bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Genomic Analysis of Two Phlebotomine Sand Fly Vectors of *Leishmania* from the New and Old World

Short title: Sand fly genomics

Authors

Frédéric Labbé[1,2], Maha Abdeladhim[3], Jenica Abrudan[4], Alejandra Saori Araki[5], Ricardo N. Araujo[6], Peter Arensburger[7], Joshua B. Benoit[8], Reginaldo Pecanha Brazil[9], Rafaela V. Bruno[5], Gustavo Bueno da Silva Rivas[5,10], Vinicius Carvalho de Abreu[11], Jason Charamis[12,13], Iliano V. Coutinho-Abreu[14], Samara G. da Costa-Latgé[5], Alistair Darby[15], Viv M. Dillon[15], Scott Emrich[16], Daniela Fernandez-Medina[17], Nelder Figueiredo Gontijo[6], Catherine M. Flannley[1], Derek Gatherer[18], Fernando A. Genta[5], Sandra Gesing[19], Gloria I. Giraldo-Calderón[1,20], Bruno Gomes[5], Eric Roberto Guimaraes Rocha Aguiar[11], Omar Hamarsheh[21], Mallory Hawksworth[1], Jacob M. Hendershot[8], Paul V. Hickner[22], Jean-Luc Imler[23], Panagiotis Ioannidis[13], Emily C. Jennings[8], Shaden Kamhawi[3], Charikleia Karageorgiou[13,24], Ryan C. Kennedy[1], Andreas Krueger[25,26], José M Latorre-Estivalis[27], Petros Ligoxgakis[28], Antonio Carlos A. Meireles-Filho[5], Jose Carlos Miranda[29], Michael Montague[30], Ronald J. Nowling[31], Fabiano Oliveira[3], João Ortigão-Farias[32], Marcio G. Pavan[5,33], Marcos Horacio Pereira[6], Andre Nobrega Pitaluga[34], Roenick Proveti Olmo[11], Marcelo Ramalho-Ortigao[35], Jose Marcos Ribeiro[3], Andrew J. Rosendale[10], Mauricio R.V. Sant'Anna[6], Steven E. Scherer[36], Nagila F. C. Secundino[37], Douglas A. Shoue[1], Caroline da Silva Moraes[5], João Silveira Moledo Gesto[5], Nataly Araujo Souza[38], Zainulabueddin Syed[39], Samuel Tadros[1], Rayane Teles-de-Freitas[5], Erich L. Telleria[32,40], Chad Tomlinson[41], Yara Traub-Cseko[32], João Trindade Marques[10], Zhijian Tu[42], Maria F Unger[43], Jesus Valenzuela[3], Flavia Viana Ferreira[44], Karla Pollyanna Vieira de Oliveira[11], Felipe M Vigoder[45], John Vontas[13,46], Lihui

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Wang[29], Gareth Weedel[47,48], Elyes Zhioua[49], Stephen Richards[36], Wesley C Warren[50], Robert M. Waterhouse[51], Rod J. Dillon[18], and Mary Ann McDowell*[1].

Affiliations

[1]Eck Institute for Global Health, Department of Biological Sciences, University of Notre dame, Notre Dame, IN 46556, USA; [2]Current Address: CIRAD, UMR PVBMT, F-97410 Saint Pierre, La Réunion, France; [3]Vector Molecular Biology Section, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland, USA; [4]Genomic Sciences & Precision Medicine Center (GSPMC), Medical College of Wisconsin, 8701 W. Watertown Plank Road, Milwaukee, WI 53226 USA; [5]Laboratório de Bioquímica e Fisiologia de Insetos, IOC, FIOCRUZ, Rio de Janeiro, Brazil; [6]Laboratório de Fisiologia de Insetos Hematófagos, Universidade Federal de Minas Gerais, Instituto de Ciencias Biológicas, Departamento de Parasitologia, Pampulha, Belo Horizonte, Minas Gerais, Brazil; [7]Department of Biological Sciences, California State Polytechnic University, Pomona, 3801 West Temple Avenue, Pomona, CA 91768, USA; [8]Department of Biological Sciences, University of Cincinnati, 318 College Drive, Cincinnati, OH 45221-0006, USA; [9]Laboratório de Doenças Parasitárias, Instituto Oswaldo Cruz, Rio de Janeiro, Brazil; [10]Department of Biology and Center for Biological Clocks Research, Texas A&M University, College Station, TX 77843 USA; [11]Department of Biochemistry and Immunology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, CEP 30270-901, Brazil; [12]Department of Biology, University of Crete, Voutes University Campus, GR-714 09, Heraklion, Crete, Greece; [13]Molecular Entomology Lab, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas (FORTH), N. Plastira 100, GR 700-13, Vassilika Vouton, Heraklion, Crete, Greece; [14]Division of Biological Sciences, Section of Cell and Developmental Biology, University of California, San Diego, La Jolla, CA 92093, USA; [15]Institute of Integrative Biology, The University of Liverpool, Liverpool, L69 7ZB UK [16]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA; [17]School of Applied Mathematics, Getulio Vargas Foundation, Brazil, Rio de Janeiro, Brazil; [18]Division of Biomedical & Life Sciences, Faculty of Health & Medicine, Lancaster University, Lancaster, LA1 4YT, UK; [19]Discovery Partners Institute, University of Illinois Chicago, Chicago, IL, USA; [20]Dept. Ciencias Biológicas & Dept.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Ciencias Básicas Médicas, Universidad Icesi, Calle 18 No 122-135, Cali, Colombia; [21]Department of Life Sciences, Faculty of Science and Technology, Al-Quds University, Jerusalem, Palestine. [22]USDA-ARS Knipling-Bushland U.S. Livestock Insects Research Laboratory and Veterinary Pest Genomics Center, Kerrville, TX, USA; [23]CNRS-UPR9022 Institut de Biologie Moléculaire et Cellulaire and Faculté des Sciences de la Vie-Université de Strasbourg, Strasbourg Cedex, France; [24]Genomics Group – Bioinformatics and Evolutionary Biology Lab, Department of Genetics and Microbiology, Autonomous University of Barcelona, Barcelona, Spain; [25]Medical Entomology Branch, Dept. Microbiology, Bundeswehr Hospital, Bernhard-Nocht-Str. 74, 20359 Hamburg, Germany; [26]Medical Zoology Branch, Dept. Microbiology, Central Bundeswehr Hospital, Andernacher Str. 100, 56070 Koblenz, Germany; [27]Laboratorio de Insectos Sociales, Instituto de Fisiología, Biología Molecular y Neurociencias, Universidad de Buenos Aires - CONICET, Buenos Aires, Argentina; [28]Laboratory of Cell Biology, Development and Genetics, Department of Biochemistry, University of Oxford, Oxford, United Kingdom; [29]Laboratório de Imunoparasitologia, CPqGM, Fundação Oswaldo Cruz, Bahia, Brazil; [30]Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; [31]Department of Electrical Engineering and Computer Science, Milwaukee School of Engineering, Milwaukee, WI, 53202, USA; [32]Instituto Oswaldo Cruz - Fiocruz, Av. Brasil, 4365, Pav. Leonidas Deane, lab 304, Manguinhos - Rio de Janeiro - RJ - Brazil, CEP: 21040-360; [33]Laboratório de Transmissores de Hematozoários, IOC, FIOCRUZ, Rio de Janeiro, Brazil; [34]Laboratório de Biologia Molecular de Parasitas e Vetores, Instituto Oswaldo Cruz/FIOCRUZ, Rio de Janeiro, Brazil; [35]F. Edward Hebert School of Medicine, Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences (USUHS), Bethesda, MD, 20814, USA; [36]Human Genome Sequencing Center, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030, USA; [37]Laboratory of Medical Entomology, René Rachou Institute-FIOCRUZ, Minas Gerais, Brazil; [38]Laboratory Interdisciplinar em Vigilancia Entomologia em Diptera e Hemiptera, Fiocruz, Rio de Janeiro, Brazil; [39]Department of Entomology, University of Kentucky, Lexington, KY 40546, USA; [40]Department of Parasitology, Faculty of Science, Charles University, Vinicna 7, 128 44, Prague, Czech Republic; [41]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108 USA; [42]Fralin Life Science Institute and Department of Biochemistry, Virginia Tech, 303 Fralin, Blacksburg, VA, 24061, USA; [43]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Dame, IN 46556, USA; [44]Department of Microbiology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, CEP 30270-901, Brazil; [45]Universidade Federal do Rio de Janeiro, Instituto de Biologia. Avenida Carlos Chagas Filho, 373, Block A / A2-075.Cidade Universitária /CCS - Centro de Ciências da Saúde, 21941902 - Rio de Janeiro, Brazil; [46]Pesticide Science Lab, Department of Crop Science, Agricultural University of Athens, Iera Odos 75, GR-118 55, Athens Greece; [47]Vector Biology Department, Liverpool School of Tropical Medicine (LSTM), Pembroke Place, Liverpool, United Kingdom; [48]School of Biological and Environmental Sciences, Liverpool John Moores University, Liverpool, United Kingdom; [49]Vector Ecology Unit, Institut Pasteur de Tunis, Tunis, Tunisia; [50]Department of Animal Sciences, Department of Surgery, Institute for Data Science and Informatics, University of Missouri, Columbia, MO 63201; [51]Department of Ecology & Evolution and Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland

*Correspondence: mcdowell.11@nd.edu

Abstract

Phlebotomine sand flies are of global significance as important vectors of human disease, transmitting bacterial, viral, and protozoan pathogens, including the devastating kinetoplastid parasites of the genus *Leishmania*, the causative agents of diseases collectively termed leishmaniasis. More than 40 pathogenic *Leishmania* species are transmitted to humans by approximately 35 sand fly species in 98 countries with hundreds of millions of people at risk around the world. As no approved efficacious vaccine exists, available drugs are expensive and/or toxic, and resistance is emerging, management of sand fly populations to break transmission is currently the most effective disease control strategy. To better understand the biology of sand flies, including the mechanisms involved in their vectorial capacity, insecticide resistance, and population structures we sequenced the genomes of two of the most important sand fly species: *Phlebotomus papatasi*, a cutaneous leishmaniasis vector, (distributed in the Middle East and

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

North Africa) and *Lutzomyia longipalpis,* a visceral leishmaniasis vector (distributed across Central and South America). We categorized and curated genes involved in processes important to their roles as disease vectors, including chemosensation, blood feeding, circadian rhythm, immunity, and detoxification, as well as mobile genetic elements. We also defined gene orthology and observed micro-synteny among the genomes. Finally, we present the genetic diversity and population structure of these species in their respective geographical areas. These genomes will be a foundation on which to base future efforts to prevent vector-borne transmission of *Leishmania* parasites.

Author Summary

The leishmaniases are a group of neglected tropical diseases caused by protist parasites from the Genus *Leishmania*. Different *Leishmania* species present a wide clinical profile, ranging from mild, often self-resolving cutaneous lesions that can lead to protective immunity, to severe metastatic mucosal disease, to visceral disease that is ultimately fatal. *Leishmania* parasites are transmitted by the bites of sand flies, and as no approved vaccine exists, available drugs are toxic and/or expensive and resistance is emerging, new dual control strategies to combat these diseases must be developed, combining interventions on human infections and integrated sand fly population management. Effective vector control requires a good understanding of the biology of sand flies. To this end, we sequenced and annotated the genomes of two sand fly species that are important leishmaniasis vectors from the Old and New Worlds. These genomes allow us to better understand, at the genetic level, processes important in the vector biology of these species, such as finding hosts, blood-feeding, immunity, and detoxification. These genomic resources highlight the driving forces of evolution of two major *Leishmania* vectors and provide foundations for future research on how to better prevent leishmaniasis by control of the sand fly vectors.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Introduction

Phlebotomine sand flies are a group of vectors that vary widely in their geographic distribution, ecology, and the pathogens they transmit. Sand flies serve as vectors for several established, emerging, and re-emerging infectious diseases, transmitting protist, bacterial and viral pathogens. The most devastating of the sand fly transmitted diseases is leishmaniasis, accounting for an estimated 2.4 million disability-adjusted life-years (DALYs) [1] and 40,000 deaths annually [2]. These statistics are largely underestimated due to misdiagnosis and lacking surveillance systems in many of the affected countries. Political instability, urbanization, and climate change are expanding *Leishmania*-endemic regions and increasing the propensity for epidemics world-wide [3]. These factors are coupled with the increase of visceral disease due to HIV co-infection, spurring the World Health Organization to classify leishmaniasis as one of the world's epidemic-prone diseases [4].

Leishmaniasis is worldwide, occurring in 98 countries and spread over five continents, putting 310 million people at risk of contracting leishmaniasis [2]. Leishmaniasis is a collective term for a group of distinct clinical manifestations ranging from mild, often self-resolving cutaneous lesions that can lead to protective immunity, to disseminated lesions that do not heal spontaneously, to destruction of the mucous membranes of the nose, mouth, and pharynx, to life-threatening visceral disease. The clinical profile depends on a variety of factors, including vector biology, host immunity, and parasite characteristics; the *Leishmania* species initiating the infection being the primary determinant. The two primary clinical forms are cutaneous leishmanisis (CL) and visceral leishmaniasis (VL). The primary *Leishmania* species that cause CL are *Leishmania major*, *Leishmania infantum, Leishmania tropica,* and *Leishmania aethipica* in the Old World and *Leishmania amazonensis, Leishmania braziliensis, Leishmania guyanensis, Le. infantum, Leishmania mexicana,* and *Leishmania panamensis* in the New World. VL is primarily caused by *Leishmania donovani* in Asia and Africa and *Le. infantum* in the Middle East, central Asia, South and Central America, and the Mediterranean Basin.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

There are approximately 35 proven, and an additional 63 suspected, vectors of at least 40 different *Leishmania* species to humans [5, 6]. *Phlebotomus* species are the primary *Leishmania* vectors in the Old World and *Lutzomyia* species are responsible for transmitting leishmaniasis throughout the Americas [7]. There is a close ecological association, if not co-evolutionary relationship [8, 9], between *Leishmania* species and their specific vectors such that generally a single sand fly species transmits a single *Leishmania* species under natural conditions. Some sand flies, however, can transmit a range of *Leishmania* species under experimental conditions [10]. This difference has given rise to the concept of “restricted” and “permissive” vectors [11]. For example, *Phlebotomus papatasi* is a restrictive vector, transmitting only *Le. major* parasites [12]. *Lutzomyia longipalpis* is considered a permissive vector in laboratory conditions, but only transmits *Le. infantum* naturally [12].

These vectors are part of the Diptera which is an extremely species-rich and ecologically diverse order of insects and contains the vectors of many of the most important pathogens of man and his domesticated animals. Both phlebotomine sand flies (family Psychodidae) and mosquitoes (Culicidae) are specified as members of distinct infra-orders within the suborder Nematocera. While the Nematocera grouping is paraphyletic, the relationships between infra-orders remains to be elucidated [13]. Some studies generated topologies with Psychodomorpha (sand flies) and Culicomorpha (mosquitoes and black flies) as sister groups [14], whereas, others place sand flies nearer to the muscoid flies (Ephydroidea) [15]. The internal relationships within the assemblage that includes Psychodidae also remains a matter of debate [16].

It is postulated that this close evolutionary relationship between sand fly species and the *Leishmania* species that they transmit may have epidemiological implications for leishmaniasis [17]. For example, there are three primary zymodemes of *Le. major* that have limited geographical distributions such that the prevalent zymodeme in a particular area overlaps with the distribution of one primary population of *Ph. papatasi* [18]. *Ph. papatasi* has a wide geographical distribution, ranging from Morocco to the Indian subcontinent and from southern Europe to central and eastern

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Africa. Given the wide ecological and geographic distribution of *Ph. papatasi* populations [19], coupled with the low dispersal capacity of these sand flies [12], it is likely that there is limited gene flow between populations and significant genetic structuring between populations. While previous studies demonstrated relatively low genetic differentiation between *Ph. papatasi* populations separated by large geographical distances [9, 20], more recent studies have identified genetic differentiation between geographically separated populations [18, 21-24] and local differentiation [25]. Microsatellite analysis, in particular, revealed two distinct genetic clusters of *Ph. papatasi* (A & B) with further substructure within each population that correlated with geographical origin (A1-5 and B1 &2) [18, 23].

While elucidating the drivers leading to reproductive isolation and speciation remains a challenge, there is strong evidence that *Lu. longipalpis* is undergoing incipient speciation in Brazil with various levels of differentiation between siblings of the complex [26]. The Brazilian populations of *Lu. longipalpis* can be divided into two primary groups. The males of one group produce Burst-type mating songs and a diterpenoid-like pheromone, Cembrene-1, recently reclassified as sobralene (pentadeca-E-3,4-Z-8,9-triene) [27]. The second group is more heterogeneous, with populations producing different subtypes of Pulse-type copulation songs and different sex pheromones [C_{16} H_{32} types Himachalene (Hima), Germacrene (Germ), and Sobralene [28]. Crossing experiments of sympatric populations of the two different types revealed reproductive isolation due to both pre-mating and copulatory mechanisms [29]. Hickner *et al.* 2020 provided genomic insights into the chemoreceptor genome repertoire underlying behavioral evolution of sexual communication in the *Lu. longipalpis* populations, but whole-genome analyses could improve the identification of loci related to critical traits such as vectorial capacity, host preference, and insecticide resistance.

