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

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.

52 Ciencias Básicas Médicas, Universidad Icesi, Calle 18 No 122-135, Cali, Colombia; [21]Department of Life
53 Sciences, Faculty of Science and Technology, Al-Quds University, Jerusalem, Palestine. [22]USDA-ARS
54 Knipling-Bushland U.S. Livestock Insects Research Laboratory and Veterinary Pest Genomics Center,
55 Kerrville, TX, USA; [23]CNRS-UPR9022 Institut de Biologie Moléculaire et Cellulaire and Faculté des
56 Sciences de la Vie-Université de Strasbourg, Strasbourg Cedex, France; [24]Genomics Group –
57 Bioinformatics and Evolutionary Biology Lab, Department of Genetics and Microbiology, Autonomous
58 University of Barcelona, Barcelona, Spain; [25]Medical Entomology Branch, Dept. Microbiology, Bundeswehr
59 Hospital, Bernhard-Nocht-Str. 74, 20359 Hamburg, Germany; [26]Medical Zoology Branch, Dept.
60 Microbiology, Central Bundeswehr Hospital, Andernacher Str. 100, 56070 Koblenz, Germany; [27]Laboratorio
61 de Insectos Sociales, Instituto de Fisiología, Biología Molecular y Neurociencias, Universidad de Buenos Aires -
62 CONICET, Buenos Aires, Argentina; [28]Laboratory of Cell Biology, Development and Genetics, Department of
63 Biochemistry, University of Oxford, Oxford, United Kingdom; [29]Laboratório de Imunoparasitologia, CPqGM,
64 Fundação Oswaldo Cruz, Bahia, Brazil; [30]Department of Neuroscience, Perelman School of Medicine,
65 University of Pennsylvania, Philadelphia, PA 19104, USA; [31]Department of Electrical Engineering and
66 Computer Science, Milwaukee School of Engineering, Milwaukee, WI, 53202, USA; [32]Instituto Oswaldo
67 Cruz - Fiocruz, Av. Brasil, 4365, Pav. Leonidas Deane, lab 304, Manguinhos - Rio de Janeiro - RJ - Brazil,
68 CEP: 21040-360; [33]Laboratório de Transmissores de Hematozoários, IOC, FIOCRUZ, Rio de Janeiro,
69 Brazil; [34]Laboratório de Biologia Molecular de Parasitas e Vetores, Instituto Oswaldo Cruz/FIOCRUZ, Rio
70 de Janeiro, Brazil; [35]F. Edward Hebert School of Medicine, Department of Preventive Medicine and
71 Biostatistics, Uniformed Services University of the Health Sciences (USUHS), Bethesda, MD, 20814, USA;
72 [36]Human Genome Sequencing Center, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030,
73 USA; [37]Laboratory of Medical Entomology, René Rachou Institute-FIOCRUZ, Minas Gerais, Brazil;
74 [38]Laboratory Interdisciplinar em Vigilancia Entomologia em Diptera e Hemiptera, Fiocruz, Rio de Janeiro, Brazil;
75 [39]Department of Entomology, University of Kentucky, Lexington, KY 40546, USA; [40]Department of
76 Parasitology, Faculty of Science, Charles University, Vinicna 7, 128 44, Prague, Czech Republic;
77 [41]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108 USA;
78 [42]Fralin Life Science Institute and Department of Biochemistry, Virginia Tech, 303 Fralin, Blacksburg, VA,
79 24061, USA; [43]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre

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

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.

131 **Introduction**

132 Phlebotomine sand flies are a group of vectors that vary widely in their geographic distribution,
133 ecology, and the pathogens they transmit. Sand flies serve as vectors for several established,
134 emerging, and re-emerging infectious diseases, transmitting protist, bacterial and viral pathogens.
135 The most devastating of the sand fly transmitted diseases is leishmaniasis, accounting for an
136 estimated 2.4 million disability-adjusted life-years (DALYs) [1] and 40,000 deaths annually [2].
137 These statistics are largely underestimated due to misdiagnosis and lacking surveillance systems
138 in many of the affected countries. Political instability, urbanization, and climate change are
139 expanding *Leishmania*-endemic regions and increasing the propensity for epidemics world-wide
140 [3]. These factors are coupled with the increase of visceral disease due to HIV co-infection,
141 spurring the World Health Organization to classify leishmaniasis as one of the world's epidemic-
142 prone diseases [4].

143 Leishmaniasis is worldwide, occurring in 98 countries and spread over five continents, putting
144 310 million people at risk of contracting leishmaniasis [2]. Leishmaniasis is a collective term for
145 a group of distinct clinical manifestations ranging from mild, often self-resolving cutaneous lesions
146 that can lead to protective immunity, to disseminated lesions that do not heal spontaneously, to
147 destruction of the mucous membranes of the nose, mouth, and pharynx, to life-threatening
148 visceral disease. The clinical profile depends on a variety of factors, including vector biology, host
149 immunity, and parasite characteristics; the *Leishmania* species initiating the infection being the
150 primary determinant. The two primary clinical forms are cutaneous leishmanisis (CL) and visceral
151 leishmaniasis (VL). The primary *Leishmania* species that cause CL are *Leishmania major*,
152 *Leishmania infantum, Leishmania tropica,* and *Leishmania aethipica* in the Old World and
153 *Leishmania amazonensis, Leishmania braziliensis, Leishmania guyanensis, Le. infantum,*
154 *Leishmania mexicana,* and *Leishmania panamensis* in the New World. VL is primarily caused by
155 *Leishmania donovani* in Asia and Africa and *Le. infantum* in the Middle East, central Asia, South
156 and Central America, and the Mediterranean Basin.

157 There are approximately 35 proven, and an additional 63 suspected, vectors of at least 40
158 different *Leishmania* species to humans [5, 6]. *Phlebotomus* species are the primary *Leishmania*
159 vectors in the Old World and *Lutzomyia* species are responsible for transmitting leishmaniasis
160 throughout the Americas [7]. There is a close ecological association, if not co-evolutionary
161 relationship [8, 9], between *Leishmania* species and their specific vectors such that generally a
162 single sand fly species transmits a single *Leishmania* species under natural conditions. Some
163 sand flies, however, can transmit a range of *Leishmania* species under experimental conditions
164 [10]. This difference has given rise to the concept of "restricted" and "permissive" vectors [11].
165 For example, *Phlebotomus papatasi* is a restrictive vector, transmitting only *Le. major* parasites
166 [12]. *Lutzomyia longipalpis* is considered a permissive vector in laboratory conditions, but only
167 transmits *Le. infantum* naturally [12].

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

177 It is postulated that this close evolutionary relationship between sand fly species and the
178 *Leishmania* species that they transmit may have epidemiological implications for leishmaniasis
179 [17]. For example, there are three primary zymodemes of *Le. major* that have limited geographical
180 distributions such that the prevalent zymodeme in a particular area overlaps with the distribution
181 of one primary population of *Ph. papatasi* [18]. *Ph. papatasi* has a wide geographical distribution,
182 ranging from Morocco to the Indian subcontinent and from southern Europe to central and eastern

183 Africa. Given the wide ecological and geographic distribution of *Ph. papatasi* populations [19],
184 coupled with the low dispersal capacity of these sand flies [12], it is likely that there is limited gene
185 flow between populations and significant genetic structuring between populations. While previous
186 studies demonstrated relatively low genetic differentiation between *Ph. papatasi* populations
187 separated by large geographical distances [9, 20], more recent studies have identified genetic
188 differentiation between geographically separated populations [18, 21-24] and local differentiation
189 [25]. Microsatellite analysis, in particular, revealed two distinct genetic clusters of *Ph. papatasi* (A
190 & B) with further substructure within each population that correlated with geographical origin (A1-5
191 and B1 &2) [18, 23].

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

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

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

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

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

235 chemosensation, detoxification, and circadian biology to provide a basis for studying and
236 understanding sand flies as *Leishmania* vectors. Moreover, as a better understanding of the
237 population structure of geographically separated vector populations is necessary, we also
238 assessed the population structure of *Ph. papatasi* and *Lu. Longipalpis* by collecting and
239 sequencing individual field-collected specimens sampled over a large geographical range in the
240 Middle East and North Africa, and Brazil, respectively. Our results provide significant advances in
241 our understanding of the genetics underlying the population structure and provide a great
242 foundation for future molecular comparative studies of these two medically important vectors.
243

244 **Methods**

245 Fly Maintenance

246 *Phlebotomus papatasi*

247 To avoid confounding effects due to genetic polymorphisms, we used an inbred colony of *Ph.*
248 *papatasi* (Israeli strain) for the genome assembly. This colony was originally established in the
249 1970s and given to Walter Reed Army Institute of Research (WRAIR) in 1983 from the Hebrew
250 University, Jerusalem and transferred to the University of Notre Dame in 2006. At several times
251 since establishment in the laboratory, the colony has fluctuated in population size and has been
252 expanded from a relatively low number of files. Sand flies were reared by the method of Modi
253 and Tesh [40].

254 *Lutzomyia longipalpis*

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

259 Flies were reared under standardized laboratory conditions [40], *i.e.* under controlled temperature
260 (27 ± 2°C), humidity (>80%), and photoperiod (8 hours light/16 hours darkness) [40].

261 Ethics Statement

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

264 Field Collections

265 *Phlebotomus papatasi*

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

275 *Lutzomyia longipalpis*

276 *Lu. longipalpis* were collected in 2014 from six different locations in Brazil (Fig. 1). Samples
277 were collected from three allopatric populations: Jacobina, Bahia State (11º10'S 40º31'W),
278 Lapinha Cave, Minas Gerais State (19º38'S 43º53'W), Marajó Island, Pará State (0°56'S
279 49°38'W), and two sympatric populations (called 1S and 2S) from Sobral, Ceará State (34º41'S
280 40º20'W). For comparison of male copulatory courtship songs, flies were also collected from
281 Olindina (11º 29' S 38° 22' W) and Araci (11º 09' S 39° 01' W), sites near Jacobina. Sand flies
282 were trapped using CDC-style light traps baited with CO_2 between 18:00 and 06:00 and
283 transported to the laboratory. Analyses of male copulatory courtship songs was carried out by as
284 previously described [41]. The recordings were performed by using males and females from

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

287 **Figure 1. *Lutzomyia longipalpis* site locations for copulation songs and pheromone types.**
288 Samples were collected from three allopatric populations: Marajó (Pará State; 0°56'S 49°38'W),
289 Jacobina (Bahia State; 11º 10'S 40º 31'W), and Lapinha Cave (Minas Gerais State; 19º 33'S 43º
290 57'W); and two sympatric populations (1S and 2S) from Sobral (Ceará State; 3º 41'S 40º 21'W).
291 Copulation songs: Burst-type (B) and Pulse-types (P1, P2, and P3). Pheromone types: cembrene
292 (Cemb), 9-methylgermacrene-B (Germ), and himachalene (Hima). For each location, the number
293 of single nucleotide polymorphisms (SNPs) identified in each population with respect to the
294 reference genome (VectorBase, LlonJ1) is indicated. A total of 1,937,819 SNPs were identified
295 among all the populations. Map Source: Google Imagery TerraMetrics.

296 Nucleic Acid Isolation

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

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

306 Genome Sequencing and Assembly

307 *Phlebotomus papatasi*

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

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

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

325 *Lutzomyia longipalpis*

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

336 Individual population sequencing

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

351 RNA-sequencing

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

362 deoxynucleoside triphosphate (dNTP) mix containing deoxyuridine triphosphate was used to
363 introduce strand specificity. For Illumina paired-end library construction, the resultant cDNA was
364 processed through end repair and A-tailing, ligated with Illumina PE adapters, and digested with
365 10 units of uracil–DNA glycosylase (New England Biolabs, Ipswich, MA; cat. no. M0280L).
366 Amplification of the libraries was performed for 13 PCR cycles using the Phusion High-Fidelity
367 PCR Master Mix (New England Biolabs, cat. no. M0531L); 6-bp molecular barcodes were also
368 incorporated during this PCR amplification. These libraries were then purified with Agencourt
369 AMPure XP beads after each enzymatic reaction, and after quantification using the Agilent
370 Bioanalyzer 2100 DNA Chip 7500 (cat. no. 5067-1506), libraries were pooled in equimolar
371 amounts for sequencing. Sequencing was performed on Illumina HiSeq2000s, generating 100-bp
372 paired-end reads. Sequenced reads were deposited in the NCBI SRA, under BioProject
373 accession PRJNA81043 for *Lu longipalpis* and PRJNA20293 of *PH. papatasi*.

