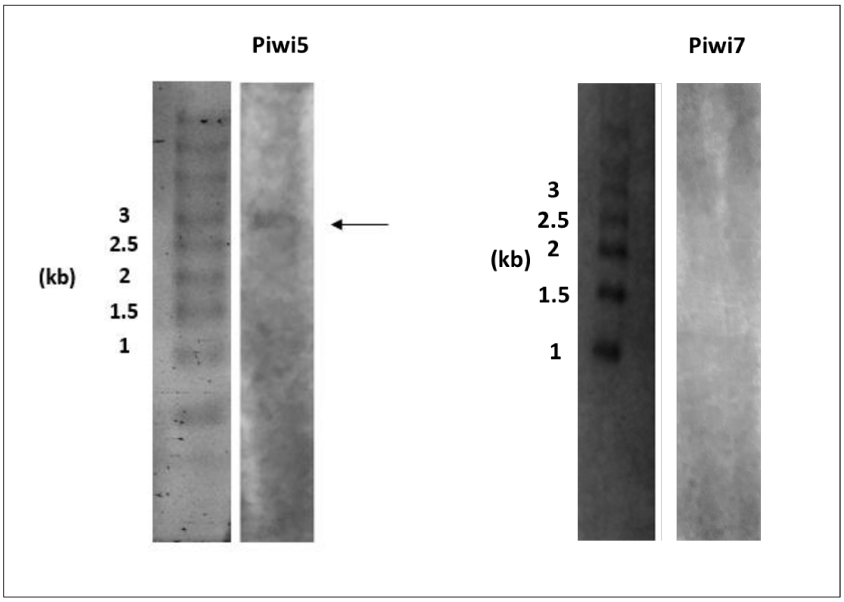


Figure 2 – Northern Blot results for *Piwi5* and *Piwi7* (adapted from [335]).

Acknowledgments

First, I would like to express my sincere gratitude to my advisor Prof. Lino Ometto for his knowledge, patience and motivation throughout my PhD research studies. Without his assistance and dedicated involvement, this thesis would have never been accomplished.

I would like to thank all members of my PhD committee for the guidance and for providing interesting and valuable feedback.

I am pleased to thank Professors Mariangela Bonizzoni, Anna R. Malacrida, Giuliano Gasperi and Ludvik Gomulski for sharing insightful suggestions and support for my work.

I also wish to thank my former and current lab mates: Davide, Elisa, Umberto, Alessandro, Annamaria, Michele, Alejandro, Leila and all the other people in the lab for the stimulating discussions and for all the fun we have had over these past years.

Special thanks to Francesca whose helpful advice, willingness to answer questions whenever needed, superlative knowledge of fruit fly biology and profound belief in my abilities have been crucial.

I am grateful to my beloved Stefano for being a never-ending source of love, inspiration, motivation and for making me enjoy every single moment together. To my best friend, Linda, for always being by my side and for the invaluable support. To my friend Alessandra and my "little" cousin Rossella, thank you for being there for me despite the kilometres between us.

Finally, I would like to thank my amazing parents and my brother for the unconditional and unparalleled love, encouragement and freedom. To them, I dedicate this thesis.

List of original manuscripts

Published:

1. Pischedda E, Scolari F, **Valerio F**, Carballar-Lejarazú R, Catapano PL, Waterhouse RM, Bonizzoni M.
   Insights into an unexplored component of the mosquito repeatome: Distribution and variability of viral sequences integrated into the genome of the arboviral vector *Aedes albopictus*. *Frontiers in genetics*. 2019;10 FEB. doi: 10.3389/fgene.2019.00093

2. Marconcini M, Hernández L, Iovino G, Houé V, **Valerio F**, Palatini U, Pischedda E, Crawford J, Carballar- Lejarazú R, Ometto L, Forneris F, Failloux AB, Bonizzoni M.
   Polymorphism analyses and protein modelling inform on functional specialization of Piwi clade genes in the arboviral vector *Aedes albopictus*. *PLOS Neglected Tropical Diseases*. 2019;13:e0007919. doi: 10.1371/journal.pntd.0007919.

3. Carraretto D, Aketarawong N, Di Cosimo A, Manni M, Scolari F, **Valerio F**, Malacrida A.R, Gomulski L and Gasperi G.
   Transcribed sex-specific markers on the Y chromosome of the oriental fruit fly, *Bactrocera dorsalis*. *BMC Genetics*. 2020;21:125. doi: 10.1186/s12863-020-00938-z.

In preparation:

4. **Valerio F**, Zadra N, Rota Stabelli O, Ometto L. Fast and recent radiation of *Bactrocera* revealed by multi-locus phylogenetic analyses.

5. **Valerio F**, Scolari F, Ometto L. The effect of host selection in the evolution of the chemosensory gene families of *Bactrocera*.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 February 2019
doi: 10.3389/fgene.2019.00093

Insights Into an Unexplored Component of the Mosquito Repeatome: Distribution and Variability of Viral Sequences Integrated Into the Genome of the Arboviral Vector *Aedes albopictus*

*Elisa Pischedda[1], Francesca Scolari[1], Federica Valerio[1], Rebeca Carballar-Lejarazú[2], Paolo Luigi Catapano[1], Robert M. Waterhouse[3] and Mariangela Bonizzoni[1]**

[1] Department of Biology and Biotechnology, University of Pavia, Pavia, Italy, [2] Department of Microbiology & Molecular Genetics, University of California, Irvine, Irvine, CA, United States, [3] Department of Ecology and Evolution, University of Lausanne and Swiss Institute of Bioinformatics, Lausanne, Switzerland

OPEN ACCESS

Edited by:
Fulvio Cruciani,
Sapienza University of Rome, Italy

Reviewed by:
Joao Trindade Marques,
Federal University of Minas Gerais, Brazil
Jinbao Gu,
Southern Medical University, China

***Correspondence:**
Mariangela Bonizzoni
m.bonizzoni@unipv.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *17 October 2018*
Accepted: *29 January 2019*
Published: *12 February 2019*

Citation:
Pischedda E, Scolari F, Valerio F, Carballar-Lejarazú R, Catapano PL, Waterhouse RM and Bonizzoni M (2019) Insights Into an Unexplored Component of the Mosquito Repeatome: Distribution and Variability of Viral Sequences Integrated Into the Genome of the Arboviral Vector Aedes albopictus. Front. Genet. 10:93. doi: 10.3389/fgene.2019.00093

The Asian tiger mosquito *Aedes albopictus* is an invasive mosquito and a competent vector for public-health relevant arboviruses such as Chikungunya (*Alphavirus*), Dengue and Zika (*Flavivirus*) viruses. Unexpectedly, the sequencing of the genome of this mosquito revealed an unusually high number of integrated sequences with similarities to non-retroviral RNA viruses of the *Flavivirus* and *Rhabdovirus* genera. These Non-retroviral Integrated RNA Virus Sequences (NIRVS) are enriched in piRNA clusters and coding sequences and have been proposed to constitute novel mosquito immune factors. However, given the abundance of NIRVS and their variable viral origin, their relative biological roles remain unexplored. Here we used an analytical approach that intersects computational, evolutionary and molecular methods to study the genomic landscape of mosquito NIRVS. We demonstrate that NIRVS are differentially distributed across mosquito genomes, with a core set of seemingly the oldest integrations with similarity to *Rhabdoviruses*. Additionally, we compare the polymorphisms of NIRVS with respect to that of fast and slow-evolving genes within the *Ae. albopictus* genome. Overall, NIRVS appear to be less polymorphic than slow-evolving genes, with differences depending on whether they occur in intergenic regions or in piRNA clusters. Finally, two NIRVS that map within the coding sequences of genes annotated as *Rhabdovirus* RNA-dependent RNA polymerase and the nucleocapsid-encoding gene, respectively, are highly polymorphic and are expressed, suggesting exaptation possibly to enhance the mosquito's antiviral responses. These results greatly advance our understanding of the complexity of the mosquito repeatome and the biology of viral integrations in mosquito genomes.

Keywords: mosquitoes, viral integrations, immunity, piRNA pathway, domestication, repeatome

Abbreviations: FC, fold-change; FGs, fast-evolving genes; F-NIRVS, NIRVS with similarity to *Flaviviruses;* LoP, level of polymorphism; N-Gs, genes harboring NIRVS; NIRVS, non-retroviral integrated RNA virus sequences; R-Gs, Genes of the RNAi pathway; R-NIRVS, NIRVS with similarity to *Rhabdoviruses;* SGs, slow-evolving genes; SSM, single-sequenced mosquitoes.

INTRODUCTION

The amount and the type of repeated DNA sequences, collectively called the "repeatome," affect the size, organization and evolution of eukaryotic genomes (Maumus and Quesneville, 2014). Transposable elements (TEs) are the major and most-studied components of the repeatome because of their potential mutagenic effects (Gilbert and Feschotte, 2018). TEs evolve through a "burst and decay model" whereby newly acquired TEs can multiply rapidly in a genome. The "burst" phase is followed by low amplification periods, the "decay" moments, when TEs tend to accumulate mutations and become inactive (Maumus and Quesneville, 2016). In eukaryotes, TE mobilization during germline formation is counterbalanced by the activity of the PIWI-interacting RNA (piRNA) pathway, the most recently identified of three small RNA-based silencing mechanisms (Brennecke et al., 2007; Guzzardo et al., 2013; Gainetdinov et al., 2017). Briefly, Argonaute proteins of the PIWI-subfamily associate with small RNAs of 25-30 nucleotides, called PIWI-interacting RNAs (piRNAs), and together they silence TEs based on sequence complementarity (Tóth et al., 2016). piRNAs arise from genomic regions called piRNA clusters, which contain fragmented sequences of previously acquired TEs.

Unexpectedly, besides TE fragments, piRNA clusters contain sequences from non-retroviral RNA viruses, which produce piRNAs, in the genome of arboviral vectors like the mosquitoes *Aedes aegypti* and *Aedes albopictus* (Arensburger et al., 2011; Olson and Bonizzoni, 2017; Palatini et al., 2017; Whitfield et al., 2017). This observation is in line with recent experimental evidence that extend the role of the piRNA pathway to immunity against viruses in *Aedes* mosquitoes, differently than in *Drosophila melanogaster* (Miesen et al., 2016; Petit et al., 2016) and show that piRNAs from integrated viral sequences are differentially expressed following viral challenge of *Ae. albopictus* (Wang et al., 2018). As such, NIRVS have been proposed as novel immunity factors of arboviral vectors (Olson and Bonizzoni, 2017; Palatini et al., 2017; Whitfield et al., 2017). However, the organization, stability and mode of action of NIRVS in mosquito genomes are poorly understood.

The landscape of viral integrations in the genome of *Ae. aegypti* and *Ae. albopictus* mosquitoes is rather complex. *Aedes* species are rare examples within the animal kingdom because they harbor dozens of NIRVS from different viruses, such as *Flaviviruses* and Mononegavirales, primarily *Rhabdoviruses* and poorly characterized *Chuviruses* (Katzourakis and Gifford, 2010; Fort et al., 2012; Palatini et al., 2017; Whitfield et al., 2017). In all other animals in which NIRVS have been identified, including mammals, birds and ticks, NIRVS appear to be mainly from one viral type and tend to be found in low numbers (<20) (Belyi et al., 2010; Katzourakis and Gifford, 2010; Holmes, 2011; Kryukov et al., 2018). NIRVS identified in the *Ae. aegypti* and *Ae. albopictus* genomes are not homologous, indicating independent integration events. However, NIRVS of both species encompass fragmented viral open reading frames (ORFs). In *Ae. albopictus,* we characterized 32 NIRVS with similarities to *Flaviviruses* (F-NIRVS) and 40 NIRVS similar to *Rhabdoviruses* (R-NIRVS). These NIRVS are enriched in piRNA clusters and within coding sequences (Palatini et al., 2017). Taken together these findings support the hypothesis that NIRVS contribute to host biology. However, because NIRVS have been identified by *in silico* analyses of the currently available *Ae. albopictus* genome assembly, which was built from the DNA of a single pupa of the Foshan strain (Chen et al., 2015) and we verified the overall conservation of NIRVS within this strain (Palatini et al., 2017), their widespread occurrence in wild mosquitoes, whether all NIRVS or some are functionally active elements, and what is the relative importance of each of them, are all still unexplored questions.

Here we addressed the following questions: is the pattern of NIRVS within mosquitoes of the Foshan strain the same as across geographic samples? If the landscape of NIRVS is variable, could NIRVS be co-opted as novel molecular markers for population genetic studies? Does NIRVS age differ depending on their viral origin? How does the LoP of NIRVS compare with that of fast- and slow-evolving mosquito genes?

Using an analytical approach that intersects computational, evolutionary, and molecular approaches we show that NIRVS are a dynamic component of the *Ae. albopictus* repeatome. The landscape of NIRVS is variable within mosquitoes of the Foshan strains and among geographic samples. The LoP of NIRVS is heterogeneous. R-NIRVS appear more widespread and older integrations than those with similarities F-NIRVS. NIRVS annotated in intergenic regions appear more variable than those mapping within piRNA clusters or gene exons. Among NIRVS identified within gene exons, six are fixed and stably expressed, albeit showing different levels of polymorphism and domestication cannot be excluded for AlbRha52 and AlbRha12, which are part of genes annotated as RNA-dependent RNA polymerase and nucleocapsid-encoding genes of *Rhabdovirus*, respectively.

Overall these results greatly advance our understanding of the widespread occurrence of NIRVS in nature. Additionally, a detailed analysis of NIRVS distribution and polymorphism within the *Ae. albopictus* genome paves the way for choosing candidate NIRVS for functional studies.

MATERIALS AND METHODS

Mosquitoes

Mosquitoes of the Foshan strain have been reared at the insectary of the University of Pavia since 2013 (Palatini et al., 2017). Upon arrival in Pavia, mosquitoes were checked for infection using *Flavivirus* degenerate primers (Crochu et al., 2004). No infection was detected. Mosquitoes are reared at 28°C and 70–80% relative humidity with 12/12 h light/dark cycles. Larvae are reared in pans and fed on finely ground fish food (Tetramin, Tetra Werke, Melle, Germany). Adults are kept in 30-cm^3 cages and allowed access to a cotton wick soaked in 0.2 g/ml sucrose as a carbohydrate source. Adult females are blood-fed using a membrane feeding apparatus and commercially available mutton blood. Sixteen Foshan mosquitoes, eight males and eight females, were sampled and forced in single mating. Progeny from each single mating was collected. DNA was extracted from single individuals, including

parents and their progeny, using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany).

Southern Blotting

Genomic DNA (∼10 mg) from pools of 10–20 adult mosquitoes of the Foshan strain were digested with restriction enzymes (Thermo Scientific, Vilnius, Lithuania) chosen to specifically target individual F-NIRVS and separated on a 0.8% agarose gel. DNA was transferred to nylon membranes (Hybond-N+) (Amersham, Buckinghamshire, United Kingdom) and immobilized by UV irradiation. Random-primed DNA probes (**Supplementary Data 1**) were labeled with [α-^{32}P] dATP/ml and [α-^{32}P] dCTP/ml (3,000 Ci/mmol; 1 Ci- 37 GBq) by using the Megaprime labeling kit (Amersham, Buckinghamshire, United Kingdom). Hybridizations were carried out at 65°C.

Real Time PCR (qPCR) to Test for NIRVS Copy Number

PCR primers were designed using PRIMER3 (Rozen and Skaletksy, 2000) within F-NIRVS to test for their copy number based on real-time PCR (Bubner and Baldwin, 2004; Yuan et al., 2007) (**Supplementary Data 2**). Reaction mixtures were prepared containing 10 μL QuantiNova Sybr Green PCR Master Mix (Qiagen, Hilden, Germany), 1 μL of each 10 μM primer, and template DNA diluted in distilled H_2O up to 20 μL total reaction volume. Template genomic DNA used in the reactions was extracted from individual adult mosquitoes following a standard protocol (Baruffi et al., 1995). Real-time PCR reactions were performed in a two-step amplification protocol consisting of 2 min at 95°C, followed by 40 cycles of 95°C for 5 s and 60°C for 10 s. Reactions were run on a RealPlex Real-Time PCR Detection System (Eppendorf, Hamburg, Germany). The single-copy gene *piwi6* (AALF016369) was used as reference after having verified the region of the primers does not harbor variability. F-NIRVS copy numbers were estimated comparing the relative quantification of NIRVS loci with respect to that of the reference genes using the ΔCt method (Pfaffl, 2006), after having verified that the efficiencies of PCR reactions with primers for F-NIRVS and the reference gene were the same. Support for using relative quantification without an internal calibrator came from a preliminary test where we cloned one NIRVS (AlbFlavi4) and we verified that estimates of its copy number by absolute vs. relative quantification were the same.

qPCR to Estimate NIRVS Expression Levels

Total RNA was extracted using TRIzol (Life Technologies, Madrid, Spain) from pools including 10–20 mosquitoes at different developmental stages such as larvae, pupae, adult males, sugar-fed females and females sampled 48 h after blood feeding. After DNaseI (Sigma-Aldrich, Schnelldorf, Germany) treatment, a total of 100 ng of RNA from each pool was used for reverse transcription using the qScript cDNA SuperMix (Quanta Biosciences, Leuven, Belgium). Expression of the eight N-Gs and always detected in Foshan (i.e., AALF005432, AALF025780, AALF000476, AALF000477, AALF020122, AALF004130, and AALF025779) was quantified using real-time qPCRs following the protocol described above. Expression values were normalized to mRNA abundance levels of the *Ae. albopictus nap* gene (Reynolds et al., 2012) (**Supplementary Data 3**). The qbase+software (Hellemans et al., 2007) was used to compare expression profiles across samples, and Morpheus[1] was used to visualize the data.

Selection of Genes With Slow and High Evolutionary Rates

Orthologous genes across 27 insect species within the Nematocera sub-order were identified in OrthoDB v9.1 (Zdobnov et al., 2016). Levels of sequence divergence were computed for each orthologous group as the average of interspecies amino acid identified normalized to the average identity of all interspecies best-reciprocal-hits, computed from pairwise Smith-Waterman alignments of protein sequences (**Supplementary Table 1**). We selected the 0.1% of the genes (n = 14, number comparable to that of our NIRVS groups) at each tail of the distribution as representative of the conserved and variable categories, the left and right tails respectively. Orthologs of these genes were searched in the *Ae. albopictus* genome (AaloF1 assembly).

NIRVS in Natural Populations

PCR primers were designed using PRIMER3 (Rozen and Skaletksy, 2000) to test for NIRVS polymorphism in *Ae. albopictus* geographic samples (**Supplementary Data 4**). Considering the level of NIRVS sequence similarity, their copy number and heterogeneous presence in Foshan mosquitoes, we selected seven F-NIRVS (AlbFlavi2, AlbFlavi4, AlbFlavi8-41, AlbFlavi10, AlbFlavi36, AlbFlavi1, and AlbFlavi12-17) and six R-NIRVS (AlbRha1, AlbRha7, AlbRha14, AlbRha36, AlbRha52, AlbRha85) that gave unambiguous PCR results, have similarities to different viral ORFs and are distributed in different genomic regions including piRNA clusters, intergenic or coding regions. Natural mosquito samples derive from a world-wide collection available at the University of Pavia and previously analyzed with microsatellite markers (Manni et al., 2017). PCR reactions were performed in a final volume of 25 μL using DreamTaq^{TM} Green PCR Master Mix 2x (Thermo Scientific, Vilnius, Lithuania) and the following cycle conditions: 94°C for 3 min, 40 cycles at 94°C for 30 s, 58–60°C for 45 s, 72°C for 1 min, and a final extension at 72°C for 10 min. Amplification products were electrophoresed on 1–1.5% agarose gels and purified using ExoSAP-IT^{M} PCR product Cleanup Reagent (Thermo Scientific, Vilnius, Lithuania). When the NIRVS were detected, at least five amplification products per population per locus were sent to be sequenced by Macrogen (Barcellona, Spain), following the company's requirements.

Non-retroviral Integrated RNA Virus Sequences alleles were first scored based on their occurrence in each population and their size. A Neighbor-joining tree was built after 1000 bootstrap resampling of the original data set and the calculation of a matrix

[1] https://software.broadinstitute.org/morpheus

of shared allele distances (DAS) using POPULATIONS version 1.2.31 (Langella, 1999).