Despite the potential importance of the sand fly immune system for influencing *Leishmania* development and survival in the sand fly gut the immune response is poorly studied. To date, this work has been largely restricted to defensins [30-33]. However, gene depletion via RNAi of the

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

negative regulator of the IMD pathway caspar [34] led to a reduction in *Leishmania* population in the gut of *Lu. longipalpis*. While the knockout of relish, the transcription factor of IMD pathway, resulted in the increase of *Leishmania* and bacteria in *Ph. papatasi* [35].

Adaptation to hematophagy presents many challenges to insects, including avoiding the physiological responses of the host that interfere with obtaining a blood meal, digestion of the blood, and excretion of the excess water contained in the blood meal. Sand flies have evolved a complex cocktail of pharmacologically active salivary molecules to facilitate blood feeding that have been extensively characterized [36]. Although hematophagy and host seeking are controlled by the circadian clock, the molecular regulation of circadian rhythms is poorly understood in sand flies. However, Yuan *et al.* 2007 proposed three clock models based on the presence of the cryptochrome (CRY) proteins, CRY1 and CRY2 [37]. In the *Drosophila* clock model, only CRY1, which acts as a blue-light photoreceptor [38], is present. In the butterfly model, CRY1 also acts as a photoreceptor and CRY2, which is a mammalian–like transcriptional repression, dimers with PER to repress CLK/CYC activity. In the bee model, there is only CRY2, which seems to act as a repressor together with PER and some other molecule that is not CRY1 that acts as photoreceptor.

A central inquiry of evolutionary biology is the elucidating drivers of speciation, however, defining species boundaries and identifying the genetic architecture that leads to reproductive isolation has been a challenge. Understanding of the mechanisms of vectorial capacity, adaptation to changing ecological environments, and insecticide resistance has epidemiological consequences for the integrated management of sand fly populations that is the cornerstone of leishmaniasis control [39]. To explore the driving forces of evolution of two major phlebotomine sand flies from the *Psychodidae* family (*Phlebotominae* subfamily), *Ph. papatasi* and *Lu. longipalpis*, that exhibit distinct distributions, behavior, and pathogen specificity, we sequenced and analyzed their whole-genomes using comparative genomics approaches. We manually curated a number of gene families with key roles in processes such as immunity, blood-feeding,

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

chemosensation, detoxification, and circadian biology to provide a basis for studying and understanding sand flies as *Leishmania* vectors. Moreover, as a better understanding of the population structure of geographically separated vector populations is necessary, we also assessed the population structure of *Ph. papatasi* and *Lu. Longipalpis* by collecting and sequencing individual field-collected specimens sampled over a large geographical range in the Middle East and North Africa, and Brazil, respectively. Our results provide significant advances in our understanding of the genetics underlying the population structure and provide a great foundation for future molecular comparative studies of these two medically important vectors.

Methods

Fly Maintenance

Phlebotomus papatasi

To avoid confounding effects due to genetic polymorphisms, we used an inbred colony of *Ph. papatasi* (Israeli strain) for the genome assembly. This colony was originally established in the 1970s and given to Walter Reed Army Institute of Research (WRAIR) in 1983 from the Hebrew University, Jerusalem and transferred to the University of Notre Dame in 2006. At several times since establishment in the laboratory, the colony has fluctuated in population size and has been expanded from a relatively low number of files. Sand flies were reared by the method of Modi and Tesh [40].

Lutzomyia longipalpis

Lu. longipalpis Jacobina strain was used for the genome assembly. This colony was originally established at the Liverpool School of Tropical Medicine by Richard Ward in 1988 from flies caught in Jacobina, Bahia State, Brazil. At several times since establishment in the laboratory, the colony has fluctuated in population size and has been expanded from a small number of flies.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Flies were reared under standardized laboratory conditions [40], *i.e.* under controlled temperature (27 ± 2°C), humidity (>80%), and photoperiod (8 hours light/16 hours darkness) [40].

Ethics Statement

The study protocol was approved by the Institutional Animal Care and Use Committee at the University of Notre Dame (#07-052).

Field Collections

Phlebotomus papatasi

Ph. papatasi were collected from three different locations: Tunisia, Egypt, and Afghanistan. In Tunisia, samples were collected in 2013 from the village of Felta located in an arid biogeographical area in Central Tunisia (35°16'N, 9°26'E). In North Sinai Egypt, samples were collected in Om Shikhan (30°50'N, 34°10'E), located approximately 340 km east of Cairo, 80 km inland from the Mediterranean coast, and 30 km west of the Israeli border in 2007. In Afghanistan, samples were collected in 2010 in and around a German military camp located near the airport of Mazar-e Sharif (36°43'N, 67°14'E). This site is located at 400m altitude north of the Hindukush Mountains and approximately 50km south of the Uzbekistan border. Sand flies were trapped using CDC-style light traps between 17:00 and 07:00.

Lutzomyia longipalpis

Lu. longipalpis were collected in 2014 from six different locations in Brazil (Fig. 1). Samples were collected from three allopatric populations: Jacobina, Bahia State (11º10'S 40º31'W), Lapinha Cave, Minas Gerais State (19º38'S 43º53'W), Marajó Island, Pará State (0°56'S 49°38'W), and two sympatric populations (called 1S and 2S) from Sobral, Ceará State (34º41'S 40º20'W). For comparison of male copulatory courtship songs, flies were also collected from Olindina (11º 29' S 38° 22' W) and Araci (11º 09' S 39° 01' W), sites near Jacobina. Sand flies were trapped using CDC-style light traps baited with CO_2 between 18:00 and 06:00 and transported to the laboratory. Analyses of male copulatory courtship songs was carried out by as previously described [41]. The recordings were performed by using males and females from

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

laboratory colonies established from wild-collected flies from Lapinha and Sobral and from Araci and Olindina.

Figure 1. *Lutzomyia longipalpis* site locations for copulation songs and pheromone types. Samples were collected from three allopatric populations: Marajó (Pará State; 0°56'S 49°38'W), Jacobina (Bahia State; 11º 10'S 40º 31'W), and Lapinha Cave (Minas Gerais State; 19º 33'S 43º 57'W); and two sympatric populations (1S and 2S) from Sobral (Ceará State; 3º 41'S 40º 21'W). Copulation songs: Burst-type (B) and Pulse-types (P1, P2, and P3). Pheromone types: cembrene (Cemb), 9-methylgermacrene-B (Germ), and himachalene (Hima). For each location, the number of single nucleotide polymorphisms (SNPs) identified in each population with respect to the reference genome (VectorBase, LlonJ1) is indicated. A total of 1,937,819 SNPs were identified among all the populations. Map Source: Google Imagery TerraMetrics.

Nucleic Acid Isolation

Genomic DNA from female sand flies was isolated from pools of flies or from single insects for population genetics analysis. For pooled insects, DNA was suspended in 50 µl of the hydration solution using a Tissue DNA isolation kit (GE HealthCare LifeSciences).

Total RNA was obtained from the four larval instars, pupae, and adult males and females one, three, and ten days post emergence (adults fed only on sugar). Additional RNA was included from *Ph. papatasi* females six, 24, and 96 hours post feeding on uninfected and *Le. major* (strain Friedlin V1) infected mouse blood and *Lu. longipalpis* females 6, 24, and 144 hours post feeding on uninfected and *Le. infantum* infected mouse blood. Total RNA was extracted using a RNAeasy Mini Kit (Qiagen).

Genome Sequencing and Assembly

Phlebotomus papatasi

Sequencing and assembly for *Ph. papatasi* were performed by the Genome Institute, Washington University School of Medicine. The assembly was built with the –het option, using

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

the Newbler assembler test release 2.6RC02 from an input of ~22.5X total sequence coverage with Sanger and 454 reads including 15.1X of whole-genome shotgun reads, 4.4X 3 kb clone inserts, 3.0X 8 kb inserts and 0.01X BAC-end read pairs. The fragment and 3 kb data were generated from a single sand fly after whole-genome amplification, while the 8kb data were derived from multiple flies. The 0.1X of Sanger 3,730 BAC end sequences (28,902 reads) were also derived from multiple flies.

Prior to submission to NCBI, this assembly was screened for contamination and 247 contigs were removed. Additional contigs were removed or merged by the McDonnell Genome Institute Washington University School of Medicine, PolyGraph Assembler, which collapses heterozygous contigs and reduced the assembled genome size from 364 Mb to 345 Mb. A total of 5,661 gaps were closed and nearly 6.8 Mb of sequence was added by another in house post-assembly program, PyGap. This program detects and merges overlaps of adjoining contigs and attempts to close gaps between non-overlapping adjoining contigs with Illumina data. The same Illumina data used in gap closure was aligned to the assembly to correct 89,378 presumed 454 insertion/deletion errors.

Lutzomyia longipalpis

Three types of *Lu. longipalpis* whole-genome shotgun (WGS) libraries were used: a 454 Titanium fragment library and paired end libraries generated from 3 kb and 8 kb inserts. The 454 data (11.5 million reads; ~24.4X coverage) was derived from the same individual while mate pair reads (7.4 million 3kb reads, 9.6X; 3.7 million 8kb reads, 4.9X) were derived from a pool of individuals. In total, approximately 22.6 million reads were generated at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) using the Celera CABOG assembler (version 6.1, 2010/03/22) and represents 38.9X coverage of this sand fly genome. These initial results were used as a backbone for longer superscaffolds using ATLAS-link [42]. Finally, discernible gaps were filled (see [43]) with ATLAS-gapfill. The total length of all contigs is 142.7 Mb; however, the total span of the assembly is 154.2 Mb after gaps are included.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Individual population sequencing

To prepare short insert libraries, an Illumina gel-cut paired-end library protocol was used. Briefly, DNA was extracted from individual adult males or females from inbred lines using the Qiagen DNAeasy Blood and Tissue kit following the manufacturer's supplementary protocol for purification of total DNA from insects. Purified DNA was sheared using a Covaris S-2 system (Covaris, Inc. Woburn, MA). Sheared DNA fragments were purified with Agencourt AMPure XP beads, end-repaired, dA-tailed, and ligated to Illumina universal adapters. After adapter ligation, DNA fragments were further size-selected by agarose gel elution and PCR amplified for 6 to 8 cycles using Illumina P1 and Index primer pair and Phusion® High-Fidelity PCR Master Mix (New England Biolabs). The final library was purified using Agencourt AMPure XP beads and quality assessed by Agilent Bioanalyzer 2100 (DNA 7500 kit) determining library quantity and fragment size distribution before sequencing. Sequencing was performed on an Illumina HiSeq2000 platform generating 100 bp paired end reads. Sequenced reads sequence reads were deposited in the NCBI SRA under Bioproject accession number PRJNA20279 for *Lu. longipalpis* and PRJNA20293 for *Ph. papatasi*.

RNA-sequencing

RNA-sequencing (RNAseq) was conducted to improve resources available for gene prediction. RNAseq was performed following standard protocols on an Illumina HiSeq 2000 platform. RNA was obtained from both sexes, during development, and following blood feeding to generate an extensive RNA-seq coverage to allow for quality gene prediction. Briefly, poly-A^+ messenger RNA (mRNA) was extracted from 1 µg total RNA using Oligo(dT)25 Dynabeads (Life Technologies, cat. no. 61002) followed by fragmentation of the mRNA by heating to 94°C for 3 min [for samples with RNA Integrity Number (RNI) = 3–6] or 4 min (for samples with RIN of >6.0). First-strand complementary DNA (cDNA) was synthesized using the Superscript III reverse transcriptase (Life Technologies, cat. no. 18080-044) and purified using Agencourt RNAClean XP beads (Beckman Coulter, cat. no. A63987). During second-strand cDNA synthesis,

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

deoxynucleoside triphosphate (dNTP) mix containing deoxyuridine triphosphate was used to introduce strand specificity. For Illumina paired-end library construction, the resultant cDNA was processed through end repair and A-tailing, ligated with Illumina PE adapters, and digested with 10 units of uracil–DNA glycosylase (New England Biolabs, Ipswich, MA; cat. no. M0280L). Amplification of the libraries was performed for 13 PCR cycles using the Phusion High-Fidelity PCR Master Mix (New England Biolabs, cat. no. M0531L); 6-bp molecular barcodes were also incorporated during this PCR amplification. These libraries were then purified with Agencourt AMPure XP beads after each enzymatic reaction, and after quantification using the Agilent Bioanalyzer 2100 DNA Chip 7500 (cat. no. 5067-1506), libraries were pooled in equimolar amounts for sequencing. Sequencing was performed on Illumina HiSeq2000s, generating 100-bp paired-end reads. Sequenced reads were deposited in the NCBI SRA, under BioProject accession PRJNA81043 for *Lu longipalpis* and PRJNA20293 of *PH. papatasi*.

Annotation

The genome assemblies were initially annotated *ab initio* with gene models derived from VectorBase annotation MAKER2 [44] pipelines [45]. The automated analyses identified 12,678 gene models for *Ph. papatasi* and 10,429 for *Lu. longipalpis*. Expert curators manually annotated several gene families of interest (Supplementary Methods) resulting in 11,849 and 10,796 gene models for *Ph. papatasi* and *Lu. longipalpis*, respectively.

Orthology Delineation

OrthoDB [46] orthology delineation was employed to define orthologous groups of genes descended from each last common ancestor of the species phylogeny across 43 insects including the two sand flies - Hemipterodea: *Pediculus humanus* and *Rhodnius prolixus*; Hymenoptera: *Apis mellifera* and *Linepithema humile*; Coleoptera: *Tribolium castaneum*; Lepidoptera: *Bombyx mori* and *Danaus plexippus*; Diptera: *Lu. longipalpis*, *Ph. papatasi* and *Glossina morsitans*, 12 *Drosophila* species (*D. grimshawi, D. mojavensis, D. virilis, D. willistoni, D. persimilis, D. pseudoobscura, D. ananassae, D. erecta, D. yakuba, D. melanogaster, D. sechellia, and D.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

simulans), two culicine mosquitoes (*Aedes aegypti* and *Culex quinquefasciatus*), and 19 *Anopheles* species (*An. darlingi, An. albimanus, An. sinensis, An. atroparvus, An. farauti, An. dirus, An. funestus, An. minimus, An. culicifacies, An. maculatus, An. stephensi* (SDA-500), *An. stephensi* (INDIAN), *A. epiroticus, An. christyi, An. melas, An. quadriannulatus, An. arabiensis, An. merus,* and *An. gambiae* (PEST). The orthology delineation was performed as part of the *Anopheles* Genomes Cluster Consortium analyses of 16 newly-sequenced *Anopheles* mosquitoes [47, 48]. From the complete set of species, the two sand flies were compared to a symmetrical set of five representative mosquitoes and five representative flies, together with four outgroup species representing four insect orders. The species compositions of all orthologous groups defined at the dipteran root were analyzed with custom Perl scripts to count the numbers of groups and genes shared among the sand flies, mosquitoes, and flies. Pairwise percent amino acid identities between single-copy and/or multi-copy orthologs among the sand flies, *An. gambiae* and *D. melanogaster* were extracted from all-against-all protein sequence comparisons performed with SWIPE [49] as part of the OrthoDB orthology delineation procedure.

Maximum likelihood species phylogeny

To establish species relationships, the maximum likelihood species phylogeny was determined from concatenated protein sequence alignments [aligned with default MUSCLE [50] parameters and trimmed with the 'automated1' trimAl [51] setting of 1,627 relaxed single-copy orthologs (no more than three paralogs in up to three species, longest protein selected)] from the two sand flies, five mosquitoes, five flies, and four outgroup insect species. These orthologs were selected from a total of 2,160 orthologous groups and were each required to have an alignment of more than 50 amino acid columns after trimming and a relative tree certainty (see [52]) of more than 50% as implemented in RAxML [53]. The concatenated alignment contained 1,065,440 amino acid columns with 627,808 distinct alignment patterns and was used to estimate the maximum likelihood species phylogeny with RAxML [54] employing the PROTGAMMAJTT model over 100 bootstrap samples and setting *Pe. humanus* as the outgroup species. The RAxML

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

phylogenies of individual ortholog groups were analyzed with custom Perl scripts and the Newick Utilities [55] to partition the phylogenies into the three relevant topologies – i) sand flies with mosquitoes, ii) sand flies with flies, or iii) sand flies as outgroup to mosquitoes and flies - and all branch lengths were subsequently averaged.

Synteny Analysis

Macrosynteny

The identifiers, scaffolds, locations, and orientations were extracted from the FASTA headers for each protein sequence. The protein identifiers were cross-referenced with the OrthoDB v8 Diptera [56] catalog to group the proteins into ortholog groups. Sequences without ortholog information or not present in all genomes were discarded. Additionally, ortholog groups with many-to-many and one-to-many relationships were not included in the analysis. Proteins were sorted along each scaffold by their starting coordinates, while scaffolds were ordered arbitrarily. Scatter plots were generated by drawing dots at the positions of orthologous proteins.