374 Annotation

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

380 Orthology Delineation

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

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

402 Maximum likelihood species phylogeny

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

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

418 Synteny Analysis

419 *Macrosynteny*

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

427 *Synteny Blocks*

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

435 *Orientation in genome by prediction of protein function*

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

440 similarity searches were run using the default algorithm parameters. Two genes were considered
441 match if they located in the corresponding pair of synteny blocks, were in the top 5 hits for each
442 other, and 3 criteria were met for the alignment: e-value < 1e-3 and query coverage and identity
443 > 20%. Relative alignment of the blocks were determined based on the corresponding
444 orientations of gene pairs; since some gene pairs had conflicting orientations, block orientation
445 was determined based on that which gave the greatest agreement across the gene pairs.

446 Population Genetics Analysis

447 *SNP Calling*

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

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

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

468 *SNP filtering*

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

478 *Genomics structure*

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

492 Individual ancestry was estimated using Admixture v.1.9 [74]. The analysis was performed for *K*
493 values (ranging from two to seven with 30 iterations per *K*). In order to better understand the
494 different solutions reported by Admixture, post processing of the Admixture results was performed
495 in CLUMPAK v.1.1 [75]. Principal component analysis (PCA) was performed in scikit-allel v.1.1.10
496 [76], following the methods described in [77]. Weir and Cockerham's F_{ST}, Nei's $D_{xy,}$ and Tajima's
497 D were calculated using VCFtools, and using the python script popGenWindows.py
498 (https://github.com/simonhmartin/genomics_general). Single marker FLK test [79] was performed
499 using HapFLK v.1.4 [80].

500 *Phylogenetic analysis*

501 We explored ancestral phylogenetic relationships between individuals by building a neighbor-
502 joining (NJ) tree across the genome using the R packages *adegenet v.2.1.1* [68, 69], *ape v.5.1*
503 [70], *poppr v.2.7.1 [71]*, and *vcfR v.1.7.0 [72]* . For *Ph. papatasi,* we included both *Ph. bergeroti*
504 and *Ph. duboscqi* and used the later to root the trees. For *Lu. longipalpis* phylogenetic analysis
505 we included both *N. intermedia* and *M. migonei* and used *M. migonei* to root the NJ trees. We
506 evaluated node support using 1,000 bootstrap replicates [73].

507 *dN/dS*

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

517

518 **Results and Discussion**

519 **Sequencing and Genome Characteristics**

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

533 <u>Orthology</u>

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

543 of individual gene phylogenies, however, shows great uncertainty with almost equal proportions
544 of phylogenies supporting clustering of sand flies with mosquitoes and with muscoids (S1 Fig.).
545
546 **Figure 2: Molecular species phylogeny and ortholog sharing.** (A) The quantitative maximum
547 likelihood species phylogeny computed from the concatenated superalignment of 1,627
548 orthologous protein-coding genes places the sand flies (Psychodomorpha) as a sister group to
549 the mosquitoes (Culicomorpha) rather than the flies (Muscomorpha), with all branches showing
550 100% bootstrap support. The Culicomorpha are represented by four *Anopheles* mosquito species
551 and *Culex quinquefasciatus* and the Muscomorpha include four *Drosophila* fruit fly species and
552 the tsetse fly, *Glossina morsitans*. Outgroup species represent Lepidoptera (*Bombyx mori*),
553 Coleoptera (*Tribolium castaneum*), Hymenoptera (*Apis mellifera*), and the phylogeny is rooted
554 with the phthirapteran human body louse, *Pe. humanus*. The inset boxplots show that single-copy
555 (1:1) and multi-copy (X:X) ortholog amino acid percent identity is higher between each sand fly
556 (Ll, *Lu. longipalpis*; Pp, *Ph. papatasi*) and *An. gambiae* (Ag) than *D. melanogaster* (Dm). Boxplots
557 show median values with boxes extending to the first and third quartiles of the distributions. (B)
558 The Venn diagram summarizes the numbers of orthologous groups and mean number of genes
559 per species (in parentheses) shared among the two sand flies (L. lon., *Lu. longipalpis*; P. pap.,
560 *Ph. papatasi*) and/or the Culicomorpha and/or the Muscomorpha. Analysis of ortholog sharing
561 shows that the sand flies share more than three times as many orthologous groups exclusively
562 with the Culicomorpha (*Anopheles* and *Culex*) compared to the Muscomorpha (*Drosophila,*
563 *Glossina*) (subsets highlighted with thin and thick dashed lines). Numbers of unique genes are in
564 italics. Colors in panel A and panel B match species and sets of species analyzed.
565 Synteny
566 Comparisons of the relative ordering of ortholog genome positions between *Lu. longipalpis*
567 and *Ph. papatasi* did not indicate evidence of macrosynteny (S2 Fig.). The comparison of ortholog
568 genome positions is complicated by the fragmented nature of the assembled genomes, with *Lu.*

569 *longipalpis*'s 10,331 genes spread across 11,532 scaffolds and the 11,216 genes of *Ph. papatasi*
570 spread across 106,826 scaffolds. Synteny analysis is likely also affected by the high rate of
571 evolutionary divergence among insects [85] and *Ph. papatasi* and *Lu. longipalpis* are separated
572 by at least 100 million years [86]. Previous studies of anopheline species have likewise shown
573 that synteny is conserved over relatively short time spans compared with other organisms [48].
574 Microsynteny, however, was observed between the sand flies and between each sand fly
575 species with *An. gambiae* (S4 Table). Microsynteny between *Lu. longipalpis* and *Ph. papatasi* is
576 similar to what was previously reported for *An. gambiae* and *D. melanogaster* [87]. We identified
577 499 synteny blocks with a maximum of 24 genes, average of 4.2 genes, and a median of 3 genes.
578 By comparison, we found 1,037 synteny blocks with a maximum of 10 genes, an average of 2.5
579 genes, and a median of 2 genes between *An. gambiae* and *D. melanogaster.* Previous studies
580 reported 948 synteny blocks with a maximum of 39 genes between *An. gambiae* and *D.*
581 *melanogaster* [87].
582 Transposable Elements
583 Transposable elements (TEs) are ubiquitous repetitive sequences present in eukaryotic
584 genomes that can be an important factor affecting genome sizes and are thought to be one of the
585 driving forces of evolution [45]. Some insect genomes have less than 3% of TEs, while others
586 contain as much as 50% or more of TEs, associated with large genomic size differences [88]. Our
587 analysis indicated that the *Ph. papatasi* genome is composed of 5.65% of TE derived sequences
588 while the *Lu. longipalpis* genome contains only 0.57%, corresponding to the genome size
589 difference between two sand fly species. This difference in TE-derived sequence could be due to
590 the result of divergent evolutionary dynamics of some TE families or superfamilies, affecting either
591 their distribution (presence or absence of specific superfamilies) or their abundance (copy number
592 per superfamily) in the genome. Higher abundance of TE derived sequences, presence of full-
593 length TEs and the genome size expansion in the *Ph. papatasi* genome also could be due to
594 recently active TEs. Alternatively, genomic differences in TE content might be the result of intrinsic

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

598 The comparison of the TE content in both sand fly genomes indicated an expansion of all the
599 TE classes and orders in the *Ph. papatasi* genome. This multiplication was more pronounced in
600 elements belonging to the class II, or "cut-and-paste" TEs, and especially in non-autonomous
601 miniature inverted-repeat transposable elements (MITEs), representing up to 29-fold differences
602 between the two genomes. Expansion of MITEs suggests the recent activity of class II TEs in the
603 *Ph. papatasi* genome. On the other hand, class I elements, or "copy-paste" elements, including
604 the Long Terminal Repeat (LTRs) and non-LTRs, which traditionally are accountable for the
605 genome expansion, showed more subtle changes between the two sand fly genomes,
606 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%

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%**

607

608 Immunity Genes

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

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

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

629 Blood Feeding Genes

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

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

643 Circadian Rhythm Genes

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

658 Chemosensory Receptors

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

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

678 G-Protein Coupled Receptors

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

684 and 92 GPCRs from *Ph. papatasi* and *Lu. longipalpis,* respectively, were compared with other
685 insects with well characterized GPCRs such as *D. melanogaster, An. gambiae, Ae. aegypti and*
686 *Pe. humanus* (S23 Table). Class A (rhodopsin-like) is the most numerous class with ~50 genes
687 in each sand fly, and includes the opsins that are thought to function in visual processes and
688 circadian rhythm. Both sand flies have one opsin gene for each functional group, the long-
689 wavelength, short-wavelength, ultraviolet, rh7-like, and pteropsin. Classes B (secretin-like), C
690 (metabotropic glutamate-like) and D (atypical GPCRs) have fewer members, with ~20, ~10 and
691 ~10 in each sand fly, respectively. Sand flies include GPCR genes absent from *D. melanogaster*
692 (ocular albinism) and absent in *Ae. aegypti* and *An. gambiae* (parathyroid hormone receptor);
693 both genes from class B.

694 Cytochrome P450 Monoxygenase Genes

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

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

708 Other Groups

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

715 **Population Structure**

716 Genetic structure across the range of *Ph. papatasi*

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

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

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

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

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

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

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

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

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

754 Consistent with the phylogeny based on the chemoreceptor repertoire [112], the full genome
755 phylogeny clustered the populations into two clades based on song and pheromone type, where
756 Marajó and Sobral 2S (Burst, Sobralene) cluster together and Lapinha and Sobral 1S (Pulse,
757 Germ) cluster together (Fig. 3A). An analysis of the male copulatory courtship songs of males
758 collected from Lapinha and Sobral are in agreement with those previously recorded [41]. In these
759 three resampled populations, we observed the sub-type P2 in Lapinha, the sub-type P3 in Sobral
760 1S, and the burst-type in Sobral 2S.

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

772 Interestingly, the phylogenetic analysis separated the Jacobina population into two groups.
773 As expected, because flies from Jacobina are known to express a C_{16} H_{32} pheromone and pulse-
774 like copulatory songs, some individuals clustered with Sobral 1S and Lapinha. Unexpectedly,
775 however, six individuals clustered with the diterpenoid-like pheromone and burst song expressing
776 individuals, suggesting that there is more than one population living in sympatry at the Jacobina
777 site. Male copulatory songs were not recorded for the Jacobina samples. However, sand flies
778 collected from two localities near to Jacobina, Araci and Olindina (Bahia state) (S14A Fig.) exhibit
779 different copulatory song patterns, suggesting the possible existence of two groups in Jacobina,
780 as observed in molecular data. Males from Araci exhibit the P1 copulation song pattern (S14B
781 Fig.), composed of train of similar pulses as previously described in males collected in Jacobina
782 [41]. Males from the nearby locality, Olindina, produced burst-type songs (S14B Fig.) with similar
783 pattern as Sobral 2S males [41]. The mean values of all song parameters observed from these
784 flies (S35 Table) are similar as previously reported [118].

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

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

789 The PCA based on the whole-genome clustered the individuals into six groups as well (Fig.
790 3B). Contrary to the phylogenetic analysis, however, the six Jacobina individuals were closely
791 clustered, but separate from the Marajó and Sobral 2S populations, which were indistinguishable.
792 PC1 explained 5.3% of the variation and separated the individuals collected from Jacobina into
793 two populations. Consistent with the NJ tree, the Sobral 1S population also exhibited some
794 population structure, the two clusters distinguishable through PC1 and PC2. PC2 accounted for
795 4.6% of the total variation and distinguished Lapinha from the other populations. The sympatric
796 Sobral 1S and 2S populations separate by both PC1 and PC2. Interestingly, while consistent with
797 Hickner *et al.* 2020, the whole-genome PCA allowed higher discriminating power among clusters
798 than the PCA based on the chemoreceptor repertoire which only identified 3-4 discrete clusters
799 [112].