Estimates of Integration Time

Non-retroviral Integrated RNA Virus Sequences sequences from geographic samples were aligned in Ugene version 1.26.1 (Okonechnikov et al., 2012) with MAFFT (Yamada et al., 2016). Default parameters with five iterative refinements were applied for the alignment. Alignments were manually curated to verify frameshifts, truncations, deletions, and insertions. All positions including gaps were filtered out from the analysis. The following formula was used to estimate the time of integration in years assuming that all mutations are neutral:

$$\text{Mean Mutations/Seq} = \frac{\text{Tot. Obs. Mutations}}{\text{N. Seqs* Seq. Length}}$$

$$\text{Age in Years} = \frac{\text{Mean Mutations/Seq}}{\text{(MR* Seq. Length* GpY)}}$$

Mutation rate (MR) were assumed to be comparable to those of *D. melanogaster* genes in range 3.5-8.4 $\times$ 10^{-09} (Haag-Liautard et al., 2007; Keightley et al., 2009). A range of 4–17 number of generations per year (GpY) was tested considering mosquitoes of temperate or tropical environments (Manni et al., 2017).

Phylogenetic Inference and Timetrees

Deduced NIRVS protein sequences were aligned with subsets of corresponding proteins from *Flavivirus* and *Rhabdovirus* genomes using MUSCLE (Edgar et al., 2004). The timetrees were generated using the RelTime method (Tamura et al., 2012) after having generated the maximum likelihood tree, with 100 bootstrap replicates. Divergence times for all branching points in the topology were calculated using the maximum likelihood method and implementing the best fitted amino acids substitution model. Phylogenies were estimated in MEGA7 (Kumar et al., 2016). The JTT matrix-based model was used for the L protein of *Rhabdovirus* (Jones et al., 1992). In this case, the estimated log likelihood value was −116005.08. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 0.8331)]. The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 0.21% sites). The analysis involved 49 amino acid sequences. There was a total of 2319 positions in the final dataset. For the G protein of *Rhabdoviruses*, the Whelan and Goldman model was implemented (Whelan and Goldman, 2001). In this case, the estimated log likelihood value was −3719.06. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 2.1095)]. The analysis involved 40 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There was a total of 56 positions in the final dataset. The LG model was used for the NS3 protein of *Flaviviruses* (Le and Gascuel, 2008). In this case, the estimated log likelihood value is −6360.35. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 0.8640)]. The analysis involved 30 amino acid sequences. All positions containing gaps and missing data were eliminated. There was a total of 180 positions in the final dataset. For NIRVS with similarities to the NS5 protein of *Flaviviruses* the JTT matrix-based model was used (Jones et al., 1992). The estimated log likelihood value of the topology shown was −26019.44. A discrete Gamma distribution was used to model evolutionary rate differences among sites [2 categories (+G, parameter = 1.0058)]. The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 10.15% sites). The analysis involved 33 amino acid sequences. There was a total of 984 positions in the final dataset. In each case, trees were drawn to scale, with branch lengths measured in the relative number of substitutions per site. The coding sequences for proteins of the Potato Yellow Dwarf virus (PYDV) were used as outgroup for trees of R-NIRVS, considering PYDV belongs to the highly divergent *Nucleorhabdovirus* genus (Dietzgen et al., 2017). To derive the genealogy of F-NIRVS, outgroups were protein sequences from Tamana Bat Virus (TABV) (de Lamballerie et al., 2002).

Bioinformatic Pipeline to Study the Polymorphisms of NIRVS, Fast- and Slow-Evolving Genes

Whole genome sequencing data of 16 singly sequenced (i.e., single-sequenced mosquitoes or SSMs) as previously described (Palatini et al., 2017) was used for the analyses of NIRVS polymorphism. NIRVS presence in a sample was established imposing a more stringent criteria than previously used in Palatini et al. (2017). Here to the "minimum of five reads of depth of coverage," we added a minimum of 30 consecutive nucleotides with that depth of coverage (**Supplementary Table 2**). This more stringent criteria resulted in a difference of one in the number of NIRVS called as absent (AlbRha43). We molecularly validated bioinformatic predictions based on this criterion (**Supplementary Figure 1**). The ratio between the number of R-NIRVS present in a sample and the total R-NIRVS of Foshan (40) was used to estimate R-NIRVS prevalence. The same calculation was done for F-NIRVS. The polymorphism of NIRVS and that of selected FGs and slow-evolving genes (i.e., SGs) was then estimated using a custom pipeline organized into different steps. In the first step, the DepthOfCoverage function of the GATK tool (Mckenna et al., 2010) is used to evaluate the coverage of the region of interest limiting to reads with Phred mapping quality greater than 20. Following read coverage analyses, four different Variant Callers i.e., GATK UnifiedGenotyper (Mckenna et al., 2010), Freebayes (Garrison and Marth, 2012), Platypus (Rimmer et al., 2014), and Vardict (Lai et al., 2016), were implemented to identify SNPs and INDELs within the regions of interest. The search of SNPs and INDELS by different variant callers allowed to increase the pool of variants and reduce the number of false positive. Custom scripts were then used to filter data, retain only variants having allele frequency higher than 0.1 or variants called by at least two

programs. The LoP of the region of interest was calculated as the total number of SNPs and INDELs identified averaged based on its length.

Follow-up statistical analyses were computed and visualized in R studio (RStudio Team, 2015). RStudio: Integrated Development for R. RStudio, Inc. (Boston, 2015). The Kolmogorov–Smirnov test was used to test the significance of the difference of LoP distributions of NIRVS, RNAi genes (R-Gs), N-Gs and FGs with respect to that of SGs (**Supplementary Table 3**). SG LoP was the median of the LoPs of the tested SGs. The threshold of significance was adjusted with the Bonferroni correction and loci were separated according to the adjusted significance of the test. Results of ratio between the LoP of each locus and the median LoP of SGs (fold change [FC]) that were different from 0 were visualized in a volcano plot. For each locus, FC was calculated as the ratio of the median LoP of the locus and that of the SG. The hypergeometric test was applied to test whether the group of NIRVS always identified across SSMs was enriched in (1) F- or R-NIRVS; (2) any viral ORFs; (3) NIRVS shorter or longer than 500 bp; (4) NIRVS mapping in exons, piRNA clusters or intergenic regions.

Search for Novel Viral Integrations

Sequences supported by the presence of soft-clipped reads were molecularly tested by PCR assays using DNA from individual mosquitoes of the Foshan strain (**Supplementary Data 5**). The Vy-PER pipeline (Forster et al., 2015) was applied to WGS data from the 16 SSMs to search for viral integrations that had not been previously identified in genome of the Foshan strain (AaloF1 assembly). Vy-PER was run using 540 viral genomes from VIPERdb (Carrillo-Tripp et al., 2009), including species of the Togaviridae, Flaviviridae, Bunyaviridae, Rhabdoviridae, Orthomyxoviridae, Reoviridae, Bornaviridae, Filoviridae, Nyamiviridae, Paramyxoviridae families. Paired-end reads identified by Vy-PER in which one read maps to the reference mosquito genome (i.e., AaloF1) and its pair maps to one of the tested viral genomes were manually inspected. Candidates including low complexity sequence (i.e., sequence showing more than 80% in mono- and di-nucleotides) or with viral portion shorter than 50 nucleotides were considered false positive and were filtered out.

RESULTS

We use read depth of coverage and variant calling tools to study NIRVS (Non-retroviral Integrated RNA Virus Sequences) widespread occurrence and their polymorphism within the genomes of mosquitoes of the Foshan strain and to look for novel viral integrations. Additionally, we studied the distribution of a selected subset of NIRVS in geographic samples.

NIRVS Are Variably Distributed in SSMs

We used the sequenced genomes of 16 mosquitoes (i.e., single-sequenced mosquitoes or SSMs) (Palatini et al., 2017) and we compared their NIRVS pattern with the list of viral integrations characterized from the Foshan genome assembly (AaloF1). Eleven NIRVS (i.e., AlbFlavi19, AlbFlavi31, AlbFlavi32, AlbFlavi33, AlbFlavi38, AlbFlavi39, AlbFlavi40, AlbRha43, AlbRha79, AlbRha80, AlbRha95) were absent in all 16 SSMs (**Supplementary Table 2**). A total of 20 NIRVS were found in all SSMs, with a statistical enrichment for NIRVS with similarities to *Rhabdovirus* (R-NIRVS) (Hypergeometric test, $p = 0.022$) and NIRVS mapping in gene exons (Hypergeometric test, $p = 0.006$) (**Figure 1A**). This "core" of 20 NIRVS included R-NIRVS identified within the coding sequence of genes (i.e., AlbRha12, AlbRha15, AlbRha28, AlbRha52, AlbRha85 and AlbRha9) and piRNA clusters (i.e., AlbRha14 and AlbRha36). Conversely, NIRVS with similarities to *Flaviviruses* (F-NIRVS) were variably distributed among SSMs. Of note is AlbFlavi4, a 512bp sequence with similarity to the capsid gene of *Aedes flavivirus* (Palatini et al., 2017). AlbFlavi4 is annotated within the second exon of AALF003313 and is also included in piRNA cluster 95 (Liu et al., 2016). AlbFlavi4 produces vepi4730383, a piRNA that is upregulated upon dengue infection (Wang et al., 2018). In SSMs and *Ae. albopictus* geographic samples, variants were identified for AALF003313, only one of which includes AlbFlavi4 (**Figures 1B,C**).

Overall, mean base pairs (bp) occupied by F-NIRVS and R-NIRVS are 12095 and 19293 bp, respectively (**Figure 1D**). Taken together, these results demonstrate that, with an average genome occupancy of 31389 bp, NIRVS represent quantitatively a limited fraction of the mosquito repeatome. However, the enrichment of NIRVS in piRNA clusters (Palatini et al., 2017) and the fact that the pattern of NIRVS is variable in host genomes support the hypothesis that NIRVS are a dynamic component of the repeatome.

NIRVS Distribution in Geographic Populations

To verify if NIRVS are variably distributed in natural samples besides in the Foshan strain, we choose seven F-NIRVS (AlbFlavi2, AlbFlavi4, AlbFlavi8-41, AlbFlavi10, AlbFlavi36, AlbFlavi1, and AlbFlavi12-17) and six R-NIRVS (AlbRha1, AlbRha7, AlbRha14, AlbRha36, AlbRha52, AlbRha85) based on their unique occurrence in different regions of the mosquito genome and their similarity to various viral ORFs. AlbRha52 and AlbRha85 are annotated as unique exons of AALF020122 and AALF004130, respectively. We tested the presence of these NIRVS in native (China and Thailand), old (La Reunion Island) and new (United States and Italy) *Ae. albopictus* populations (Manni et al., 2017). NIRVS alleles were differentially distributed across geographic populations so that a tree built from a matrix of shared-allele distances (DAS) proved able to differentiate mosquito populations in accordance with the historical records of *Ae. albopictus* invasive process when considering all thirteen NIRVS, only F-NIRVS or NIRVS mapping in intergenic regions (**Figures 2A–C** and **Supplementary Data 6**). On the contrary, when data from exclusively R-NIRVS or NIRVS identified in piRNA clusters, were analyzed, bootstrap values differentiating populations were below 50% (**Figures 2D,E**). This result agrees with the observation that the higher abundant R-NIRVS are also more prevalent than F-NIRVS.

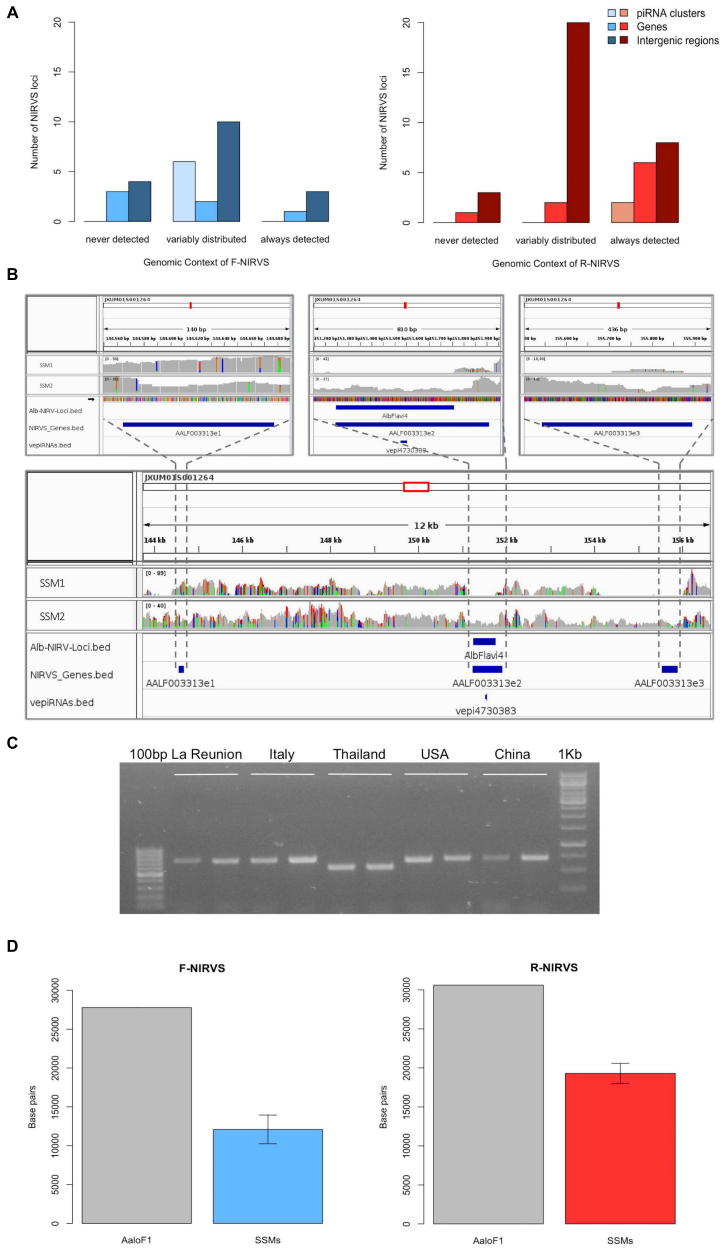


FIGURE 1 | NIRVS are variably distributed in SSMs. **(A)** Number of *Flavivirus* (F-NIRVS) and *Rhabdovirus* (R-NIRVS) loci mapping within genes, piRNA clusters or intergenic regions, classified on the basis of read-coverage across SSMs. **(B)** IGV screen shot showing read-coverage at AALF003313 in SSM1 and SSM2. Positions of the three AALF003313 exons, AlbFlavi4 and vepi4730383 are indicated by blue bars. **(C)** PCR amplification of AALF003313 exon2 in ten *Ae. albopictus* geographic samples. **(D)** F-NIRVS and R-NIRVS loci occupancy in the 16 single-sequenced mosquitoes (SSMs) of the Foshan strain is about half of that expected based on the annotated sequences of the reference genome assembly (AaloF1). F-NIRVS are in blue, R-NIRVS are in red.

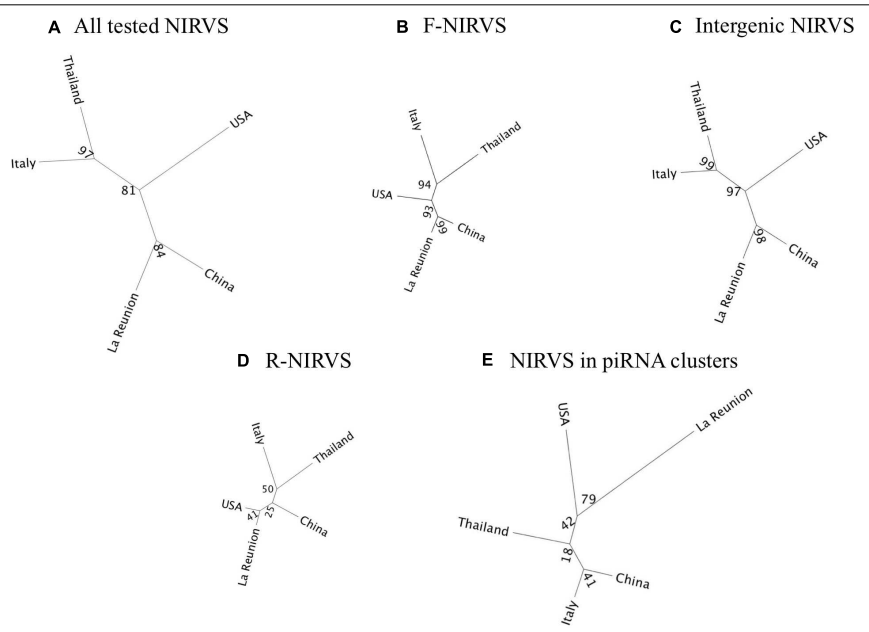


FIGURE 2 | Phylogeographic distribution of NIRVS. Genetic relationships among five mosquito populations shown by a Neighbor-joining trees based on shared-allele distance using data from all 13 genotyped NIRVS **(A)**, only F-NIRVS **(B)**, NIRVS mapping in intergenic regions **(C)**, only R-NIRVS **(D)** and NIRVS mapping in piRNA clusters **(E)**. Bootstrap values are shown at the corresponding nodes.

R-NIRVS Appear to Be Older Integrations Than F-NIRVS

The higher prevalence of R-NIRVS with respect to F-NIRVS suggests R-NIRVS are older integrations (**Figure 3**). To verify this hypothesis, we sequenced alleles of NIRVS identified in the five tested populations and we estimated integration times, assuming comparable mutation rates between *Ae. albopictus* and *D. melanogaster*, that is 3.5–8.4 $\times$ 10^{-9} per site per generation (Haag-Liautard et al., 2007; Keightley et al., 2009), and a range of generations per year between 4 and 17, accounting for mosquitoes from temperate and tropical environments, respectively (Manni et al., 2017). Under these conditions, R-NIRVS integrated between 36 thousand and 2.7 million years ago (mya) and F-NIRVS between 7.4 thousand and 2.4 mya (**Figure 4**). This large window supports the conclusion that integration of viral sequence is a dynamic process occurring occasionally at different times. As shown in **Figure 4**, estimates of integration times varied greatly depending on the genomic context of NIRVS. NIRVS annotated within gene exons appear statistically more recent than NIRVS of piRNA clusters (ANOVA, $^{***}P < 0.001$). Besides reflecting a different integration time, this result is consistent with the hypothesis that integrations within exons are under rapid evolution, a hallmark of domestication (Frank and Feschotte, 2017).

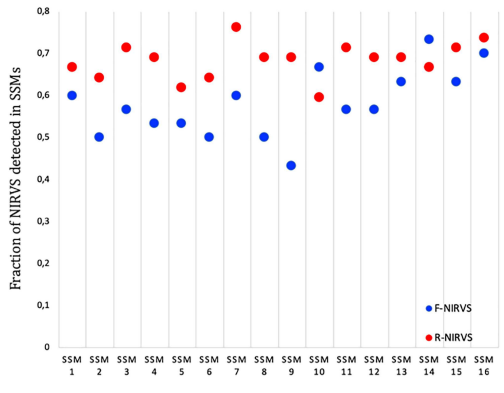


FIGURE 3 | Prevalence of R-NIRVS (red) and F-NIRVS (blue) in each SSMs. Prevalence was calculated in each sample as the ration between detected NIRVS and annotated NIRVS for both R- and F-NIRVS.