Synteny Blocks

For each genome, gene identifiers, protein sequences, orientations, scaffolds, and locations were extracted from FASTA headers and reformatted as input for the programs CHROnicle and SynChro [57]. SynChro (Δ = 5) was run pairwise for all species. Synteny blocks (lists of gene identifiers, scaffolds, and positions) were extracted from the OrthBlocks synt files. Three-way synteny blocks for *An. gambiae*, *Lu. longipalpis*, and *Ph. papatasi* were constructed by finding all pairs of synteny blocks for *An. gambiae* and *Lu. longipalpis* and *Lu. longipalpis* and *Ph. papatasi* that overlapped by at least one gene.

Orientation in genome by prediction of protein function

While SynChro predicts synteny, it does not predict the orientation of the blocks in the genome. For each syntony block, the genes from one species were matched to genes in the second species using blastp [58] against local BLAST databases created from *An. gambiae*, *Ph. papatasi*, and *Lu. lutzomyia* protein sequences obtained from VectorBase [60]. The blastp

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

similarity searches were run using the default algorithm parameters. Two genes were considered match if they located in the corresponding pair of synteny blocks, were in the top 5 hits for each other, and 3 criteria were met for the alignment: e-value < 1e-3 and query coverage and identity > 20%. Relative alignment of the blocks were determined based on the corresponding orientations of gene pairs; since some gene pairs had conflicting orientations, block orientation was determined based on that which gave the greatest agreement across the gene pairs.

Population Genetics Analysis

SNP Calling

We performed alignments and variant calling on the raw reads of whole-genome samples of *Ph. papatasi* collected from Tunisia (N=6), Egypt (N=6), and Afghanistan (N=5) to the *Ph. papatasi* reference genome (Ppap_1.0). We also aligned and called variants for *Phlebotomus bergeroti* (N=2), and *Phlebotomus duboscqui* (N=1) as outgroups. In addition, we called variants from the raw reads of whole-genome samples of *Lu. longipalpis* collected from locations in Brazil [Marajó (N=9), Lapinha (N=13), Jacobina (N=14), and Sobral (1S N=13; 2S N=16)] and *Nyssomyia intermedia* (N=2) and *Migonemyia migonei* (N=2) as outgroups aligned to the *Lu. longipalpis* reference genome (Llon_1.0).

All genomic reads were pre-processed by removing duplicate reads with Picard (v1.113), and paired-end reads were aligned to the reference genome using bwa-mem [61]. Base position differences (SNV) were based on the unique convergence from two variant calling software tools, SAMtools [62] and VarScan 2 [63], using standard variant calling and filtering parameters that are optimized for whole genome data with moderate coverage (10X-40X). These parameters included a *P*-value of 0.1, a map quality of 10, a minimum coverage of three reads, and parameters for filtering by false positives. After alignment and variant detection, we implemented a filter to exclude variants that were clustered in groups of more than five variants per 500 bp. We finally implemented backfilling to include homozygous reference calls for each site where a variant is called in the final multi-sample variant call format (VCF) file for each individual when the coverage

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

exceeded three reads. Sites that did not exceed this threshold were included as missing diploid genotypes.

SNP filtering

To aid in the quality assessment of variants, we excluded the genotypes having a genotype quality (GQ) lower than 30 (i.e., minimum accuracy of 99.9%). We also applied hard filters on the variants, excluding any variants having an average depth lower than 10 or higher than 200, a Hardy-Weinberg equilibrium (HWE) *P*-value lower than 0.001, levels of missing genotypes higher than 20%, and having a minor allele frequency (MAF) lower than 1%. The dataset used in population structure inferences was pruned for linkage disequilibrium, excluding variants above an r^2 threshold of 0.5 in sliding windows of 50 variants with a step size of 5 variants using PLINK v.1.90 [64, 65]. Variants in linkage disequilibrium were pruned from the 6,390,876 sites using a sliding window of 500 kb and a linkage disequilibrium threshold of 0.2 using SNPRelate v.x [66].

Genomics structure

Although low powered due to limited sampling, we made an initial attempt to identify regions in the genome that may be contributing to differentiation between the populations. For the *Ph. papatasi* samples, VCFtools v.0.1.15 [67] was used to run a sliding window analysis with a 5 kb sliding window size, a 500 bp step size, and at least 10 variants per window [67]. After calculating Tajima's D for each window within each population [Tunisia (TUN), Egypt (EGP); Afghanistan (AFG)], we calculated pairwise population divergence using Wright's fixation index (F_{ST}). We made three pairwise comparisons: i) TUN to NS; ii) TUN to AFG; and iii) NS to AFG. The distributions of these results were not normal, so we relied on a percentile approach and selected all 5 kb windows that met the 1st percentile for Tajima's D and the 99th percentile for F_{ST}. Windows with fewer than 10 SNPs and windows with coordinates from 1-500 were eliminated. We then searched for 5 kb windows that passed the following thresholds: i) low within-population Tajima's D and ii) high F_{ST}. We looked for direct overlapping windows of high F_{ST} with low Tajima's D scores and indirect overlap, allowing for a 10kb buffer on either end of each window we identified.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Individual ancestry was estimated using Admixture v.1.9 [74]. The analysis was performed for *K* values (ranging from two to seven with 30 iterations per *K*). In order to better understand the different solutions reported by Admixture, post processing of the Admixture results was performed in CLUMPAK v.1.1 [75]. Principal component analysis (PCA) was performed in scikit-allel v.1.1.10 [76], following the methods described in [77]. Weir and Cockerham's F_{ST}, Nei's $D_{xy,}$ and Tajima's D were calculated using VCFtools, and using the python script popGenWindows.py (https://github.com/simonhmartin/genomics_general). Single marker FLK test [79] was performed using HapFLK v.1.4 [80].

Phylogenetic analysis

We explored ancestral phylogenetic relationships between individuals by building a neighbor-joining (NJ) tree across the genome using the R packages *adegenet v.2.1.1* [68, 69], *ape v.5.1* [70], *poppr v.2.7.1 [71]*, and *vcfR v.1.7.0 [72]* . For *Ph. papatasi,* we included both *Ph. bergeroti* and *Ph. duboscqi* and used the later to root the trees. For *Lu. longipalpis* phylogenetic analysis we included both *N. intermedia* and *M. migonei* and used *M. migonei* to root the NJ trees. We evaluated node support using 1,000 bootstrap replicates [73].

dN/dS

Selective constraints on gene sequence evolution were estimated using the dN/dS statistic calculated for orthologous group multiple sequence alignments. Protein sequences were assigned to ortholog groups by cross-referencing the OrthoDB v8 catalog [56]. For ortholog groups with one-to-many and many-to-many orthologs, a single protein sequence was chosen for each species by choosing randomly, with uniform probabilities, from the sequences for each species. Protein sequence multiple alignments were generated first using Clustal-Ω [81], and then used to inform CDS alignments with the codon-aware PAL2NAL alignment program [82]. The yn00 program from PAML v4.8 [83] was used to calculate dN/dS ratios for each pair of sequences in each aligned orthologous group.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Results and Discussion

Sequencing and Genome Characteristics

The genome of *Ph. papatasi* is ~350 Mb and was completed in 2012 for community analysis and population comparisons (S1 Table). The assembly was built from an input of ~22.5X total sequence coverage and resulted in 139,199 contigs with an N50 of 5.8 kb and 106,826 scaffolds with an N50 of 28 kb. The draft genome of *Lu. longipalpis* (Llon_1.0) was also completed in 2012 and is approximately 154.2 Mb, more than two time smaller than the size of the *Ph. papatasi* genome, representing 38.9X coverage (S1 Table). There are 35,696 contigs with an N50 of 7.5 kb and 11,532 scaffolds with an N50 of 85.1 kb. Based on automated and manual annotations, the *Ph. papatasi* and *Lu. longipalpis* genomes are estimated to contain 11,216 and 10,311 protein-coding genes, respectively. The BUSCO analysis [84] indicated 86.5% and 86.1% completeness for the *Ph. papatasi* and *Lu. longipalpis* genomes respectively (S2 Table). Annotation was augmented with RNA-seq expression evidence from different life-cycle stages, multiple days post adult emergence, and after blood-feeding in uninfected and *Le. major*-infected blood for *Ph. papatasi* and *Le. infantum*-infected blood for *Lu. longipalpis* (S3 Table).

<u>Orthology</u>

To improve our understanding of the phylogenetic relationships, we generated a maximum likelihood phylogenetic tree using orthologous genes selected from an orthology dataset comprising 43 insect species, including 36 dipterans. Consistent with [55], the phylogenetic tree supported clustering of sand flies with Culicomorpha infraorder (mosquitoes and black flies) rather than with the Muscomorpha infraorder (*Drosophila* and *Glossina)* (Fig 2A). In addition, percent identity between orthologs is higher between sand flies and mosquitoes than between sand flies and fruit files, in agreement with the maximum likelihood phylogeny (Fig 2A). Sand flies and culicines have more than three times as many exclusively-shared orthologous groups than sand flies do with muscoids, also consistent with the maximum likelihood phylogeny (Fig 2B). Analysis

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

of individual gene phylogenies, however, shows great uncertainty with almost equal proportions of phylogenies supporting clustering of sand flies with mosquitoes and with muscoids (S1 Fig.).

Figure 2: Molecular species phylogeny and ortholog sharing. (A) The quantitative maximum likelihood species phylogeny computed from the concatenated superalignment of 1,627 orthologous protein-coding genes places the sand flies (Psychodomorpha) as a sister group to the mosquitoes (Culicomorpha) rather than the flies (Muscomorpha), with all branches showing 100% bootstrap support. The Culicomorpha are represented by four *Anopheles* mosquito species and *Culex quinquefasciatus* and the Muscomorpha include four *Drosophila* fruit fly species and the tsetse fly, *Glossina morsitans*. Outgroup species represent Lepidoptera (*Bombyx mori*), Coleoptera (*Tribolium castaneum*), Hymenoptera (*Apis mellifera*), and the phylogeny is rooted with the phthirapteran human body louse, *Pe. humanus*. The inset boxplots show that single-copy (1:1) and multi-copy (X:X) ortholog amino acid percent identity is higher between each sand fly (Ll, *Lu. longipalpis*; Pp, *Ph. papatasi*) and *An. gambiae* (Ag) than *D. melanogaster* (Dm). Boxplots show median values with boxes extending to the first and third quartiles of the distributions. (B) The Venn diagram summarizes the numbers of orthologous groups and mean number of genes per species (in parentheses) shared among the two sand flies (L. lon., *Lu. longipalpis*; P. pap., *Ph. papatasi*) and/or the Culicomorpha and/or the Muscomorpha. Analysis of ortholog sharing shows that the sand flies share more than three times as many orthologous groups exclusively with the Culicomorpha (*Anopheles* and *Culex*) compared to the Muscomorpha (*Drosophila, Glossina*) (subsets highlighted with thin and thick dashed lines). Numbers of unique genes are in italics. Colors in panel A and panel B match species and sets of species analyzed.

<u>Synteny</u>

Comparisons of the relative ordering of ortholog genome positions between *Lu. longipalpis* and *Ph. papatasi* did not indicate evidence of macrosynteny (S2 Fig.). The comparison of ortholog genome positions is complicated by the fragmented nature of the assembled genomes, with *Lu.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

longipalpis's 10,331 genes spread across 11,532 scaffolds and the 11,216 genes of *Ph. papatasi* spread across 106,826 scaffolds. Synteny analysis is likely also affected by the high rate of evolutionary divergence among insects [85] and *Ph. papatasi* and *Lu. longipalpis* are separated by at least 100 million years [86]. Previous studies of anopheline species have likewise shown that synteny is conserved over relatively short time spans compared with other organisms [48].

Microsynteny, however, was observed between the sand flies and between each sand fly species with *An. gambiae* (S4 Table). Microsynteny between *Lu. longipalpis* and *Ph. papatasi* is similar to what was previously reported for *An. gambiae* and *D. melanogaster* [87]. We identified 499 synteny blocks with a maximum of 24 genes, average of 4.2 genes, and a median of 3 genes. By comparison, we found 1,037 synteny blocks with a maximum of 10 genes, an average of 2.5 genes, and a median of 2 genes between *An. gambiae* and *D. melanogaster.* Previous studies reported 948 synteny blocks with a maximum of 39 genes between *An. gambiae* and *D. melanogaster* [87].

Transposable Elements

Transposable elements (TEs) are ubiquitous repetitive sequences present in eukaryotic genomes that can be an important factor affecting genome sizes and are thought to be one of the driving forces of evolution [45]. Some insect genomes have less than 3% of TEs, while others contain as much as 50% or more of TEs, associated with large genomic size differences [88]. Our analysis indicated that the *Ph. papatasi* genome is composed of 5.65% of TE derived sequences while the *Lu. longipalpis* genome contains only 0.57%, corresponding to the genome size difference between two sand fly species. This difference in TE-derived sequence could be due to the result of divergent evolutionary dynamics of some TE families or superfamilies, affecting either their distribution (presence or absence of specific superfamilies) or their abundance (copy number per superfamily) in the genome. Higher abundance of TE derived sequences, presence of full-length TEs and the genome size expansion in the *Ph. papatasi* genome also could be due to recently active TEs. Alternatively, genomic differences in TE content might be the result of intrinsic

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

genomic deletion patterns in *Lu. longipalpis*, due to the effective recognition and elimination machinery removing these foreign sequences from the genome, as has been shown to occur in *Drosophila* species [89].

The comparison of the TE content in both sand fly genomes indicated an expansion of all the TE classes and orders in the *Ph. papatasi* genome. This multiplication was more pronounced in elements belonging to the class II, or "cut-and-paste" TEs, and especially in non-autonomous miniature inverted-repeat transposable elements (MITEs), representing up to 29-fold differences between the two genomes. Expansion of MITEs suggests the recent activity of class II TEs in the *Ph. papatasi* genome. On the other hand, class I elements, or "copy-paste" elements, including the Long Terminal Repeat (LTRs) and non-LTRs, which traditionally are accountable for the genome expansion, showed more subtle changes between the two sand fly genomes, representing up to 4-fold difference. (Table 1).

Table 1: Transposable Elements

	Phlebotomus papatasi **% genome**	***Lutzomyia longipalpis*** **% genome**
LTR retrotransposons	**0.41%**	**0.21%**
Bel	0.17%	0.09%
Mag	0.07%	0.04%
Pao	0.06%	0.01%
Mdg3	0.05%	0.02%
Gypsy	0.03%	0.02%
Mdg1	0.02%	0.01%
Osvaldo	0.01%	0.01%
Copia	0.01%	0.00%

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Non-LTR retrotransposons	**0.95%**	**0.22%**
L2	0.25%	0.04%
RTE	0.21%	0.02%
Jck	0.18%	0.04%
CR1	0.16%	0.03%
LOA	0.07%	0.02%
I	0.05%	0.05%
Loner	0.03%	0.01%
Ocas	0.01%	0.01%
DNA transposons	**1.13%**	**0.05%**
Tc1/mariner	0.96%	0.04%
piggyBac	0.18%	0.00%
Helitron	0.00%	0.01%
MITEs	**4.11%**	**0.14%**
mTA	2.63%	0.01%
m2bp	0.54%	0.01%
m8bp	0.29%	0.10%
m3bp	0.24%	0.00%
otherMITEs	0.23%	0.00%
m4bp	0.18%	0.03%
Unclassified TE sequences	**0.00%**	**0.01%**
Total, percent TE in genome	**5.65%**	**0.57%**

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Immunity Genes

Several immune pathways are conserved among insects. These include the Toll, Immune Deficiency (IMD), Janus Kinase/Signal Transducer and Activator of Transcription (JAK/STAT), lectin, and encapsulation pathways. We found the Toll signaling pathway highly conserved in the genomes of both sand fly species, including homologues of the upstream peptidoglycan recognition proteins (PGRPs), and glucan binding protein (GNBPs) (S5 Table). Similarly, the IMD and JAK/STAT pathways appear to be conserved among dipterans, including *Drosophila*, *Aedes*, *Anopheles* and both sand fly species analysed in this study (S6 and S7 Tables).

Galactose-binding proteins (galectins) are a diverse family of proteins playing roles in development and immunity [90]. Comparing the sand flies' galectin protein sequences with other Diptera, both shared and independent orthologs were identified (S3 Fig.; S8 Table). Future analyses evaluating *Leishmania* parasite interactions with the *Ph. Papatasi* and *Lu. longipalpis* galectins may provide a better understanding of the mechanisms that influence restrictive versus permissive vectorial competence due to the key role some galectins play in parasite establishment and survival [91].