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

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

813 values indicate that genome wide differentiation is greater among population comparisons of
814 different pheromone and song types (Lapinha- Marajó, 0.214; Lapinha-Sobral 2S, 0.211; Sobral
815 1S-Sobral 2S, 0.116 compared to Lapinha – Sobral 1S, 0.154; Marajó-Sobral 2S, 0.114) and
816 allopatric populations (Lapinha- Marajó, 0.214; Lapinha-Sobral 2S, 0.211; Lapinha – Sobral 1S,
817 0.154 compared to Sobral 1S-Sobral 2S, 0.116) (S15 Fig.).

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

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

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

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

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

853 **Figure 5. Measures of divergence in 1 kb non-overlapping genomic windows between the**
854 **different populations of *Lutzomyia longipalpis*.** (A) Box plots of D_{xy} and F_{ST} values in the top
855 2.5% quantile (outlier) of each population comparison. (B) Box plots of D_{xy} and F_{ST} values for
856 windows having both high D_{xy} and high F_{ST} (differentiation islands). (C) Box plots of Tajimas' D
857 values for the differentiation islands. (D) Box plots of FLK values for sites within differentiation
858 islands. Box plots show the medians (lines) and interquartile ranges (boxes); the whiskers extend
859 out from the box plots to 1.5 times the interquartile range, and values outside this limit are
860 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**

# 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

874 homolog, LLOJ009447 a rRNA adenine N(6)-methyltransferase, and LLOJ009732 a Lipase
875 maturation factor. No enrichment of gene ontology terms was identified using the *An. gambiae*
876 orthologs.
877

878 **Conclusions**

879 Our study provides the genome assembly and annotation of two divergent sand fly species that
880 will facilitate molecular and comparative studies of these medically important vectors. These
881 results provide a foundation for annotating and analyzing future chromosome length assemblies
882 generated from single sand flies. Global comparisons between sand fly vectors will greatly inform
883 the evolutionary relationships among these species and lead to advances in our understanding
884 of genes involved in important phenomena such as vectorial capacity, host-specificity, blood-
885 feeding, insecticide resistance, and immune system modulation.
886

887 **Acknowledgments**

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

897
898

899 **References**

900 1. Desjeux P. Leishmaniasis: current situation and new perspectives. Comparative
901 immunology, microbiology and infectious diseases. 2004;27(5):305-18. doi:
902 10.1016/j.cimid.2004.03.004. PubMed PMID: 15225981.
903 2. Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide
904 and global estimates of its incidence. PloS one. 2012;7(5):e35671. doi:
905 10.1371/journal.pone.0035671. PubMed PMID: 22693548; PubMed Central PMCID:
906 PMC3365071.
907 3. Jones CM, Welburn SC. Leishmaniasis Beyond East Africa. Front Vet Sci.
908 2021;8:618766. Epub 20210226. doi: 10.3389/fvets.2021.618766. PubMed PMID: 33732738;
909 PubMed Central PMCID: PMCPMC7959848.
910 4. Organization WH. WHO report on global surveillance of epidemic-prone infectious
911 diseases. 2000.
912 5. Ramalho-Ortigao M, Saraiva EM, Traub-Cseko YM. Sand fly- interactions: long
913 relationships are not necessarily easy. The open parasitology journal. 2010;4:195-204. doi:
914 10.2174/1874421401004010195. PubMed PMID: 24159365; PubMed Central PMCID:
915 PMC3805382.
916 6. Maroli M, Feliciangeli MD, Bichaud L, Charrel RN, Gradoni L. Phlebotomine sandflies
917 and the spreading of leishmaniases and other diseases of public health concern. Med Vet
918 Entomol. 2013;27(2):123-47. Epub 20120827. doi: 10.1111/j.1365-2915.2012.01034.x. PubMed
919 PMID: 22924419.
920 7. Sacks D, Kamhawi S. Molecular aspects of parasite-vector and vector-host interactions
921 in leishmaniasis. Annu Rev Microbiol. 2001;55:453-83. PubMed PMID: 11544364.
922 8. Esseghir S, Ready PD. Speciation of *Phlebotomus* sandflies of the subgenus
923 *Larroussius* coincided with the late Miocene-Pliocene aridification of the Mediterranean
924 subregion. Biological Journal of the Linnean Society. 2000;70:189-219.

925 9. Esseghir S, Ready PD, Killick-Kendrick R, Ben-Ismail R. Mitochondrial haplotypes and
926 phylogeography of Phlebotomus vectors of Leishmania major. Insect Mol Biol. 1997;6(3):211-
927 25. PubMed PMID: 9272439.
928 10. Lainson R, Ward RD, Shaw JJ. Experimental transmission of Leishmania chagasi,
929 causative agent of neotropical visceral leishmaniasis, by the sandfly Lutzomyia longipalpis.
930 Nature. 1977;266(5603):628-30. PubMed PMID: 859627.
931 11. Myskova J, Svobodova M, Beverley SM, Volf P. A lipophosphoglycan-independent
932 development of Leishmania in permissive sand flies. Microbes Infect. 2007;9(3):317-24. doi:
933 10.1016/j.micinf.2006.12.010. PubMed PMID: 17307009; PubMed Central PMCID:
934 PMC2839925.
935 12. Killick-Kendrick R. The biology and control of phlebotomine sand flies. Clin Dermatol.
936 1999;17(3):279-89. PubMed PMID: 10384867.
937 13. Yeates DK, Wiegmann BM. Congruence and controversy: toward a higher-level
938 phylogeny of Diptera. Annu Rev Entomol. 1999;44:397-428. PubMed PMID: 15012378.
939 14. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, et al. Episodic
940 radiations in the fly tree of life. Proc Natl Acad Sci U S A. 2011;108(14):5690-5. doi:
941 10.1073/pnas.1012675108. PubMed PMID: 21402926; PubMed Central PMCID: PMC3078341.
942 15. Simon S, Narechania A, Desalle R, Hadrys H. Insect phylogenomics: exploring the
943 source of incongruence using new transcriptomic data. Genome biology and evolution.
944 2012;4(12):1295-309. doi: 10.1093/gbe/evs104. PubMed PMID: 23175716; PubMed Central
945 PMCID: PMC3542558.
946 16. Curler GR, Moulton JK. Phylogeny of psychodid subfamilies (Diptera: Psychodidae)
947 inferred from nuclear DNA sequences with a review of morphological evidence for relationships.
948 Syst Entomol. 2012;37(3):603-16. doi: 10.1111/j.1365-3113.2012.00634.x. PubMed PMID:
949 WOS:000305616300009.

950 17. Akhoundi M, Kuhls K, Cannet A, Votypka J, Marty P, Delaunay P, et al. A Historical
951 Overview of the Classification, Evolution, and Dispersion of Leishmania Parasites and
952 Sandflies. PLoS Negl Trop Dis. 2016;10(3):e0004349. Epub 2016/03/05. doi:
953 10.1371/journal.pntd.0004349. PubMed PMID: 26937644; PubMed Central PMCID:
954 PMCPMC4777430.
955 18. Hamarsheh O. Distribution of Leishmania major zymodemes in relation to populations of
956 Phlebotomus papatasi sand flies. Parasites & vectors. 2011;4:9. doi: 10.1186/1756-3305-4-9.
957 PubMed PMID: 21266079; PubMed Central PMCID: PMC3035596.
958 19. Colacicco-Mayhugh MG, Masuoka PM, Grieco JP. Ecological niche model of
959 Phlebotomus alexandri and P. papatasi (Diptera: Psychodidae) in the Middle East. Int J Health
960 Geogr. 2010;9. doi: Artn 2

961 10.1186/1476-072x-9-2. PubMed PMID: WOS:000274782300001.
962 20. Depaquit J, Lienard E, Verzeaux-Griffon A, Ferte H, Bounamous A, Gantier JC, et al.
963 Molecular homogeneity in diverse geographical populations of Phlebotomus papatasi (Diptera,
964 Psychodidae) inferred from ND4 mtDNA and ITS2 rDNA Epidemiological consequences. Infect
965 Genet Evol. 2008;8(2):159-70. PubMed PMID: 18243814.
966 21. Hamarsheh O, Presber W, Abdeen Z, Sawalha S, Al-Lahem A, Schonian G. Genetic
967 structure of Mediterranean populations of the sandfly Phlebotomus papatasi by mitochondrial
968 cytochrome b haplotype analysis. Med Vet Entomol. 2007;21(3):270-7. PubMed PMID:
969 17897368.
970 22. Hamarsheh O, Presber W, Al-Jawabreh A, Abdeen Z, Amro A, Schonian G. Molecular
971 markers for Phlebotomus papatasi (Diptera: Psychodidae) and their usefulness for population
972 genetic analysis. Trans R Soc Trop Med Hyg. 2009;103(11):1085-6. doi:
973 10.1016/j.trstmh.2009.02.011. PubMed PMID: 19303124.

974 23. Hamarsheh O, Presber W, Yaghoobi-Ershadi MR, Amro A, Al-Jawabreh A, Sawalha S,
975 et al. Population structure and geographical subdivision of the Leishmania major vector
976 Phlebotomus papatasi as revealed by microsatellite variation. Med Vet Entomol. 2009;23(1):69-
977 77. doi: 10.1111/j.1365-2915.2008.00784.x. PubMed PMID: 19239616.
978 24. Flanley CM, Ramalho-Ortigao M, Coutinho-Abreu IV, Mukbel R, Hanafi HA, El-Hossary
979 SS, et al. Population genetics analysis of Phlebotomus papatasi sand flies from Egypt and
980 Jordan based on mitochondrial cytochrome b haplotypes. Parasites & vectors. 2018;11(1):214.
981 Epub 2018/03/29. doi: 10.1186/s13071-018-2785-9. PubMed PMID: 29587873; PubMed Central
982 PMCID: PMCPMC5872541.
983 25. Khalid NM, Aboud MA, Alrabba FM, Elnaiem DE, Tripet F. Evidence for genetic
984 differentiation at the microgeographic scale in Phlebotomus papatasi populations from Sudan.
985 Parasites & vectors. 2012;5:249. doi: 10.1186/1756-3305-5-249. PubMed PMID: 23146340;
986 PubMed Central PMCID: PMC3503571.
987 26. Araki AS, Ferreira GE, Mazzoni CJ, Souza NA, Machado RC, Bruno RV, et al. Multilocus
988 analysis of divergence and introgression in sympatric and allopatric sibling species of the
989 Lutzomyia longipalpis complex in Brazil. PLoS neglected tropical diseases. 2013;7(10):e2495.
990 Epub 2013/10/23. doi: 10.1371/journal.pntd.0002495. PubMed PMID: 24147172; PubMed
991 Central PMCID: PMCPMC3798421.
992 27. Palframan MJ, Bandi KK, Hamilton JGC, Pattenden G. Sobralene, a new sex-
993 aggregation pheromone and likely shunt metabolite of the taxadiene synthase cascade,
994 produced by a member of the sand fly Lutzomyia longipalpis species complex. Tetrahedron Lett.
995 2018;59(20):1921-3. Epub 2018/05/22. doi: 10.1016/j.tetlet.2018.03.088. PubMed PMID:
996 29780183; PubMed Central PMCID: PMCPMC5937913.
997 28. Araki AS, Vigoder FM, Bauzer LG, Ferreira GE, Souza NA, Araujo IB, et al. Molecular
998 and behavioral differentiation among Brazilian populations of Lutzomyia longipalpis (Diptera:
999 Psychodidae: Phlebotominae). PLoS neglected tropical diseases. 2009;3(1):e365. Epub