Additionally, we tested the genealogy of R-NIRVS and F-NIRVS in comparison to circulating *Rhabdoviruses* and *Flaviviruses*. Relative timetrees were generated for (i) F-NIRVS and corresponding NS3 and NS5 viral proteins from representative *Flaviviruses*, and (ii) R-NIRVS and corresponding L and G proteins of representative *Rhabdoviruses*. Timetrees

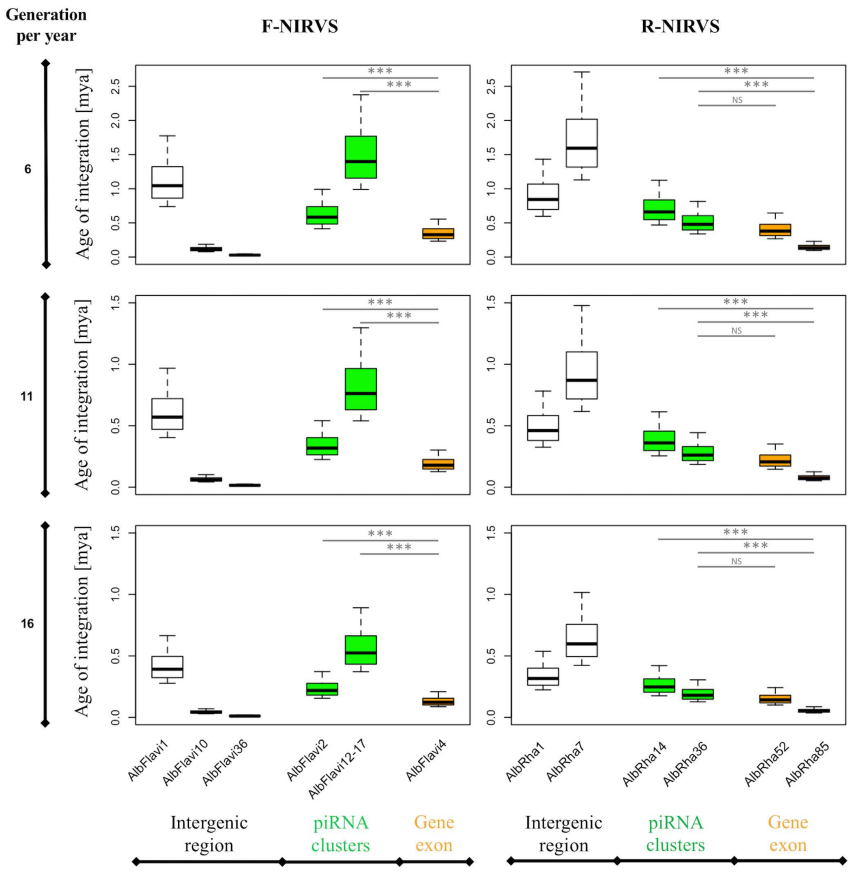


FIGURE 4 | NIRVS integration times. Boxplots showing the integration times for the NIRVS whose variability was studied across five geographic populations. Estimates are based on the *D. melanogaster* mutation rate, i.e., 3.5–8.4 $\times$ 10^{-9} per site per generation (Haag-Liautard et al., 2007; Keightley et al., 2009), and a range of generations per year between 6 and 16 to include mosquitoes from temperate and tropical regions (Manni et al., 2017). NIRVS of piRNA clusters are statistically older than NIRVS mapping in gene exons (ANOVA, ***$P < 0.001$).

showed shorter divergence times between F-NIRVS and viral proteins than R-NIRVS and viral proteins. This clearly indicated multiple integration events and a tendency of R-NIRVS to be older integrations (**Supplementary Figure 2**).

NIRVS Are Heterogeneously Polymorphic at the Sequence Level, With the Majority Being Less Variable Than Slow-Evolving Genes

We selected genes having low and high evolutionary rates in *Ae. albopictus* and we compared their levels of polymorphism (LoP) with that of NIRVS in WGS data from our 16 SSMs. LoP was evaluated as the ratio between the number of total mutations (SNPs and INDELs) found in the locus and its length. We expanded the analyses to include also R-Gs, for which intraspecific rapid evolution has been observed in *Ae. aegypti* (Bernhardt et al., 2012), and the 13 N-Gs in their coding sequence or UTRs (Palatini et al., 2017).

Estimates of gene evolutionary rates were derived from comparisons of levels of protein sequence divergence within groups of orthologous genes across 27 insect species of the Nematocera sub-order (**Supplementary Table 1**). The first and last 0.1% of the genes from the evolutionary rate distribution were selected as slow and fast evolving genes, respectively, and their single-copy orthology status with respect to *Ae. albopictus* was verified. We did not select genes based on their length because of the wide length range of NIRVS, which includes partial viral ORFs of between 151 and 3206 bp (Palatini et al., 2017). SGs that met the above

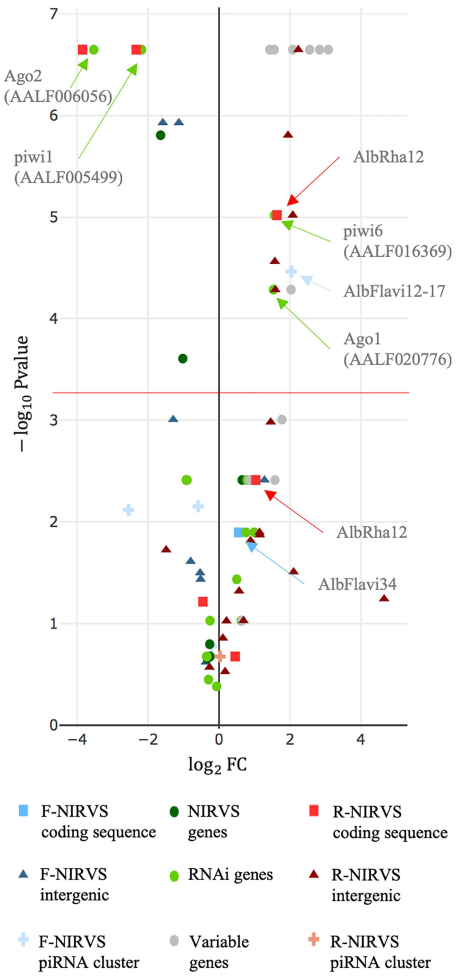


FIGURE 5 | NIRVS polymorphism. Volcano plot showing LoP comparison between SGs and NIRVS, genes encompassing NIRVS (Palatini et al., 2017), genes of the RNAi pathway and FGs. Entities with LoPs statistically different than that of conserved genes are above the red line [adjusted significance of the test (–log10 0.05/99 = 3.32)]. Entities on the left side of the panel (log2FC < 0) have smaller LoP than conserved genes. The opposite for entities on right side of the panel (log2FC > 0).

criteria include genes with hypothetical protein transporter or vesicle-mediated transport activity (*i.e.*, AALF003606, AALF014156, AALF014287; AALF014448; AALF004102), structural activity (AALF005886, annotated as tubulin alpha chain), signal transducer activity (AALF026109), protein and DNA binding activity (AALF027761, AALF028431), SUMO transferase activity (AALF020750), the homothorax homeobox encoding gene AALF019476, the tropomyosin invertebrate gene (AALF0082224), the Protein yippee-like (AALF018378) and autophagy (AALF018476). FGs include genes with unknown functions (AALF004733, AALF009493, AALF009839, AALF012271, AALF026991, AALF014993, AALF017064, AALF018679), proteolysis functions (AALF010748) a gene associated with transcriptional (AALF022019), DNA-binding (AALF019413, AALF024551), structural (AALF028390) and proteolytic (AALF010877) activities. Median LoP of SGs within mosquitoes of the Foshan strain is 0.0071, a value higher than that observed across 63.3% of the detected NIRVS (**Supplementary Figure 3**). Eleven out of fourteen FGs were more variable than SGs, with seven appearing also statistically more polymorphic than SGs (Kolmogorov-Smirnov test, $^*P < 0.05$) (**Figure 5** and **Supplementary Table 3**). This result further supports our selection of SGs and FGs.

Genes of the RNAi pathway are heterogeneously polymorphic (**Figure 5**), with *Ago1* (AALF020776) and *piwi6* (AALF016369) being statistically more polymorphic than SGs; the opposite result was obtained for *piwi1* and *3* (AALF005499, AALF005498), and *Ago2* (AALF006056) (**Figure 5**). LoP of NIRVS is heterogeneous both among SSMs and. NIRVS identified within piRNA clusters (Liu et al., 2016) are all less polymorphic than SGs, with the exception of AlbFlavi12-17 that has a median LoP value of 0.0258. This large LoP may be due by the fact that AlbFlavi12-17 is composed of four small viral sequences nested one next to the other (Palatini et al., 2017). Unlike NIRVS from piRNA clusters, NIRVS spanning gene exons are more heterogeneous; three (i.e., AlbFlavi34, AlbRha12, and AlbRha52) have LoP values higher than those of SGs, while others (*i.e.*, AlbFlavi24, AlbRha28, AlbRha85) are less polymorphic than SGs. AlbFlavi24, AlbFlavi34, AlbRha12, and AlbRha28 are annotated as the only exons of AALF023281, AALF005432, AALF025780, AALF000478, respectively.

NIRVS Identified Within Coding Sequences Are Expressed

The observed LoP for AALF020122 with AlbRha52, AALF025780 with AlbRha12 and AALF005432 with AlbFlavi34 is analogous to that of rapidly evolving genes, suggesting co-option for immunity functions (Frank and Feschotte, 2017). Because domestication of exogenous sequences is a multi-step process, including persistence, immobilization and stable expression of the newly acquired sequences besides rapid evolution (Joly-Lopez and Bureau, 2018), we analyzed the distribution and expression pattern of these genes. Expression analyses were extended to all other N-Gs (AALF025779 with a unique exon containing AlbRha9, AALF000476 with a unique exon corresponding to AlbRha15, AALF000477, and AALF004130 in which the unique exons are contained within AlbRha18 and AlbRha85, respectively) that are fixed within the Foshan strain, but have LoP levels comparable to or lower than those of conserved genes (**Figure 5**). AlbFlavi34 had been previously studied and showed to be expressed in pupae and adult males more than in larvae (Palatini et al., 2017). Genes with NIRVS (N-Gs) form

TABLE 1 | Characteristics of genes with NIRVS in their coding sequence.

Gene ID	NIRVS	Viral ORF	PfamID	Median LoP
AALF000476[a]	AlbRha15	*Rhabdovirus* nucleocapsid protein	PF00945	0.0086
AALF000477[a]	AlbRha18	*Rhabdovirus* nucleocapsid protein	PF00945	0.0052
AALF000478[a,c]	AlbRha28	*Rhabdovirus* nucleocapsid protein	PF00945	0.0004
AALF025780[a]	AlbRha12	*Rhabdovirus* nucleocapsid protein	PF00945	0.0129
AALF025779[a]	AlbRha9	*Rhabdovirus* nucleocapsid protein	PF00945	0.0031
AALF004130[b]	AlbRha85	*Rhabdovirus* RNA dependent RNA polymerase	PF00946	0.0020
AALF020122[b]	AlbRha52	*Rhabdovirus* RNA dependent RNA polymerase	PF00946	0.0196
AALF005432	AlbFlavi34	*Flavivirus* NS2A, NS2B, NS3	PF00949, PF00271, PF07652	0.0099

[a]*Paralogous genes;* [b]*paralogous genes;* [c]*no expression data.*

two groups of paralogs, with similarity to the *Rhabdovirus* RNA-dependent RNA polymerase (RdRPs) and the nucleocapsid-encoding gene, respectively (**Table 1**). As shown in **Figure 6**, apart from AALF00477, all other genes are expressed throughout *Ae. albopictus* development with a similar profile, but at different levels. None of the genes showed sex-biased expression or tissue-specific expression in the ovaries; on the contrary highest expression was observed in sugar- and blood-fed females.

Additional NIRVS Variants Are Found in the Genome of Foshan Mosquitoes

We verified the presence of novel NIRVS alleles by investigating soft-clipped reads. Soft-clipped reads support the contiguity of AlbFlavi6 and AlbFlavi7, that were annotated in separated regions of the same contig (**Figure 7A**). This newly resolved arrangement revealed the existence of a viral ORF of 1191 bp, corresponding to a partial non-structural protein 5 (NS5) of *Flaviviruses*. Additionally, soft-clipped reads supported longer than annotated alleles in AlbFlavi10, AlbFlavi2 and AlbRha4 (**Figure 7B**). We further looked for the presence of novel viral integrations using Vy-PER (Forster et al., 2015). No viral integrations different than the ones identified *in silico* from the Foshan genome (Palatini et al., 2017) were found in the 16 SSMs.

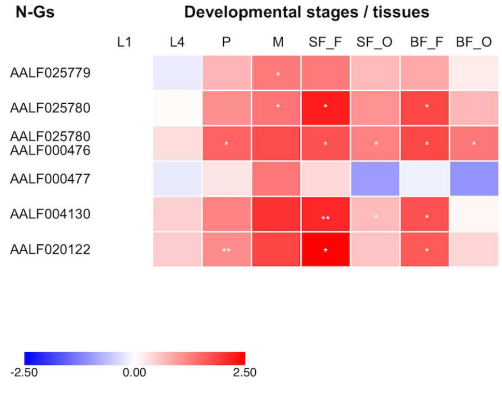


FIGURE 6 | Expression of NIRVS mapping in coding sequences. Heatmap of the expression profiles of genes with NIRVS in the coding sequence (N-Gs) across developmental stages and body tissues (L1-L4: 1st-4th instar larvae; P: pupae; M: male whole body; SF_F/BF_F: sugar/blood fed female whole body; SF_O/BF_O: ovaries from sugar/blood fed females). Color key expresses the log10-fold change relative to larva 1st instar (calibrator). Asterisks indicate significant differences in transcript abundances (Unpaired two-tailed *t*-tests, $*P < 0.05$, $**P < 0.01$).

DISCUSSION

Repetitive DNA is a major source of genome variability and there are also examples of repetitive sequences being co-opted for cellular functions (Gilbert and Feschotte, 2018). Besides having an impact on the evolution, organization and behavior of eukaryotic genomes, variations in repeat sequences or their copy numbers have been exploited for taxonomic and phylogenetic studies (Djadid et al., 2006; Wang-Sattler et al., 2007; Lammers et al., 2017). Therefore, knowledge of the repeatome assists in understanding the plasticity of eukaryotic genomes and may provide markers for population genetic studies (Goubert et al., 2016, 2017). Repetitive DNA represents 68% of the genome of *Ae. albopictus* and include dozens of still-poorly characterized sequences from non-retroviral RNA viruses (Chen et al., 2015; Palatini et al., 2017). Here we studied the widespread occurrence of NIRVS in relation to the geographic distribution of the species and the variability of NIRVS in comparison to that of mosquito genes showing slow and high evolutionary rates. We clearly show that the landscape of viral integrations is variable within and across geographic populations, with a core set of seemingly the oldest integrations from *Rhabdoviruses*. Additionally, the polymorphism of viral integrations is heterogeneous and depends primarily on their location within the genome. Overall, results of this study emphasize the complexity of the composition and structure of the mosquito repeatome and provide an objective strategy to identify viral integrations that most probably affect mosquito biology.

Biological Significance of NIRVS Variable Genomic Landscape and Their Polymorphism in SSMs

The landscape of viral integrations is variable among SSMs, longer than annotated alleles are identified through the analyses

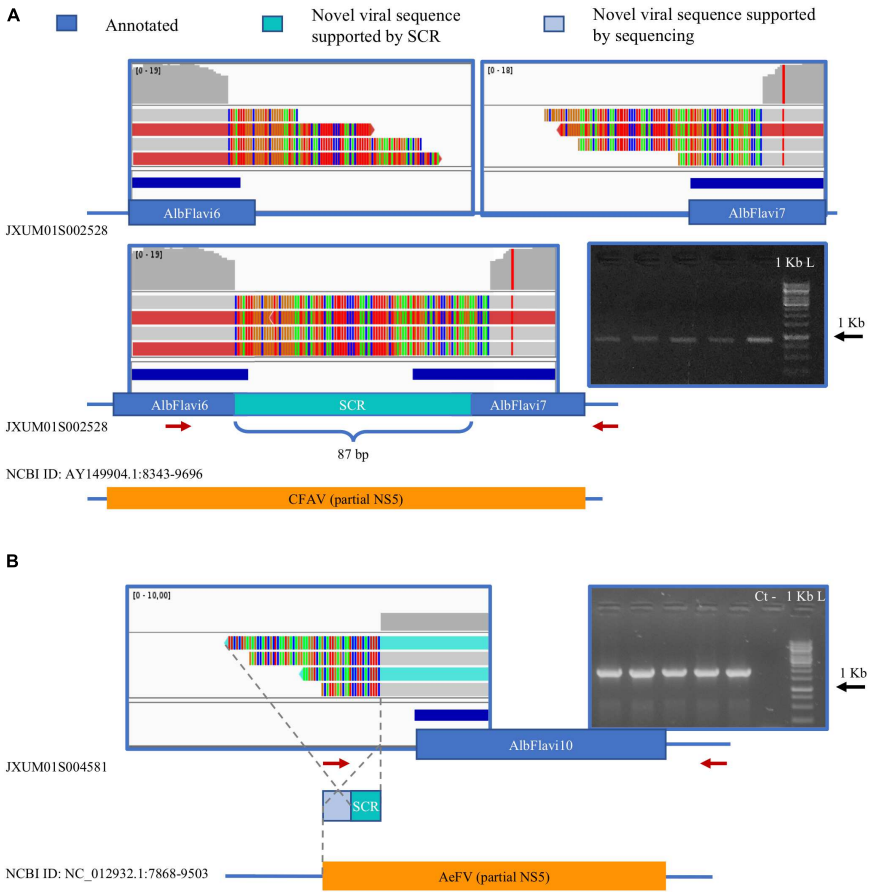


FIGURE 7 | Soft clipped reads (SCR) support novel arrangements and longer than annotated viral integrations. **(A)** AlbFlavi6 and AlbFlavi7 are annotated on the same contig, but 5000 bp apart from each other. Soft-clipped reads (in light blue) and PCR experiments support their contiguity, with a unique ORF with similarity to the *Flavivirus* NS5. **(B)** A sequence of 212 bp extending AlbFlavi10 on its left side was identified investigating SCRs and confirmed by further sequencing. Red arrows indicate primer positions.

of soft-clipped reads, but no additional viral sequences, different than the ones characterized from the Foshan-based genome assembly, are found in SSMs. NIRVS are considered rare events following viral infections. In *Aedes* spp. mosquitoes and mosquito cells short segments of cDNA of viral origin (vDNA) are synthetized upon infection with arboviruses of different genera (*i.e., Flavivirus*, *Alphavirus* and *Bunyavirus*) by the reverse transcriptase of endogenous retrotransposons (Goic et al., 2016; Nag and Kramer, 2017). These vDNAs are composed of fragmented viral sequences, from different regions of the viral genome, next to sequences of TEss (Goic et al., 2016; Nag and Kramer, 2017). Because the composition of vDNAs is analogous to that of NIRVS, vDNAs have been proposed to be the substrate for NIRVS (Olson and Bonizzoni, 2017; Palatini et al., 2017). The SSMs analyzed in this study are from the Foshan strain. The Foshan strain derives from mosquitoes collected in the Chinese city of Foshan in the early '1980 and have since been kept in laboratory settings with no viral exposure (Chen et al., 2015). Under this scenario, the absence of novel viral integrations in SSMs is not unexpected. However, the identification of a variable landscape among SSMs with a core set of NIRVS, which is enriched for integrations with similarity to *Rhabdoviruses* and NIRVS mapping in coding sequences, is significant because it demonstrates that viral integrations are a dynamic component of the repeatome and not all viral integrations are dispensable genomic elements. Interestingly, when compared to fast- and slow-evolving mosquito genes, NIRVS polymorphism was not homogeneous. NIRVS identified

within piRNA clusters were less polymorphic than SGs. Selection constraints on sequences within piRNA clusters have been previously identified in both flies and mice (Chirn et al., 2015). This is despite piRNAs have an incredible sequence diversity and their biogenesis and processing do not appear to be linked to common sequences or structural motifs (Huang et al., 2017). In *D. melanogaster*, piRNA clusters are dynamic loci and their composition has been linked to their regulatory abilities. For instance, the ability of the *D. melanogaster* master piRNA locus *flamenco* to control transposons such as *gypsy*, *ZAM* and *Idefix* was shown to be dependent on frequent chromosomal rearrangements, loss or gain of fragmented TE sequences (Zanni et al., 2013; Guida et al., 2016). Additionally, variations in the composition of subtelomeric piRNA clusters were observed upon adaptation to laboratory conditions of *D. melanogaster* wild collected flies (Asif-Laidin et al., 2017). Importantly, structural differences in subtelomeric piRNA clusters did not impair host genome integrity and occurred with the maintenance of conserved groups of sequences, which could be alternatively distributed among different strains (Asif-Laidin et al., 2017). Data on the geographic distribution of NIRVS mapping in piRNA clusters studied here (i.e., AlbFlavi2, AlbFlavi4, AlbFlavi12-17, AlbRha14, and AlbRha36) show a situation analogous to that identified with TE fragments of the *flamenco* locus in *D. melanogaster*. On this basis, it is tempting to propose that the analogy between *D. melanogaster* and *Ae. albopictus* in the dynamic composition of piRNA clusters extends to their function so that the pattern of viral integrations within piRNA clusters influence mosquito susceptibly to viral infection. If proven, this hypothesis may help explain the observed variability in vector competence across mosquito populations and could be adapted into novel genetic-based strategies of vector control.

Among NIRVS encompassing gene exons, three appeared more variable than FGs and are also expressed; two of these (AlbRha52 and AlbRha12) are also persistent suggesting exaptation (Joly-Lopez and Bureau, 2018). AlbRha52 and AlbRha12 have similarity to the RdRPs and nucleocapsid-encoding genes of *Rhabdovirus*, respectively. RdRPs are ancient enzymes, essential for RNA viruses (de Farias et al., 2017). While the existence of RdRP genes in insects is still debated, cellular RdRP activity has been observed in plants, fungi and *Caenorhabditis elegans* in association with RNA silencing functions (Zong et al., 2009; de Farias et al., 2017; Pinzon et al., 2018). An RdRP of viral origin was recently described in a bat species of the *Eptesicus* clade (Horie et al., 2016) and exaptation of a viral nucleocapsid gene was shown in Afrotherians (Kobayashi et al., 2016). On this basis, further experiments to characterize the functions of the *Ae. albopictus* genes AALF020122 and AALF025780 are on-going.