Fourteen genes related to TGF-beta or TGF-beta pathways were found in each of the sand fly genomes (S9 Table) and 16 and 15 different MAPK gene loci were identified in the genome of *Lu. longipalpis* and *Ph. Papatasi* genome, respectively (S10 Table). Interestingly, two prophenoloxidase homologues were identified in each species (S11 Table). A TEP-1-like protein, and a COX-like ortholog were also found in the genomes of both *Lu. longipalpis* and *Ph. Papatasi* (S6 Table).

Blood Feeding Genes

We mapped *Ph. Papatasi* and *Lu. longipalpis* putative salivary genes deposited at the NCBI to the sand fly assemblies (S12 Table). Equally well studied in phlebotomine sand flies are the

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

genes associated with digestive properties [93-102]. Here, we characterized the following digestive gene families: Peptidases (S13 Table); Glycoside Hydrolase Family 13 (S14 Table); Chitinase and Chitinase-like protein family (S15 Table); N-acetylhexosaminidases (S16 Table; S4 Fig.); Chitin deacetylases (S17 Table; S5 Fig.); and Peritrophin-like proteins (S18 Table; S6 Fig.). A detailed analysis of GH13 genes, including amylases, maltases and sucrases has been published elsewhere [103]. Aquaporins (AQPs) are required for the transportation of water and other small solutes across cell membranes and are important for excreting water from the blood meal. We have identified six AQP genes from both species of sand flies (S19 Table; S7 Fig.). This is similar to the number present in mosquitoes (N = 6), but two and four less than *Drosophila* and *Glossina*, respectively [104]. Members of each AQP group previously identified from insects are present in the sand fly genomes.

Circadian Rhythm Genes

Many important aspects in sand fly biology such as hematophagy and host seeking are controlled by the biological clock [105]. In *Lu. longipalpis*, the main clock genes and their expression pattern throughout the day have been previously characterized [106, 107]. Orthologues of all the core circadian clock genes were found in the genome of both *Lu. longipalpis* (S20 Table) and *Ph. papatasi* (S21 Table). Interestingly, cryptochrome evolution has been a matter of great interest [105] and we similarly found compelling features in sand fly CRY gene structure. We found both *CRY1 and CRY2* genes in *Ph. papatasi* but, surprisingly, we did not find a *cry1* gene in *Lu. longipalpis* genome assembly (S8 Fig.). Although both sand fly species are closely related, these data suggest that whereas *Ph. papatasi* seems to have a functional mammalian-like clock closer to butterflies, mosquitoes, and other dipterans, with *CRY1* and *CRY2* genes, *Lu. longipalpis* may have a circadian clock working with a mechanism more similar to that found in triatomines, bees and beetles, presenting only *CRY2* gene. We can speculate that the possible loss of *CRY1* in *Lu. longipalpis* genome may be related to a better adaptation of these insects to living in caves and dark places or alternatively, is just missing in the current assembly.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Chemosensory Receptors

The sand fly olfactory receptor (OR) repertoires comprise 139 canonical ORs in *Lu. longipalpis* and *Ph. papatasi*, plus one copy each of the odorant receptor co-receptor, *Orco* (S22 Table). We identified 82 genes encoding 91 gustatory receptors (GRs) in *Lu. longipalpis*, and 77 genes encoding 88 GRs in *Ph. papatasi* (S22 Table). Based on homology to *D. melanogaster*, these include seven to eight sugar receptors and two carbon dioxide receptors [108]. Among the other conserved GRs are orthologs to *DmelGr28a/b*, *DmelGr33a,* and *DmelGr66a.* The ionotropic (IR) repertoires are relatively small, with 23 and 28 in *Lu. longipalpis* and *Ph. papatasi*, respectively (S22 Table). Among these are orthologs of 13 of the 16 IRs expressed in *D. melanogaster* antennae [109]. We focused on antennal IRs that are conserved among Diptera [110, 111] and did not attempt to identify glutamate receptors or highly divergent IRs, therefore, the true number of IRs could be greater than reported here. Three ORs and three IRs suspected to be missing in the *Lu. longipalpis* references assembly were found in *de novo* assemblies of the field isolates Jacobina 14 and/or Jacobina 23. A comprehensive analysis of the sand fly chemoreceptors was published elsewhere [112].

Nine and ten members of the transient receptor potential (TRP) cation channel family are found in *Lu. longipalpis* and *Ph. papatasi* genomes, respectively, and the phylogenetic tree showed a separation of the different TRP subfamilies (S9 Fig.). The pickpocket (PPK) receptor phylogenetic tree demonstrated a division of the six different PPK subfamilies (S10 Fig.) with 14 and 13 family members in *Lu. longipalpis* and *Ph. papatasi* genomes, respectively.

G-Protein Coupled Receptors

G-Protein Coupled Receptors (GPCRs) are a large family of membrane-bound proteins that operate in cellular signal transduction and interact with a wide variety of chemistries including small molecules, neuro-peptides, and proteins. These proteins play roles in essential invertebrate functions [113]. We utilized a novel classifier to identify insect GPCRs [114] in both *Ph. papatasi* and *Lu. longipalpis*, followed by validation and manual annotation of identified genes. Ninety-four

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

and 92 GPCRs from *Ph. papatasi* and *Lu. longipalpis,* respectively, were compared with other insects with well characterized GPCRs such as *D. melanogaster, An. gambiae, Ae. aegypti and Pe. humanus* (S23 Table). Class A (rhodopsin-like) is the most numerous class with ~50 genes in each sand fly, and includes the opsins that are thought to function in visual processes and circadian rhythm. Both sand flies have one opsin gene for each functional group, the long-wavelength, short-wavelength, ultraviolet, rh7-like, and pteropsin. Classes B (secretin-like), C (metabotropic glutamate-like) and D (atypical GPCRs) have fewer members, with ~20, ~10 and ~10 in each sand fly, respectively. Sand flies include GPCR genes absent from *D. melanogaster* (ocular albinism) and absent in *Ae. aegypti* and *An. gambiae* (parathyroid hormone receptor); both genes from class B.

Cytochrome P450 Monoxygenase Genes

Cytochrome P450s (CYPs or P450s) constitute a conserved enzyme superfamily with a diverse array of functions, ranging from core developmental pathways to the detoxification of xenobiotics [115]. The CYP gene repertoire (CYPome) plays an important role in insect physiology and in the development of resistance to insecticides used for vector control. Here we identified and manually curated 104 CYP genes in *Lu. longipalpis* (Data File S1) and 93 CYP genes in *Ph. papatasi* (Data File S2). These numbers are similar to the number of CYPs identified in the mosquito *An. gambiae* (n =100). In *Lu. longipalpis* all 104 CYPs are full-length genes, compared to 34 full-length and 59 fragmented genes in *Ph. papatasi*, likely reflecting the more fragmented genome assembly of *Ph. papatasi* compared to *Lu. longipalpis*.

The identified sand fly CYP genes belong to the four clans typically found in insects; mitochondrial (Mito), CYP2, CYP3, and CYP4 clan [115]. Remarkably, both sand fly species have an expanded CYP3 clan compared to *An. gambiae* (S11 Fig.). This expansion is mostly caused by gains in the CYP9J/9L, CYP6AG, and CYP6AK subfamilies (S11 Fig.).

Other Groups

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

In addition, we identified core genes as well as non-coding RNAs in the siRNA, miRNA, and piRNA pathways, suggesting that these regulatory mechanisms are fully functional in sand flies (S24 Table). We have also annotated heat shock and hypoxia proteins (S25 Table), cuticular proteins (S26 Table), hormonal signaling (S27 Table), insulin signaling (S28 Table), and antioxidant (S29 Table) genes, as well as genes involved in vitamin metabolism (S30 Table). Additional information about annotated gene families can be found in the Supplementary Results.

Population Structure

Genetic structure across the range of *Ph. papatasi*

Average genome coverage ranged from 8X-16X (mean=12X; S31 Table). A total of 6,390,876 sites passed the thresholds using variant calling methods, where at least one sample displayed a variant at a reference coordinate. As expected, the *Ph. papatasi* samples showed the lowest count of Single Nucleotide Variants (SNVs) (1.84-1.99M SNVs) while the two *Ph. bergeroti* samples (mean SNVs = 3.26M), and the *Ph. duboscqi* sample (4.01M SNVs) contained a higher variant count (S32 Table). We found a small percentage of singletons (unique SNV's) in the *Ph. papatasi* samples (3.0%-4.3%) and 3,482 shared variant alleles among the *Ph. papatasi* samples. We also calculated the transition: transversion ratios, inbreeding coefficients, and pairwise relatedness (S32 Table).

For phylogenetic analysis the dataset was filtered by keeping only variants of the highest quality, leaving 1,084,952 total variants: 284,341 for Afghanistan, 435,972 for Egypt, and 439,446 for Tunisia. The dataset used in population structure inferences was further pruned for linkage disequilibrium, creating a final dataset containing 423,236 total variants.

We explored ancestral phylogenetic relationships between individuals by building a NJ tree across the genome. The NJ tree clustered the *Ph. papatasi* individuals into three clades that correlated to geographical location, with bootstrap values of 100 (S12A Fig.).

Admixture was used to estimate the individual ancestries. Admixture cross-validation errors (CV) suggest that the number of genetic clusters that best explains the observed population

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

structure as *K*=2 (S13A Fig.), where the Afghanistan samples were distinct from the Tunisian and Egyptian samples (S12C Fig.).

We next performed a PCA, which does not depend on any model assumption and can thus provide a useful validation of the results of Admixture analysis. The PCA supported the phylogenetic analysis, separating the individuals into three distinct clusters, with all individuals from the collection site clustering together. Principal components 1 and 2 accounted for 20.1% of the total variation (S12B Fig.).

We found no direct overlapping windows of high F_{ST} with low Tajima's D scores for any of the *Ph*. *papatasi* populations. Next, we searched for windows that met the above criteria but included a 20 kb (10 kb on either side of the window) to identify indirect overlaps. We identified 29 genes that fell within in the indirect overlapping windows (S33 Table). Functional annotation revealed 3 tRNAs, 3 putative transcription factors, and a snoRNA as possibly under selective pressure, as well as 9 genes involved in metabolic pathways.

Genomic evidence of cryptic species within *Lu. longipalpis sensu lato*

The average genome coverage ranged from 8X to 105X (mean = 47X; S34 Table). We identified 4,821,847 variants across all the individuals. To aid in quality assessment of variants, filtration was performed as described for *Ph. papatasi.* After filtration, 1,937,819 variants remained, ranging from 206,588 for Marajó to 633,519 for Jacobina (Fig. 1). After filtration and LD pruning, the dataset contained 1,059,627 variants.

Consistent with the phylogeny based on the chemoreceptor repertoire [112], the full genome phylogeny clustered the populations into two clades based on song and pheromone type, where Marajó and Sobral 2S (Burst, Sobralene) cluster together and Lapinha and Sobral 1S (Pulse, Germ) cluster together (Fig. 3A). An analysis of the male copulatory courtship songs of males collected from Lapinha and Sobral are in agreement with those previously recorded [41]. In these three resampled populations, we observed the sub-type P2 in Lapinha, the sub-type P3 in Sobral 1S, and the burst-type in Sobral 2S.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure 3. *Lutzomyia longipalpis* population structure. Inferred population structure of *Lu. longipalpis* individuals collected from Marajó (MAR; pink), Lapinha (LAP; blue), from Jacobina (JAC; red), and Sobral, including Sobral 1S (S1S; orange) and 16 Sobral 2S (S2S; green). (A) Rooted neighbor joining (NJ) radial tree. We included both *N. intermedia* (INT; yellow) and *M. migonei* (MIG; purple) and used *M. migonei* to root the trees. Bootstrap values represent the percentage of 1,000 replicates. (B) Principal component analysis (PCA). Individuals were plotted according to their coordinates on the first two principal components (PC1 and PC2). (C) Admixture analysis. Ancestry proportions for Admixture models from K = 2 to K = 7 ancestral populations. Each individual is represented by a thin vertical line, partitioned into K coloured segments representing the individual's estimated membership fractions to the K clusters. These data are the average of the major q-matrix clusters derived by CLUMPAK analysis.

Interestingly, the phylogenetic analysis separated the Jacobina population into two groups. As expected, because flies from Jacobina are known to express a C_{16} H_{32} pheromone and pulse-like copulatory songs, some individuals clustered with Sobral 1S and Lapinha. Unexpectedly, however, six individuals clustered with the diterpenoid-like pheromone and burst song expressing individuals, suggesting that there is more than one population living in sympatry at the Jacobina site. Male copulatory songs were not recorded for the Jacobina samples. However, sand flies collected from two localities near to Jacobina, Araci and Olindina (Bahia state) (S14A Fig.) exhibit different copulatory song patterns, suggesting the possible existence of two groups in Jacobina, as observed in molecular data. Males from Araci exhibit the P1 copulation song pattern (S14B Fig.), composed of train of similar pulses as previously described in males collected in Jacobina [41]. Males from the nearby locality, Olindina, produced burst-type songs (S14B Fig.) with similar pattern as Sobral 2S males [41]. The mean values of all song parameters observed from these flies (S35 Table) are similar as previously reported [118].

In addition, the phylogenetic analysis indicated sub-structure within the Sobral 1S population. The song tracings, however, did not suggest any sort of split. Although the analysis suggests

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

seven distinct populations, there is not enough statistical support to separate the six Jacobina individuals from the Sobral 2S population, resulting in support for six populations.

The PCA based on the whole-genome clustered the individuals into six groups as well (Fig. 3B). Contrary to the phylogenetic analysis, however, the six Jacobina individuals were closely clustered, but separate from the Marajó and Sobral 2S populations, which were indistinguishable. PC1 explained 5.3% of the variation and separated the individuals collected from Jacobina into two populations. Consistent with the NJ tree, the Sobral 1S population also exhibited some population structure, the two clusters distinguishable through PC1 and PC2. PC2 accounted for 4.6% of the total variation and distinguished Lapinha from the other populations. The sympatric Sobral 1S and 2S populations separate by both PC1 and PC2. Interestingly, while consistent with Hickner *et al.* 2020, the whole-genome PCA allowed higher discriminating power among clusters than the PCA based on the chemoreceptor repertoire which only identified 3-4 discrete clusters [112].

Seven groups are clearly distinguishable from the Admixture analysis at K=7, consistent with the PCA, NJ tree (Fig. 3C), and [112]. However, the cross-validation error analysis indicates 3-4 populations (S13B Fig.), one population consisting of all Marajó and Sobral 2S individuals and six Jacobina individuals, one population made up of 8 Jacobina individuals and another population with 7 Sobral 1S individuals. In contrast to the NJ tree that suggests that the individuals from Lapinha make up a single population, the Admixture analysis indicates that all Lapinha individuals and six Sobral 1S individuals are of similar ancestry. The analysis suggests no introgression between the sympatric Sobral 1S and 2S individuals.

To identify candidate genomic regions contributing to reproductive isolation and to distinguish between the two models of speciation, that with and without gene flow, pairwise measures of divergence were calculated for Marajó, Sobral 1S, Sobral 2S, and Lapinha. Relative (Weir and Cockerham's F_{ST}) and absolute (Nei's D_{xy}) measures of divergence were calculated for 1 kb non-overlapping windows for all population comparisons, excluding Jacobina. Mean weighted F_{ST}

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

values indicate that genome wide differentiation is greater among population comparisons of different pheromone and song types (Lapinha- Marajó, 0.214; Lapinha-Sobral 2S, 0.211; Sobral 1S-Sobral 2S, 0.116 compared to Lapinha – Sobral 1S, 0.154; Marajó-Sobral 2S, 0.114) and allopatric populations (Lapinha- Marajó, 0.214; Lapinha-Sobral 2S, 0.211; Lapinha – Sobral 1S, 0.154 compared to Sobral 1S-Sobral 2S, 0.116) (S15 Fig.).

We identified genomic regions possibly contributing to population differentiation as F_{ST} outlier windows that were in the top 2.5% quantile for each sympatric and allopatric comparison of differing pheromone/song phenotype (S16 Fig.). There were 170 differentiation regions in common among all of the different pheromone and song type comparisons (Fig. 4A). The mean F_{ST} estimates were higher in the genomic regions shared by more than one comparison than in those unique to each comparison, suggesting that these regions are being targeted by selection in each case. Supporting the hypothesis that the Sobral populations have more recently diverged from one another, the F_{ST} outlier windows had a mean value less than the allopatric populations (Fig. 4B).

Figure 4. Genomic regions with high pairwise F_{ST} between the different populations of *Lutzomyia longipalpis*. (A) Venn diagram depicting the number of 1 kb non-overlapping genomic windows having F_{ST} values in the top 2.5% quantile (outlier) among the different population comparisons. (B) Box plots of outlier F_{ST} windows shared with another population comparison (blue) or unique to a population comparison (pink). Box plots show the medians (lines) and interquartile ranges (boxes); the whiskers extend out from the box plots to 1.5 times the interquartile range, and values outside this limit are represented by dots. Mean F_{ST} values are represented by open diamonds.