1000 2009/01/28. doi: 10.1371/journal.pntd.0000365. PubMed PMID: 19172187; PubMed Central
1001 PMCID: PMCPMC2628317.
1002 29. Souza NA, Andrade-Coelho CA, Vigoder FM, Ward RD, Peixoto AA. Reproductive
1003 isolation between sympatric and allopatric Brazilian populations of Lutzomyia longipalpis s.l.
1004 (Diptera: Psychodidae). Mem Inst Oswaldo Cruz. 2008;103(2):216-9. Epub 2008/04/22.
1005 PubMed PMID: 18425278.
1006 30. Boulanger N, Lowenberger C, Volf P, Ursic R, Sigutova L, Sabatier L, et al.
1007 Characterization of a defensin from the sand fly Phlebotomus duboscqi induced by challenge
1008 with bacteria or the protozoan parasite Leishmania major. Infect Immun. 2004;72(12):7140-6.
1009 PubMed PMID: 15557638.
1010 31. Telleria EL, Sant'Anna MR, Alkurbi MO, Pitaluga AN, Dillon RJ, Traub-Cseko YM.
1011 Bacterial feeding, Leishmania infection and distinct infection routes induce differential defensin
1012 expression in Lutzomyia longipalpis. Parasites & vectors. 2013;6:12. doi: 10.1186/1756-3305-6-
1013 12. PubMed PMID: 23311993; PubMed Central PMCID: PMC3573903.
1014 32. Telleria EL, Tinoco-Nunes B, Leštinová T, de Avellar LM, Tempone AJ, Pitaluga AN, et
1015 al. Antimicrobial Peptides: Differential Expression during Development and Potential
1016 Involvement in Vector Interaction with Microbiota and. Microorganisms. 2021;9(6). Epub
1017 20210611. doi: 10.3390/microorganisms9061271. PubMed PMID: 34207941; PubMed Central
1018 PMCID: PMCPMC8230673.
1019 33. Kykalová B, Tichá L, Volf P, Loza Telleria E. Antimicrobial Peptides in Larvae and
1020 Females and a Gut-Specific Defensin Upregulated by. Microorganisms. 2021;9(11). Epub
1021 20211106. doi: 10.3390/microorganisms9112307. PubMed PMID: 34835433; PubMed Central
1022 PMCID: PMCPMC8625375.
1023 34. Telleria EL, Sant'Anna MR, Ortigao-Farias JR, Pitaluga AN, Dillon VM, Bates PA, et al.
1024 Caspar-like gene depletion reduces Leishmania infection in sand fly host Lutzomyia longipalpis.

1025 J Biol Chem. 2012;287(16):12985-93. doi: 10.1074/jbc.M111.331561. PubMed PMID:
1026 22375009; PubMed Central PMCID: PMC3339954.
1027 35. Louradour I, Ghosh K, Inbar E, Sacks DL. CRISPR/Cas9 Mutagenesis in Phlebotomus
1028 papatasi: the Immune Deficiency Pathway Impacts Vector Competence for Leishmania major.
1029 mBio. 2019;10(4). Epub 20190827. doi: 10.1128/mBio.01941-19. PubMed PMID: 31455654;
1030 PubMed Central PMCID: PMCPMC6712399.
1031 36. Abdeladhim M, Kamhawi S, Valenzuela JG. What's behind a sand fly bite? The profound
1032 effect of sand fly saliva on host hemostasis, inflammation and immunity. Infect Genet Evol.
1033 2014;28:691-703. doi: 10.1016/j.meegid.2014.07.028. PubMed PMID: 25117872.
1034 37. Yuan Q, Metterville D, Briscoe AD, Reppert SM. Insect cryptochromes: gene duplication
1035 and loss define diverse ways to construct insect circadian clocks. Mol Biol Evol. 2007;24(4):948-
1036 55. doi: 10.1093/molbev/msm011. PubMed PMID: 17244599.
1037 38. Ceriani MF, Darlington TK, Staknis D, Mas P, Petti AA, Weitz CJ, et al. Light-dependent
1038 sequestration of TIMELESS by CRYPTOCHROME. Science. 1999;285(5427):553-6. PubMed
1039 PMID: 10417378.
1040 39. Thomas MB. Biological control of human disease vectors: a perspective on challenges
1041 and opportunities. Biocontrol (Dordr). 2018;63(1):61-9. Epub 2018/02/03. doi: 10.1007/s10526-
1042 017-9815-y. PubMed PMID: 29391855; PubMed Central PMCID: PMCPMC5769823.
1043 40. Modi GB, Tesh RB. A simple technique for mass rearing Lutzomyia longipalpis and
1044 Phlebotomus papatasi (Diptera: Psychodidae) in the laboratory. J Med Entomol.
1045 1983;20(5):568-9. PubMed PMID: 6644754.
1046 41. Souza NA, Vigoder FM, Araki AS, Ward RD, Kyriacou CP, Peixoto AA. Analysis of the
1047 copulatory courtship songs of Lutzomyia longipalpis in six populations from Brazil. J Med
1048 Entomol. 2004;41(5):906-13. Epub 2004/11/13. doi: 10.1603/0022-2585-41.5.906. PubMed
1049 PMID: 15535620.

1050 42. Deng J, Worley KC. Atlas-Link 2010. Available from:
1051 https://www.hgsc.bcm.edu/software/atlas-link.
1052 43. Song X, Liu Y, Qu J, Gibbs RA, Worley KC. ATLAS gapfill 2.2. 2012.
1053 44. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management
1054 tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491. Epub
1055 20111222. doi: 10.1186/1471-2105-12-491. PubMed PMID: 22192575; PubMed Central
1056 PMCID: PMCPMC3280279.
1057 45. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of
1058 Aedes aegypti, a major arbovirus vector. Science. 2007;316(5832):1718-23. PubMed PMID:
1059 17510324.
1060 46. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: a
1061 hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res.
1062 2013;41(Database issue):D358-65. doi: 10.1093/nar/gks1116. PubMed PMID: 23180791;
1063 PubMed Central PMCID: PMC3531149.
1064 47. Neafsey DE, Christophides GK, Collins FH, Emrich SJ, Fontaine MC, Gelbart W, et al.
1065 The evolution of the Anopheles 16 genomes project. G3. 2013;3(7):1191-4. doi:
1066 10.1534/g3.113.006247. PubMed PMID: 23708298; PubMed Central PMCID: PMC3704246.
1067 48. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, et al.
1068 Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles
1069 mosquitoes. Science. 2015;347(6217):1258522. doi: 10.1126/science.1258522. PubMed PMID:
1070 25554792.
1071 49. Rognes T. Faster Smith-Waterman database searches with inter-sequence SIMD
1072 parallelisation. BMC bioinformatics. 2011;12:221. doi: 10.1186/1471-2105-12-221. PubMed
1073 PMID: 21631914; PubMed Central PMCID: PMC3120707.

1074 50. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and
1075 space complexity. BMC bioinformatics. 2004;5:113. doi: 10.1186/1471-2105-5-113. PubMed
1076 PMID: 15318951; PubMed Central PMCID: PMC517706.
1077 51. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated
1078 alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3.
1079 doi: 10.1093/bioinformatics/btp348. PubMed PMID: 19505945; PubMed Central PMCID:
1080 PMC2712344.
1081 52. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong
1082 phylogenetic signals. Nature. 2013;497(7449):327-31. doi: 10.1038/nature12130. PubMed
1083 PMID: 23657258.
1084 53. Salichos L, Stamatakis A, Rokas A. Novel information theory-based measures for
1085 quantifying incongruence among phylogenetic trees. Mol Biol Evol. 2014;31(5):1261-71. doi:
1086 10.1093/molbev/msu061. PubMed PMID: 24509691.
1087 54. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1088 large phylogenies. Bioinformatics. 2014;30(9):1312-3. doi: 10.1093/bioinformatics/btu033.
1089 PubMed PMID: 24451623; PubMed Central PMCID: PMC3998144.
1090 55. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree
1091 processing in the UNIX shell. Bioinformatics. 2010;26(13):1669-70. doi:
1092 10.1093/bioinformatics/btq243. PubMed PMID: 20472542; PubMed Central PMCID:
1093 PMC2887050.
1094 56. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao FA, Pozdnyakov IA, et
1095 al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software.
1096 Nucleic Acids Res. 2015;43(Database issue):D250-6. doi: 10.1093/nar/gku1220. PubMed
1097 PMID: 25428351; PubMed Central PMCID: PMC4383991.
1098 57. Drillon G, Carbone A, Fischer G. SynChro: a fast and easy tool to reconstruct and
1099 visualize synteny blocks along eukaryotic chromosomes. PloS one. 2014;9(3):e92621. doi:

1100 10.1371/journal.pone.0092621. PubMed PMID: 24651407; PubMed Central PMCID:
1101 PMC3961402.
1102 58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool.
1103 J Mol Biol. 1990;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PubMed PMID:
1104 2231712.
1105 59. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-
1106 redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res.
1107 2007;35(Database issue):D61-5. doi: 10.1093/nar/gkl842. PubMed PMID: 17130148; PubMed
1108 Central PMCID: PMC1716718.
1109 60. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, et al.
1110 VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Res.
1111 2007;35(Database issue):D503-5. PubMed PMID: 17145709.
1112 61. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1113 arXiv:1303.3997v1 [q-bio.GN]; 2013.
1114 62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
1115 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi:
1116 papers2://publication/doi/10.1093/bioinformatics/btp352.
1117 63. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2:
1118 Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genes
1119 Dev. 2012;22(3):568-76. doi: papers2://publication/doi/10.1101/gr.129684.111.
1120 64. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation
1121 PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. Epub
1122 2015/02/28. doi: 10.1186/s13742-015-0047-8. PubMed PMID: 25722852; PubMed Central
1123 PMCID: PMCPMC4342193.
1124 65. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a
1125 tool set for whole-genome association and population-based linkage analyses. Am J Hum

1126 Genet. 2007;81(3):559-75. Epub 2007/08/19. doi: 10.1086/519795. PubMed PMID: 17701901;
1127 PubMed Central PMCID: PMCPMC1950838.
1128 66. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance
1129 computing toolset for relatedness and principal component analysis of SNP data.
1130 Bioinformatics. 2012;28(24):3326-8. doi:
1131 papers3://publication/doi/10.1093/bioinformatics/bts606.
1132 67. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
1133 call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. doi:
1134 10.1093/bioinformatics/btr330.
1135 68. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers.
1136 Bioinformatics (Oxford, England). 2008;24(11):1403-5. Epub 2008/04/10. doi:
1137 10.1093/bioinformatics/btn129. PubMed PMID: 18397895.
1138 69. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP
1139 data. Bioinformatics (Oxford, England). 2011;27(21):3070-1. Epub 2011/09/20. doi:
1140 10.1093/bioinformatics/btr521. PubMed PMID: 21926124; PubMed Central PMCID:
1141 PMCPMC3198581.
1142 70. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R
1143 language. Bioinformatics (Oxford, England). 2004;20(2):289-90. Epub 2004/01/22. PubMed
1144 PMID: 14734327.
1145 71. Kamvar ZN, Tabima JF, Grunwald NJ. Poppr: an R package for genetic analysis of
1146 populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2:e281. Epub
1147 2014/04/02. doi: 10.7717/peerj.281. PubMed PMID: 24688859; PubMed Central PMCID:
1148 PMCPMC3961149.
1149 72. Knaus BJ, Grunwald NJ. vcfr: a package to manipulate and visualize variant call format
1150 data in R. Mol Ecol Resour. 2017;17(1):44-53. Epub 2016/07/13. doi: 10.1111/1755-
1151 0998.12549. PubMed PMID: 27401132.

1152 73. Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proc
1153 Natl Acad Sci U S A. 1996;93(23):13429-34. Epub 1996/11/12. PubMed PMID: 8917608;
1154 PubMed Central PMCID: PMCPMC24110.
1155 74. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
1156 unrelated individuals. Genome research. 2009;19(9):1655-64. Epub 2009/08/04. doi:
1157 10.1101/gr.094052.109. PubMed PMID: 19648217; PubMed Central PMCID:
1158 PMCPMC2752134.
1159 75. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program
1160 for identifying clustering modes and packaging population structure inferences across K. Mol
1161 Ecol Resour. 2015;15(5):1179-91. Epub 2015/02/17. doi: 10.1111/1755-0998.12387. PubMed
1162 PMID: 25684545; PubMed Central PMCID: PMCPMC4534335.
1163 76. Suite 2011: LigPrep, version 2.5, Schrödinger, LLC, New York, NY, 2011.
1164 77. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet.
1165 2006;2(12):e190. Epub 2006/12/30. doi: 10.1371/journal.pgen.0020190. PubMed PMID:
1166 17194218; PubMed Central PMCID: PMCPMC1713260.
1167 78. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
1168 call format and VCFtools. Bioinformatics (Oxford, England). 2011;27(15):2156-8. Epub
1169 2011/06/10. doi: 10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central
1170 PMCID: PMCPMC3137218.
1171 79. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting
1172 selection in population trees: the Lewontin and Krakauer test extended. Genetics.
1173 2010;186(1):241-62. Epub 2010/09/22. doi: 10.1534/genetics.104.117275. PubMed PMID:
1174 20855576; PubMed Central PMCID: PMCPMC2940290.
1175 80. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of
1176 selection through haplotype differentiation among hierarchically structured populations.