Biological Significance of NIRVS Variable Genomic Landscape in Geographic Populations

To start gaining insights into the natural widespread occurrence of NIRVS, a set of 13 viral integrations representative of both R- and F-NIRVS and mapping within piRNA clusters, intergenic regions and gene exons were selected and both their occurrence and their sequence polymorphism was analyzed in mosquitoes from five geographic populations. Populations were selected following the invasion history of *Ae. albopictus* out of its native home range in south East Asia and included samples from China, Thailand, La Reunion island and newly colonized areas such as Italy and United States. Distributions of NIRVS in these populations was consistent with results from SSMs as R-NIRVS were more frequently detected than F-NIRVS. Additionally, R-NIRVS appeared on overage older integrations than F-NIRVS.

The difference in the number and age of the integration events among sequences from *Rhabdoviruses* and *Flaviviruses* is intriguing because Mononegavirales, including *Rhabdoviruses*, are considered evolutionary more recent than *Flaviviridae* (Koonin et al., 2015). The *Rhabdovirus* genus contains viruses that are extremely variable in both their genomic organization and host preferences, with viruses infecting vertebrates, invertebrates and plants (Dietzgen et al., 2017; Geoghegan et al., 2017). Additionally, *Rhabdoviruses* have been shown to frequently transfer horizontally among host species based on their ecological and geographic proximity (Geoghegan et al., 2017). Thus, the ecological diversity and the wide geographic distribution range of *Rhabdoviruses* may favor their integrations into mosquito genomes. Alternatively, the promiscuous nature of *Rhabdoviruses* with frequent horizontal transfers could select for the emergence of generalist protection mechanisms, of which integrations could be part of.

The variable landscape of NIRVS across geographic populations should be interpreted with caution. The rapid global invasion of *Ae. albopictus* from South-East Asia, which happened over the past 50–60 years, was human-mediated and occurred through the movement of propagules (Manni et al., 2017), creating a situation of genetic admixture. Mosquito populations from newly invaded areas, such as Italy and United States, lack isolation by distance and appear genetically mixed (Kotsakiozi et al., 2017; Manni et al., 2017; Maynard et al., 2017). The occurrence of frequent bottlenecks followed by interbreeding can partly explain the variable NIRVS landscape observed here. However, the enrichment for R-NIRVS, the variable distribution of NIRVS within piRNA clusters and their heterogenous polymorphism indicate that evolutionary forces other than genetic drift and gene flow have played a role in the distribution of NIRVS and suggests a multifaceted impact of NIRVS on mosquito physiology.

DATA AVAILABILITY STATEMENT

Whole Genome Sequencing data alignments have been deposited to the SRA archive under accession number from SAMN09759672 to SAMN09759687.

AUTHOR CONTRIBUTIONS

MB and EP conceived and designed the study, analyzed the data, and drafted the manuscript. EP and RW contributed to

bioinformatic analyses, analyzed the results, and revised the manuscript. FS, FV, PC, and RC-L collected and analyzed molecular data and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was funded by a European Research Council Consolidator Grant (ERC-CoG) under the European Union's Horizon 2020 Program (Grant Number ERC-CoG 682394) to MB, by the Italian Ministry of Education, University and Research FARE-MIUR project R1623HZAH5 to MB, by the Italian Ministry of Education, University and Research (MIUR): Dipartimenti di Eccellenza Program (2018–2022) Department of Biology and Biotechnology "L. Spallanzani," University of Pavia, and by the Swiss National Science Foundation grant PP00P3_170664 to RW.

ACKNOWLEDGMENTS

The authors thank Ruth Monica Waghchoure for mosquito maintenance and Lino Ometto for fruitful discussions. A previous version of the study was published as a preprint here: https://www.biorxiv.org/content/early/2018/08/06/385666.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00093/full#supplementary-material

REFERENCES

Arensburger, P., Hice, R. H., Wright, J. A., Craig, N. L., and Atkinson, P. W. (2011). The mosquito *Aedes aegypti*has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics* 12:606. doi: 10.1186/1471-2164-12-606

Asif-Laidin, A., Delmarre, V., Laurentie, J., Miller, W. J., Ronsseray, S., and Teysset, L. (2017). Short and long-term evolutionary dynamics of subtelomeric piRNA clusters in *Drosophila*. *DNA Res.* 24, 459–472. doi: 10.1093/dnares/dsx017

Baruffi, L., Damiani, G., Guglielmino, C. R., Bandi, C., Malacrida, A. R., and Gasperi, G. (1995). Polymorphism within and between populations of Ceratitis: comparison between RAPD and multilocus enzyme electrophoresis data. *Heredity* 74, 425–437. doi: 10.1038/hdy.1995.60

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010). Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate Genomes. *PLoS Pathog.* 6:e1001030. doi: 10.1371/journal.ppat.1001030

Bernhardt, S. A., Simmons, M. P., Olson, K. E., Beaty, B. J., Blair, C. D., and Black, W. C. (2012). Rapid intraspecific evolution of miRNA and siRNA Genes in the mosquito *Aedes aegypti*. *PLoS One* 7:e44198. doi: 10.1371/journal.pone.0044198

Boston, M. (2015). *R Studio*. Available at: https://www.rstudio.com

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., et al. (2007). Discrete small RNA-Generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128, 1089–1103. doi: 10.1016/j.cell.2007.01.043

Bubner, B., and Baldwin, I. T. (2004). Use of real-time PCR for determining copy number and zygosity in transgenic plants. *Plant Cell Rep.* 23, 263–271. doi: 10.1007/s00299-004-0859-y

Carrillo-Tripp, M., Shepherd, C. M., Borelli, I. A., Venkataraman, S., Lander, G., Natarajan, P., et al. (2009). VIPERdb 2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res.* 37, 436–442. doi: 10.1093/nar/gkn840

Chen, X.-G., Jiang, X., Gu, J., Xu, M., Wu, Y., Deng, Y., et al. (2015). Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5907–E5915. doi: 10.1073/pnas.1516410112

Chirn, G. W., Rahman, R., Sytnikova, Y. A., Matts, J. A., Zeng, M., Gerlach, D., et al. (2015). Conserved piRNA expression from a distinct set of piRNA cluster loci in Eutherian mammals. *PLoS Genet.* 11:e1005652. doi: 10.1371/journal.pgen.1005652

Crochu, S., Cook, S., Attoui, H., Charrel, R. N., De Chesse, R., Belhouchet, M., et al. (2004). Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. *mosquitoes. J. Gen. Virol.* 85, 1971–1980. doi: 10.1099/vir.0.79850-0

de Farias, S. T., dos Santos, A. P. Jr., Rêgo, T. G., and José, M. V. (2017). Origin and evolution of RNA-dependent RNA Polymerase. *Front. Genet.* 8:125. doi: 10.3389/fgene.2017.00125

de Lamballerie, X., Crochu, S., Billoir, F., Neyts, J., de Micco, P., Holmes, E. C., et al. (2002). Genome sequence analysis of Tamana bat virus and its relationship with the genus *Flavivirus*. *J. Gen. Virol.* 83, 2443–2454. doi: 10.1099/0022-1317-83-10-2443

Dietzgen, R. G., Kondob, H., Goodinc, M. M., Kurathd, G., and Vasilakise, N. (2017). The family *Rhabdoviridae*: mono - and bipartite negative-sense RNA viruses with diverse genome organization and common evolutionary origins. *Virus Res.* 227, 158–170. doi: 10.1016/j.virusres.2016.10.010.

Djadid, N. D., Gholizadeh, S., Aghajari, M., Zehi, A. H., Raeisi, A., and Zakeri, S. (2006). Genetic analysis of rDNA-ITS2 and RAPD loci in field populations of the malaria vector, Anopheles stephensi (Diptera: Culicidae): implications for the control program in Iran. *Acta Trop.* 97, 65–74. doi: 10.1016/j.actatropica.2005.08.003

Edgar, R. C., Drive, R. M., and Valley, M. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Forster, M., Szymczak, S., Ellinghaus, D., Hemmrich, G., Rühlemann, M., Kraemer, L., et al. (2015). Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci. Rep.* 5:11534. doi: 10.1038/srep11534

Fort, P., Albertini, A., Van-Hua, A., Berthomieu, A., Roche, S., Delsuc, F., et al. (2012). Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. *Mol. Biol. Evol.* 29, 381–390. doi: 10.1093/molbev/msr226

Frank, J. A., and Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. *Curr. Opin. Virol.* 25, 81–89. doi: 10.1016/j.coviro.2017.07.021

Gainetdinov, I., Skvortsova, Y., Kondratieva, S., Funikov, S., and Azhikina, T. (2017). Two modes of targeting transposable elements by piRNA pathway in human testis. *RNA* 23, 1614–1625. doi: 10.1261/rna.060939.117

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*:1207.3907 [Preprint].

Geoghegan, J. L., Duchêne, S., and Holmes, E. C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* 13:e1006215. doi: 10.1371/journal.ppat.1006215

Gilbert, C., and Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Curr. Opin. Genet. Dev.* 49, 15–24. doi: 10.1016/j.gde.2018.02.007

Goic, B., Stapleford, K. A., Frangeul, L., Doucet, A. J., Gausson, V., Blanc, H., et al. (2016). Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nat. Commun.* 7:12410. doi: 10.1038/ncomms12410

Goubert, C., Henri, H., Minard, G., Valiente Moro, C., Mavingui, P., Vieira, C., et al. (2017). High-throughput sequencing of transposable element insertions suggests adaptive evolution of the invasive Asian tiger mosquito towards temperate environments. *Mol. Ecol.* 26, 3968–3981. doi: 10.1111/mec.14184

Goubert, C., Minard, G., Vieira, C., and Boulesteix, M. (2016). Population genetics of the Asian tiger mosquito Aedes albopictus, an invasive vector of human diseases. *Heredity* 117, 125–134. doi: 10.1038/hdy.2016.35

Guida, V., Cernilogar, F. M., Filograna, A., De Gregorio, R., Ishizu, H., Siomi, M. C., et al. (2016). Production of small noncoding RNAs from the *flamenco*locus is regulated by the *gypsy*retrotransposon of *Drosophila melanogaster*. *Genetics* 204, 631–644. doi: 10.1534/genetics.116.187922

Guzzardo, P. M., Muerdter, F., and Hannon, G. J. (2013). The piRNA pathway in flies: highlights and future directions. *Curr. Opin. Genet. Dev.* 23, 44–52. doi: 10.1016/j.gde.2012.12.003

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B., et al. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* 445, 82–85. doi: 10.1038/nature05388

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* 8:R19. doi: 10.1186/gb-2007-8-2-r19

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe* 10, 368–377. doi: 10.1016/j.chom.2011.09.002

Horie, M., Kobayashi, Y., Honda, T., Fujino, K., Akasaka, T., Kohl, C., et al. (2016). An RNA-dependent RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus. *Sci. Rep.* 6:25873. doi: 10.1038/srep25873

Huang, X., Fejes Tóth, K., and Aravin, A. A. (2017). piRNA biogenesis in *Drosophila melanogaster*. *Trends Genet.* 33, 882–894. doi: 10.1016/j.tig.2017.09.002

Joly-Lopez, Z., and Bureau, T. E. (2018). Exaptation of transposable element coding sequences. *Curr. Opin. Genet. Dev.* 49, 34–42. doi: 10.1016/j.gde.2018.02.011

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matricies from protein sequences. *Comput. Appl. Biosci.* 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Katzourakis, A., and Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet.* 6:e1001191. doi: 10.1371/journal.pgen.1001191

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster*spontaneous mutation accumulation lines. *Genome Res.* 19, 1195–1201. doi: 10.1101/gr.091231.109

Kobayashi, Y., Horie, M., Nakano, A., Murata, K., Itou, T., and Suzuki, Y. (2016). Exaptation of bornavirus-Like nucleoprotein elements in Afrotherians. *PLoS Pathog.* 12:e1005785. doi: 10.1371/journal.ppat.1005785

Koonin, E. V., Dolja, V. V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 47, 2–25. doi: 10.1016/j.virol.2015.02.039

Kotsakiozi, P., Richardson, J. B., Pichler, V., Favia, G., Martins, A. J., Urbanelli, S., et al. (2017). Population genomics of the Asian tiger mosquito, *Aedes albopictus*: insights into the recent worldwide invasion. *Ecol. Evol.* 7, 10143–10157. doi: 10.1002/ece3.3514

Kryukov, K., Ueda, M. T., Imanishi, T., and Nakagawa, S. (2018). Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. *Virus Res.* [Epub ahead of print] doi: 10.1016/j.virusres.2018.02.002,

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., Mcewen, R., et al. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 44:e108. doi: 10.1093/nar/gkw227

Lammers, F., Gallus, S., Janke, A., and Nilsson, M. A. (2017). Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes. *Genome Biol. Evol.* 9, 2862–2878. doi: 10.1093/gbe/evx170

Langella, O. (1999). *Population 1.2.31*. Available at: http://bioinformatics.org/populations/

Le, S. Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307–1320. doi: 10.1093/molbev/msn067

Liu, P., Dong, Y., Gu, J., Puthiyakunnon, S., Wu, Y., and Chen, X. G. (2016). Developmental piRNA profiles of the invasive vector mosquito *Aedes albopictus*. *Parasit. Vectors* 9:524. doi: 10.1186/s13071-016-1815-8

Manni, M., Guglielmino, C. R., Scolari, F., Vega-Rúa, A., Failloux, A. B., Somboon, P., et al. (2017). Genetic evidence for a worldwide chaotic dispersion pattern of the arbovirus vector, *Aedes albopictus*. *PLoS Negl. Trop. Dis.* 11:e0005332. doi: 10.1371/journal.pntd.0005332

Maumus, F., and Quesneville, H. (2014). Deep investigation of *Arabidopsis thaliana*junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS One* 9:e94101. doi: 10.1371/journal.pone.0094101

Maumus, F., and Quesneville, H. (2016). Impact and insights from ancient repetitive elements in plant genomes. *Curr. Opin. Plant Biol.* 30, 41–46. doi: 10.1016/j.pbi.2016.01.003

Maynard, A. J., Ambrose, L., Cooper, R. D., Chow, W. K., Davis, J. B., Muzari, M. O., et al. (2017). Tiger on the prowl: invasion history and spatio-temporal genetic structure of the Asian tiger mosquito *Aedes albopictus*(Skuse 1894) in the Indo-Pacific. *PLoS Negl. Trop. Dis.* 11:e0005546. doi: 10.1371/journal.pntd.0005546

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit?: a MapReduce framework for analyzing next-generation DNA sequencing data the genome analysis toolkit?: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Miesen, P., Joosten, J., and van Rij, R. P. (2016). PIWIs go viral: arbovirus-derived piRNAs in vector mosquitoes. *PLoS Pathog.* 12:e1006017. doi: 10.1371/journal.ppat.1006017

Nag, D. K., and Kramer, L. D. (2017). Patchy DNA forms of the Zika virus RNA genome are generated following infection in mosquito cell cultures and in mosquitoes. *J. Gen. Virol.* 98, 2731–2737. doi: 10.1099/jgv.0.000945

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., et al. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Olson, K. E., and Bonizzoni, M. (2017). Nonretroviral integrated RNA viruses in arthropod vectors: an occasional event or something more? *Curr. Opin. Insect Sci.* 22, 45–53. doi: 10.1016/j.cois.2017.05.010

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., et al. (2017). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti*and *Aedes albopictus*. *BMC Genomics* 18:512. doi: 10.1186/s12864-017-3903-3

Petit, M., Mongelli, V., Frangeul, L., Blanc, H., Jiggins, F., and Saleh, M.-C. (2016). piRNA pathway is not required for antiviral defense in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 113, E4218–E4227. doi: 10.1073/pnas.1607952113

Pfaffl, M. W. (2006). "Relative quantification' in real-time PCR," in *"Relative Quantification" in Real-Time PCR*, ed. T. Dorak (La Jolla, CA: International University Line), 63–82.

Pinzon, N., Bertrand, S., Subirana, L., Busseau, I., Escriva, H., and Seitz, H. (2018). Functional lability of RNA-dependent RNA polymerases in animals. *bioRXiv* [Preprint]. doi: 10.1101/339820

Reynolds, J. A., Poelchau, M. F., Rahman, Z., Armbruster, P. A., and Denlinger, D. L. (2012). Transcript profiling reveals mechanisms for lipid conservation during diapause in the mosquito, *Aedes albopictus*. *J. Insect Physiol.* 58, 966–973. doi: 10.1016/j.jinsphys.2012.04.013

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., and Twigg, S. R. F. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918. doi: 10.1038/ng.3036

Rozen, S., and Skaletksy, H. (2000). Primer3 on the WWW for general uses and for biologist programmers. *Methods Mol. Biol.* 132, 365–386.

RStudio Team (2015). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Tamura, K., Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A., and Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19333–19338. doi: 10.1073/pnas.1213199109

Tóth, K. F., Pezic, D., Stuwe, E., and Webster, A. (2016). The piRNA pathway guards the germline genome against transposable elements. *Adv. Exp. Med. Biol.* 886, 51–77. doi: 10.1007/978-94-017-7417-8_4

Wang, Y., Jin, B., Liu, P., Li, J., Chen, X., and Gu, J. (2018). PiRNA profiling of dengue virus type 2-infected Asian tiger mosquito and midgut tissues. *Viruses* 10:E213. doi: 10.3390/v10040213

Wang-Sattler, R., Blandin, S., Ning, Y., Blass, C., Dolo, G., Touré, Y. T., et al. (2007). Mosaic genome architecture of the anopheles gambiae species complex. *PLoS One* 2:e1249. doi: 10.1371/journal.pone.0001249

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691–699. doi: 10.1093/oxfordjournals.molbev.a003851

Whitfield, Z. J., Dolan, P. T., Kunitomi, M., Tassetto, M., Seetin, M. G., Oh, S., et al. (2017). The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti*genome. *Curr. Biol.* 27, 3511–3519. doi: 10.1016/j.cub.2017.09.067

Yamada, K. D., Tomii, K., and Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees. *Bioinformatics* 32, 3246–3251. doi: 10.1093/bioinformatics/btw412

Yuan, J. S., Burris, J., Stewart, N. R., Mentewab, A., and Stewart, C. N. (2007). Statistical tools for transgene copy number estimation based on real-time PCR. *BMC Bioinformatics* 8:S6. doi: 10.1186/1471-2105-8-S7-S6

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., and Luyten, I. (2013). Retrotransposons at the *flamenco*locus reflect the regulatory properties of piRNA clusters. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19842–19847. doi: 10.1073/pnas.1313677110

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simao, F. A., Ioannidis, P., et al. (2016). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45, D744–D749. doi: 10.1093/nar/gkw1119

Zong, J., Yao, X., Yin, J., Zhang, D., and Ma, H. (2009). Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* 447, 29–39. doi: 10.1016/j.gene.2009.07.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PLOS | NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

Polymorphism analyses and protein modelling inform on functional specialization of *Piwi* clade genes in the arboviral vector *Aedes albopictus*

Michele Marconcini[1], Luis Hernandez[1], Giuseppe Iovino[1], Vincent Houé[2], Federica Valerio[1], Umberto Palatini[1], Elisa Pischedda[1], Jacob E. Crawford[3], Bradley J. White[3], Teresa Lin[3], Rebeca Carballar-Lejarazu[1¤], Lino Ometto[1], Federico Forneris[1], Anna-Bella Failloux[2], Mariangela Bonizzoni[1]*

1 Department of Biology and Biotechnology, University of Pavia, Pavia, Italy, **2** Arbovirus and Insect Vectors Units, Department of Virology, Institut Pasteur, Paris, France, **3** Verily Life Sciences, South San Francisco, California, United States of America

¤ Current address: Department of Molecular Biology and Biochemistry, University of California, Irvine, California, United States of America
* mariangela.bonizzoni@unipv.it

OPEN ACCESS

Citation: Marconcini M, Hernandez L, Iovino G, Houé V, Valerio F, Palatini U, et al. (2019) Polymorphism analyses and protein modelling inform on functional specialization of *Piwi* clade genes in the arboviral vector *Aedes albopictus*. PLoS Negl Trop Dis 13(12): e0007919. https://doi.org/10.1371/journal.pntd.0007919

Editor: Fabiano Oliveira, National Institutes of Health, UNITED STATES

Received: July 1, 2019

Accepted: November 11, 2019

Published: December 2, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files

Funding: This research was funded by a European Research Council Consolidator Grant (ERCCoG) under the European Union's Horizon 2020 Programme (Grant Number ERCCoG 682394) to M.B.; by the Italian Ministry of Education, University and Research FARE-MIUR project R1623HZAH5 to M.B.; by the Italian Ministry of

Abstract

Current knowledge of the piRNA pathway is based mainly on studies on *Drosophila melanogaster* where three proteins of the Piwi subclade of the Argonaute family interact with PIWI-interacting RNAs to silence transposable elements in gonadal tissues. In mosquito species that transmit epidemic arboviruses such as dengue and chikungunya viruses, *Piwi* clade genes underwent expansion, are also expressed in the soma and cross-talk with proteins of recognized antiviral function cannot be excluded for some Piwi proteins. These observations underscore the importance of expanding our knowledge of the piRNA pathway beyond the model organism *D. melanogaster*. Here we focus on the emerging arboviral vector *Aedes albopictus* and we couple traditional approaches of expression and adaptive evolution analyses with most current computational predictions of protein structure to study evolutionary divergence among Piwi clade proteins. Superposition of protein homology models indicate possible high structure similarity among all Piwi proteins, with high levels of amino acid conservation in the inner regions devoted to RNA binding. On the contrary, solvent-exposed surfaces showed low conservation, with several sites under positive selection. Analysis of the expression profiles of *Piwi* transcripts during mosquito development and following infection with dengue serotype 1 or chikungunya viruses showed a concerted elicitation of all *Piwi* transcripts during viral dissemination of dengue viruses while maintenance of infection relied on expression of primarily *Piwi5*. Opposite, establishment of persistent infection by chikungunya virus is accompanied by increased expression of all *Piwi* genes, particularly *Piwi4* and, again, *Piwi5*. Overall these results are consistent with functional specialization and a general antiviral role for *Piwi5*. Experimental evidences of sites under positive selection in *Piwi1/3*, *Piwi4* and *Piwi6*, that have complex expression profiles, provide useful knowledge to design tailored functional experiments.