We further characterized the genomic regions by computing additional statistics in each window. We tested if these regions were enriched for signatures of selection by computing Tajima's D in the 1 kb non-overlapping windows, negative values of Tajima's D indicating a potential selective sweep. As with the F_{ST} values, we considered outlier windows as those that

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

were in the lower or upper 2.5% quantiles (S17 Fig.). The vast majority of Tajima's D outlier windows were unique to each population (S18 Fig.). No positive outlier windows overlapped among the four populations (S18A Fig.) and only four negative outlier windows were shared among all of the populations (S18B Fig.), of which only one contained a gene, LLOJ005792 of unknown function. None of the Tajima's D outlier windows overlapped with the outlier F_{ST} windows.

As absolute measures of divergence are less affected by within population levels of polymorphism than relative measures of divergence, like F_{ST} [119], we calculated Nei's measure of absolute divergence, D_{xy}, as an additional signature of selection. As expected, because these populations are thought to have recently diverged from one another, the top 2.5% of D_{xy} values were substantially lower than the outlier F_{ST} windows (Fig. 5A). The majority of outlier F_{ST} values did not fall in the upper quantile of D_{xy} values (Table 2) and the windows with the highest D_{xy} values did not overlap with the F_{ST} outlier windows (Fig. 5B), suggesting that there may be varying levels of genetic diversity within each population.

Figure 5. Measures of divergence in 1 kb non-overlapping genomic windows between the different populations of *Lutzomyia longipalpis*. (A) Box plots of D_{xy} and F_{ST} values in the top 2.5% quantile (outlier) of each population comparison. (B) Box plots of D_{xy} and F_{ST} values for windows having both high D_{xy} and high F_{ST} (differentiation islands). (C) Box plots of Tajimas' D values for the differentiation islands. (D) Box plots of FLK values for sites within differentiation islands. Box plots show the medians (lines) and interquartile ranges (boxes); the whiskers extend out from the box plots to 1.5 times the interquartile range, and values outside this limit are represented by dots. Mean values are represented by open diamonds.

Table 2: *Lutzomyia longipalpis* Differentiation Island (DI) Statistics

	Lapinha-Marajó	Lapinha-Sobral 2S	Marajó-Sobral 1S	Sobral 1S-Sobral 2S

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

# 1 kb Windows	127,065	129,513	127,065	131,430
# F_{ST} Outlier Windows	3,176	3,237	3,176	3,285
# Regions of Interest	729	841	740	1,023
% F_{ST} Outlier Non-DI	77.05	74.03	76.70	68.87
Mean F_{ST} DI	0.93	0.85	0.85	0.86
Mean D_{xy} DI	0.30	0.29	0.29	0.29

There were 92 1 kb windows that fell in the upper 2.5% of F_{ST} and D_{xy} values and shared among all the comparisons. These windows were defined as Differentiation Islands (DI).

To identify genomic loci that may be contributing to the reproductive isolation of these populations [29], we defined 'regions of interest' as those windows that fell in both the upper 2.5% quantile of F_{ST} and D_{xy} values. There were 729, 841, 740, and 1023 regions of interest between Lapinha-Marajó, Lapinha-Sobral 2S, Marajó-Sobral 1S, and Sobral 1S-Sobral 2S, respectively (Table 2). The 92 regions of interest shared among all the population comparisons we interpreted as 'differentiation islands' (DI).

We tested whether the DIs were enriched for signatures of selection by calculating Tajima's D for these windows and performing a single marker FLK test [79] with HapFLK v. 1.4 [80]. The Tajima's D (Fig. 5C) and FLK (Fig. 5D) values do not provide evidence that selection (either balancing or positive) has led to the genomic divergence in these regions.

The genes present in the DIs are candidates that might explain the reproductive isolation of the populations. The 92 DIs contained 35 genes, 25 of which had orthologues in *An. gambiae* (S36 Table). Thirty-two of these genes were uncharacterized, LLOJ001208 is a protein MAK16

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

homolog, LLOJ009447 a rRNA adenine N(6)-methyltransferase, and LLOJ009732 a Lipase maturation factor. No enrichment of gene ontology terms was identified using the *An. gambiae* orthologs.

Conclusions

Our study provides the genome assembly and annotation of two divergent sand fly species that will facilitate molecular and comparative studies of these medically important vectors. These results provide a foundation for annotating and analyzing future chromosome length assemblies generated from single sand flies. Global comparisons between sand fly vectors will greatly inform the evolutionary relationships among these species and lead to advances in our understanding of genes involved in important phenomena such as vectorial capacity, host-specificity, blood-feeding, insecticide resistance, and immune system modulation.

Acknowledgments

Two of our collaborators, Dr Alexandre A. Peixoto (Feb. 2013) and Dr. Hector M. Diaz Albiter (Feb. 2021) were influential in this project but died before the work could be completed. We dedicate this manuscript to them in recognition of contributions to both this work and the vector disease community as a whole. In addition, gratitude goes to Sandra Clifton and Patrick Minx, who were both instrumental in the work but retired before completion and to Daniel Lawson who was instrumental in the initial stages of the project while a member of the VectorBase staff. The authors thank the sequencing team members of the Washington University Genome Center and Baylor College of Medicine Human Genome Sequencing Center for their efforts in generating the raw sequence data.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

References

1. Desjeux P. Leishmaniasis: current situation and new perspectives. Comparative immunology, microbiology and infectious diseases. 2004;27(5):305-18. doi: 10.1016/j.cimid.2004.03.004. PubMed PMID: 15225981.

2. Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide and global estimates of its incidence. PloS one. 2012;7(5):e35671. doi: 10.1371/journal.pone.0035671. PubMed PMID: 22693548; PubMed Central PMCID: PMC3365071.

3. Jones CM, Welburn SC. Leishmaniasis Beyond East Africa. Front Vet Sci. 2021;8:618766. Epub 20210226. doi: 10.3389/fvets.2021.618766. PubMed PMID: 33732738; PubMed Central PMCID: PMCPMC7959848.

4. Organization WH. WHO report on global surveillance of epidemic-prone infectious diseases. 2000.

5. Ramalho-Ortigao M, Saraiva EM, Traub-Cseko YM. Sand fly- interactions: long relationships are not necessarily easy. The open parasitology journal. 2010;4:195-204. doi: 10.2174/1874421401004010195. PubMed PMID: 24159365; PubMed Central PMCID: PMC3805382.

6. Maroli M, Feliciangeli MD, Bichaud L, Charrel RN, Gradoni L. Phlebotomine sandflies and the spreading of leishmaniases and other diseases of public health concern. Med Vet Entomol. 2013;27(2):123-47. Epub 20120827. doi: 10.1111/j.1365-2915.2012.01034.x. PubMed PMID: 22924419.

7. Sacks D, Kamhawi S. Molecular aspects of parasite-vector and vector-host interactions in leishmaniasis. Annu Rev Microbiol. 2001;55:453-83. PubMed PMID: 11544364.

8. Esseghir S, Ready PD. Speciation of *Phlebotomus* sandflies of the subgenus *Larroussius* coincided with the late Miocene-Pliocene aridification of the Mediterranean subregion. Biological Journal of the Linnean Society. 2000;70:189-219.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

9. Esseghir S, Ready PD, Killick-Kendrick R, Ben-Ismail R. Mitochondrial haplotypes and phylogeography of Phlebotomus vectors of Leishmania major. Insect Mol Biol. 1997;6(3):211-25. PubMed PMID: 9272439.

10. Lainson R, Ward RD, Shaw JJ. Experimental transmission of Leishmania chagasi, causative agent of neotropical visceral leishmaniasis, by the sandfly Lutzomyia longipalpis. Nature. 1977;266(5603):628-30. PubMed PMID: 859627.

11. Myskova J, Svobodova M, Beverley SM, Volf P. A lipophosphoglycan-independent development of Leishmania in permissive sand flies. Microbes Infect. 2007;9(3):317-24. doi: 10.1016/j.micinf.2006.12.010. PubMed PMID: 17307009; PubMed Central PMCID: PMC2839925.

12. Killick-Kendrick R. The biology and control of phlebotomine sand flies. Clin Dermatol. 1999;17(3):279-89. PubMed PMID: 10384867.

13. Yeates DK, Wiegmann BM. Congruence and controversy: toward a higher-level phylogeny of Diptera. Annu Rev Entomol. 1999;44:397-428. PubMed PMID: 15012378.

14. 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(14):5690-5. doi: 10.1073/pnas.1012675108. PubMed PMID: 21402926; PubMed Central PMCID: PMC3078341.

15. Simon S, Narechania A, Desalle R, Hadrys H. Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. Genome biology and evolution. 2012;4(12):1295-309. doi: 10.1093/gbe/evs104. PubMed PMID: 23175716; PubMed Central PMCID: PMC3542558.

16. Curler GR, Moulton JK. Phylogeny of psychodid subfamilies (Diptera: Psychodidae) inferred from nuclear DNA sequences with a review of morphological evidence for relationships. Syst Entomol. 2012;37(3):603-16. doi: 10.1111/j.1365-3113.2012.00634.x. PubMed PMID: WOS:000305616300009.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

17. Akhoundi M, Kuhls K, Cannet A, Votypka J, Marty P, Delaunay P, et al. A Historical Overview of the Classification, Evolution, and Dispersion of Leishmania Parasites and Sandflies. PLoS Negl Trop Dis. 2016;10(3):e0004349. Epub 2016/03/05. doi: 10.1371/journal.pntd.0004349. PubMed PMID: 26937644; PubMed Central PMCID: PMCPMC4777430.

18. Hamarsheh O. Distribution of Leishmania major zymodemes in relation to populations of Phlebotomus papatasi sand flies. Parasites & vectors. 2011;4:9. doi: 10.1186/1756-3305-4-9. PubMed PMID: 21266079; PubMed Central PMCID: PMC3035596.

19. Colacicco-Mayhugh MG, Masuoka PM, Grieco JP. Ecological niche model of Phlebotomus alexandri and P. papatasi (Diptera: Psychodidae) in the Middle East. Int J Health Geogr. 2010;9. doi: Artn 2

10.1186/1476-072x-9-2. PubMed PMID: WOS:000274782300001.

20. Depaquit J, Lienard E, Verzeaux-Griffon A, Ferte H, Bounamous A, Gantier JC, et al. Molecular homogeneity in diverse geographical populations of Phlebotomus papatasi (Diptera, Psychodidae) inferred from ND4 mtDNA and ITS2 rDNA Epidemiological consequences. Infect Genet Evol. 2008;8(2):159-70. PubMed PMID: 18243814.

21. Hamarsheh O, Presber W, Abdeen Z, Sawalha S, Al-Lahem A, Schonian G. Genetic structure of Mediterranean populations of the sandfly Phlebotomus papatasi by mitochondrial cytochrome b haplotype analysis. Med Vet Entomol. 2007;21(3):270-7. PubMed PMID: 17897368.

22. Hamarsheh O, Presber W, Al-Jawabreh A, Abdeen Z, Amro A, Schonian G. Molecular markers for Phlebotomus papatasi (Diptera: Psychodidae) and their usefulness for population genetic analysis. Trans R Soc Trop Med Hyg. 2009;103(11):1085-6. doi: 10.1016/j.trstmh.2009.02.011. PubMed PMID: 19303124.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

23. Hamarsheh O, Presber W, Yaghoobi-Ershadi MR, Amro A, Al-Jawabreh A, Sawalha S, et al. Population structure and geographical subdivision of the Leishmania major vector Phlebotomus papatasi as revealed by microsatellite variation. Med Vet Entomol. 2009;23(1):69-77. doi: 10.1111/j.1365-2915.2008.00784.x. PubMed PMID: 19239616.

24. Flanley CM, Ramalho-Ortigao M, Coutinho-Abreu IV, Mukbel R, Hanafi HA, El-Hossary SS, et al. Population genetics analysis of Phlebotomus papatasi sand flies from Egypt and Jordan based on mitochondrial cytochrome b haplotypes. Parasites & vectors. 2018;11(1):214. Epub 2018/03/29. doi: 10.1186/s13071-018-2785-9. PubMed PMID: 29587873; PubMed Central PMCID: PMCPMC5872541.

25. Khalid NM, Aboud MA, Alrabba FM, Elnaiem DE, Tripet F. Evidence for genetic differentiation at the microgeographic scale in Phlebotomus papatasi populations from Sudan. Parasites & vectors. 2012;5:249. doi: 10.1186/1756-3305-5-249. PubMed PMID: 23146340; PubMed Central PMCID: PMC3503571.

26. Araki AS, Ferreira GE, Mazzoni CJ, Souza NA, Machado RC, Bruno RV, et al. Multilocus analysis of divergence and introgression in sympatric and allopatric sibling species of the Lutzomyia longipalpis complex in Brazil. PLoS neglected tropical diseases. 2013;7(10):e2495. Epub 2013/10/23. doi: 10.1371/journal.pntd.0002495. PubMed PMID: 24147172; PubMed Central PMCID: PMCPMC3798421.

27. Palframan MJ, Bandi KK, Hamilton JGC, Pattenden G. Sobralene, a new sex-aggregation pheromone and likely shunt metabolite of the taxadiene synthase cascade, produced by a member of the sand fly Lutzomyia longipalpis species complex. Tetrahedron Lett. 2018;59(20):1921-3. Epub 2018/05/22. doi: 10.1016/j.tetlet.2018.03.088. PubMed PMID: 29780183; PubMed Central PMCID: PMCPMC5937913.

28. Araki AS, Vigoder FM, Bauzer LG, Ferreira GE, Souza NA, Araujo IB, et al. Molecular and behavioral differentiation among Brazilian populations of Lutzomyia longipalpis (Diptera: Psychodidae: Phlebotominae). PLoS neglected tropical diseases. 2009;3(1):e365. Epub

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

2009/01/28. doi: 10.1371/journal.pntd.0000365. PubMed PMID: 19172187; PubMed Central PMCID: PMCPMC2628317.

29. Souza NA, Andrade-Coelho CA, Vigoder FM, Ward RD, Peixoto AA. Reproductive isolation between sympatric and allopatric Brazilian populations of Lutzomyia longipalpis s.l. (Diptera: Psychodidae). Mem Inst Oswaldo Cruz. 2008;103(2):216-9. Epub 2008/04/22. PubMed PMID: 18425278.

30. Boulanger N, Lowenberger C, Volf P, Ursic R, Sigutova L, Sabatier L, et al. Characterization of a defensin from the sand fly Phlebotomus duboscqi induced by challenge with bacteria or the protozoan parasite Leishmania major. Infect Immun. 2004;72(12):7140-6. PubMed PMID: 15557638.

31. Telleria EL, Sant'Anna MR, Alkurbi MO, Pitaluga AN, Dillon RJ, Traub-Cseko YM. Bacterial feeding, Leishmania infection and distinct infection routes induce differential defensin expression in Lutzomyia longipalpis. Parasites & vectors. 2013;6:12. doi: 10.1186/1756-3305-6-12. PubMed PMID: 23311993; PubMed Central PMCID: PMC3573903.

32. Telleria EL, Tinoco-Nunes B, Leštinová T, de Avellar LM, Tempone AJ, Pitaluga AN, et al. Antimicrobial Peptides: Differential Expression during Development and Potential Involvement in Vector Interaction with Microbiota and. Microorganisms. 2021;9(6). Epub 20210611. doi: 10.3390/microorganisms9061271. PubMed PMID: 34207941; PubMed Central PMCID: PMCPMC8230673.

33. Kykalová B, Tichá L, Volf P, Loza Telleria E. Antimicrobial Peptides in Larvae and Females and a Gut-Specific Defensin Upregulated by. Microorganisms. 2021;9(11). Epub 20211106. doi: 10.3390/microorganisms9112307. PubMed PMID: 34835433; PubMed Central PMCID: PMCPMC8625375.

34. Telleria EL, Sant'Anna MR, Ortigao-Farias JR, Pitaluga AN, Dillon VM, Bates PA, et al. Caspar-like gene depletion reduces Leishmania infection in sand fly host Lutzomyia longipalpis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

J Biol Chem. 2012;287(16):12985-93. doi: 10.1074/jbc.M111.331561. PubMed PMID: 22375009; PubMed Central PMCID: PMC3339954.

35. Louradour I, Ghosh K, Inbar E, Sacks DL. CRISPR/Cas9 Mutagenesis in Phlebotomus papatasi: the Immune Deficiency Pathway Impacts Vector Competence for Leishmania major. mBio. 2019;10(4). Epub 20190827. doi: 10.1128/mBio.01941-19. PubMed PMID: 31455654; PubMed Central PMCID: PMCPMC6712399.

36. Abdeladhim M, Kamhawi S, Valenzuela JG. What's behind a sand fly bite? The profound effect of sand fly saliva on host hemostasis, inflammation and immunity. Infect Genet Evol. 2014;28:691-703. doi: 10.1016/j.meegid.2014.07.028. PubMed PMID: 25117872.

37. Yuan Q, Metterville D, Briscoe AD, Reppert SM. Insect cryptochromes: gene duplication and loss define diverse ways to construct insect circadian clocks. Mol Biol Evol. 2007;24(4):948-55. doi: 10.1093/molbev/msm011. PubMed PMID: 17244599.