1177 Genetics. 2013;193(3):929-41. Epub 2013/01/12. doi: 10.1534/genetics.112.147231. PubMed
1178 PMID: 23307896; PubMed Central PMCID: PMCPMC3584007.
1179 81. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable
1180 generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular
1181 systems biology. 2011;7:539. doi: 10.1038/msb.2011.75. PubMed PMID: 21988835; PubMed
1182 Central PMCID: PMC3261699.
1183 82. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence
1184 alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(Web Server
1185 issue):W609-12. doi: 10.1093/nar/gkl315. PubMed PMID: 16845082; PubMed Central PMCID:
1186 PMC1538804.
1187 83. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol.
1188 2007;24(8):1586-91. doi: 10.1093/molbev/msm088. PubMed PMID: 17483113.
1189 84. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
1190 assessing genome assembly and annotation completeness with single-copy orthologs.
1191 Bioinformatics (Oxford, England). 2015;31(19):3210-2. Epub 2015/06/11. doi:
1192 10.1093/bioinformatics/btv351. PubMed PMID: 26059717.
1193 85. Zdobnov EM, Bork P. Quantification of insect genome divergence. Trends Genet.
1194 2007;23(1):16-20. Epub 2006/11/14. doi: 10.1016/j.tig.2006.10.004. PubMed PMID: 17097187.
1195 86. Filho JD, Brazil RP. Relationships of new world phlebotomine sand flies (Diptera:
1196 Psychodidae) based on fossil evidence. Mem Inst Oswaldo Cruz. 2003;98 Suppl 1:145-9.
1197 PubMed PMID: 12687775.
1198 87. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, et al.
1199 Comparative genome and proteome analysis of Anopheles gambiae and Drosophila
1200 melanogaster. Science. 2002;298(5591):149-59. PubMed PMID: 12364792.

1201 88. Gilbert C, Peccoud J, Cordaux R. Transposable Elements and the Evolution of Insects.
1202 Annu Rev Entomol. 2021;66:355-72. Epub 20200915. doi: 10.1146/annurev-ento-070720-
1203 074650. PubMed PMID: 32931312.
1204 89. Petrov DA, Hartl DL. High rate of DNA loss in the Drosophila melanogaster and
1205 Drosophila virilis species groups. Mol Biol Evol. 1998;15(3):293-302. Epub 1998/03/21. doi:
1206 10.1093/oxfordjournals.molbev.a025926. PubMed PMID: 9501496.
1207 90. Vasta GR. Roles of galectins in infection. Nature reviews Microbiology. 2009;7(6):424-
1208 38. doi: 10.1038/nrmicro2146. PubMed PMID: 19444247; PubMed Central PMCID:
1209 PMC3759161.
1210 91. Kamhawi S, Ramalho-Ortigao M, Pham VM, Kumar S, Lawyer PG, Turco SJ, et al. A
1211 role for insect galectins in parasite survival. Cell. 2004;119(3):329-41. PubMed PMID:
1212 15543683.
1213 92. Grimaldi D, Engel MS. Evolution of the Insects. New York, NY: Cambridge University
1214 Press; 2005.
1215 93. Abrudan J, Ramalho-Ortigao M, O'Neil S, Stayback G, Wadsworth M, Bernard M, et al.
1216 The characterization of the Phlebotomus papatasi transcriptome. Insect Mol Biol.
1217 2013;22(2):211-32. doi: 10.1111/imb.12015. PubMed PMID: 23398403; PubMed Central
1218 PMCID: PMC3594503.
1219 94. Coutinho-Abreu IV, Sharma NK, Robles-Murguia M, Ramalho-Ortigao M. Targeting the
1220 midgut secreted PpChit1 reduces Leishmania major development in its natural vector, the sand
1221 fly Phlebotomus papatasi. PLoS neglected tropical diseases. 2010;4(11):e901. Epub
1222 2010/12/15. doi: 10.1371/journal.pntd.0000901. PubMed PMID: 21152058; PubMed Central
1223 PMCID: PMCPMC2994919.
1224 95. Coutinho-Abreu IV, Sharma NK, Robles-Murguia M, Ramalho-Ortigao M.
1225 Characterization of Phlebotomus papatasi peritrophins, and the role of PpPer1 in Leishmania
1226 major survival in its natural vector. PLoS Negl Trop Dis. 2013;7(3):e2132. doi:

1227 10.1371/journal.pntd.0002132. PubMed PMID: 23516661; PubMed Central PMCID:
1228 PMC3597473.
1229 96. Ortigao-Farias JR, Di-Blasi T, Telleria EL, Andorinho AC, Lemos-Silva T, Ramalho-
1230 Ortigao M, et al. Alternative splicing originates different domain structure organization of
1231 Lutzomyia longipalpis chitinases. Mem Inst Oswaldo Cruz. 2018;113(2):96-101. Epub
1232 2017/12/14. doi: 10.1590/0074-02760170179. PubMed PMID: 29236932; PubMed Central
1233 PMCID: PMCPMC5722264.
1234 97. Pitaluga AN, Beteille V, Lobo AR, Ortigao-Farias JR, Davila AM, Souza AA, et al. EST
1235 sequencing of blood-fed and Leishmania-infected midgut of Lutzomyia longipalpis, the principal
1236 visceral leishmaniasis vector in the Americas. Mol Genet Genomics. 2009;282(3):307-17. Epub
1237 2009/07/01. doi: 10.1007/s00438-009-0466-2. PubMed PMID: 19565270.
1238 98. Pruzinova K, Sadlova J, Seblova V, Homola M, Votypka J, Volf P. Comparison of
1239 Bloodmeal Digestion and the Peritrophic Matrix in Four Sand Fly Species Differing in
1240 Susceptibility to Leishmania donovani. PloS one. 2015;10(6):e0128203. Epub 2015/06/02. doi:
1241 10.1371/journal.pone.0128203. PubMed PMID: 26030610; PubMed Central PMCID:
1242 PMCPMC4452187.
1243 99. Ramalho-Ortigao JM, Kamhawi S, Rowton ED, Ribeiro JM, Valenzuela JG. Cloning and
1244 characterization of trypsin- and chymotrypsin-like proteases from the midgut of the sand fly
1245 vector Phlebotomus papatasi. Insect Biochem Mol Biol. 2003;33(2):163-71. PubMed PMID:
1246 12535675.
1247 100. Ramalho-Ortigao JM, Temporal P, de Oliveira SM, Barbosa AF, Vilela ML, Rangel EF, et
1248 al. Characterization of constitutive and putative differentially expressed mRNAs by means of
1249 expressed sequence tags, differential display reverse transcriptase-PCR and randomly
1250 amplified polymorphic DNA-PCR from the sand fly vector Lutzomyia longipalpis. Mem Inst
1251 Oswaldo Cruz. 2001;96(1):105-11. PubMed PMID: 11285481.

1252 101. Ramalho-Ortigao JM, Traub-Cseko YM. Molecular characterization of Llchit1, a midgut
1253 chitinase cDNA from the leishmaniasis vector Lutzomyia longipalpis. Insect Biochem Mol Biol.
1254 2003;33(3):279-87. PubMed PMID: 12609513.
1255 102. Vale VF, Moreira BH, Moraes CS, Pereira MH, Genta FA, Gontijo NF. Carbohydrate
1256 digestion in Lutzomyia longipalpis' larvae (Diptera - Psychodidae). J Insect Physiol.
1257 2012;58(10):1314-24. Epub 2012/07/31. doi: 10.1016/j.jinsphys.2012.07.005. PubMed PMID:
1258 22841889.
1259 103. da Costa-Latgé SG, Bates P, Dillon R, Genta FA. Characterization of Glycoside
1260 Hydrolase Families 13 and 31 Reveals Expansion and Diversification of α-Amylase Genes in
1261 the Phlebotomine. Front Physiol. 2021;12:635633. Epub 20210409. doi:
1262 10.3389/fphys.2021.635633. PubMed PMID: 33897451; PubMed Central PMCID:
1263 PMCPMC8063059.
1264 104. Benoit JB, Hansen IA, Szuter EM, Drake LL, Burnett DL, Attardo GM. Emerging roles of
1265 aquaporins in relation to the physiology of blood-feeding arthropods. Journal of comparative
1266 physiology B, Biochemical, systemic, and environmental physiology. 2014;184(7):811-25. doi:
1267 10.1007/s00360-014-0836-x. PubMed PMID: 24942313.
1268 105. Meireles-Filho AC, Kyriacou CP. Circadian rhythms in insect disease vectors. Mem Inst
1269 Oswaldo Cruz. 2013;108 Suppl 1:48-58. doi: 10.1590/0074-0276130438. PubMed PMID:
1270 24473802; PubMed Central PMCID: PMCPMC4109179.
1271 106. Meireles-Filho AC, da S Rivas GB, Gesto JS, Machado RC, Britto C, de Souza NA, et al.
1272 The biological clock of an hematophagous insect: locomotor activity rhythms, circadian
1273 expression and downregulation after a blood meal. FEBS Lett. 2006;580(1):2-8. Epub
1274 20051201. doi: 10.1016/j.febslet.2005.11.031. PubMed PMID: 16337945.
1275 107. Meireles-Filho AC, Amoretty PR, Souza NA, Kyriacou CP, Peixoto AA. Rhythmic
1276 expression of the cycle gene in a hematophagous insect vector. BMC Mol Biol. 2006;7:38. doi:
1277 10.1186/1471-2199-7-38. PubMed PMID: 17069657; PubMed Central PMCID: PMC1636064.

1278 108. Freeman EG, Dahanukar A. Molecular neurobiology of Drosophila taste. Curr Opin
1279 Neurobiol. 2015;34:140-8. Epub 2015/06/24. doi: 10.1016/j.conb.2015.06.001. PubMed PMID:
1280 26102453; PubMed Central PMCID: PMCPMC4577450.
1281 109. Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. Variant ionotropic glutamate
1282 receptors as chemosensory receptors in Drosophila. Cell. 2009;136(1):149-62. Epub
1283 2009/01/13. doi: 10.1016/j.cell.2008.12.001. PubMed PMID: 19135896; PubMed Central
1284 PMCID: PMCPMC2709536.
1285 110. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, et al. Ancient
1286 protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect
1287 taste and olfaction. PLoS Genet. 2010;6(8):e1001064. Epub 2010/09/03. doi:
1288 10.1371/journal.pgen.1001064. PubMed PMID: 20808886; PubMed Central PMCID:
1289 PMCPMC2924276.
1290 111. Rytz R, Croset V, Benton R. Ionotropic receptors (IRs): chemosensory ionotropic
1291 glutamate receptors in Drosophila and beyond. Insect biochemistry and molecular biology.
1292 2013;43(9):888-97. Epub 2013/03/06. doi: 10.1016/j.ibmb.2013.02.007. PubMed PMID:
1293 23459169.
1294 112. Hickner PV, Timoshevskaya N, Nowling RJ, Labbé F, Nguyen AD, McDowell MA, et al.
1295 Molecular signatures of sexual communication in the phlebotomine sand flies. PLoS Negl Trop
1296 Dis. 2020;14(12):e0008967. Epub 20201228. doi: 10.1371/journal.pntd.0008967. PubMed
1297 PMID: 33370303; PubMed Central PMCID: PMCPMC7793272.
1298 113. Liu N, Li T, Wang Y, Liu S. G-Protein Coupled Receptors (GPCRs) in Insects-A Potential
1299 Target for New Insecticide Development. Molecules. 2021;26(10). Epub 20210518. doi:
1300 10.3390/molecules26102993. PubMed PMID: 34069969; PubMed Central PMCID:
1301 PMCPMC8157829.
1302 114. Nowling RJ, Abrudan JL, Shoue DA, Abdul-Wahid B, Wadsworth M, Stayback G, et al.
1303 Identification of novel arthropod vector G protein-coupled receptors. Parasites & vectors.

