Sex differentiation, divergence, and introgression in the leishmaniasis vector *Lutzomyia longipalpis*

Authors:
Adam MM Stuckert[1]*, David Turissini[2], Emmanuel RR D'Agostino[3], David Peede[4], Maya Parvathaneni[2], Jonathan A. Rader[2], Carlos Muskus[5], Eduar Bejarano[6], Jeremy R. Wang[7], Rafael Vivero[5,8], Daniel R. Matute[2]*

[1]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA
[2]Biology Department, University of North Carolina, Chapel Hill, NC 27514, USA
[3]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA
[4]Department of Ecology, Evolution, and Organismal Biology, Brown University, Providence, RI 02912, USA
[5]PECET (Programa de Estudio y Control de Enfermedades Tropicales), Unidad de Entomología Médica, Universidad de Antioquia, SIU-Sede de Investigación Universitaria
[6]Grupo de investigaciones Biomédicas, Línea de Entomología Médica, Universidad de Sucre, Sincelejo, Colombia
[7]Department of Pathology, University of North Carolina, Chapel Hill, NC 27514, USA
[8]Grupo de Microdiversidad and Bioprospección, Facultad de Ciencias, Universidad Nacional de Colombia, Sede Medellín, Medellín, Colombia

Email: astuckert@uh.edu and dmatute@email.unc.edu

Keywords: Diptera, Psychodidae, sex-determination, genomics, disease vector

SIGNIFICANCE: Leishmaniasis has emerged as an important public health issue. Transmission occurs exclusively through Phlebotomine sandflies of the genera *Lutzomyia* and *Phlebotomus*, yet our ability to prevent its spread is constrained by limited knowledge of the biological and ecological factors that make these insects effective vectors. In this manuscript, we leverage whole-genome sequencing to investigate the evolutionary forces shaping genetic diversity in *Lutzomyia longipalpis*, a key vector of leishmaniasis.

ABSTRACT

An estimated one billion people are at risk of acquiring the disease leishmaniasis annually. This disease, among others, is transmitted to humans via sandfly vectors of the family Psychodidae. In the Neotropics, *Lutzomyia longipalpis* is one of the most common vectors of leishmaniasis, but little is known about its genetics and population structure, hampering control efforts. We generated sex-specific genome assemblies of *Lu. longipalpis* from Colombia, and examined population genomics of this species. We found that sex accounts for ~11% of genome-wide variation, primarily driven by differentiation between the sexes in an XY sex chromosome. Analyses of transposable elements indicate that the size difference in the sex chromosome is driven by more transposable elements in the male chromosome. Further, we examined population genomics and structure across the South American continent. We identified significant genetic divergence between Colombian and Brazilian sandflies and population structure within *Lu. longipalpis sensu lato* samples collected from the same sites in Brazil. These results strongly suggest the presence of multiple, sympatric species within a cryptic species complex, concordant with other published phenotypic evidence.

INTRODUCTION

Phlebotomine sandflies of the genera *Lutzomyia* and *Phlebotomus* are the only known vectors of *Leishmania*, a protozoan parasite that is transmitted to humans[1,2]. *Leishmania* infections can lead to the clinical disease leishmaniasis, which can cause skin ulcers and scarring, destruction of mucous membranes, extreme organ damage (e.g., liver, spleen, and bone marrow), and can be fatal in some cases. *Leishmania* is endemic to 98 countries and an estimated one billion people are at risk of infection[3,4]. Despite its high prevalence and disease burden, leishmaniasis is poorly understood, and thus is classified as a neglected tropical disease[5,6]. Although leishmaniasis is largely regarded as a tropical disease, *Lutzomyia* ranges are expanding northward[7] and the U.S. is experiencing increasing caseloads of leishmaniasis[7–9]. Ecological niche modeling predicts that leishmaniasis is likely to become more common in the U.S. due to climate change[10], and this is already evident via increasing clinical cases[11]. Treating leishmaniasis is difficult and often prohibitively expensive for individuals in underserved areas with endemic *Leishmania,* thus, preventing the spread of leishmaniasis is critically important[12]. Perhaps the most effective way to prevent the transmission of vector-transmitted diseases like leishmaniasis is to control the vector itself[13]. Despite the clear importance of *Leishmania'*s vectors for human well-being, preventive efforts are hampered by our lack of knowledge of the genetic and ecological factors that make sandflies effective vectors of *Leishmania*. Compounding the problem, functional studies of leishmaniasis vector biology are hindered by a variety of gaping holes in scientific knowledge, including: taxonomy and phylogenetics of sandflies, genetic characteristics of sandfly susceptibility to leishmaniasis, and a lack of genomic resources.