38. Ceriani MF, Darlington TK, Staknis D, Mas P, Petti AA, Weitz CJ, et al. Light-dependent sequestration of TIMELESS by CRYPTOCHROME. Science. 1999;285(5427):553-6. PubMed PMID: 10417378.

39. Thomas MB. Biological control of human disease vectors: a perspective on challenges and opportunities. Biocontrol (Dordr). 2018;63(1):61-9. Epub 2018/02/03. doi: 10.1007/s10526-017-9815-y. PubMed PMID: 29391855; PubMed Central PMCID: PMCPMC5769823.

40. Modi GB, Tesh RB. A simple technique for mass rearing Lutzomyia longipalpis and Phlebotomus papatasi (Diptera: Psychodidae) in the laboratory. J Med Entomol. 1983;20(5):568-9. PubMed PMID: 6644754.

41. Souza NA, Vigoder FM, Araki AS, Ward RD, Kyriacou CP, Peixoto AA. Analysis of the copulatory courtship songs of Lutzomyia longipalpis in six populations from Brazil. J Med Entomol. 2004;41(5):906-13. Epub 2004/11/13. doi: 10.1603/0022-2585-41.5.906. PubMed PMID: 15535620.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

42. Deng J, Worley KC. Atlas-Link 2010. Available from: https://www.hgsc.bcm.edu/software/atlas-link.

43. Song X, Liu Y, Qu J, Gibbs RA, Worley KC. ATLAS gapfill 2.2. 2012.

44. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491. Epub 20111222. doi: 10.1186/1471-2105-12-491. PubMed PMID: 22192575; PubMed Central PMCID: PMCPMC3280279.

45. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of Aedes aegypti, a major arbovirus vector. Science. 2007;316(5832):1718-23. PubMed PMID: 17510324.

46. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 2013;41(Database issue):D358-65. doi: 10.1093/nar/gks1116. PubMed PMID: 23180791; PubMed Central PMCID: PMC3531149.

47. Neafsey DE, Christophides GK, Collins FH, Emrich SJ, Fontaine MC, Gelbart W, et al. The evolution of the Anopheles 16 genomes project. G3. 2013;3(7):1191-4. doi: 10.1534/g3.113.006247. PubMed PMID: 23708298; PubMed Central PMCID: PMC3704246.

48. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, et al. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science. 2015;347(6217):1258522. doi: 10.1126/science.1258522. PubMed PMID: 25554792.

49. Rognes T. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. BMC bioinformatics. 2011;12:221. doi: 10.1186/1471-2105-12-221. PubMed PMID: 21631914; PubMed Central PMCID: PMC3120707.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

50. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics. 2004;5:113. doi: 10.1186/1471-2105-5-113. PubMed PMID: 15318951; PubMed Central PMCID: PMC517706.

51. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3. doi: 10.1093/bioinformatics/btp348. PubMed PMID: 19505945; PubMed Central PMCID: PMC2712344.

52. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013;497(7449):327-31. doi: 10.1038/nature12130. PubMed PMID: 23657258.

53. Salichos L, Stamatakis A, Rokas A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol Biol Evol. 2014;31(5):1261-71. doi: 10.1093/molbev/msu061. PubMed PMID: 24509691.

54. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3. doi: 10.1093/bioinformatics/btu033. PubMed PMID: 24451623; PubMed Central PMCID: PMC3998144.

55. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26(13):1669-70. doi: 10.1093/bioinformatics/btq243. PubMed PMID: 20472542; PubMed Central PMCID: PMC2887050.

56. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao FA, Pozdnyakov IA, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 2015;43(Database issue):D250-6. doi: 10.1093/nar/gku1220. PubMed PMID: 25428351; PubMed Central PMCID: PMC4383991.

57. Drillon G, Carbone A, Fischer G. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PloS one. 2014;9(3):e92621. doi:

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

10.1371/journal.pone.0092621. PubMed PMID: 24651407; PubMed Central PMCID: PMC3961402.

58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PubMed PMID: 2231712.

59. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(Database issue):D61-5. doi: 10.1093/nar/gkl842. PubMed PMID: 17130148; PubMed Central PMCID: PMC1716718.

60. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, et al. VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Res. 2007;35(Database issue):D503-5. PubMed PMID: 17145709.

61. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]; 2013.

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi: papers2://publication/doi/10.1093/bioinformatics/btp352.

63. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genes Dev. 2012;22(3):568-76. doi: papers2://publication/doi/10.1101/gr.129684.111.

64. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. Epub 2015/02/28. doi: 10.1186/s13742-015-0047-8. PubMed PMID: 25722852; PubMed Central PMCID: PMCPMC4342193.

65. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Genet. 2007;81(3):559-75. Epub 2007/08/19. doi: 10.1086/519795. PubMed PMID: 17701901; PubMed Central PMCID: PMCPMC1950838.

66. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326-8. doi: papers3://publication/doi/10.1093/bioinformatics/bts606.

67. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. doi: 10.1093/bioinformatics/btr330.

68. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics (Oxford, England). 2008;24(11):1403-5. Epub 2008/04/10. doi: 10.1093/bioinformatics/btn129. PubMed PMID: 18397895.

69. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics (Oxford, England). 2011;27(21):3070-1. Epub 2011/09/20. doi: 10.1093/bioinformatics/btr521. PubMed PMID: 21926124; PubMed Central PMCID: PMCPMC3198581.

70. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics (Oxford, England). 2004;20(2):289-90. Epub 2004/01/22. PubMed PMID: 14734327.

71. Kamvar ZN, Tabima JF, Grunwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2:e281. Epub 2014/04/02. doi: 10.7717/peerj.281. PubMed PMID: 24688859; PubMed Central PMCID: PMCPMC3961149.

72. Knaus BJ, Grunwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour. 2017;17(1):44-53. Epub 2016/07/13. doi: 10.1111/1755-0998.12549. PubMed PMID: 27401132.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

73. Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A. 1996;93(23):13429-34. Epub 1996/11/12. PubMed PMID: 8917608; PubMed Central PMCID: PMCPMC24110.

74. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009;19(9):1655-64. Epub 2009/08/04. doi: 10.1101/gr.094052.109. PubMed PMID: 19648217; PubMed Central PMCID: PMCPMC2752134.

75. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour. 2015;15(5):1179-91. Epub 2015/02/17. doi: 10.1111/1755-0998.12387. PubMed PMID: 25684545; PubMed Central PMCID: PMCPMC4534335.

76. Suite 2011: LigPrep, version 2.5, Schrödinger, LLC, New York, NY, 2011.

77. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190. Epub 2006/12/30. doi: 10.1371/journal.pgen.0020190. PubMed PMID: 17194218; PubMed Central PMCID: PMCPMC1713260.

78. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics (Oxford, England). 2011;27(15):2156-8. Epub 2011/06/10. doi: 10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central PMCID: PMCPMC3137218.

79. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics. 2010;186(1):241-62. Epub 2010/09/22. doi: 10.1534/genetics.104.117275. PubMed PMID: 20855576; PubMed Central PMCID: PMCPMC2940290.

80. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Genetics. 2013;193(3):929-41. Epub 2013/01/12. doi: 10.1534/genetics.112.147231. PubMed PMID: 23307896; PubMed Central PMCID: PMCPMC3584007.

81. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;7:539. doi: 10.1038/msb.2011.75. PubMed PMID: 21988835; PubMed Central PMCID: PMC3261699.

82. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(Web Server issue):W609-12. doi: 10.1093/nar/gkl315. PubMed PMID: 16845082; PubMed Central PMCID: PMC1538804.

83. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586-91. doi: 10.1093/molbev/msm088. PubMed PMID: 17483113.

84. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics (Oxford, England). 2015;31(19):3210-2. Epub 2015/06/11. doi: 10.1093/bioinformatics/btv351. PubMed PMID: 26059717.

85. Zdobnov EM, Bork P. Quantification of insect genome divergence. Trends Genet. 2007;23(1):16-20. Epub 2006/11/14. doi: 10.1016/j.tig.2006.10.004. PubMed PMID: 17097187.

86. Filho JD, Brazil RP. Relationships of new world phlebotomine sand flies (Diptera: Psychodidae) based on fossil evidence. Mem Inst Oswaldo Cruz. 2003;98 Suppl 1:145-9. PubMed PMID: 12687775.

87. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, et al. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science. 2002;298(5591):149-59. PubMed PMID: 12364792.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

88. Gilbert C, Peccoud J, Cordaux R. Transposable Elements and the Evolution of Insects. Annu Rev Entomol. 2021;66:355-72. Epub 20200915. doi: 10.1146/annurev-ento-070720-074650. PubMed PMID: 32931312.

89. Petrov DA, Hartl DL. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol Biol Evol. 1998;15(3):293-302. Epub 1998/03/21. doi: 10.1093/oxfordjournals.molbev.a025926. PubMed PMID: 9501496.

90. Vasta GR. Roles of galectins in infection. Nature reviews Microbiology. 2009;7(6):424-38. doi: 10.1038/nrmicro2146. PubMed PMID: 19444247; PubMed Central PMCID: PMC3759161.

91. Kamhawi S, Ramalho-Ortigao M, Pham VM, Kumar S, Lawyer PG, Turco SJ, et al. A role for insect galectins in parasite survival. Cell. 2004;119(3):329-41. PubMed PMID: 15543683.

92. Grimaldi D, Engel MS. Evolution of the Insects. New York, NY: Cambridge University Press; 2005.

93. Abrudan J, Ramalho-Ortigao M, O'Neil S, Stayback G, Wadsworth M, Bernard M, et al. The characterization of the Phlebotomus papatasi transcriptome. Insect Mol Biol. 2013;22(2):211-32. doi: 10.1111/imb.12015. PubMed PMID: 23398403; PubMed Central PMCID: PMC3594503.

94. Coutinho-Abreu IV, Sharma NK, Robles-Murguia M, Ramalho-Ortigao M. Targeting the midgut secreted PpChit1 reduces Leishmania major development in its natural vector, the sand fly Phlebotomus papatasi. PLoS neglected tropical diseases. 2010;4(11):e901. Epub 2010/12/15. doi: 10.1371/journal.pntd.0000901. PubMed PMID: 21152058; PubMed Central PMCID: PMCPMC2994919.

95. Coutinho-Abreu IV, Sharma NK, Robles-Murguia M, Ramalho-Ortigao M. Characterization of Phlebotomus papatasi peritrophins, and the role of PpPer1 in Leishmania major survival in its natural vector. PLoS Negl Trop Dis. 2013;7(3):e2132. doi:

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

10.1371/journal.pntd.0002132. PubMed PMID: 23516661; PubMed Central PMCID: PMC3597473.

96. Ortigao-Farias JR, Di-Blasi T, Telleria EL, Andorinho AC, Lemos-Silva T, Ramalho-Ortigao M, et al. Alternative splicing originates different domain structure organization of Lutzomyia longipalpis chitinases. Mem Inst Oswaldo Cruz. 2018;113(2):96-101. Epub 2017/12/14. doi: 10.1590/0074-02760170179. PubMed PMID: 29236932; PubMed Central PMCID: PMCPMC5722264.

97. Pitaluga AN, Beteille V, Lobo AR, Ortigao-Farias JR, Davila AM, Souza AA, et al. EST sequencing of blood-fed and Leishmania-infected midgut of Lutzomyia longipalpis, the principal visceral leishmaniasis vector in the Americas. Mol Genet Genomics. 2009;282(3):307-17. Epub 2009/07/01. doi: 10.1007/s00438-009-0466-2. PubMed PMID: 19565270.

98. Pruzinova K, Sadlova J, Seblova V, Homola M, Votypka J, Volf P. Comparison of Bloodmeal Digestion and the Peritrophic Matrix in Four Sand Fly Species Differing in Susceptibility to Leishmania donovani. PloS one. 2015;10(6):e0128203. Epub 2015/06/02. doi: 10.1371/journal.pone.0128203. PubMed PMID: 26030610; PubMed Central PMCID: PMCPMC4452187.

99. Ramalho-Ortigao JM, Kamhawi S, Rowton ED, Ribeiro JM, Valenzuela JG. Cloning and characterization of trypsin- and chymotrypsin-like proteases from the midgut of the sand fly vector Phlebotomus papatasi. Insect Biochem Mol Biol. 2003;33(2):163-71. PubMed PMID: 12535675.

100. Ramalho-Ortigao JM, Temporal P, de Oliveira SM, Barbosa AF, Vilela ML, Rangel EF, et al. Characterization of constitutive and putative differentially expressed mRNAs by means of expressed sequence tags, differential display reverse transcriptase-PCR and randomly amplified polymorphic DNA-PCR from the sand fly vector Lutzomyia longipalpis. Mem Inst Oswaldo Cruz. 2001;96(1):105-11. PubMed PMID: 11285481.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

101. Ramalho-Ortigao JM, Traub-Cseko YM. Molecular characterization of Llchit1, a midgut chitinase cDNA from the leishmaniasis vector Lutzomyia longipalpis. Insect Biochem Mol Biol. 2003;33(3):279-87. PubMed PMID: 12609513.

102. Vale VF, Moreira BH, Moraes CS, Pereira MH, Genta FA, Gontijo NF. Carbohydrate digestion in Lutzomyia longipalpis' larvae (Diptera - Psychodidae). J Insect Physiol. 2012;58(10):1314-24. Epub 2012/07/31. doi: 10.1016/j.jinsphys.2012.07.005. PubMed PMID: 22841889.

103. da Costa-Latgé SG, Bates P, Dillon R, Genta FA. Characterization of Glycoside Hydrolase Families 13 and 31 Reveals Expansion and Diversification of α-Amylase Genes in the Phlebotomine. Front Physiol. 2021;12:635633. Epub 20210409. doi: 10.3389/fphys.2021.635633. PubMed PMID: 33897451; PubMed Central PMCID: PMCPMC8063059.

104. Benoit JB, Hansen IA, Szuter EM, Drake LL, Burnett DL, Attardo GM. Emerging roles of aquaporins in relation to the physiology of blood-feeding arthropods. Journal of comparative physiology B, Biochemical, systemic, and environmental physiology. 2014;184(7):811-25. doi: 10.1007/s00360-014-0836-x. PubMed PMID: 24942313.

105. Meireles-Filho AC, Kyriacou CP. Circadian rhythms in insect disease vectors. Mem Inst Oswaldo Cruz. 2013;108 Suppl 1:48-58. doi: 10.1590/0074-0276130438. PubMed PMID: 24473802; PubMed Central PMCID: PMCPMC4109179.

106. Meireles-Filho AC, da S Rivas GB, Gesto JS, Machado RC, Britto C, de Souza NA, et al. The biological clock of an hematophagous insect: locomotor activity rhythms, circadian expression and downregulation after a blood meal. FEBS Lett. 2006;580(1):2-8. Epub 20051201. doi: 10.1016/j.febslet.2005.11.031. PubMed PMID: 16337945.

107. Meireles-Filho AC, Amoretty PR, Souza NA, Kyriacou CP, Peixoto AA. Rhythmic expression of the cycle gene in a hematophagous insect vector. BMC Mol Biol. 2006;7:38. doi: 10.1186/1471-2199-7-38. PubMed PMID: 17069657; PubMed Central PMCID: PMC1636064.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

108. Freeman EG, Dahanukar A. Molecular neurobiology of Drosophila taste. Curr Opin Neurobiol. 2015;34:140-8. Epub 2015/06/24. doi: 10.1016/j.conb.2015.06.001. PubMed PMID: 26102453; PubMed Central PMCID: PMCPMC4577450.

109. Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila. Cell. 2009;136(1):149-62. Epub 2009/01/13. doi: 10.1016/j.cell.2008.12.001. PubMed PMID: 19135896; PubMed Central PMCID: PMCPMC2709536.

110. 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(8):e1001064. Epub 2010/09/03. doi: 10.1371/journal.pgen.1001064. PubMed PMID: 20808886; PubMed Central PMCID: PMCPMC2924276.

111. Rytz R, Croset V, Benton R. Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in Drosophila and beyond. Insect biochemistry and molecular biology. 2013;43(9):888-97. Epub 2013/03/06. doi: 10.1016/j.ibmb.2013.02.007. PubMed PMID: 23459169.

112. Hickner PV, Timoshevskaya N, Nowling RJ, Labbé F, Nguyen AD, McDowell MA, et al. Molecular signatures of sexual communication in the phlebotomine sand flies. PLoS Negl Trop Dis. 2020;14(12):e0008967. Epub 20201228. doi: 10.1371/journal.pntd.0008967. PubMed PMID: 33370303; PubMed Central PMCID: PMCPMC7793272.

113. Liu N, Li T, Wang Y, Liu S. G-Protein Coupled Receptors (GPCRs) in Insects-A Potential Target for New Insecticide Development. Molecules. 2021;26(10). Epub 20210518. doi: 10.3390/molecules26102993. PubMed PMID: 34069969; PubMed Central PMCID: PMCPMC8157829.