1304 2013;6:150. doi: 10.1186/1756-3305-6-150. PubMed PMID: 23705687; PubMed Central
1305 PMCID: PMC3680159.
1306 115. Fevereisen, R. Insect CYP, genes and P450 enzymes. In: Gilbert, LI, editors.
1307 Insect Molecular Biology and Biochemistry. Amsterdam: Elsevier; 2012. p. 236-316.
1308 116. Bates PA, Depaquit J, Galati EA, Kamhawi S, Maroli M, McDowell MA, et al. Recent
1309 advances in phlebotomine sand fly research related to leishmaniasis control. Parasites &
1310 vectors. 2015;8:131. Epub 2015/04/18. doi: 10.1186/s13071-015-0712-x. PubMed PMID:
1311 25885217; PubMed Central PMCID: PMCPMC4352286.
1312 117. Mccoy KD. The population genetic structure of vectors and our understanding of disease
1313 epidemiology. Parasite-Journal De La Societe Francaise De Parasitologie. 2008;15(3):444-8.
1314 doi: DOI 10.1051/parasite/2008153444. PubMed PMID: WOS:000258945000042.
1315 118. Vigoder FM, Souza NA, Brazil RP, Bruno RV, Costa PL, Ritchie MG, et al. Phenotypic
1316 differentiation in love song traits among sibling species of the Lutzomyia longipalpis complex in
1317 Brazil. Parasites & vectors. 2015;8:290. Epub 2015/05/29. doi: 10.1186/s13071-015-0900-8.
1318 PubMed PMID: 26017472; PubMed Central PMCID: PMCPMC4456791.
1319 119. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are
1320 due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23(13):3133-57. Epub
1321 2014/05/23. doi: 10.1111/mec.12796. PubMed PMID: 24845075.

1322

1323

1324 **Supporting Information**

1325 **Supplemental Methods: Methods used for manual annotation**

1326 **Supplemental Results: Detailed information of annotated gene families**

1327 **S1 Table: Assembly statistics**

1328 **S2 Table: BUSCO analysis**

1329 **S3 Table: RNAseq samples**

1330 **S4 Table: Microsyntenny**

1331 **S5 Table: Toll pathway annotation**

1332 **S6 Table: Insect immune deficiency pathway annotation**

1333 **S7 Table: JakStat pathway annotation**

1334 **S8 Table: Galectin family annotation**

1335 **S9 Table: Transforming growth factor-beta family annotation**

1336 **S10 Table: Mitogen activated protein kinase family annotation**

1337 **S11 Table: Prophenoloxidase family annotation**

1338 **S12 Table: Salivary protein annotation**

1339 **S13 Table: Peptidase annotation**

1340 **S14 Table: Glycosidase Hydrolase family 13 annotation**

1341 **S15 Table: Chitinase family annotation**

1342 **S16 Table: Hexosaminidase family annotation**

1343 **S17 Table: Chitinase deacetylase family annotation**

1344 **S18 Table: Peritrophin family annotation**

1345 **S19 Table: Aquoporin family annotation**

1346 **S20 Table: *Lutzomyia longipalpis* circadian rhythm pathway annotation**

1347 **S21 Table: *Phlebotomus papatasi* circadian rhythm pathway annotation**

1348 **S22 Table: Number of chemoreceptor family genes**

1349 **S23 Table: G-protein coupled receptor family annotation**

1350 **S24 Table: MicroRNA annotation**

1351 **S25 Table: Heat shock and hypoxia gene family annotation**

1352 **S26 Table: Cuticular protein gene family annotation**

1353 **S27 Table: Juvenile hormone family annotation**

1354 **S28 Table: Insulin signaling pathway annotation**

1355 **S29 Table: Antioxidant family annotation**

1356 **S30 Table: Vitamin metabolism pathway annotation**

1357 **S31 Table: *Phlebotomus papatasi* population sequencing median coverage depth**

1358 **S32 Table: *Phlebotomus papatasi* population variant summary statistics**

1359 **S33 Table: *Phlebotomus papatasi* F_{ST}-Tajima's D overlap (including 10kb upstream and**

1360 **downstream)**

1361 **S34 Table: *Lutzomyia longipalpis* population sequencing median coverage depth**

1362 **S35 Table: Parameter values of male copulatory songs from *Lutzomyia longipalpis* from**

1363 **Araci and Olindina**

1364 **S36 Table: *Lutzomyia longipalpis* genes within differentiation islands**

1365 **S1 Figure: Conflicting phylogenetic signals.** Analysis of the gene phylogenies of individual

1366 orthologous groups identified three major topologies with sand fly-mosquito (41%), sand fly-fly

1367 (37%), or mosquito-fly (22%) sister clades. Comparisons of average branch lengths for each

1368 topology suggest that, although substitution rates in flies are always higher, orthologs that support

1369 the sand fly-mosquito topology show the lowest substitution rates in flies and the smallest

1370 differences in substitution rates among the fly, sand fly, and mosquito clades. In contrast, the

1371 sand fly-fly and mosquito-fly topologies show much higher substitution rates in flies and much

1372 greater differences in substitution rates among the three clades.

1373

1374 **S2 Figure: Scatter plots of ortholog locations in each pair of genomes.** 1- to -1 ortholog

1375 groups were associated utilizing OrthoDB and scaffolds with only one gene were removed prior

1376 to generating plots. (A) *P. papatasi* vs. *Lu. longipalpis*. (B) *Lu. longipalpis* vs. *An. gambiae*. (C) *P.*
1377 *papatasi* vs. *An. gambiae*. (D) *Lu. longipalpis* vs. *D. melanogaster*. (E) *P. papatasi* vs. *D.*
1378 *melanogaster*. (F) *D. melanogaster* vs. *D. simulans*. (F) *D. melanogaster* vs. *An. gambiae*.
1379

1380 **S3 Figure: Clustering of sand fly galectin protein sequences.** Condensed Neighbor-Joining
1381 tree depicting clustering among galectin protein sequences of sand flies (*P. papatasi* and *Lu.*
1382 *longipalpis*; open and filled squares, respectively), mosquitoes (*Ae. aegypti* and *An. gambiae*;
1383 open and filled circles, respectively), fly (*D. melanogaster;* filled triangle), eastern oyster (*C.*
1384 *virginica*; upside-down open triangle), and freshwater snail (*B. glabrata*; upside-down filled
1385 triangle). Branches encompassing shared orthologs are highlighted by blue shades. Sand fly
1386 specific clusters and genes are highlighted by orange shades. The evolutionary distances were
1387 computed using the p-distance method and are in the units of the number of amino acid
1388 differences per site. One thousand bootstrap replicates were performed, and only branches
1389 displaying at least 50% confidence are shown.

1390 **S4 Figure: Condensed Neighbor-Joining tree depicting clustering among n-**
1391 **acetylhexosaminidase** protein sequences of sand flies (*P. papatasi* and *Lu. longipalpis*; open
1392 and filled squares, respectively), mosquitoes (*Ae. aegypti* and *An. gambiae*; open and filled
1393 circles, respectively), fly (*D. melanogaster*; filled triangle), and beetle (*T. castaneum*; filled
1394 diamond). Branches encompassing sequences belonging to group I-IV n-acetylhexosaminidases
1395 are highlighted by a blue shade. The sand fly specific cluster is highlighted by an orange shade.
1396 The evolutionary distances were computed using the p-distance method and are in the units of
1397 the number of amino acid differences per site. One thousand bootstrap replicates were performed,
1398 and only branches displaying at least 50% confidence are shown.
1399

1400 **S5 Figure: Condensed Neighbor-Joining tree depicting clustering among chitin**
1401 **deacetylase catalytic domain sequences** of sand flies (*P. papatasi* and *Lu. longipalpis*; open
1402 and filled squares, respectively), mosquitoes (*Ae. aegypti* and *An. gambiae*; open and filled
1403 circles, respectively), fly (*D. melanogaster*; filled triangle), and beetle (*T. castaneum*; filled
1404 diamond). Branches encompassing sequences belonging to group 1-5 and 9 CDA are highlighted
1405 by blue shades. The evolutionary distances were computed using the p-distance method and are
1406 in the units of the number of amino acid differences per site. One thousand bootstrap replicates
1407 were performed, and only branches displaying at least 50% confidence are shown.
1408
1409 **S6 Figure: Condensed Maximum likelihood tree depicting peritrophin CBD domain**
1410 **similarities** among the sand flies *P. papatasi* and *Lu. longipalpis* and the red flour beetle *T.*
1411 *castaneum*. Open squares, filled squares, and filled diamonds represent *P. papatasi*, *Lu.*
1412 *longipalpis*, and *T. castaneum* domains, respectively. Branches exclusive to *T. castaneum* were
1413 color-coded in magenta; those specific to sand flies were highlighted in blue. The branch
1414 encompassing the CBD-like domain "CBDput" is highlighted in green. The branches shared by
1415 sand flies and RFB CBD domains are color-coded in orange. Maximum likelihood tree was
1416 constructed using the Whelan and Goldman (WAG) model with Gamma distributed among
1417 Invariant sites (G+I), as suggested by the Model test function of the Mega6 software. One
1418 thousand bootstrap replicates were performed, and only branches displaying at least 50%
1419 confidence are shown.
1420
1421 **S7 Figure: Comparison of predicted aquaporins from other flies.** Neighbor-joining tree was
1422 produced using MEGA6 using Dayhoff Model and pairwise matching; branch values indicate
1423 support following 3000 bootstraps; values below 50% are omitted.
1424

1425 **S8 Figure: Molecular phylogenetic analysis of vertebrate and invertebrate photolyases**
1426 **containing *Lu. longipalpis* and *P. papatasi* gene models.** The different photoyases are
1427 displayed on the right. The evolutionary history was inferred by using the Maximum Likelihood
1428 method based on the Jones-Taylor-Thorton + four gamma categories with 1000 bootstrap
1429 replicates (showing only above 65). Sequences with squares are vertebrate cryptochromes (black
1430 – cry-4, white – cry-1, cry-2, and cry-3); sequences with black traingles represent (6-4) insect
1431 photolyases; sequences with inverted black triangles are reprenting all insect photolyase repir
1432 proteins; and sequences with a dot symbol show insect cryptochromes (black –cry-1, white –cry-
1433 2). Dashed arrows point to *P. papatasi* photolyase sequences and straight arrows to *Lu.*
1434 *longipalpois* photolyase sequences.
1435
1436 **S9 Figure: Molecular phylogenetic analysis of *Lu. longipalpis, P. papatasi* and *D.***
1437 ***melanogaster* TRP channel sequences.** The different TRP subfamilies are displayed on the
1438 right. The evolutionary history was inferred by using the Maximum Likelihood method based on
1439 the Whelan and Goldman +Freq. model with 1000 bootstrap replicates.
1440
1441 **S10 Figure: Molecular phylogenetic analysis of *Lu. longipalpis, P. papatasi* and *D.***
1442 ***melanogaster* PPK sequences.** The different PPK subfamilies are displayed on the right. The
1443 evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan
1444 and Goldman +Freq. model with 1000 bootstrap replicates.
1445
1446 **S11 Figure: Comparison of predicted aquaporins from other flies.** Neighbor-joining tree was
1447 produced using MEGA6 using Dayhoff Model and pairwise matching; branch values indicate
1448 support following 3000 bootstraps; values below 50% are omitted.
1449