Education, University and Research (MIUR): Dipartimenti Eccellenza Program (2018–2022) Dept. of Biology and Biotechnology "L. Spallanzani", University of Pavia. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Competing interests: The authors have declared that no competing interests exist. Jacob E. Crawford, Bradley White and Teresa Lin are employed by a commercial company, Verily Life Sciences LLC. They have no competing interests.

Author summary

Argonautes are ancient proteins involved in many cellular processes, including innate immunity. Early in eukaryote evolution, Argonautes separated into Ago and Piwi clades, which maintain a dynamic evolutionary history with frequent duplications and losses. The use *of Drosophila melanogaste*r as a model organism proved fundamental to understand the function of Argonautes. However, recent studies showed that the patterns and observations made in *D. melanogaster*, including the number of Argonautes, their expression profile and their function, are a rarity among Dipterans. In vectors of epidemic arboviruses such as dengue and chikungunya viruses, *Piwi* genes underwent expansion, are expressed in the soma, and some of them appear to have antiviral functions. Besides being an important basic question, the identification of which (and how) *Piwi* genes have antiviral functions may be used for the development of novel genetic-based strategies of vector control. Here we coupled population genetics models with computational predictions of protein structure and expression analyses to investigate the evolution and function of *Piwi* genes of the emerging vector *Aedes albopictus*. Our data support a general antiviral role for *Piwi*5. Instead, the detection of complex expression profiles with the presence of sites under positive selection in *Piwi*1/3, *Piwi*4 and *Piwi*6 requires tailored functional experiments to clarify their antiviral role.

Introduction

First discovered for their role in plant development, proteins of the Argonaute family were found in all domains of life, where they are essential for a wide variety of cellular processes, including innate immunity [1,2].

Recent studies provided evidence of evolutionary expansion and functional divergence of Argonautes in Dipterans, including examples in both the Ago and Piwi subclades [3]. Differences in function and copy number have also been found in other taxa such as nematodes [4], oomycetes [5] and higher plants [6], showing that this protein family is subject to a dynamic evolutionary history. In eukaryotes, Argonautes are key components of RNA interference (RNAi) mechanisms, which can be distinguished in three main pathways: the small interfering RNA (siRNA), microRNA (miRNA) and the PIWI-interacting RNA (piRNA) pathways.

The siRNA pathway is the cornerstone of antiviral defense in insects. The canonical activity of this pathway is the Argonaute 2 (Ago2)-dependent cleavage of viral target sequences. Ago2 is guided to its target through an RNA-induced silencing complex (RISC) loaded with 21-nucleotide (nt)-long siRNAs. siRNAs are produced from viral double-strand RNAs intermediates by the RNAase-III endonuclease activity of Dicer-2 (Dcr2) and define the target based on sequence complementarity [7]. Dcr2 also possesses a DExD/H helicase domain that mediates the synthesis of viral DNA (vDNA) fragments [8]. vDNAs appear to further modulate antiviral immunity [8]. vDNA fragments are synthesized in both circular and linear forms, in complex arrangements with sequences from retrotransposons, but details of their mode of action have not been elucidated yet [8,9]. We and others recently showed that the genomes of *Aedes spp*. mosquitoes harbor fragmented viral sequences, which are integrated next to transposon sequences, are enriched in piRNA clusters and produced PIWI-interacting RNAs (piRNAs) [10,11]. The similar organization between vDNAs and viral integrations, along with the production of piRNAs of viral origin (vpiRNAs) following arboviral infection of *Aedes* spp.

mosquitoes, led to the hypothesis that the piRNA pathway function cooperatively with the siRNA pathway in the acquisition of tolerance to infection [10,12,13].

Current knowledge on the piRNA pathway in insects is based mainly on studies on *Drosophila melanogaster* where three proteins of the Piwi subclade, namely Argonaute-3 (AGO3), PIWI and Aubergine (AUB), interact with piRNAs to silence transposable elements (TEs) in gonadal tissues [14]. Interestingly, the piRNA pathway of *D. melanogaster* does not have antiviral activity and no viral integrations have been detected [15]. Additional differences exist between the piRNA pathway of *D. melanogaster* and that of mosquitoes, suggesting that *D. melanogaster* cannot be used as a model to unravel the molecular cross-talk between the siRNA and piRNA pathways leading to antiviral immunity in *Aedes* spp. mosquitoes. For instance, in *Aedes aegypti*, Piwi subclade has undergone expansion with seven proteins (i.e. Ago3, Piwi2, Piwi3, Piwi4, Piwi5, Piwi6 and Piwi7), which are alternatively expressed in somatic and germline cells and interact with both endogenous and vpiRNAs [12,16,17]. Gonadal- or embryonic-specific expression is found for *Piwi1/3* and *Piwi7*, respectively [16]. On the contrary, *Ago3*, *Piwi4*, *Piwi5* and *Piwi6* are highly expressed in *Ae. aegypti* soma and Aag2 cells and all contribute to the production of transposon-derived piRNAs [16,18]. Ago3 and Piwi5 also regulate biogenesis of piRNAs from the replication-dependent histone gene family [19]. Production of vpiRNAs is dependent on Piwi5 and Ago3 during infection of Aag2 cells with the *Alphavirus* CHIKV, Sindbis and Semliki Forest (SF) viruses, but relies also on Piwi6 following infection with the *Flavivirus* DENV2 [18,20–22]. Piwi4 does not bind piRNAs and its knock-down does not alter vpiRNA production upon infection of Aag2 cells with either SFV or DENV2 [18,23]. On the contrary Piwi4 coimmunoprecipate with Ago2, Dcr2, Piwi5, Piwi6 and Ago3, suggesting a bridging role between the siRNA and piRNA pathways [21]. These studies support an antiviral role for Piwi proteins in *Aedes* spp. mosquitoes but given the number of *Piwi* genes in these species, it is a challenge to uncover their distinct physiological roles, if any. In duplicated genes, the presence of sites under positive selection is usually a sign of the acquisition of novel functions [24]. Additionally, under the "arm-race theory", rapid intraspecific evolution is expected for genes with immunity functions because their products should act against fast evolving viruses [25].

Besides being an important basic question, the understanding of functional divergence among Piwi proteins has applied perspectives for the development of novel genetic-based methods of transmission-blocking vector control strategies.

In recent years, the Asian tiger mosquito *Aedes albopictus* has emerged as a novel global arboviral threat. This species is a competent vector for a number of arboviruses, such as chikungunya (CHIKV), dengue (DENV), yellow fever (YFV) and Zika (ZIKV) viruses and is now present in every continent except Antarctica following its quick spread out of its native home range of South East Asia [26]. Establishment of *Ae. albopictus* in temperate regions of the world fostered the re-emergence or the new introduction of arboviruses [27]. For instance, chikungunya outbreaks occurred in Italy in 2007 and 2017 [28,29]; France and Croatia suffered from autochthonous cases of dengue and chikungunya in several occasions since 2010 [30–33] and dengue is remerging in some regions of the United States due to the presence of *Ae. albopictus* [34]. Knowledge on *Ae. albopictus* biology and the molecular mechanisms underlying its competence to arboviruses are still limited in comparison to *Ae. aegypti* despite its increasing public-health relevance.

Here we elucidate the molecular organization, polymorphism and expression of *Piwi* clade genes of *Ae. albopictus* in an evolutionary framework using a combination of molecular, population genomics and computational protein modelling approaches. We show that the genome of *Ae. albopictus* harbours seven *Piwi* genes, namely *Ago3*, *Piwi1/3*, *Piwi2*, *Piwi4*, *Piwi5*, *Piwi6* and *Piwi7*. For the first time in mosquitoes, we show sign of adaptive evolution in *Piwi1/3*,

Piwi4, *Piwi5* and *Piwi6*, including sites in the MID and PAZ domains. Additionally, expression profiles during mosquito development and following infection with the dengue or chikungunya viruses support functional specialization of Piwi proteins, with a prominent and general antiviral role for the transcript of *Piwi5*.

Results

Seven *Piwi* genes are present in the genome of *Ae. albopictus*

Bioinformatic analyses of the current genome assemblies of *Ae. albopictus* (AaloF1) and the C6/36 cell line (canu_80X_arrow2.2), followed by copy number validation, confirmed the presence of seven *Piwi* genes (i.e. *Ago3*, *Piwi1/3*, *Piwi2*, *Piwi4*, *Piwi5*, *Piwi6 and Piwi7*) in *Ae. albopictus* (S1 Table). Genomic DNA sequences were obtained for each exon-intron boundaries, confirming in all *Piwi* genes the presence of the PAZ, MID and PIWI domains, the hallmarks of the Piwi subfamily of Argonaute proteins [35]. For *Ago3*, *Piwi1/3*, *Piwi2*, *Piwi4* and *Piwi6*, single transcript sequences that correspond to predictions based on the identified DNA sequences were retrieved (S1 Dataset). Sequencing results of the transcript from *Piwi5* showed a sequence 27 bp shorter than predicted on the reference genome, due to a 45bp gap followed by a 18b insertion, 110 and 333 bases after the ATG starting codon, respectively. This transcript still includes the PAZ, MID and PIWI domains. The presence of this transcript was further validated by northern-blot (Fig 1). For *Piwi7*, the transcript sequence also appears shorter than predicted (Fig 1). Alignment and phylogenetic analyses, in the context of currently annotated *Piwi* transcripts of Culicinae and Anophelinae mosquitoes, confirmed one-to-one orthologous pairing between *Ae. albopictus Piwi* gene transcripts and those of Ae. aegypti (S2 Table,

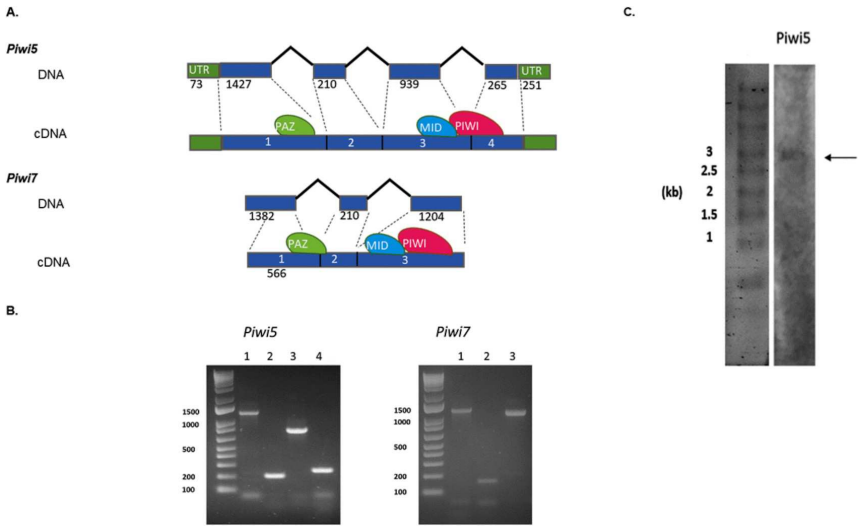


Fig 1. Gene and transcript structure of *Ae. albopictus Piwi5* and *Piwi7*. A) Schematic representation of the DNA structure of *Piwi5* and *Piwi7* genes and their corresponding transcripts as obtained from cDNA amplification of single sugar-fed mosquito samples. Exons and introns are shown by blue boxes and black lines, respectively, with corresponding length in nucleotide below each. The positions of the predicted PAZ, MID and PIWI domains are shown by green, blue and magenta ovals, respectively. Exon numbers correspond to lane numbers. B) Amplification of each exon of Piwi5 and Piwi7 on genomic DNA. Exon numbers correspond to lane numbers. C) Northern-blot results of *Piwi5* indicate the presence of a transcript of 3 kb.

https://doi.org/10.1371/journal.pntd.0007919.g001

S1 Fig). Interestingly, *Piwi5*, *Piwi6* and *Piwi7* transcripts group together and appear more similar to one of the two Aubergine-like transcripts annotated in different Anophelinae species than to *Aedes Piwi2*, *Piwi1/3* and *Piwi4* transcripts. Regarding the latter, *Piwi2* and *Piwi1/3* form a species-specific clade, rather than follow a speciation pattern. Rather, the two genes, which on *Ae. aegypti* map on chromosome 1 and are ~20 kb apart [17], may have originated as a duplication in the ancestor of *Ae. aegypti* and *Ae. albopictus* and be subsequently undergoing interlocus gene conversion. This mechanism causes nonreciprocal recombination, whereby one locus (i.e. part of a gene copy) replaces the homologous sequence of the other copy. The result is concerted evolution of the gene duplicates [36], which in this case eliminates divergence between *Piwi2* and *Piwi1/3* within each species.

Piwi genes display high levels of polymorphism across populations and show signs of adaptive evolution

Across *Drosophila* phylogeny, genes of the piRNA pathway display elevated rates of adaptive evolution [37], with rapidly evolving residues not clustering at the RNA binding site, but being distributed across the proteins [3]. The RNA binding site is found within the PAZ domain, at the amino-terminal part of Piwi proteins [35,38]. On the opposite side, at the carboxyl terminus, the PIWI domain resides. The PIWI domain belongs to the RNase H family of enzymes and the catalytic site is formed by three conserved amino acids (usually aspartate-aspartate-glutamate, DDE or aspartate-aspartate-histidine, DDH) [35,39]. Between the PAZ and PIWI domains the MID domain resides. MID specifies strand- and nucleotide-biases of piRNAs, including their Uridine 5' bias [40,41]. To evaluate the selective pressures acting along these genes, we analysed the polymorphism pattern in *Ae. albopictus* samples from wild-collected populations and from the Foshan reference strain. Synonymous and non-synonymous mutations were found for each gene in all populations (Fig 2), with *Piwi1/3* displaying the lowest polymorphism (Table 1).

As expected, the laboratory strain Foshan showed the lowest levels of variability and Tajima's D values that contrast (in sign) from those of the other populations and from the pooled

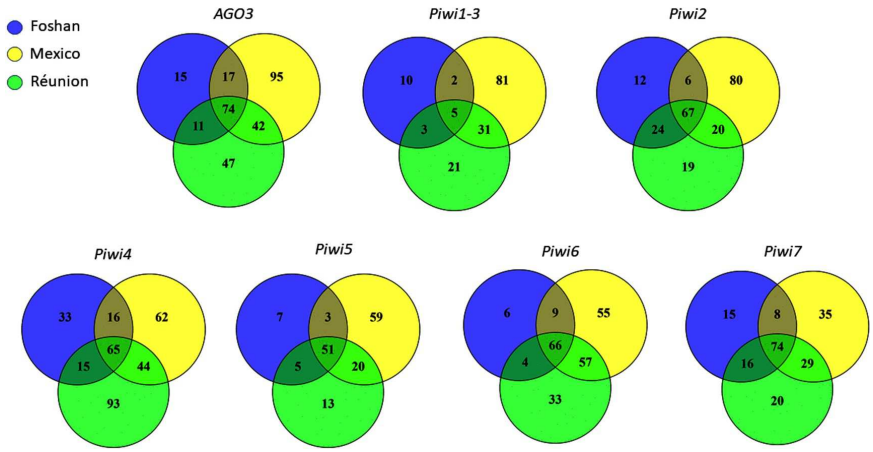


Fig 2. Venn diagrams showing the number of positions harbouring synonymous and non-synonymous mutations in tested samples for each *Piwi* gene.

https://doi.org/10.1371/journal.pntd.0007919.g002

Table 1. Polymorphism of *Aedes albopictus Piwi* genes in mosquitoes from the Foshan strain and wild-caught mosquitoes from La Reunion (Reu) and Mexico (Mex). We report the number of sequences (*n*), as well as the number of sites (*L*), segregating sites (*S*), polymorphism measured as π and θ, and the Tajima's *D* statistic for both synonymous (s) and non-synonymous sites (a) for each gene and population (and for the pooled sample).

	n	*L*	L_s	L_a	S_s	S_a	π_s	π_a	θ_s	θ_a	π_a/π_s	D_s	D_a
Ago3													
Pooled	112	2832	680.2	2151.8	316	19	0.0699	0.0005	0.0878	0.0017	0.007	-0.68	-1.95
Foshan	32	2832	680.1	2151.9	124	5	0.0559	0.0004	0.0453	0.0006	0.007	0.89	-0.82
Mex	48	2832	680.2	2151.8	253	14	0.0780	0.0007	0.0838	0.0015	0.009	-0.25	-1.60
Reu	32	2658	643.8	2014.2	189	4	0.0678	0.0002	0.0729	0.0005	0.003	-0.27	-1.50
Piwi1/3													
Pooled	112	2658	644.3	2013.7	136	23	0.0319	0.0010	0.0399	0.0022	0.033	-0.66	-1.51
Foshan	32	2658	644.0	2014.0	10	2	0.0047	0.0003	0.0039	0.0002	0.064	0.68	0.44
Mex	48	2658	644.9	2013.1	117	21	0.0463	0.0017	0.0409	0.0024	0.037	0.48	-0.89
Reu	32	2658	643.8	2014.2	52	4	0.0188	0.0004	0.0201	0.0005	0.021	-0.23	-0.48
Piwi2													
Pooled	112	2625	644.0	1981.0	242	28	0.0760	0.0012	0.0710	0.0027	0.016	0.23	-1.65
Foshan	32	2625	644.0	1981.0	115	10	0.0663	0.0017	0.0443	0.0013	0.026	1.88	1.11
Mex	48	2625	643.9	1981.1	184	15	0.0823	0.0010	0.0644	0.0017	0.012	1.01	-1.28
Reu	32	2625	644.1	1980.9	151	6	0.0712	0.0005	0.0582	0.0008	0.007	0.85	-0.94
Piwi4													
Pooled	112	2592	620.0	1972.1	268	61	0.0729	0.0025	0.0817	0.0058	0.034	-0.36	-1.82
Foshan	32	2592	620.1	1971.9	122	18	0.0610	0.0009	0.0489	0.0023	0.015	0.94	-2.05
Mex	48	2592	619.8	1972.2	181	41	0.0692	0.0035	0.0658	0.0047	0.051	0.19	-0.87
Reu	32	2592	620.1	1971.9	161	45	0.0699	0.0029	0.0645	0.0057	0.041	0.32	-1.79
Piwi5													
Pooled	112	2745	653.1	2091.9	148	23	0.0457	0.0016	0.0428	0.0021	0.035	0.22	-0.66
Foshan	32	2793	664.5	2128.5	58	8	0.0361	0.0018	0.0217	0.0009	0.050	2.47	2.78
Mex	48	2745	652.9	2092.1	137	13	0.0470	0.0017	0.0473	0.0014	0.036	-0.02	0.65
Reu	32	2793	663.4	2129.6	89	6	0.0326	0.0008	0.0333	0.0007	0.025	-0.08	0.40
Piwi6													
Pooled	112	2661	649.0	2012.0	242	8	0.0805	0.0010	0.0705	0.0008	0.013	0.47	0.82
Foshan	32	2661	648.3	2012.8	92	3	0.0632	0.0001	0.0352	0.0004	0.002	2.99	-1.69
Mex	48	2661	649.9	2011.1	213	7	0.0840	0.0001	0.0739	0.0008	0.001	0.50	-2.33
Reu	32	2661	648.5	2012.5	163	4	0.0784	0.0001	0.0624	0.0005	0.001	0.98	-2.01
Piwi7													
Pooled	112	1977	469.8	1507.2	192	33	0.0877	0.0036	0.0772	0.0041	0.041	0.45	-0.42
Foshan	32	1977	469.8	1507.2	118	15	0.0905	0.0034	0.0624	0.0025	0.038	1.71	1.25
Mex	48	1977	469.9	1507.1	150	23	0.0905	0.0034	0.0719	0.0034	0.038	0.93	-0.04
Reu	32	1977	469.6	1507.5	137	17	0.0803	0.0030	0.0724	0.0028	0.037	0.41	0.24

https://doi.org/10.1371/journal.pntd.0007919.t001

sample, consistent with a strong bottleneck associated to the strain establishment. In *Piwi4*, between 20 and 80 non-synonymous variants could be found inside and in proximity of the PAZ, MID and PIWI domains (S2A Fig), ten of these mutations were shared across all populations (S3 Table). The 5' region of *Piwi5* harboured several indels: two in-frame variants (i.e. 94_99del; 113_118del) were shared across all populations and were present in homozygosity in at least one sample (S2B Fig), suggesting that they are not detrimental. *Ago3* and *Piwi6* have very low non-synonymous nucleotide diversity, suggesting strong constraints at the protein level. The results of the McDonald-Kreitman test [42] further revealed an excess of non-synonymous substitutions compared to the polymorphism pattern in both *Piwi1/3* and *Piwi6*,

Table 2. Insights into Evolutionary divergence of *Piwi* genes in *Ae. albopictus*. A) McDonald-Kreitman test for each *Piwi* gene using the orthologous sequences of *Ae. aegypti* as outgroup. NI = Neutrality Index; Alpha = proportion of base substitutions fixed by natural selection; *P* estimated using Fisher's exact test. B) Output of Codeml with significant results regarding sites under positive selection.