114. Nowling RJ, Abrudan JL, Shoue DA, Abdul-Wahid B, Wadsworth M, Stayback G, et al. Identification of novel arthropod vector G protein-coupled receptors. Parasites & vectors.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

2013;6:150. doi: 10.1186/1756-3305-6-150. PubMed PMID: 23705687; PubMed Central PMCID: PMC3680159.

115. Fevereisen, R. Insect CYP, genes and P450 enzymes. In: Gilbert, LI, editors. Insect Molecular Biology and Biochemistry. Amsterdam: Elsevier; 2012. p. 236-316.

116. Bates PA, Depaquit J, Galati EA, Kamhawi S, Maroli M, McDowell MA, et al. Recent advances in phlebotomine sand fly research related to leishmaniasis control. Parasites & vectors. 2015;8:131. Epub 2015/04/18. doi: 10.1186/s13071-015-0712-x. PubMed PMID: 25885217; PubMed Central PMCID: PMCPMC4352286.

117. Mccoy KD. The population genetic structure of vectors and our understanding of disease epidemiology. Parasite-Journal De La Societe Francaise De Parasitologie. 2008;15(3):444-8. doi: DOI 10.1051/parasite/2008153444. PubMed PMID: WOS:000258945000042.

118. Vigoder FM, Souza NA, Brazil RP, Bruno RV, Costa PL, Ritchie MG, et al. Phenotypic differentiation in love song traits among sibling species of the Lutzomyia longipalpis complex in Brazil. Parasites & vectors. 2015;8:290. Epub 2015/05/29. doi: 10.1186/s13071-015-0900-8. PubMed PMID: 26017472; PubMed Central PMCID: PMCPMC4456791.

119. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23(13):3133-57. Epub 2014/05/23. doi: 10.1111/mec.12796. PubMed PMID: 24845075.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Supporting Information

Supplemental Methods: Methods used for manual annotation

Supplemental Results: Detailed information of annotated gene families

S1 Table: Assembly statistics

S2 Table: BUSCO analysis

S3 Table: RNAseq samples

S4 Table: Microsyntenny

S5 Table: Toll pathway annotation

S6 Table: Insect immune deficiency pathway annotation

S7 Table: JakStat pathway annotation

S8 Table: Galectin family annotation

S9 Table: Transforming growth factor-beta family annotation

S10 Table: Mitogen activated protein kinase family annotation

S11 Table: Prophenoloxidase family annotation

S12 Table: Salivary protein annotation

S13 Table: Peptidase annotation

S14 Table: Glycosidase Hydrolase family 13 annotation

S15 Table: Chitinase family annotation

S16 Table: Hexosaminidase family annotation

S17 Table: Chitinase deacetylase family annotation

S18 Table: Peritrophin family annotation

S19 Table: Aquoporin family annotation

S20 Table: *Lutzomyia longipalpis* circadian rhythm pathway annotation

S21 Table: *Phlebotomus papatasi* circadian rhythm pathway annotation

S22 Table: Number of chemoreceptor family genes

S23 Table: G-protein coupled receptor family annotation

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

S24 Table: MicroRNA annotation

S25 Table: Heat shock and hypoxia gene family annotation

S26 Table: Cuticular protein gene family annotation

S27 Table: Juvenile hormone family annotation

S28 Table: Insulin signaling pathway annotation

S29 Table: Antioxidant family annotation

S30 Table: Vitamin metabolism pathway annotation

S31 Table: *Phlebotomus papatasi* population sequencing median coverage depth

S32 Table: *Phlebotomus papatasi* population variant summary statistics

S33 Table: *Phlebotomus papatasi* F_{ST}-Tajima's D overlap (including 10kb upstream and downstream)

S34 Table: *Lutzomyia longipalpis* population sequencing median coverage depth

S35 Table: Parameter values of male copulatory songs from *Lutzomyia longipalpis* from Araci and Olindina

S36 Table: *Lutzomyia longipalpis* genes within differentiation islands

S1 Figure: Conflicting phylogenetic signals. Analysis of the gene phylogenies of individual orthologous groups identified three major topologies with sand fly-mosquito (41%), sand fly-fly (37%), or mosquito-fly (22%) sister clades. Comparisons of average branch lengths for each topology suggest that, although substitution rates in flies are always higher, orthologs that support the sand fly-mosquito topology show the lowest substitution rates in flies and the smallest differences in substitution rates among the fly, sand fly, and mosquito clades. In contrast, the sand fly-fly and mosquito-fly topologies show much higher substitution rates in flies and much greater differences in substitution rates among the three clades.

S2 Figure: Scatter plots of ortholog locations in each pair of genomes. 1- to -1 ortholog groups were associated utilizing OrthoDB and scaffolds with only one gene were removed prior

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

to generating plots. (A) *P. papatasi* vs. *Lu. longipalpis*. (B) *Lu. longipalpis* vs. *An. gambiae*. (C) *P. papatasi* vs. *An. gambiae*. (D) *Lu. longipalpis* vs. *D. melanogaster*. (E) *P. papatasi* vs. *D. melanogaster*. (F) *D. melanogaster* vs. *D. simulans*. (F) *D. melanogaster* vs. *An. gambiae*.

S3 Figure: Clustering of sand fly galectin protein sequences. Condensed Neighbor-Joining tree depicting clustering among galectin protein sequences of sand flies (*P. papatasi* and *Lu. longipalpis*; open and filled squares, respectively), mosquitoes (*Ae. aegypti* and *An. gambiae*; open and filled circles, respectively), fly (*D. melanogaster;* filled triangle), eastern oyster (*C. virginica*; upside-down open triangle), and freshwater snail (*B. glabrata*; upside-down filled triangle). Branches encompassing shared orthologs are highlighted by blue shades. Sand fly specific clusters and genes are highlighted by orange shades. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. One thousand bootstrap replicates were performed, and only branches displaying at least 50% confidence are shown.

S4 Figure: Condensed Neighbor-Joining tree depicting clustering among n-acetylhexosaminidase protein sequences of sand flies (*P. papatasi* and *Lu. longipalpis*; open and filled squares, respectively), mosquitoes (*Ae. aegypti* and *An. gambiae*; open and filled circles, respectively), fly (*D. melanogaster*; filled triangle), and beetle (*T. castaneum*; filled diamond). Branches encompassing sequences belonging to group I-IV n-acetylhexosaminidases are highlighted by a blue shade. The sand fly specific cluster is highlighted by an orange shade. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. One thousand bootstrap replicates were performed, and only branches displaying at least 50% confidence are shown.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

S5 Figure: Condensed Neighbor-Joining tree depicting clustering among chitin deacetylase catalytic domain sequences of sand flies (*P. papatasi* and *Lu. longipalpis*; open and filled squares, respectively), mosquitoes (*Ae. aegypti* and *An. gambiae*; open and filled circles, respectively), fly (*D. melanogaster*; filled triangle), and beetle (*T. castaneum*; filled diamond). Branches encompassing sequences belonging to group 1-5 and 9 CDA are highlighted by blue shades. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. One thousand bootstrap replicates were performed, and only branches displaying at least 50% confidence are shown.

S6 Figure: Condensed Maximum likelihood tree depicting peritrophin CBD domain similarities among the sand flies *P. papatasi* and *Lu. longipalpis* and the red flour beetle *T. castaneum*. Open squares, filled squares, and filled diamonds represent *P. papatasi*, *Lu. longipalpis*, and *T. castaneum* domains, respectively. Branches exclusive to *T. castaneum* were color-coded in magenta; those specific to sand flies were highlighted in blue. The branch encompassing the CBD-like domain “CBDput” is highlighted in green. The branches shared by sand flies and RFB CBD domains are color-coded in orange. Maximum likelihood tree was constructed using the Whelan and Goldman (WAG) model with Gamma distributed among Invariant sites (G+I), as suggested by the Model test function of the Mega6 software. One thousand bootstrap replicates were performed, and only branches displaying at least 50% confidence are shown.

S7 Figure: Comparison of predicted aquaporins from other flies. Neighbor-joining tree was produced using MEGA6 using Dayhoff Model and pairwise matching; branch values indicate support following 3000 bootstraps; values below 50% are omitted.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

S8 Figure: Molecular phylogenetic analysis of vertebrate and invertebrate photolyases containing *Lu. longipalpis* and *P. papatasi* gene models. The different photoyases are displayed on the right. The evolutionary history was inferred by using the Maximum Likelihood method based on the Jones-Taylor-Thorton + four gamma categories with 1000 bootstrap replicates (showing only above 65). Sequences with squares are vertebrate cryptochromes (black – cry-4, white – cry-1, cry-2, and cry-3); sequences with black traingles represent (6-4) insect photolyases; sequences with inverted black triangles are reprenting all insect photolyase repir proteins; and sequences with a dot symbol show insect cryptochromes (black –cry-1, white –cry-2). Dashed arrows point to *P. papatasi* photolyase sequences and straight arrows to *Lu. longipalpois* photolyase sequences.

S9 Figure: Molecular phylogenetic analysis of *Lu. longipalpis, P. papatasi* and *D. melanogaster* TRP channel sequences. The different TRP subfamilies are displayed on the right. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan and Goldman +Freq. model with 1000 bootstrap replicates.

S10 Figure: Molecular phylogenetic analysis of *Lu. longipalpis, P. papatasi* and *D. melanogaster* PPK sequences. The different PPK subfamilies are displayed on the right. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan and Goldman +Freq. model with 1000 bootstrap replicates.

S11 Figure: Comparison of predicted aquaporins from other flies. Neighbor-joining tree was produced using MEGA6 using Dayhoff Model and pairwise matching; branch values indicate support following 3000 bootstraps; values below 50% are omitted.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

S12 Figure: *Phlebotomus papatasi* population structure. Inferred population structure of *P. papatasi* individuals collected from Afghanistan (PPAFG; green), North Sinai - Egypt (PPNS; purple), and Tunisia (PPTUN; orange). (A) Phylogenetic Analysis. Rooted neighbor joining (NJ) radial tree generated with the Adegenet and ape packages of R. We included both *P. bergeroti* (PBRG; black) and *P. duboscqi* (PDMA; gray), and used *P. duboscqi* to root the trees. Bootstrap values represent the percentage of 1,000 replicates. (B) Principle component analysis (PCA). Individuals are plotted according to their coordinates on the first two principal components (PC1 and PC2). (C) Admixture analysis. Ancestry proportions for Admixture models from K = 2 to K = 7 ancestral populations. Each individual is represented by a thin vertical line, partitioned into K coloured segments representing the individual's estimated membership fractions to the K clusters. These data are the average of the major q-matrix clusters derived by CLUMPAK analysis.

S13 Figure: Admixture cross validation error. Violin plot of the cross-validation error for each of 30 replicates for each K value. (A) *Phlebotomus papatasi* populations. (B) *Lutzomyia longipalpis* populations.

S14 Figure: Male copulatory courtship songs from Araci and Olinda. (A) Approximate distance of Araci and Olinda from Jacobina (B). Male copulatory courtship song tracings of *Lutzomyia longialpis* males collected from Araci and Olindina. The figure shows ~1 s of song in each case. Map Source: Google Imagery TerraMetrics.

S15 Figure : Distribution plots of the pairwise F_{ST} between the different populations of *Lutzomyia longipalpis*. Weighted F_{ST} values for 1kb non-overlapping windows were calculated across the genome for each population comparison.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

S16 Figure: Manhattan plots of the pairwise F_{ST} between the different populations of *Lutzomyia longipalpis*. The red horizontal lines indicate the upper 0.05% of F_{ST} distribution over the entire genome.

S17 Figure: Manhattan plot of Tajimas'D for each population of *Lutzomyia longipalpis*. The red and blue horizontal lines indicate the upper and lower 0.05% of Tajima's D distribution, respectively.

S18 Figure: Genomic regions with high (outlier) Tajimas'D for different populations of *Lutzomyia longipalpis*. (A) The Venn diagram summarizes the numbers of 1kb genomic windows with Tajimas'D values in the upper 2.5% of the different populations. (B) The Venn diagram summarizes the numbers of 1kb genomic windows with Tajimas'D values in the lower 2.5% of the different populations.

S1 Data File: *Phlebotomus papatasi* CYPome Fasta File

S2 Data File: *Lutzomyia longipalpis* CYPome Fasta File

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.05.510924; this version posted October 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