1450 **S12 Figure: *Phlebotomus papatasi* population structure.** Inferred population structure of *P.*
1451 *papatasi* individuals collected from Afghanistan (PPAFG; green), North Sinai - Egypt (PPNS;
1452 purple), and Tunisia (PPTUN; orange). (A) Phylogenetic Analysis. Rooted neighbor joining (NJ)
1453 radial tree generated with the Adegenet and ape packages of R. We included both *P. bergeroti*
1454 (PBRG; black) and *P. duboscqi* (PDMA; gray), and used *P. duboscqi* to root the trees. Bootstrap
1455 values represent the percentage of 1,000 replicates. (B) Principle component analysis (PCA).
1456 Individuals are plotted according to their coordinates on the first two principal components (PC1
1457 and PC2). (C) Admixture analysis. Ancestry proportions for Admixture models from *K* = 2 to *K* =
1458 7 ancestral populations. Each individual is represented by a thin vertical line, partitioned into *K*
1459 coloured segments representing the individual's estimated membership fractions to the *K* clusters.
1460 These data are the average of the major *q*-matrix clusters derived by CLUMPAK analysis.
1461
1462 **S13 Figure: Admixture cross validation error**. Violin plot of the cross-validation error for each
1463 of 30 replicates for each *K* value. (A) *Phlebotomus papatasi* populations. (B) *Lutzomyia*
1464 *longipalpis* populations.
1465
1466 **S14 Figure: Male copulatory courtship songs from Araci and Olinda.** (A) Approximate
1467 distance of Araci and Olinda from Jacobina (B). Male copulatory courtship song tracings of
1468 *Lutzomyia longialpis* males collected from Araci and Olindina. The figure shows ~1 s of song in
1469 each case. Map Source: Google Imagery TerraMetrics.
1470
1471 **S15 Figure : Distribution plots of the pairwise F_{ST} between the different populations of**
1472 ***Lutzomyia longipalpis*.** Weighted F_{ST} values for 1kb non-overlapping windows were calculated
1473 across the genome for each population comparison.
1474

1475 **S16 Figure: Manhattan plots of the pairwise F_{ST} between the different populations of**
1476 ***Lutzomyia longipalpis*.** The red horizontal lines indicate the upper 0.05% of F_{ST} distribution over
1477 the entire genome.
1478
1479 **S17 Figure: Manhattan plot of Tajimas'D for each population of *Lutzomyia longipalpis*.** The
1480 red and blue horizontal lines indicate the upper and lower 0.05% of Tajima's D distribution,
1481 respectively.
1482
1483 **S18 Figure: Genomic regions with high (outlier) Tajimas'D for different populations of**
1484 ***Lutzomyia longipalpis*.** (A) The Venn diagram summarizes the numbers of 1kb genomic
1485 windows with Tajimas'D values in the upper 2.5% of the different populations. (B) The Venn
1486 diagram summarizes the numbers of 1kb genomic windows with Tajimas'D values in the lower
1487 2.5% of the different populations.
1488
1489 **S1 Data File:** *Phlebotomus papatasi* CYPome Fasta File
1490
1491 **S2 Data File:** *Lutzomyia longipalpis* CYPome Fasta File
1492
1493

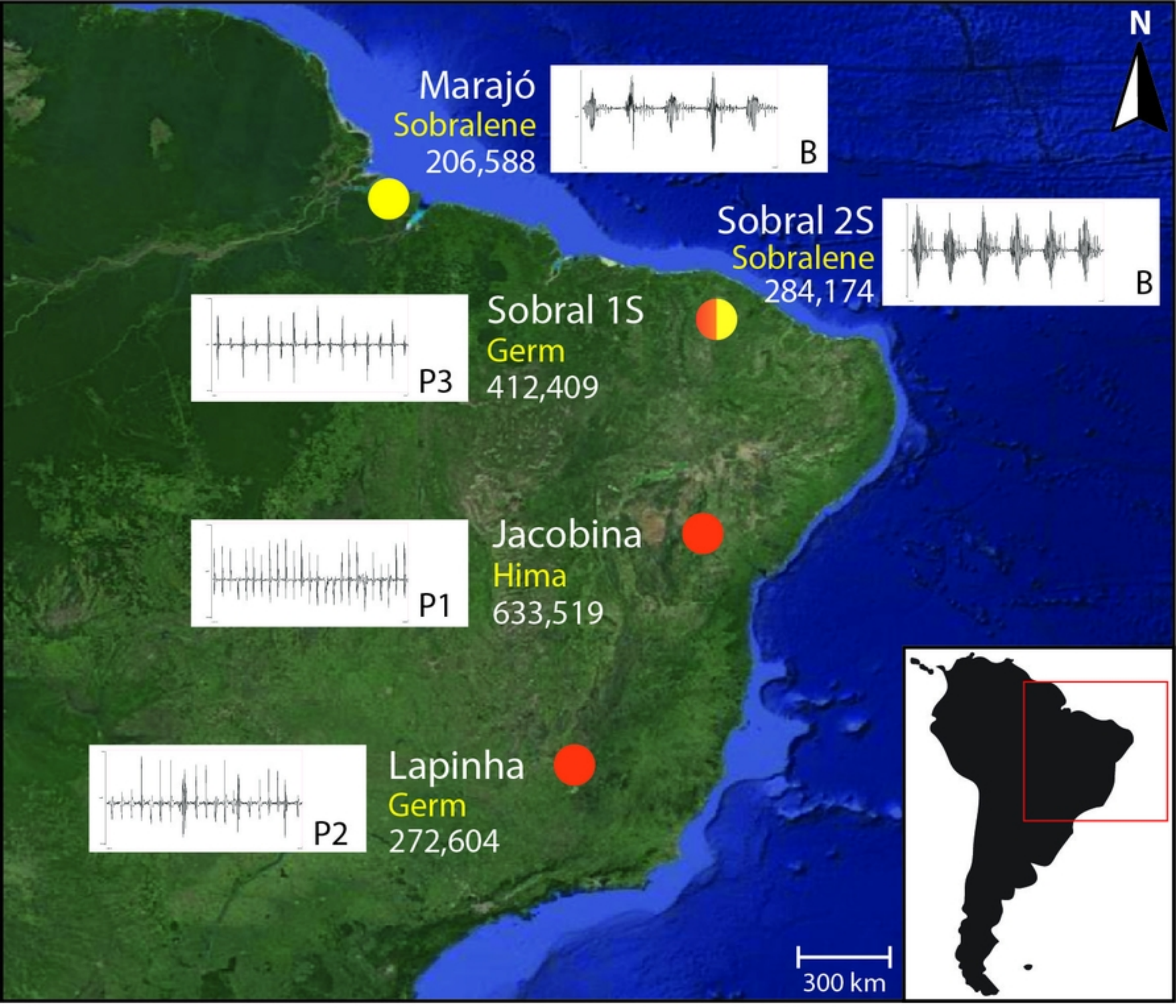
N
Marajó
Sobralene
206,588
B
Sobral 2S
Sobralene
284,174
B
Sobral 1S
Germ
412,409
P3
Jacobina
Hima
633,519
P1
Lapinha
Germ
272,604
P2
300 km

Figure 1

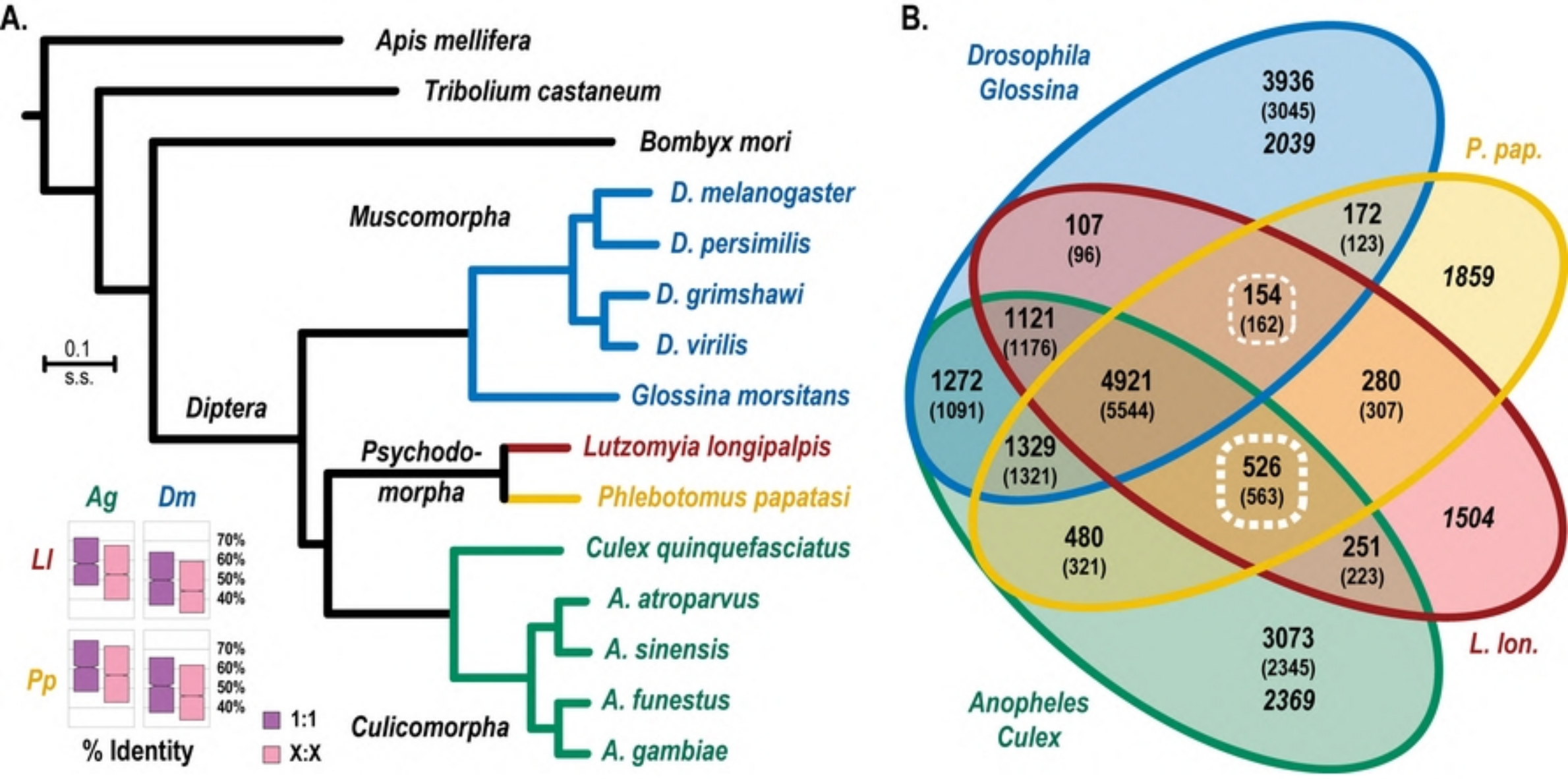
A.
Apis mellifera
Tribolium castaneum
Bombyx mori
Muscomorpha
D. melanogaster
D. persimilis
D. grimshawi
D. virilis
Glossina morsitans
0.1
s.s.
Diptera
Psychodo-
morpha
Lutzomyia longipalpis
Phlebotomus papatasi
Culex quinquefasciatus
A. atroparvus
A. sinensis
A. funestus
A. gambiae
Culicomorpha
Ag
Dm
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