A. McDonald-Kreitman test

	Ago3	*Piwi1/3*	*Piwi2*	*Piwi4*	*Piwi5*	*Piwi6*	*Piwi7*
NI	0.582	0.516	0.9	3.888	0.696	0.154	0.745
alpha	0.418	0.484	0.1	-2.888	0.304	0.846	0.255
P	0.114	0.008	0.785	< 0.001	0.18	< 0.001	0.272

B. Codeml output for sites under positive selection

Gene	Position[1]	Reference>Mutant[2]	ω[3]	P[4]	Domain[5]
AGO3	-	-	-	-	
Piwi1/3	484	E>G	3.026	0.990*	Linker2
	485	K>R	2.979	0.965*	Linker2
	548	M>I	3.014	0.984*	MID
Piwi2	-	-	-	-	
Piwi4	278	Y>D	2.522	0.993**	PAZ
	287	H>A,D,P,V	2.532	1.000**	PAZ
Piwi5	89–90	SA>PT	7.813	1.000**	Flex
	139	T>A	7.810	1.000**	Flex
Piwi6	67	A>P	3.560	0.992**	Flex
	86	G>R,S	3.460	0.957*	Flex
	258	V>I	3.581	0.999**	Linker2
Piwi7	-	-	-	-	

[1] sites where signs of positive selection (ω > 1) were found

[2] reference amino acid and alternative missense variant

[3]mean omega (ω) value

[4]probability that ω > 1 under the Bayes empirical Bayes (BEB) method (* = P > 0.95; ** = P > 0.99)

[5]protein domain based on computational predictions of molecular structures. Domains are as follows: Linker2, linker region between PAZ and MID; PAZ domain; MID domain; and Flex, the Flexible stretch at the N-terminus.

https://doi.org/10.1371/journal.pntd.0007919.t002

suggesting that they have been target of adaptive evolution (Table 2A). In contrast, *Piwi4* has a significant deficit of non-synonymous substitutions and/or excess of polymorphic non-synonymous segregating sites (Table 2A). In this gene, Tajima's D is negative but in line with the values of the other *Piwi* genes, and the high non-synonymous polymorphism may reflect selection of intraspecific diversifying selection, as expected in genes involved in immunity. Because positive selection may have acted at the level of very few sites, this not contributing to the gene-level non-synonymous substitution pattern; we explicitly tested models of codon evolution. Signs of positive selection were found at different sites, including one site in the Linker2 and one site in the MID domain of *Piwi1/3*, two sites in the PAZ domain of *Piwi4*, two sites in the Flex domain of *Piwi5* and three sites, two in the Flex and one in the Linker2 domains, of *Piwi6* (Table 2B). Haplotype reconstruction of our samples showed that these mutations can co-occur on the same gene, with the only exception of Y278D+H287P in *Piwi4* and A67P+G86S in *Piwi6*.

Finally, to gain insight on how variable *Piwi* genes are in comparison to slow- and fast -evolving genes of *Ae. albopictus*, we collected variability data of sets of genes previously identified to have slow and high evolutionary rates [43]. For each population, we compared the overall level of polymorphism (LoP) of the *Piwi* genes and of a dataset of fast-evolving genes (FGs) to that measured for a dataset of slow-evolving genes (SGs) as listed in the material and methods section "polymorphisms of *Piwi* genes" [43]. Our results indicate that *Piwi4*, *Piwi6* and

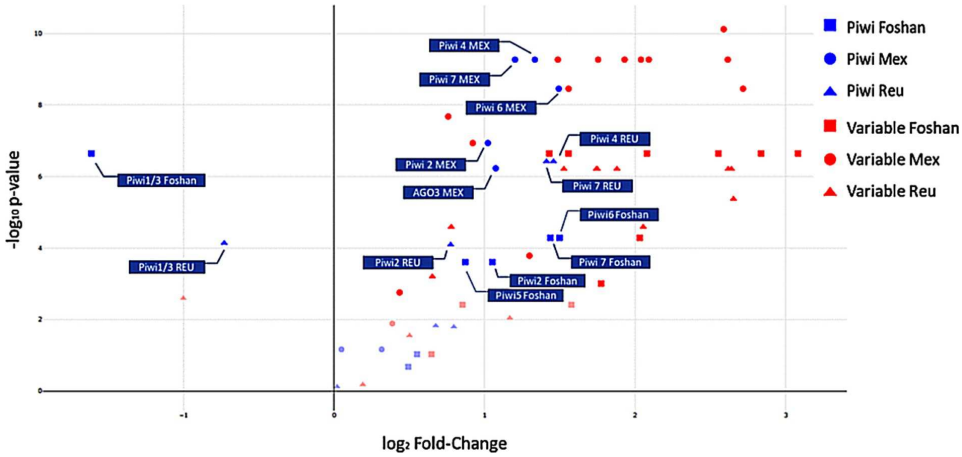


Fig 3. Volcano plot. Level of polymorphism (LoP) comparison between slow-evolving genes (SGs), fast-evolving genes (FGs) and *Piwi* genes by population. Genes on the right side of the panel have LoP values greater than those of SGs, while genes on the left side have smaller LoPs than SGs. The y-axis represents the -log10 p-values of the Kolmogorov-Smirnov test. Faint datapoints are not significant after Bonferroni correction for multiple testing (-log10 0.0024 (0.05/21 genes) = 2.62).

https://doi.org/10.1371/journal.pntd.0007919.g003

Piwi7 have LoP values comparable to those of FGs, while *Ago3* and *Piwi5* do not significantly deviate from the LoP values of SGs. *Piwi1*/3 appears to be conserved (Fig 3).

Computational predictions of molecular structures

The functional significance of the mutations under selection, as well as that of all the shared missense mutations in the PAZ and PIWI domains, was tested by computing predictions of three dimensional molecular structures of the Piwi proteins using the most-recent X-ray crystallography structure of Argonaute proteins as templates [44,45]. Homology modelling revealed high structural conservation among the seven Piwi proteins despite sequence heterogeneity (S2 Fig; Fig 4A).

Similarly, to *D. melanogaster*, the highest levels of amino acid sequence conservation were found in the regions that, based on homology modelling, define the inner pocket of Argonaute molecular assembly where the RNA binds. Significantly lower sequence conservation was found on the proteins surface (Fig 4B). Based on our computational predictions, we could not detect amino acidic polymorphisms that would affect RNA binding or processing, suggesting that all *Ae. albopictus* Piwi proteins may retain the Argonaute-like functions. Mapping of mutations under positive selection (Table 2B) on the homology models and sequence comparisons with known PIWI structural homologs showed that the identified variant amino acids are unlikely to induce severe alterations in protein folding. All mutant variants were found to localize in regions distant from the predicted RNA-binding and/or processing sites, ruling out possible effects associated to alterations in RNA recognition, but raising the intriguing possibility of regulatory roles during interactions with additional binding partners.

Developmental profile of *Ae. albopictus Piwi* genes

To further gain insights on the functional specialization of *Piwi* genes, we assessed their expression profile throughout mosquito development, namely at 4–8 hours (h) after

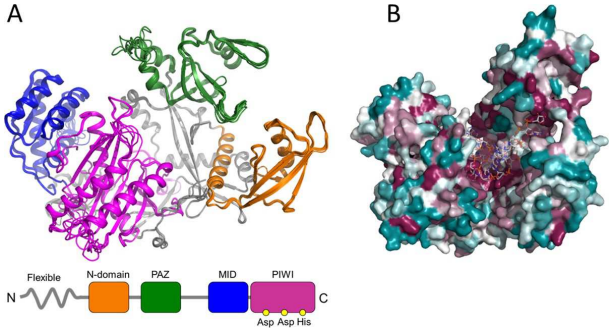


Fig 4. Computational homology models of the *Ae. Albopictus Piwi* proteins. Homology models were generated for the seven *Piwi* genes as described in the methods section. A) Superposition of cartoon representations of Piwi homology models, with highlight of domain organization: the N-terminal domain is shown in orange, the PAZ domain in green, the MID domain in blue and the PIWI domain in magenta. B) *CONSURF* [46]overview of the aminoacid sequence conservation mapped on three-dimensional homology models in a putative RNA-bound arrangement based on the structure of human Argonaute bound to a target RNA (PDB ID 4Z4D), colored from teal (very low conservation) to dark magenta (highly conserved).

https://doi.org/10.1371/journal.pntd.0007919.g004

deposition to capture the maternal-zygotic transition in expression, at late embryogenesis (i.e. 12–16 h and 16–24 h post deposition), at two time points during larval development (i.e. 1st and 4th instar larvae) and at pupal and adult stages (for the latter only we sampled separately males and females). Adult females were dissected to extract ovaries from the carcasses both from females kept on a sugar diet and 48 h after a blood meal, when a peak in *Piwi* gene expression was previously observed [47].

Expression levels of *Ago3*, *Piwi4*, *Piwi5*, *Piwi6* and *Piwi7* are at their peak in the embryonic stages, although at different time points (Fig 5A). Overall, *Ago3*, *Piwi1/3*, *Piwi2* and *Piwi6* have a similar trend during development showing a second peak of expression in adult females and their ovaries, while the expression levels of *Piwi4*, *Piwi5* and *Piwi7* remain constant. In details, *Piwi7* is mostly expressed 4-8h after deposition, *Piwi5* and *Piwi6* are mostly expressed after 8-16h and *Ago3* and *Piwi4* have their pick of expression at 16-24h. On the contrary, *Piwi1/3* and *Piwi2* are mostly expressed in ovaries extracted from blood-fed and sugar-fed females, respectively (Fig 5A, S4A Table). These results are consistent with lack of expression from published RNA-seq data from adult mosquitoes.

Overall, at the adult stages, *Ago3* and all *Piwi* genes were more expressed in females than males. Expression in ovaries was higher than in the corresponding carcasses, in both sugar- and blood-fed females. Differences in carcasses vs. ovaries expression were more pronounced after blood-meal for *Ago3*, *Piwi1/3* and *Piwi6*, while expression of *Piwi2* was doubled in sugar-fed vs. blood-fed ovaries.

Piwi genes expression following viral infection

Finally, we assessed whether the expression pattern of *Piwi* genes was altered upon DENV and CHIKV infection (Fig 5B). The expression profile of the *Piwi* genes was different following CHIKV- and DENV-infection and also when comparing samples of carcasses and ovaries. In ovaries, during CHIKV infection all *Piwi* genes were significantly up-regulated compared to both sugar- and blood-fed mosquitoes. Four days post infection (dpi), the expression of *Ago3*, *Piwi1/3*, *Piwi6* and *Piwi7* was between 4 to 10 folds higher than that of *Piwi2*, *Piwi4* and *Piwi5*, which nevertheless were upregulated with respect to ovaries of sugar- and blood-fed

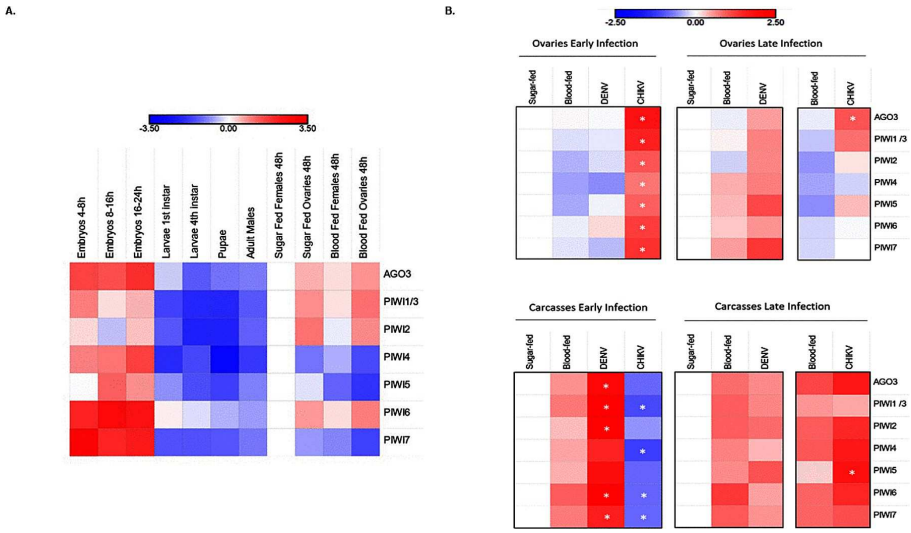


Fig 5. Expression profile of *Piwi* genes. Heatmap representations of log10 transformed fold-change expression values of each *Piwi* gene. A) Developmental expression pattern of the *Piwi* genes normalized on the expression in sugar fed females. B) Expression pattern of *Piwi* genes following viral infection normalized with respect to sugar-fed samples. Expression was verified in ovaries and carcasses separately, during the early and late stages of infections, that is 4 dpi for both viruses and 14 or 21 dpi for CHIKV and DENV, respectively. Each day post-infection was analysed with respect to sugar and blood-fed controls of the same day. * indicates significant difference ($P<0.05$) between infected samples and the corresponding blood-fed control.

https://doi.org/10.1371/journal.pntd.0007919.g005

mosquitoes. An opposite profile was seen in the carcasses, where all *Piwi* genes, particularly *Piwi1/3* and *Piwi4*, were down-regulated. At 4 dpi, CHIKV has already disseminated throughout the mosquito body, has reached the salivary glands and is able to be transmitted. CHIKV viral titer was reduced ten folds by 14 dpi and the profile of *Piwi* genes changed. Expression in the ovaries decreased between 3 (*Piwi5*) to 20 (*Piwi7*) times with respect to values observed at 4 dpi, but remained higher than the corresponding expression values in ovaries of both sugar and blood-fed mosquitoes. In carcasses all *Piwi* genes inverted their expression pattern during the infection phase, increasing up to more than 100 times in the case of *Piwi4*, *Piwi5* and *Piwi6*. At 14 dpi, expression of the *Piwi* genes was highest in CHIKV-infected carcasses than in carcasses of sugar- and blood-fed mosquitoes.

For DENV, infection progresses differently than CHIKV. At 4 dpi there is no virus in the salivary gland, where the viral titer was measured at zero. By 21 dpi, DENV has established persistent infection [48]. At 4 dpi expression of *Piwi* genes was lower in DENV- and blood-fed ovaries than in ovaries of sugar-fed mosquitoes. The only exception was *Piwi6*, which was slightly up regulated in ovaries of DENV-infected samples, but slightly down-regulated in ovaried of blood-fed mosquitoes. On the contrary, at the same time point, carcasses of DENV-infected samples showed a drastic increase in the expression of all *Piwi* genes with respect to blood-fed samples; this increase was between 7 to 87 times for *Piwi7* and *Piwi2*, respectively. By 21 dpi, expression in the ovaries increased for all *Piwi* genes, in comparison to what observed both at 4dpi and in blood-fed ovaries, suggesting the increase in expression of *Piwi* genes is related to DENV dissemination. Interestingly, if we compare levels of expression in CHIKV-infected ovaries at 4 dpi and DENV-infected samples at 21 dpi, corresponding to the

time at which both viral species have disseminated throughout the mosquito body, we observe similar levels of fold-change expression of *Piwi4* and *Piwi7*, while *Ago3*, *Piwi1/3* and *Piwi6* show higher fold-change in CHIKV compared to DENV samples. Whether this trend is dependent on the viral species or viral titer requires further investigation. The same type of comparison in carcasses shows a higher fold-change expression level of all *Piwi* genes, particularly *Piwi1/3* and *Piwi5*, in DENV- versus CHIKV-infected samples, even if viral titers are lower for DENV (S4B Table). Overall these results support the hypothesis of a concerted activity of all PIWI proteins during viral dissemination for DENV, and maintenance of infection rely on expression of primarily *Piwi5*. On the contrary, establishment of persistent CHIKV infection was accompanied by elicitation of all Piwi gene expression, particularly *Piwi4* and, again, *Piwi5*.

Discussion

The piRNA pathway does not have antiviral immunity in *D. melanogaster* [15]. In the arboviral vectors *Aedes spp*. mosquitoes, vpiRNAs are found following infections with arboviruses, piRNAs are produced in the soma besides the germline and there has been an expansion on the number of *Piwi* genes, supporting the hypothesis that the piRNA pathway has antiviral immunity[12,49,50]. Besides *Ago3*, the genome of *Ae. aegypti* harbours six *Piwi* genes (i.e. *Piwi1/3*, *Piwi2*, *Piwi4*, *Piwi5*, *Piwi6*, *Piwi7)*, some of which show tissue and development-specific expression profile and have been preferentially associated with either TE-derived or viral piRNAs, [16,20,21]. These studies were based on the knowledge of the gene structure of each *Ae. aegypti Piwi* gene and the application of *ad hoc* RNAi-based silencing experiments and *in vitro* expression assays, but lack an evolutionary perspective [18–21].

In this work we focused on the emerging arboviral vector *Ae. albopictus* and we show how the application of evolutionary and protein modelling techniques helps to unravel functional specialization of Piwi proteins. The genome of *Ae. albopictus* harbours one copy of *Ago3* and six *Piwi* genes (i.e. *Piwi1/3*, *Piwi2*, *Piwi4*, *Piwi5*, *Piwi6 and Piwi7*), each a one-to-one orthologue to the *Ae. aegypti Piwi* genes. The only exceptions are *Piwi2* and *Piwi1/3*, where the two genes from the same species cluster together. In *Ae. aegypti*, these two genes both map on Chromosome 1, separated by ~ 20kb, suggesting they may undergo frequent gene conversion.

All transcripts retain the PAZ and PIWI domains, which are the hallmarks of the Argonaute protein family [35]. By using homology modelling, we obtained predictions of molecular architectures for *Ae. albopictus Ago3* and Piwi proteins, onto which we mapped the putative boundaries of each domain. Superpositions and sequence comparisons allowed clear identification of the catalytic DDH triad within the PIWI domain of all modelled proteins. This conservation is consistent with strong sequence matching in the putative RNA binding regions of the PIWI, PAZ and MID domains and suggests the possible maintenance of slicer activity, albeit experimental validation of each isoform is necessary.