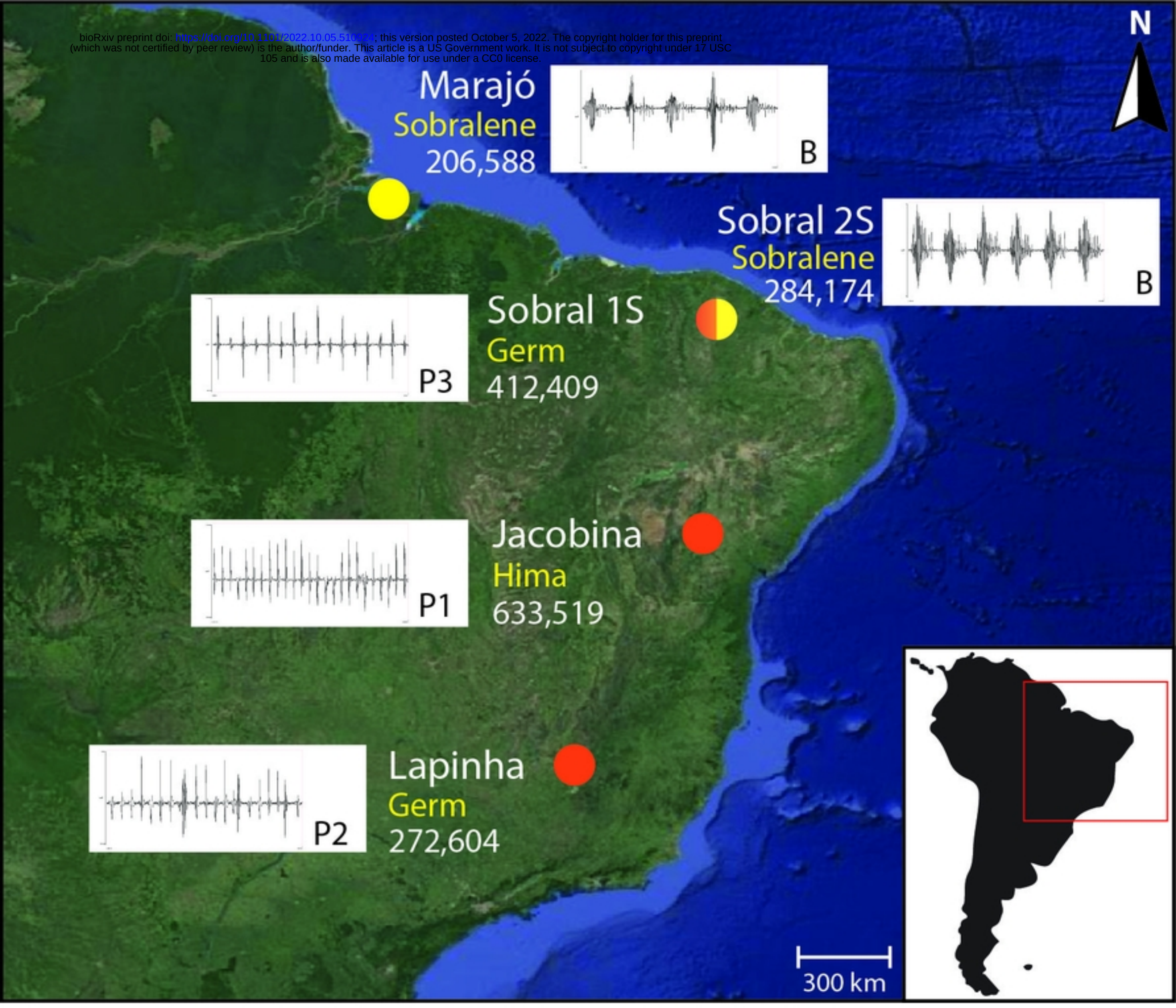

Figure 1

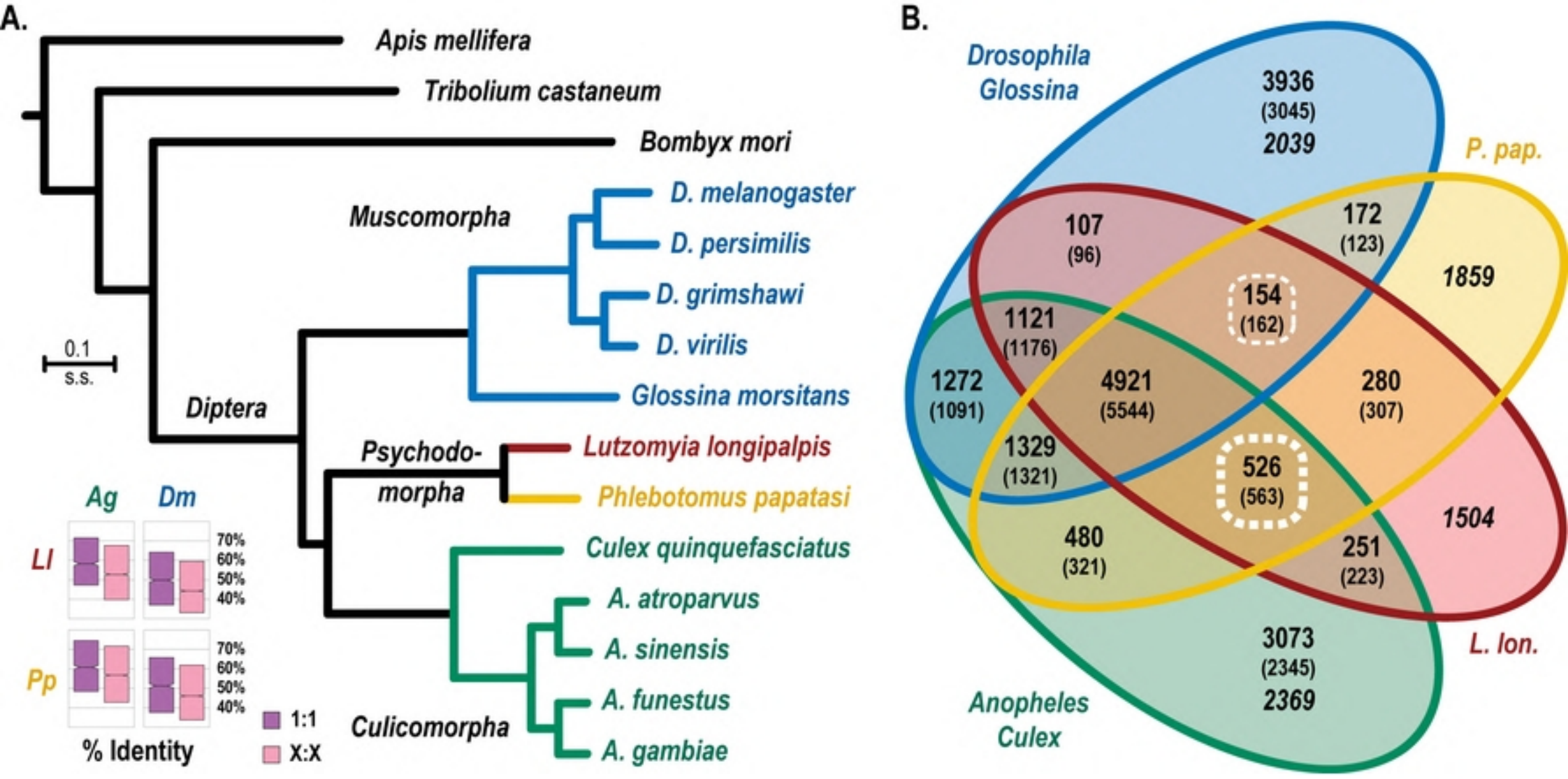
A.
Apis mellifera
Tribolium castaneum
Bombyx mori
Muscomorpha
D. melanogaster
D. persimilis
D. grimshawi
D. virilis
Glossina morsitans
0.1
s.s.
Diptera
Psychodo-
morpha
Lutzomyia longipalpis
Phlebotomus papatasi
Culex quinquefasciatus
A. atroparvus
A. sinensis
A. funestus
A. gambiae
Culicomorpha
Ag
Dm
LI
Pp
70%
60%
50%
40%
% Identity
1:1
X:X
B.
Drosophila
Glossina
3936
(3045)
2039
P. pap.
107
(96)
172
(123)
1859
154
(162)
1121
(1176)
1272
(1091)
4921
(5544)
280
(307)
1329
(1321)
526
(563)
1504
480
(321)
251
(223)
L. lon.
3073
(2345)
2369
Anopheles
Culex

Figure 2

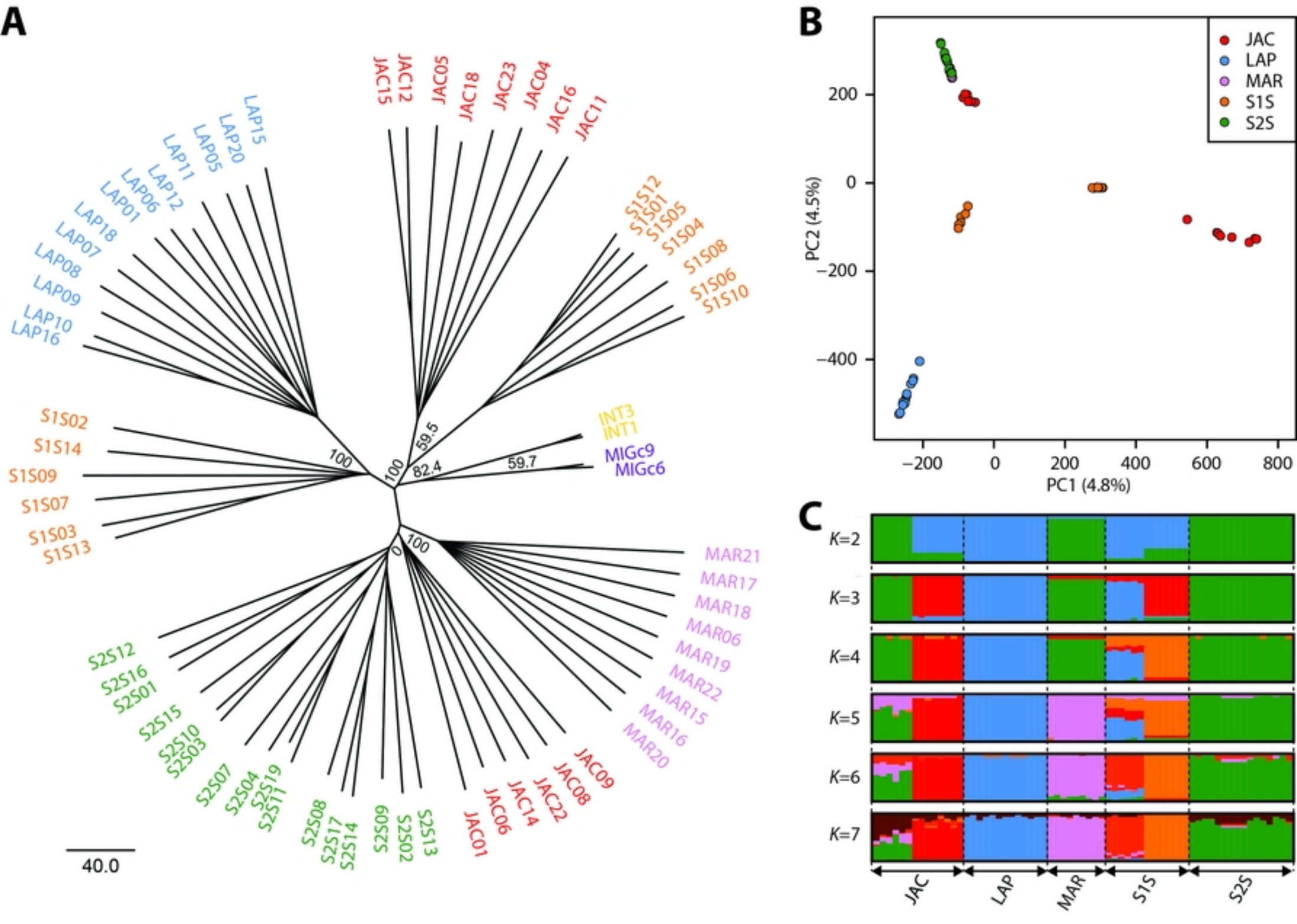
A
JAC15
JAC12
JAC05
JAC18
JAC23
JAC04
JAC16
JAC11
LAP15
LAP20
LAP05
LAP11
LAP12
LAP06
LAP01
LAP18
LAP07
LAP08
LAP09
LAP10
LAP16
S1S12
S1S01
S1S05
S1S04
S1S08
S1S06
S1S10
S1S02
S1S14
S1S09
S1S07
S1S03
S1S13
INT3
INT1
MIGc9
MIGc6
MAR21
MAR17
MAR18
MAR06
MAR19
MAR22
MAR15
MAR16
MAR20
JAC09
JAC08
JAC22
JAC14
JAC06
JAC01
S2S13
S2S02
S2S09
S2S14
S2S17
S2S08
S2S11
S2S19
S2S04
S2S07
S2S03
S2S10
S2S15
S2S01
S2S16
S2S12
100
100
59.5
82.4
59.7
0
100
40.0
B
JAC
LAP
MAR
S1S
S2S
PC2 (4.5%)
PC1 (4.8%)
200
0
−200
−400
−200
0
200
400
600
800
C
K=2
K=3
K=4
K=5
K=6
K=7
JAC
LAP
MAR
S1S
S2S

Figure 3

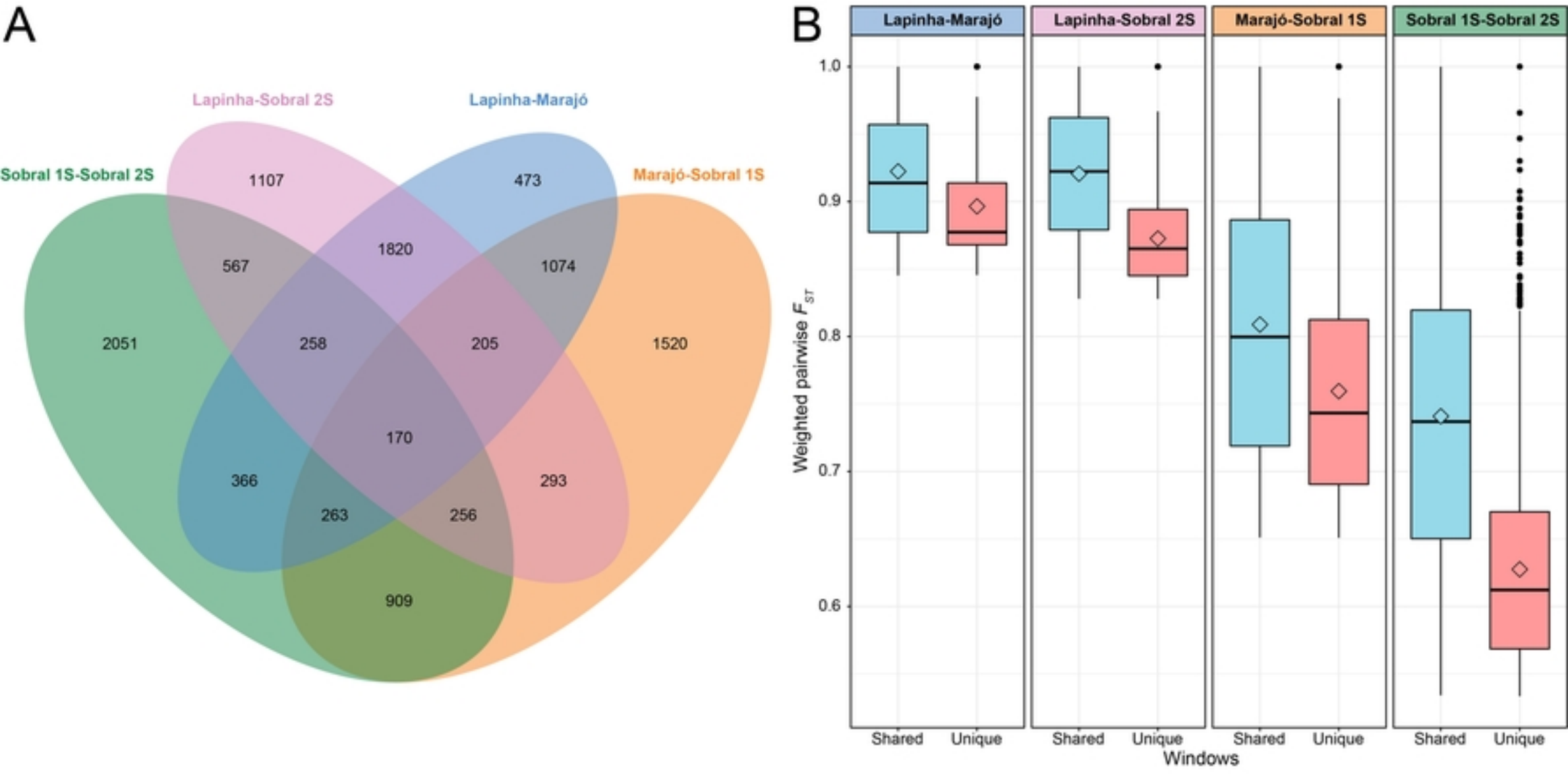
A
Lapinha-Sobral 2S
Lapinha-Marajó
Sobral 1S-Sobral 2S
Marajó-Sobral 1S
1107
473
1820
567
1074
2051
258
205
1520
170
366
293
263
256
909
B
Lapinha-Marajó
Lapinha-Sobral 2S
Marajó-Sobral 1S
Sobral 1S-Sobral 2S
Weighted pairwise F_{ST}
1.0
0.9
0.8
0.7
0.6
Shared
Unique
Windows

Figure 4

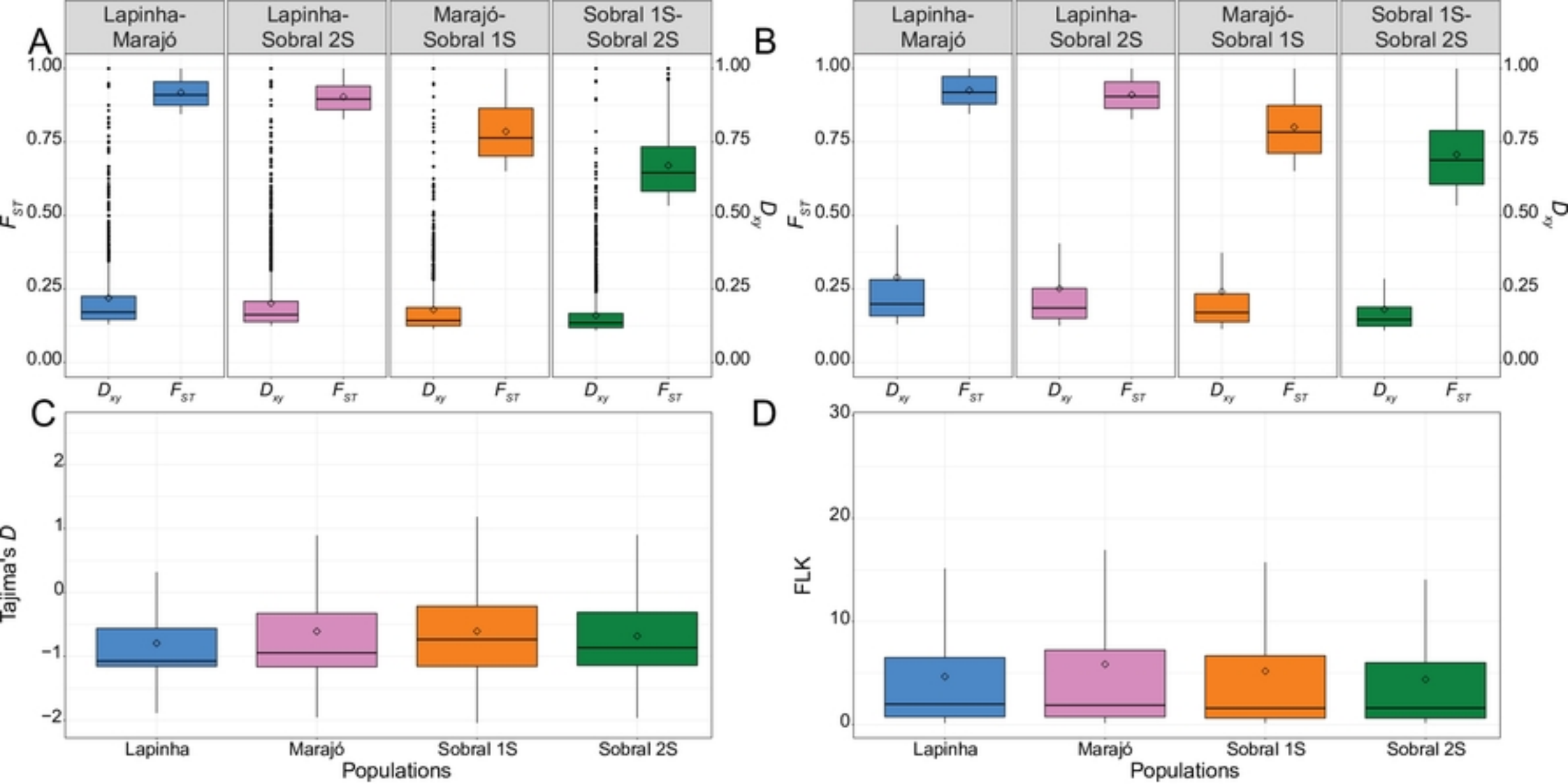
A
Lapinha-Marajó
Lapinha-Sobral 2S
Marajó-Sobral 1S
Sobral 1S-Sobral 2S
F_{ST}
D_{xy}
B
C
Tajima's D
Populations
Lapinha
Marajó
Sobral 1S
Sobral 2S
D
FLK

Figure 5