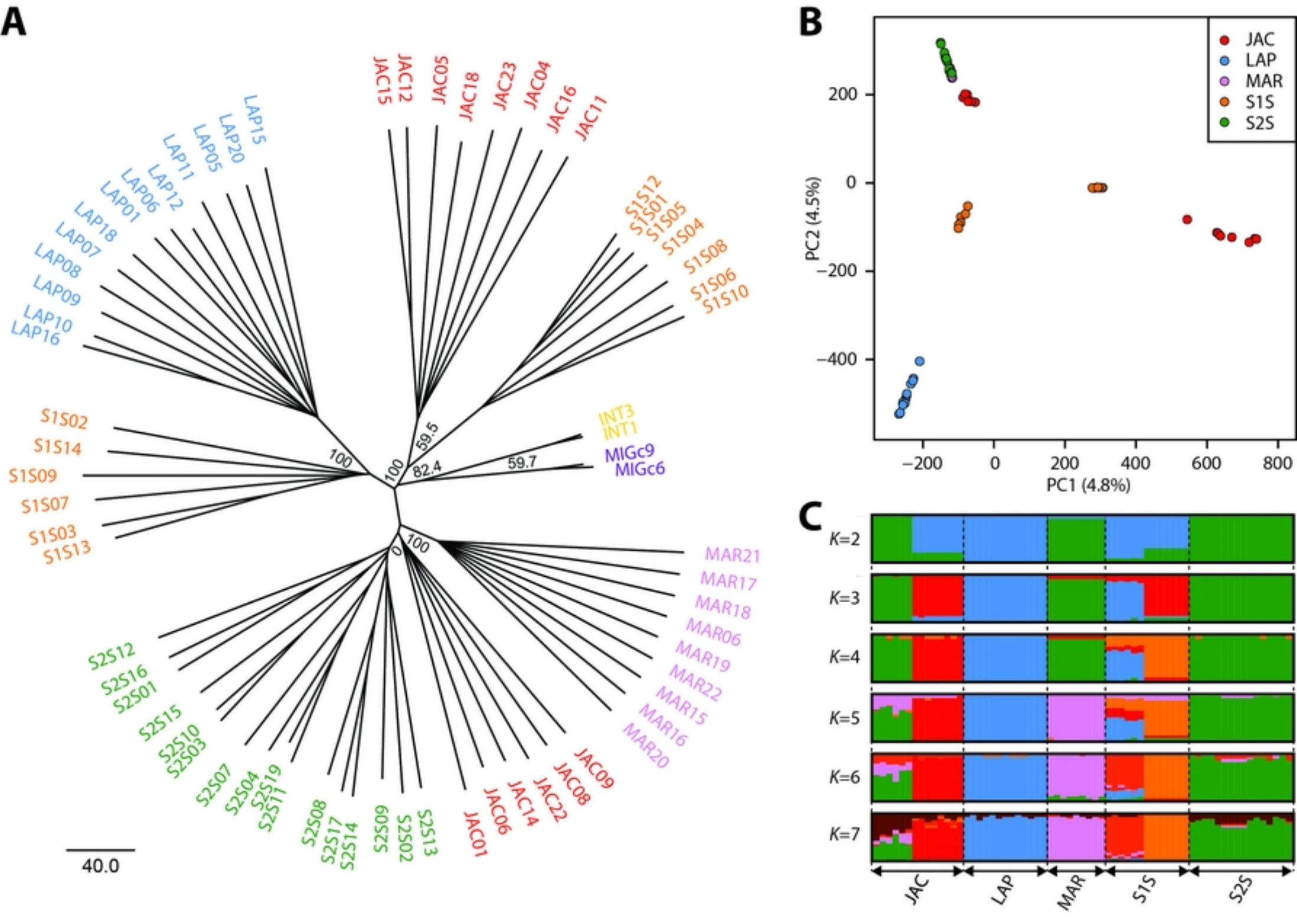
A
JAC15
JAC12
JAC05
JAC18
JAC23
JAC04
JAC16
JAC11
LAP15
LAP20
LAP05
LAP11
LAP12
LAP06
LAP01
LAP18
LAP07
LAP08
LAP09
LAP10
LAP16
S1S12
S1S01
S1S05
S1S04
S1S08
S1S06
S1S10
S1S02
S1S14
S1S09
S1S07
S1S03
S1S13
INT3
INT1
MIGc9
MIGc6
MAR21
MAR17
MAR18
MAR06
MAR19
MAR22
MAR15
MAR16
MAR20
JAC09
JAC08
JAC22
JAC14
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

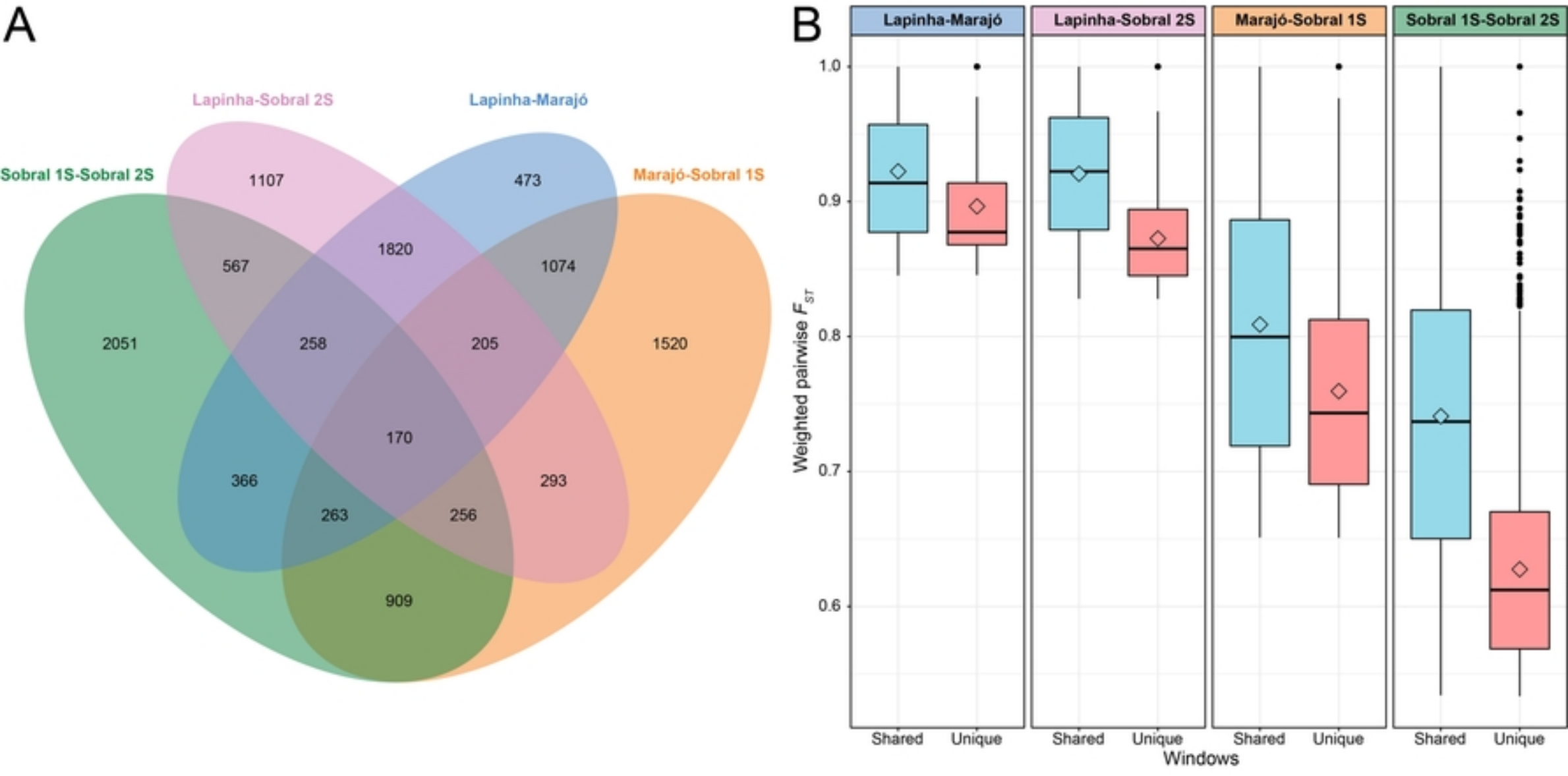
A
Lapinha-Sobral 2S
Lapinha-Marajó
Sobral 1S-Sobral 2S
Marajó-Sobral 1S
1107
473
1820
567
1074
2051
258
205
1520
170
366
293
263
256
909
B
Lapinha-Marajó
Lapinha-Sobral 2S
Marajó-Sobral 1S
Sobral 1S-Sobral 2S
Weighted pairwise F_{ST}
1.0
0.9
0.8
0.7
0.6
Shared
Unique
Windows

Figure 4

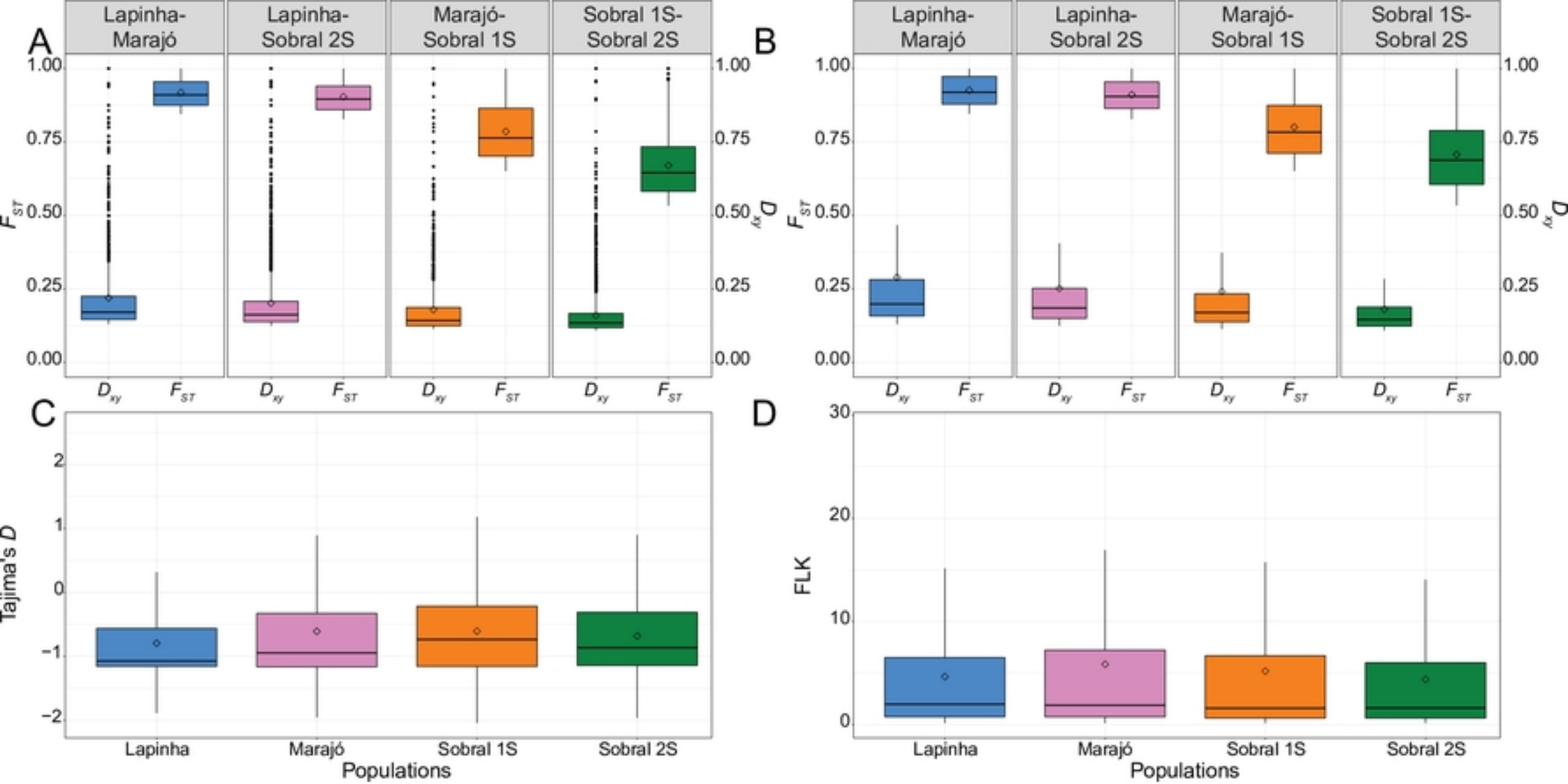
A
Lapinha-Marajó
Lapinha-Sobral 2S
Marajó-Sobral 1S
Sobral 1S-Sobral 2S
F_{ST}
D_{xy}
B
C
Tajima's D
Populations
Lapinha
Marajó
Sobral 1S
Sobral 2S
D
FLK

Figure 5