The expression of all *Piwi* genes was confirmed throughout the developmental stages and the adult life of the mosquito, both in ovaries and somatic tissues. Interestingly, *Piwi7* transcript expression starkly drops following early embryogenesis, to the point that we could detect it neither in RNA-seq analyses, nor in Northern-blot experiments (S5 Fig). The expression of *Piwi* genes was elicited upon arboviral infection, indirectly confirming the antiviral role of the piRNA pathway. The expression profile of *Piwi* genes showed differences depending on both the species of infecting virus and on when the expression was measured. In CHIKV-infected samples, expression of *Piwi* genes was mostly elicited in ovaries or carcasses at 4 or 14 dpi, respectively. On the contrary, in DENV-infected samples, the highest expression *of Piwi* genes was seen in carcasses 4 dpi. These results are concordant with the timing in piRNAs

accumulation following CHIKV or DENV infection. In *Ae*. *albopictus* mosquitoes infected with CHIKV, secondary piRNAs are not found 3 dpi, but are enriched 9 dpi [9]. In contrast, in *Ae*. *aegypt*i mosquitoes infected with DENV2, piRNAs are the dominant small RNA populations 2 dpi [50].

Overall, these observations and our expression analyses support the hypothesis of an early activation of the piRNA pathway following DENV infection, but a late activation after CHIKV infection. Additionally, our expression analysis is consistent with a generalist antiviral role for *Piwi5*, which is elicited both during DENV and CHIKV infection [20], but suggest a more prominent role for *Piwi6* and *Piwi1/3* or *Piwi4* and *Ago3* during infection with DENV and CHIKV, respectively.

Materials and methods

Mosquitoes

Aedes albopictus mosquitoes of the Foshan strain were used in this study. This strain was established in 1981 in the Center for Disease Control and Prevention of Guangdong Province in China. It has been at the University of Pavia since 2013 [10,51]. Mosquitoes are reared under constant conditions, at 28˚C and 70–80% relative humidity with a 12/12h light/dark cycle. Larvae are reared in plastic containers, at a controlled density to avoid competition for food. Food is provided daily in the form of fish food (Tetra Goldfish Gold Colour). Adults are kept in 30 cm^3 cages and fed with cotton soaked in 0.2 g/ml sucrose as a carbohydrate source. Adult females are fed with defibrinated mutton blood (Biolife Italiana) using a Hemotek blood feeding apparatus. Mosquitoes from Mexico and La Reunion island were collected in 2017 as adults and maintained in ethanol 70% before shipment to Italy. All samples were processed at the University of Pavia.

Mosquito infections

Foshan mosquitoes were infected with DENV serotype 1, genotype 1806 or CHIKV 06.21. DENV-1 (1806) was isolated from an autochthonous case from Nice, France in 2010 [52]. CHIKV 06–21 was isolated from a patient on La Reunion Island in 2005 [53]. Both strains were kindly provided by the French National Reference Center for Arboviruses at the Institut Pasteur. CHIKV 06–21 and DENV-1 1806 were passaged twice on cells to constitute the viral stocks for experimental infections of mosquitoes, on C6/36 cells for CHIKV 06–21 and on African green monkey kidney Vero cells for DENV-1 1806. Viral titers of stocks were estimated by serial dilutions and expressed in focus-forming units (FFU)/mL.

Four boxes containing 60 one-week-old females were exposed to an infectious blood-meal composed by 2 mL of washed rabbit red blood cells, 1 mL of viral suspension and 5 mM of ATP. The titer of the blood-meal was 10^7 PFU/mL for CHIKV and $10^{6.8}$ PFU/mL for DENV. Fully engorged females were placed in cardboard boxes and fed with a 10% sucrose solution. Mosquitoes were incubated at 28˚C until analysis.

In parallel, mosquitoes were fed with uninfected blood-meal or kept on a sugar-diet and grown in the same conditions. Thirty mosquitoes were killed to be analyzed at days 4 and 14 post-infection (pi) for CHIKV, and at days 4 and 21 pi for DENV. To estimate transmission, saliva was collected from individual mosquitoes as described in [54]. After removing wings and legs from each mosquito, the proboscis was inserted into a 20 μL tip containing 5 μL of Fetal Bovine Serum (FBS) (Gibco, MA, USA). After 30 min, FBS containing saliva was expelled in 45 μL of Leibovitz L15 medium (Invitrogen, CA, USA) for titration. Transmission efficiency refers to the proportion of mosquitoes with infectious saliva among tested mosquitoes (which correspond to engorged mosquitoes at day 0 pi having survived until the day of examination).

The number of infectious particles in saliva was estimated by focus fluorescent assay on C6/36 *Ae. albopictus* cells. Samples were serially diluted and inoculated into C6/36 cells in 96-well plates. After incubation at 28˚C for 3 days (CHIKV) or 5 days (DENV), plates were stained using hyperimmune ascetic fluid specific to CHIKV or DENV-1 as primary antibody. A Fluorescein-conjugated goat anti-mouse was used as the second antibody (Biorad). Viral titers were 16,266±50,446 FFU and 155±125 FFU for CHIKV at 14 dpi and DENV at 21 dpi, respectively.

At the same time points mosquitoes that had been fed a non-infectious blood or kept on a sugar diet were sampled and dissected as above.

Bioinformatic identification of *Piwi* genes in the *Ae. albopictus* genome

The sequences of the *Ae. aegypti Piwi* genes [55] were used as query to find orthologs in the reference genome of the *Ae. albopictus* Foshan strain (AaloF1 assembly) and in the genome of the *Ae. albopictus* C6/36 cell line (canu_80X_arrow2.2 assembly) using the BLAST tool in Vectorbase. Deduced coding sequences (CDS) where analysed in Prosite (Prosite.expasy.org/prosite.html) to screen for the typical PAZ and PIWI domains of Argonaute proteins [56].

Copy number of *Piwi* genes

qPCR reactions were performed using the QuantiNova SYBR Green PCR Kit (Qiagen) following the manufacturer's instructions on an Eppendorf Mastercycler RealPlex4, on genomic DNA from four mosquitoes and using gene-specific primers, after having verified their efficiency (S4 Table). DNA was extracted using DNA Isolation DNeasy Blood & Tissue Kit (Qiagen). Estimates of gene copy number were performed based on the $2^{-\Delta CT}$ method using *Piwi6* and the para sodium channel genes (AALF000723) as references [57].

Structure of *Piwi* genes

DNA extracted from whole mosquitoes and dissected ovaries [58] was used as template in PCR amplifications to confirm the presence and the genome structure of each bioinformatically-identified *Piwi* gene. Primers were designed to amplify each exon, with particular attention to detect differences between paralogous *Piwi* genes (S1 Table). The DreamTaq Green PCR Master Mix (Thermo Scientific) was used for PCR reactions with the following parameter: 94˚C for 3 minutes, 40 cycles at 94˚C for 30 sec, 55˚C-62˚C for 40 sec, 72˚C for 1–2 minutes and final extension step of 72˚C for 10 minutes. PCR products were visualized under UV light after gel electrophoresis using 1–1.5% agarose gels stained with ethidium bromide and a 100 bp or 1 kb molecular marker. PCR products were either directly sequenced or cloned using the TOPO TA Cloning Kit strategy (Invitrogen) following the manufacturer's instructions. DNA plasmids were purified using the QIAprep Spin Miniprep Kit and sequenced.

Piwi gene transcript sequences and phylogeny

RNA was extracted using a standard TRIzol protocol from pools of 5 adult female mosquitoes to verify the transcript sequence of each *Piwi* gene. Sets of primers were designed for each gene to amplify its entire transcript sequence (S4 Table). PCR reactions were performed using a High Fidelity taq-polymerase (Platinum SuperFi DNA Polymerase, Invitrogen) following manufacturer's instructions. PCR products were cloned using the TOPO TA Cloning Kit (Invitrogen) and plasmid DNA, purified using the QIAprep Spin Miniprep Kit, was sequenced. Rapid amplification of cDNA ends (RACE) PCRs were performed using First-Choice RLM-RACE Kit (Thermo Fisher Scientific) to analyse 5' and 3' ends of the transcript

sequences following manufacturer's instructions. Amplification products were cloned and sequenced as previously indicated.

Sequences of the identified *Ae. albopictus Piwi* gene transcripts were aligned to sequences of *Culicidae* and *D. melanogaster Piwi* transcripts, as downloaded from VectorBase (www.vectorbase.org), using MUSCLE [59]. Maximum-likelihood based phylogenetic inference was based on RAxML after 1000 bootstrap resampling of the original dataset. Phylogeny reconstruction was done through the CIPRESS portal (http://www.phylo.org/index.php/). Resulting tree was visualised using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Northern blot analysis

10μg of total RNA from a pool of 10 sugar-fed females was run in a 1% x 2% agarose/formaldehyde gel (1 g agarose, 10 ml 10x MOPS buffer, 5.4 ml 37% formaldehyde, 84.6 ml DEPC water). Gels were washed twice in 20x SSC for 15 minutes prior to blotting. RNA was transferred to a Amersham Hybond-N+ nylon membrane (GE healthcare) using 20x SSC and cross-linked using UV light exposure for 1 minute. Probes were labelled with biotin using Biotin-High Prime (Roche). Hybridization and detection of biotinylated probes was performed using the North2South Chemiluminescent Hybridization and Detection Kit (Thermo Fisher Scientific) following manufacturer instructions.

Polymorphisms of *Piwi* genes

We investigated *Piwi* gene polymorphism by looking at the distribution of single nucleotide polymorphism in whole genome sequence data from a total of 56 mosquitoes, of which 24 from Mexico, 16 from the island of La Reunion island and 16 from the reference Foshan strain.

Whole genome sequencing libraries were generated and sequenced on the Illumina HiSeqX platform at the Genomics Laboratory of Verily in South San Francisco, California to generate 150 basepair paired end reads. Whole Genome Sequencing data alignments have been deposited to the SRA archive (BioProjects PRJNA484104 and PRJNA562979). Libraries from Tampon and Tapachula had an average of 225216071 reads, meaning an average coverage of 15X based on the AaloF1 assembly (S4 Fig).

Illumina reads were mapped to *Piwi* gene transcript sequences using Burrows-Wheeler Aligner (BWA-MEM) [60] with custom parameters. Polymorphisms was tested by Freebayes [61]. Annotation of the detected mutations, as well counts of synonymous and non-synonymous variants, were performed in snpEff [62]. Frameshifts and non-synonymous variants were plotted using muts needle-plot [63]. Venn diagrams of positions with mutations in the three tested samples were built using Venny 2.1 [64]. Haplotype reconstruction was performed using seqPHASE [65] and PHASE [66,67]. The inferred haplotypes were analysed with DnaSP [68], which estimated the number of segregating sites and the level of nucleotide diversity π [69] in both synonymous and non-synonymous sites. We manually calculated, for synonymous and non-synonymous positions separately, the nucleotide diversity estimator theta [70] and Tajima's D statistic [71], which are a function of the total number of sites and the number of segregating sites (both estimated by DnaSP), and of sample size (see references for detailed formulas). We also tested for signatures of adaptive evolution using the McDonald-Kreitman test [42] (as implemented in DnaSP), which compares the rate of polymorphism and substitutions in synonymous and non-synonymous sites. For this analysis we used alignments that included the orthologous sequences from *Ae. aegypti*.

Haplotype sequences for each gene from each individual were also aligned in TranslatorX [72] using Clustalw [73] and used for Maximum-likelihood based phylogenetic inference

based on RAxML after 1000 bootstrap under the GTRGAMMA model. The relative rates of synonymous and nonsynonymous mutations (dN/dS = ω) averaged across sites was calculated using Codeml in PAML version 4.9 [74], as implemented in PAMLX [75]. Signs of selective pressure for each detected mutation were investigated comparing the M1a (nearly-neutral) *versus* the M2a (positive selection) site models by inferring ω estimations and posterior probabilities under the Bayes empirical Bayes (BEB) approach as implemented in Codeml [74]. This analysis was performed with the following default parameters: runmode = 0, clock = 0, Mgene = 0, CodonFreq = 0, estFreq = 0, fix_blength = 0, optimization method = 0, icode = 0, Seqtype = 1, fix α = 0, ncatG = 5, Small_Diff = 5e-7, n = 1, aaDist = 0.

The level of polymorphism (LoP) for slow-evolving genes (SGs) (AALF008224, AALF005886, AALF020750, AALF026109, AALF014156, AALF018476, AALF014287, AALF004102, AALF003606, AALF019476, AALF028431, AALF018378, AALF027761, AALF014448), fast-evolving genes (FGs) (AALF010748, AALF022019, AALF024551, AALF017064, AALF004733, AALF018679, AALF028390, AALF026991, AALF014993, AALF009493, AALF010877, AALF012271, AALF009839, AALF019413) and the *Piwi* genes was calculated for each population following the pipeline as in [43]. Briefly, SNPs and INDELs were inferred using four Variant callers (i.e. Freebayes [61], Platypus [76], Vardict [77] and GATK UnifiedGenotyper [78]) and the data merged and filtered with custom scripts. Filters include: minimum phred mapping quality = 20 (corresponding to 0.01 error rate), minimum phred base quality = 20, minimum allele frequency = 0.2, minimum allele observation = 2, minimum coverage = 8, maximum depth = 5000. The LoP for each individual was calculated as the number of variants averaged over the region length and the median value for each population was used for subsequent analyses. Statistical analyses were performed in R studio [79]. Fold-change differences were computed as the ratio of the median LoP for each *Piwi* gene and each FG gene over the median LoP of the SG genes. Statistical differences in LoP distribution was assessed via the Kolmogorov-Smirnov test and the p-value threshold was adjusted with the Bonferroni correction.

Homology modelling

Computational structural investigations were carried out initially through the identification of the closest homologs based on sequence similarity (using *NCBI Blast* [80]) and secondary structure matching (using *HHPRED* [81]), using the whole PDB as source collection. Homology model were then generated using *MODELLER* [82] by selecting only homologous RNA-bound structures as template models: *Kluyveromyces polysporus* Argonaute with a guide RNA (PDB ID 4F1N), Human Argonaute2 Bound to t1-G Target RNA (PDB ID 4z4d [83], *T. thermophilus* Argonaute complexed with DNA guide strand and 19-nt RNA target strand (PDB ID 3HM9), and silkworm PIWI-clade Argonaute Siwi bound to piRNA (PDB ID 5GUH).

Computational models were manually adjusted through the removal of non-predictable N- and C-terminal flexible regions using *COOT* [84] followed by geometry idealization in *PHENIX* [85] to adjust the overall geometry. Final model quality was assessed by evaluating average bond lengths, bond angles, clashes, and Ramachandran statistics using Molprobity [86] and the *QMEAN* server [87]. The sequence alignment was generated using EBI muscle [88] and depicted using ESPRIPT3 [89]. Structural figures were generated with *PyMol* [90].

Developmental expression profile of *Piwi* genes

Publicly available RNA-seq data (runs: SRR458468, SRR458471, SRR1663685, SRR1663700, SRR1663754, SRR1663913, SRR1812887, SRR1812889, SRR1845684) were downloaded and aligned using Burrows-Wheeler Aligner (BWA-MEM) [60] to the current *Ae. albopictus* genome assembly (AaloF1). Aligned reads were visualized in Integrative Genomics Viewer

(IGV) [91]. Total RNA was extracted from embryos, 1st and 4th instar larvae, pupae, and adults using Trizol (Thermo Fisher Scientific). Embryos consisted of two pools of 60 eggs at different time points (i.e. 4-8h, 8-16h and 16-24h). Adult samples consisted of males and females kept on a sugar-diet and females fed an uninfected blood-meal. Blood-fed females were dissected to separate ovaries from the carcasses 48 h after blood-meal. These parameters were based on the results of previous studies on *Anopheles stephensi* and *Ae. aegypti* that showed high *Piwi* gene expression during early embryogenesis or 48-72h post blood meal [47]. For each stage, RNA was extracted from pools of 10 mosquitoes, except for first instar larvae and embryos when 20 and 60 individuals were used respectively.

RNA was DNaseI-treated (Sigma-Aldrich) and reverse-transcribed in a 20 μl reaction using the qScript cDNA SuperMix (Quantabio) following the manufacturer's instructions. Quantitative RT-PCRs (qRT-PCR) were performed as previously described using two biological replicates per condition and the RPL34 gene as housekeeping [92]. Relative quantification of *Piwi* genes was determined using the delta-delta-Ct method implemented in the software qBase+ (Biogazelle). Expression values were normalized with respect to those obtained from sugar-fed females.

Expression analyses following infection

Fold-change expression values for each *Piwi* gene was assessed for non-infectious-blood-fed controls, CHIKV-infected and DENV-infected samples after normalization on sugar-fed controls. qRT-PCR experiments (S4 Table) were set up for two replicate pools of 15 ovaries and 15 carcasses at days 4, 14 and 4, 21 for CHIKV and DENV, respectively and the corresponding sugar and non-infectious-blood controls. RNA extraction, qRT-PCR and data analyses were performed as described in the previous paragraph (see "Developmental expression profile of *Piwi* genes"). Fold-change differences significance was assessed using the Analysis of Variance (ANOVA) procedure [93,94] as implemented in qBASE+.

Supporting information

S1 Table. List of the core components of the piRNA pathway in *Ae. aegypti* and their orthologous in *Ae. albopictus*.
(PDF)

S2 Table. List of Transcript IDs and abbreviations of the Culicidae and Drosophilidae species included in the phylogenetic analyses.
(PDF)

S3 Table. Number of non-synonymous mutations found in mosquitoes of the Foshan strain (Foshan) and wild-caught samples from Mexico (Mex) and the island of La Reunion (Reu) divided by type (i.e. missense [M], frameshift [F], indel [I] and nonsense [N]) and number of sites in which mutations were found in all tested samples.
(PDF)

S4 Table. Relative expression values (log10 fold-change) of Piwi genes during development (A) and following viral infection (B) normalized with respect to sugar-fed samples. Samples (2 pools per condition, 15 individuals each) were analysed at 4 days post infection (early infection) and at 14 and 21 days post infection for CHIKV and DENV, respectively (late infection). Each condition was normalized to the corresponding sugar-fed control and compared to the corresponding Blood-fed control. Ovaries and carcasses were analysed independently. * indicates statistically significant difference between infected and non-infected blood-fed samples (ANOVA framework). Relative expression values may mask differences in levels of expression.

For instance, the Ct values of Piwi6, Piwi7 and Piwi1/3 in ovaries 4 days post infection with CHIKV were 30, 33.39 and 25.20, respectively. Ovaries of blood-fed samples at the same time point showed Ct values of 30.30, 33.93 and 26.55 for Piwi6, Piwi7 and Piwi1/3. When relative expression was calculated with respect to Ct values of RPL34, fold-changes in gene expression were comparable among the three genes in both conditions, but Ct values clearly indicate that Piwi7 is less expressed than both Piwi1/3 and Piwi6. These considerations were taken into account when describing results.
(PDF)

S5 Table. List of primers used for CDS analyses, copy number estimation, qPCR experiments and Northern Blot probe design.
(PDF)

S1 Dataset. CDS of the seven Piwi genes of *Ae. albopictus*. The sequence of the PAZ, MID and PIWI domains is in bold, underline and bold-italics, respectively.
(PDF)

S1 Fig. Maximum likelihood cladogram generated from the alignment 862 of transcript sequences of annotated Piwi genes in Culicidae. Transcript IDs and species abbreviations are as listed in S2 Table. AlbPiwi3 is the same as Piwi1/3 in the text. Piwi gene transcripts from *Ae. albopictus* are in red, from *Ae. aegypti* in purple, from *Culex quinquefasciatus* in pink. Transcripts from *D. melanogaster* Ago3, Piwi and Aubergine genes are included for reference and shown in blue. All nodes were supported by bootstrap values higher than 50% with the exception of the three nodes with a black dot.
(PDF)

S2 Fig. Polymorphism of Piwi4 and Piwi5. Lollipop plots representing position, amount and type of mutation along the coding sequences of Piwi4 and Piwi5 in mosquitoes of the Foshan strain, from la Reunion Island (Reu) and Mexico (Mex) as inferred by Freebayes and SnpEFF analyses. Only missense (blue), nonsense (red) and indels (orange) and frameshift (yellow) are shown. The PAZ, MID and PIWI domains are shown in green, blue and magenta, respectively. DDH residues positions are highlighted in the PIWI domain.
(PDF)

S3 Fig. Sequence alignment of *Aedes albopictus* Piwi proteins. Domain boundaries inferred from structural predictions are highlighted by coloured lines using the same colour coding as in Fig 4 (Orange: N-terminus; Green: PAZ; Blue: MID; Magenta: PIWI). Conserved DDH residues found in PIWI are indicated by a black triangle (▲). The "acc" line indicates relative solvent accessibility, ranging from blue (accessible) to white (buried). The sequence alignment was generated using EBI muscle [88] and depicted using ESPRIPT3 [89].
(PDF)

S4 Fig. Per-site read depth of the sequenced libraries of mosquitoes from Tapachula and Tampon.
(PDF)

S5 Fig. Northern blot results for Piwi5 and Piwi7.
(PDF)

Acknowledgments

We thank Monica Ruth Waghacore for insectary work. We thank Verily for having generated libraries and sequenced the mosquitoes from Tapachula and Tampon.