In the Neotropics, there are ~40 *Lutzomyia* species that transmit *Leishmania*, but one of the most commonly implicated and widespread is *Lutzomyia longipalpis.* This species has a range from Argentina through Mexico[2,14]. Complicating our understanding of this species' biology, a variety of lines of evidence suggest that *Lu. longipalpis* is actually a species complex containing a number of cryptic species with substantial phenotypic divergence across its geographic range[15,16]. This evidence includes variation in microsatellites of key genes[17–19], courtship song[20], morphology[21,22], and pheromones[18]. Nevertheless, this "species" is thought to be the major vector of *Leishmania infantum,* the causative agent of leishmaniasis in the Neotropics[23]. This species inhabits a wide variety of microhabitats, including silvan habitats and inside of rural houses[24–29]. Like most phlebotomine sandflies, females of this species are hematophagous (feed on blood), and can spread leishmaniasis to birds, dogs, pigs, horses, humans, and a variety of other mammalian hosts during feeding, meaning that common subsistence animals or pets can be intermediate hosts and promote vector populations[30]. Dramatic environmental changes in human land use (e.g. deforestation, roads, etc.) have caused concomitant changes in *Lu. longipalpis* distribution and habitat use. Of particular note is that *Lu. longipalpis* has shown a shift in recent decades to live and breed in highly populated urban areas[27]. Given that this species is a common vector of leishmaniasis, this puts an ever-increasing number of people at risk of acquiring leishmaniasis.

In this piece, we address fundamental genetic questions in the widespread leishmaniasis vector *Lutzomyia longipalpis.* First, we present highly-contiguous chromosomal and sex-specific genome assemblies of *Lu. longipalpis.* We then use these new assemblies as tools to examine

population genomics of this sandfly from the Andean region of Colombia. We then further use publicly available genomic data from Brazilian *Lu. longipalpis* to examine population genomics of this vector species across the South American continent. As predicted by previous work suggesting *Lu. longipalpis sensu lato* is a cryptic species complex, we find extensive population differentiation between Colombian and Brazilian populations. Additionally, our population genomic analyses indicate extremely low nucleotide diversity, elevated runs of homozygosity, and lower effective population sizes in Colombia relative to any Brazilian population. Our results provide a roadmap for the implementation of genomic analyses in sandflies and to ultimately understand the molecular and genetic underpinnings of the interactions between vectors and parasites.

RESULTS

Genome sequencing:

We produced sex-specific genomes for female and male Colombian sandflies (Table 1). Table S1 lists all the accession numbers. The initial female contig assembly built from PacBio HiFi reads was 148 Mbp in length spread across 2,254 contigs and had an N50 of 1.25 Mbp. The male contig assembly built from PacBio HiFi reads was 159.7 Mbp in length spread across 1,835 contigs with an N50 of 1.14 Mbp. The inclusion of HiC data improved our genome assembly. Scaffolding our assemblies with HiC data yielded putative chromosome-level assemblies for both female and male sandflies. The female assembly was 151.2 Mbp in length and contained four chromosomes with lengths of 47.7, 37.2, 29.0, and 22.8 Mbp and represented over 95% of the total genome. The female assembly contained >95% of complete dipteran orthologs: 94.16% in complete single copies, 1.13% in complete duplicated copies, 0.46% fragmented, and 4.26% missing. The male assembly was 159.7 Mbp in length and contained four chromosomes of 49.0, 36.8, 29.3, and 22.5 Mb, representing over 95% of the total genome. The male assembly contained just under 95% of complete dipteran orthologs: 93.49% in complete single copies, 1.19% in complete duplicated copies, 0.33% fragmented, and 4.99% missing.

Next, we identified the differences between the male and female genomes (Figure 1A). We identified 640,461 unique kmers associated with sex-limited chromosomes using SRY on our Illumina data. These kmers mapped to 111,862 HiFi reads spanning 888 Mbp in the male dataset. Our initial draft “Y” chromosome assembly was 26.7 Mbp across 740 contigs. After scaffolding, we placed 691 contigs (93.4% of contigs) into 50 scaffolds in an assembly that was 26.8 Mbp in length (including placed gaps) with a largest scaffold of 26.2 Mbp (thus the other 49 scaffolds contain ~0.6 Mbp). Thus, we were able to scaffold nearly the entire section of the “Y” chromosome that is limited to males into a single scaffold, with minimal unplaced content outside of this scaffold.

Table 1. Genome statistics. The Brazilian assembly comes from[31], and is derived from a colony collected in the 1980s from Jacobina, Bahia State, Brazil. Our female and male assemblies are derived from wild-caught sandflies from Ricaurte, Colombia.