Author Contributions

Conceptualization: Rebeca Carballar-Lejarazu, Federico Forneris, Anna-Bella Failloux, Mariangela Bonizzoni.

Data curation: Michele Marconcini, Mariangela Bonizzoni.

Formal analysis: Michele Marconcini.

Funding acquisition: Mariangela Bonizzoni.

Investigation: Michele Marconcini, Luis Hernandez, Giuseppe Iovino, Vincent Houé, Federica Valerio, Umberto Palatini, Elisa Pischedda, Rebeca Carballar-Lejarazu, Lino Ometto, Federico Forneris.

Methodology: Michele Marconcini, Luis Hernandez, Giuseppe Iovino, Vincent Houé, Federica Valerio, Umberto Palatini, Bradley J. White, Teresa Lin, Rebeca Carballar-Lejarazu, Lino Ometto, Federico Forneris.

Project administration: Michele Marconcini, Anna-Bella Failloux, Mariangela Bonizzoni.

Resources: Jacob E. Crawford, Bradley J. White, Teresa Lin, Anna-Bella Failloux, Mariangela Bonizzoni.

Software: Umberto Palatini, Elisa Pischedda.

Supervision: Lino Ometto, Federico Forneris, Anna-Bella Failloux, Mariangela Bonizzoni.

Validation: Federico Forneris.

Writing – original draft: Michele Marconcini, Federico Forneris, Mariangela Bonizzoni.

Writing – review & editing: Jacob E. Crawford, Lino Ometto, Federico Forneris, Anna-Bella Failloux, Mariangela Bonizzoni.

References

1. Bohmert K, Camus I, Bellini C, Bouchez D, Caboche M, Banning C. AGO1 defines a novel locus of Arabidopsis controlling leaf development. EMBO J. 1998; https://doi.org/10.1093/emboj/17.1.170 PMID: 9427751
2. Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin E V., et al. The evolutionary journey of Argonaute proteins. Nature Structural and Molecular Biology. 2014. https://doi.org/10.1038/nsmb.2879 PMID: 25192263
3. Lewis SH, Salmela H, Obbard DJ. Duplication and diversification of dipteran argonaute genes, and the evolutionary divergence of Piwi and Aubergine. Genome Biol Evol. 2016; 8: 507–518. https://doi.org/10.1093/gbe/evw018 PMID: 26868596
4. Buck AH, Blaxter M. Functional diversification of Argonautes in nematodes: an expanding universe: Figure 1. Biochem Soc Trans. 2013; https://doi.org/10.1042/BST20130086 PMID: 23863149
5. Bollmann SR, Press CM, Tyler BM, Grünwald NJ. Expansion and divergence of argonaute genes in the oomycete genus phytophthora. Front Microbiol. 2018; https://doi.org/10.3389/fmicb.2018.02841 PMID: 30555430
6. Singh RK, Gase K, Baldwin IT, Pandey SP. Molecular evolution and diversification of the Argonaute family of proteins in plants. BMC Plant Biol. 2015; https://doi.org/10.1186/s12870-014-0364-6 PMID: 25626325
7. Bronkhorst AW, Van Rij RP. The long and short of antiviral defense: Small RNA-based immunity in insects. Current Opinion in Virology. 2014. https://doi.org/10.1016/j.coviro.2014.03.010 PMID: 24732439
8. Poirier EZ, Goic B, Tomé-Poderti L, Frangeul L, Boussier J, Gausson V, et al. Dicer-2-Dependent Generation of Viral DNA from Defective Genomes of RNA Viruses Modulates Antiviral Immunity in Insects. Cell Host Microbe. 2018; https://doi.org/10.1016/j.chom.2018.02.001 PMID: 29503180

9. Goic B, Stapleford KA, Frangeul L, Doucet AJ, Gausson V, Blanc H, et al. Virus-derived DNA drives mosquito vector tolerance to arboviral infection. Nat Commun. 2016; https://doi.org/10.1038/ncomms12410 PMID: 27580708

10. Palatini U, Miesen P, Carballar-Lejarazu R, Ometto L, Rizzo E, Tu Z, et al. Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. BMC Genomics. 2017; https://doi.org/10.1186/s12864-017-3903-3 PMID: 28676109

11. Whitfield ZJ, Dolan PT, Kunitomi M, Tassetto M, Seetin MG, Oh S, et al. The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the *Aedes aegypti* Genome. Curr Biol. 2017; https://doi.org/10.1016/j.cub.2017.09.067 PMID: 29129531

12. Miesen P, Joosten J, van Rij RP. PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector Mosquitoes. PLoS Pathog. 2016; 12: 1–17. https://doi.org/10.1371/journal.ppat.1006017 PMID: 28033427

13. Olson KE, Bonizzoni M. Nonretroviral integrated RNA viruses in arthropod vectors: an occasional event or something more? Curr Opin Insect Sci. Elsevier Inc; 2017; 22: 45–53. https://doi.org/10.1016/j.cois.2017.05.010 PMID: 28805638

14. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell. 2007; 128: 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043 PMID: 17346786

15. Petit M, Mongelli V, Frangeul L, Blanc H, Jiggins F, Saleh M-C. piRNA pathway is not required for antiviral defense in *Drosophila melanogaster*. Proc Natl Acad Sci. 2016; https://doi.org/10.1073/pnas.1607952113 PMID: 27357659

16. Akbari OS, Antoshechkin I, Amrhein H, Williams B, Diloreto R, Sandler J, et al. The Developmental Transcriptome of the Mosquito *Aedes aegypti*, an Invasive Species and Major Arbovirus Vector. G3: Genes|Genomes|Genetics. 2013; https://doi.org/doi:10.1534/g3.113.006742

17. Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, Crawford JE, et al. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. Nature. 2018; https://doi.org/10.1038/s41586-018-0692-z PMID: 30429615

18. Miesen P, Ivens A, Buck AH, van Rij RP. Small RNA Profiling in Dengue Virus 2-Infected Aedes Mosquito Cells Reveals Viral piRNAs and Novel Host miRNAs. PLoS Negl Trop Dis. 2016; https://doi.org/10.1371/journal.pntd.0004452 PMID: 26914027

19. Girardi E, Miesen P, Pennings B, Frangeul L, Saleh MC, Van Rij RP. Histone-derived piRNA biogenesis depends on the ping-pong partners Piwi5 and Ago3 in *Aedes aegypti*. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkw1368 PMID: 28115625

20. Miesen P, Girardi E, Van Rij RP. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in *Aedes aegypti* mosquito cells. Nucleic Acids Res. 2015; 43: 6545–6556. https://doi.org/10.1093/nar/gkv590 PMID: 26068474

21. Varjak M, Kean J, Vazeille M, Failloux A, Kohl A. *Aedes aegypti* Piwi4 Is a Noncanonical PIWI Protein Involved in Antiviral Responses. mSphere. 2017; 2: e00144–17. https://doi.org/10.1128/mSphere.00144-17 PMID: 28497119

22. Varjak M, Leggewie M, Schnettler E. The antiviral piRNA response in mosquitoes? 2018; 1–12. https://doi.org/10.1099/jgv.0.001157 PMID: 30372405

23. Schnettler E, Donald CL, Human S, Watson M, Siu RWC, McFarlane M, et al. Knockdown of piRNA pathway proteins results in enhanced semliki forest virus production in mosquito cells. J Gen Virol. 2013; 94: 1680–1689. https://doi.org/10.1099/vir.0.053850-0 PMID: 23559478

24. Hahn MW. Distinguishing among evolutionary models for the maintenance of gene duplicates. Journal of Heredity. 2009. https://doi.org/10.1093/jhered/esp047 PMID: 19596713

25. Obbard DJ, Jiggins FM, Halligan DL, Little TJ. Natural selection drives extremely rapid evolution in antiviral RNAi genes. Curr Biol. 2006; 16: 580–585. https://doi.org/10.1016/j.cub.2006.01.065 PMID: 16546082

26. Bonizzoni M, Gasperi G, Chen X, James AA. The invasive mosquito species *Aedes albopictus*: Current knowledge and future perspectives. Trends in Parasitology. 2013. https://doi.org/10.1016/j.pt.2013.07.003 PMID: 23916878

27. Rezza G. Dengue and chikungunya: long-distance spread and outbreaks in naïve areas. Pathog Glob Health. 2014; https://doi.org/10.1179/2047773214Y.0000000163 PMID: 25491436

28. Bonilauri P, Bellini R, Calzolari M, Angelini R, Venturi L, Fallacara F, et al. Chikungunya virus in *Aedes albopictus*, Italy. Emerging Infectious Diseases. 2008. https://doi.org/10.3201/eid1405.071144 PMID: 18439383

29. Venturi G, Di Luca M, Fortuna C, Elena Remoli M, Riccardo F, Severini F, et al. Detection of a chikungunya outbreak in Central Italy Detection of a chikungunya outbreak in Central. Euro Surveill. 2017; 22: 1–4. https://doi.org/10.2807/1560

30. Gjenero-Margan I, Aleraj B, Krajcar D, Lesnikar V, Klobučar A, Pem-Novosel I, et al. Autochthonous dengue fever in Croatia, August- September 2010. Eurosurveillance. 2011; 16: 1–4. 19805 [pii]

31. Marchand E, Prat C, Jeannin C, Lafont E, Bergmann T, Flusin O, et al. Autochthonous case of dengue in France, October 2013. Eurosurveillance. 2013; https://doi.org/10.2807/1560-7917.ES2013.18.50.20661 PMID: 24342514

32. Delisle E, Rousseau C, Broche B, Leparc-Goffart I, L'Ambert G, Cochet A, et al. Chikungunya outbreak in Montpellier, France, September to October 2014. Euro Surveill. 2015; https://doi.org/10.2807/1560-7917.ES2015.20.17.21108 PMID: 25955774

33. Calba C, Guerbois-Galla M, Franke F, Jeannin C, Auzet-Caillaud M, Grard G, et al. Preliminary report of an autochthonous chikungunya outbreak in France, July to September 2017. Eurosurveillance. 2017; https://doi.org/10.2807/1560-7917.ES.2017.22.39.17–00647

34. Bouri N, Sell TK, Franco C, Adalja AA, Henderson DA, Hynes NA. Return of epidemic dengue in the United States: Implications for the public health practitioner. Public Health Rep. 2012; https://doi.org/10.1177/003335491212700305 PMID: 22547856

35. Joshua-Tor L. The argonautes. Cold Spring Harbor Symposia on Quantitative Biology. 2006. https://doi.org/10.1101/sqb.2006.71.048 PMID: 17381282

36. Innan H, Kondrashov F. The evolution of gene duplications: Classifying and distinguishing between models. Nature Reviews Genetics. 2010. https://doi.org/10.1038/nrg2689 PMID: 20051986

37. Stone SS, Haldar JP, Tsao SC, Hwu W -m. W, Sutton BP, Liang Z-P. Accelerating advanced MRI reconstructions on GPUs. J Parallel Distrib Comput. 2008; 68: 1307–1318. https://doi.org/10.1016/j.jpdc.2008.05.013 PMID: 21796230

38. Yan KS, Yan S, Farooq A, Han A, Zeng L, Zhou M-M. Structure and conserved RNA binding of the PAZ domain. Nature. 2003; https://doi.org/10.1038/nature02129 PMID: 14615802

39. Song JJ, Smith SK, Hannon GJ, Joshua-Tor L. Crystal structure of argonaute and its implications for RISC slicer activity. Science (80-). 2004; https://doi.org/10.1126/science.1102514 PMID: 15284453

40. Cora E, Pandey RR, Xiol J, Taylor J, Sachidanandam R, McCarthy AA, et al. The MID-PIWI module of Piwi proteins specifies nucleotide- and strand-biases of piRNAs. RNA. 2014; https://doi.org/10.1261/rna.044701.114 PMID: 24757166

41. Stein CB, Genzor P, Mitra S, Elchert AR, Ipsaro JJ, Benner L, et al. Decoding the 5′ nucleotide bias of PIWI-interacting RNAs. Nat Commun. Springer US; 2019; 10: 828. https://doi.org/10.1038/s41467-019-08803-z PMID: 30783109

42. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991; https://doi.org/10.1038/351652a0 PMID: 1904993

43. Pischedda E, Scolari F, Valerio F, Carballar-Lejarazú R, Catapano PL, Waterhouse RM, et al. Insights Into an Unexplored Component of the Mosquito Repeatome: Distribution and Variability of Viral Sequences Integrated Into the Genome of the Arboviral Vector *Aedes albopictus*. Front Genet. 2019; https://doi.org/10.3389/fgene.2019.00093 PMID: 30809249

44. Schirle NT, MacRae IJ. The crystal structure of human argonaute2. Science (80-). 2012; https://doi.org/10.1126/science.1221551 PMID: 22539551

45. Matsumoto N, Nishimasu H, Sakakibara K, Nishida KM, Hirano T, Ishitani R, et al. Crystal Structure of Silkworm PIWI-Clade Argonaute Siwi Bound to piRNA. Cell. 2016; https://doi.org/10.1016/j.cell.2016.09.002 PMID: 27693359

46. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 2010; https://doi.org/10.1093/nar/gkq399 PMID: 20478830

47. Macias V, Coleman J, Bonizzoni M, James AA. piRNA pathway gene expression in the malaria vector mosquito *Anopheles stephensi*. Insect Mol Biol. 2014; 23: 579–586. https://doi.org/10.1111/imb.12106 PMID: 24947897

48. Bonizzoni M, Dunn WA, Campbell CL, Olson KE, Marinotti O, James AA. Complex Modulation of the *Aedes aegypti* Transcriptome in Response to Dengue Virus Infection. PLoS One. 2012; https://doi.org/10.1371/journal.pone.0050512 PMID: 23209765

49. Campbell CL, Keene KM, Brackney DE, Olson KE, Blair CD, Wilusz J, et al. *Aedes aegypti* uses RNA interference in defense against Sindbis virus infection. BMC Microbiol. 2008; 8: 1–12. https://doi.org/10.1186/1471-2180-8-1

50. Hess AM, Prasad AN, Ptitsyn A, Ebel GD, Olson KE, Barbacioru C, et al. Small RNA profiling of Dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. BMC Microbiol. 2011; https://doi.org/10.1186/1471-2180-11-45 PMID: 21356105

51. Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci. 2015; https://doi.org/10.1073/pnas.1516410112 PMID: 26483478

52. La Ruche G, Souarès Y, Armengaud A, Peloux-Petiot F, Delaunay P, Desprès P, et al. First two autochthonous dengue virus infections in metropolitan France, september 2010. Eurosurveillance. 2010; 19676 [pii]

53. Schuffenecker I, Iteman I, Michault A, Murri S, Frangeul L, Vaney MC, et al. Genome microevolution of chikungunya viruses causing the Indian Ocean outbreak. PLoS Med. 2006; https://doi.org/10.1371/journal.pmed.0030263 PMID: 16700631

54. Dubrulle M, Mousson L, Moutailier S, Vazeille M, Failloux AB. Chikungunya virus and Aedes mosquitoes: Saliva is infectious as soon as two days after oral infection. PLoS One. 2009; https://doi.org/10.1371/journal.pone.0005895 PMID: 19521520

55. Campbell CL, Black IV WC, Hess AM, Foy BD. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. BMC Genomics. 2008; 9. https://doi.org/10.1186/1471-2164-9-425 PMID: 18801182

56. Arensburger P, Hice RH, Wright JA, Craig NL, Atkinson PW. The mosquito *Aedes aegypti* has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. BMC Genomics. 2011; 12. https://doi.org/10.1186/1471-2164-12-606 PMID: 22171608

57. Yuan JS, Burris J, Stewart NR, Mentewab A, Neal CN. Statistical tools for transgene copy number estimation based on real-time PCR. BMC Bioinformatics. 2007; 8: 1–12. https://doi.org/10.1186/1471-2105-8-1

58. Baruffi L, Damiani G, Guglielmino CR, Bandii C, Malacrida AR, Gasperi G. Polymorphism within and between populations of ceratitis: Comparison between RAPD and multilocus enzyme electrophoresis data. Heredity (Edinb). 1995; 74: 425–437. https://doi.org/10.1038/hdy.1995.60 PMID: 7759289

59. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; https://doi.org/10.1093/nar/gkh340 PMID: 15034147

60. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; Available: https://arxiv.org/abs/1303.3997

61. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing—Free bayes—Variant Calling—Longranger. arXiv Prepr arXiv12073907. 2012; arXiv:1207.3907 [q-bio.GN]

62. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; https://doi.org/10.4161/fly.19695 PMID: 22728672

63. Schroeder MP, Lopez-Bigas N. muts-needle-plot: Mutations Needle Plot v0.8.0. 2015; https://doi.org/10.5281/ZENODO.14561

64. Oliveros JC. Venny. An interactive tool for comparing lists with Venn's diagrams. https://bioinfogp.cnb.csic.es/tools/venny/index.html. 2016;

65. Flot JF. Seqphase: A web tool for interconverting phase input/output files and fasta sequence alignments. Mol Ecol Resour. 2010; https://doi.org/10.1111/j.1755-0998.2009.02732.x PMID: 21565002

66. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001; https://doi.org/10.1086/319501 PMID: 11254454

67. Stephens M, Scheet P. Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. Am J Hum Genet. 2005; https://doi.org/10.1086/428594 PMID: 15700229

68. Librado P, Rozas J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325

69. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;

70. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975; https://doi.org/10.1016/0040-5809(75)90020-9

71. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;

72. Abascal F, Zardoya R, Telford MJ. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 2010; https://doi.org/10.1093/nar/gkq291 PMID: 20435676

73. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; https://doi.org/10.1093/nar/22.22.4673 PMID: 7984417

74. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; https://doi.org/10.1093/molbev/msm088 PMID: 17483113

75. Xu B, Yang Z. PamlX: A graphical user interface for PAML. Mol Biol Evol. 2013; https://doi.org/10.1093/molbev/mst179 PMID: 24105918

76. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014; https://doi.org/10.1038/ng.3036 PMID: 25017105

77. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, Mcewen R, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016; https://doi.org/10.1093/nar/gkw227 PMID: 27060149

78. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; https://doi.org/10.1101/gr.107524.110 PMID: 20644199

79. Rstudio Team. RStudio: Integrated Development for R. [Online] RStudio, Inc., Boston, MA. 2016. https://doi.org/10.1007/978-81-322-2340-5

80. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; https://doi.org/10.1016/S0022-2836(05)80360-2

81. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018; https://doi.org/10.1016/j.jmb.2017.12.007 PMID: 29258817

82. Webb B, Sali A. Protein structure modeling with MODELLER. Methods in Molecular Biology. 2017. https://doi.org/10.1007/978-1-4939-7231-9_4 PMID: 28986782

83. Schirle NT, Sheu-Gruttadauria J, Chandradoss SD, Joo C, MacRae IJ. Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. Elife. 2015; https://doi.org/10.7554/elife.07646 PMID: 26359634

84. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr Sect D Biol Crystallogr. 2010; https://doi.org/10.1107/s0907444910007493 PMID: 20383002

85. Adams PD, Afonine P V., Bunkóczi G, Chen VB, Echols N, Headd JJ, et al. The Phenix software for automated determination of macromolecular structures. Methods. 2011. https://doi.org/10.1016/j.ymeth.2011.07.005 PMID: 21821126

86. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 2018; https://doi.org/10.1002/pro.3330 PMID: 29067766

87. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009; https://doi.org/10.1093/nar/gkp322 PMID: 19429685

88. Madeira F, Park Y M, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;

89. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. 2014; https://doi.org/10.1093/nar/gku316 PMID: 24753421

90. Schrödinger L. The PyMOL molecular graphics system, version 1.8. https://www.pymol.org/citing. 2015;

91. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011. https://doi.org/10.1038/nbt.1754 PMID: 21221095

92. A. Reynolds J, Poelchau MF, Rahman Z, Armbruster PA, Denlinger DL. Transcript profiling reveals mechanisms for lipid conservation during diapause in the mosquito, *Aedes albopictus*. J Insect Physiol. 2012; https://doi.org/10.1016/j.jinsphys.2012.04.013 PMID: 22579567

93. Khan A, Rayner GD. Robustness to non-normality of common tests for the many-sample location problem. J Appl Math Decis Sci. 2004; https://doi.org/10.1155/S1173912603000178

94. Blanca MJ, Alarcón R, Arnau J, Bendayan R. Non-normal data: Is ANOVA still a valid option? Psicothema. 2017; https://doi.org/10.7334/psicothema2016.383 PMID: 29048317