143

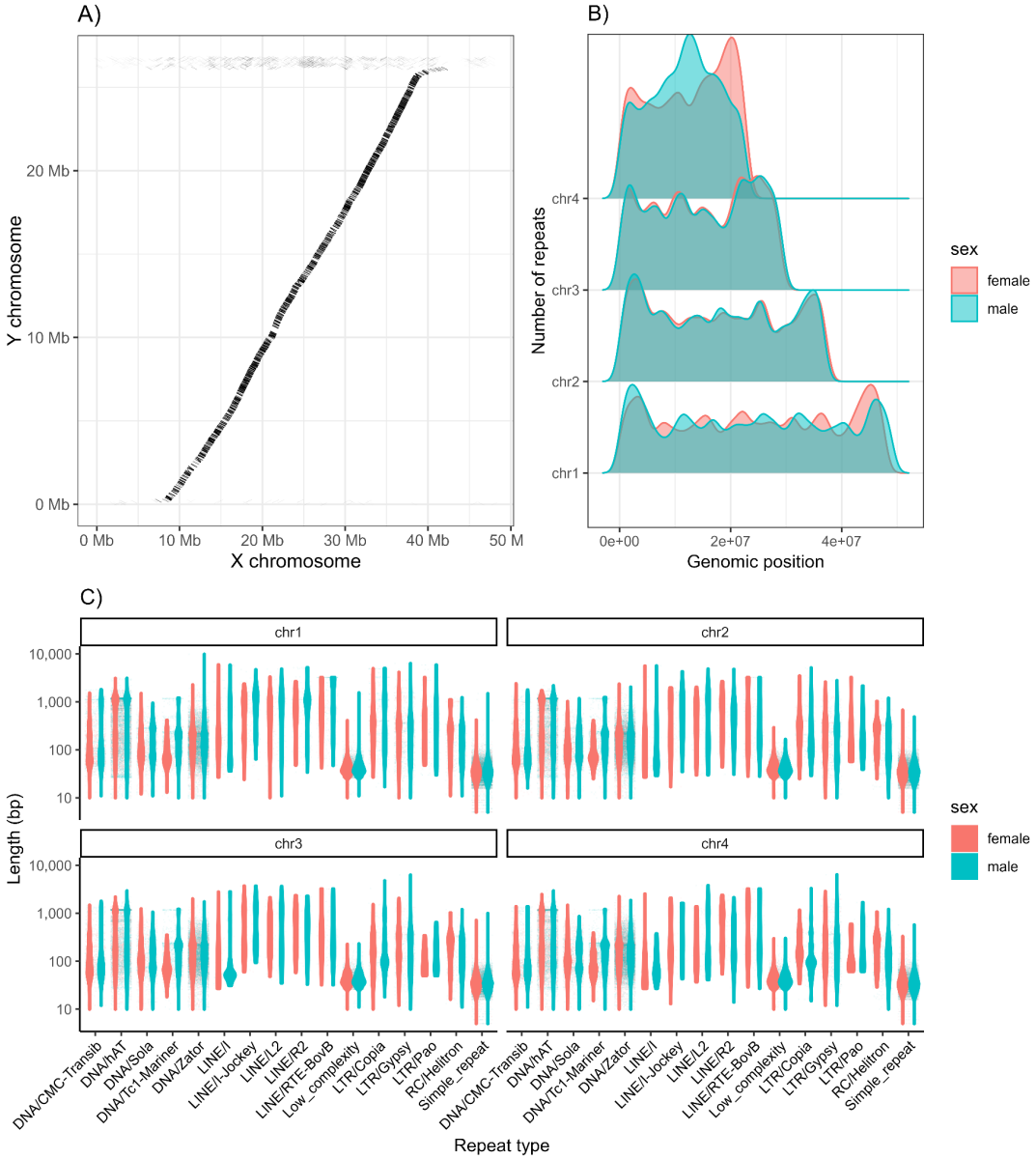
	Brazilian *Lu. longipalpis* assembly	Female contig	Female scaffolded	Male contig	Male scaffolded
Summary					
Assembled bases (Mbp)	159.7	1997	151.2	159.7	159.7
Number of contigs	383	151.2	2002	1835	1849
Contig N50 (Mbp)	1.04	1	1	1.1	1.1
Number of scaffolds	138	-	1742	-	1618
Scaffold N50 (Mbp)	40.6	-	37.2	-	36.8
Percentage of assembly in scaffolds	93.2	-	91	-	86.7
BUSCO (percentages)					
Single copies	90.5	94.12	94.16	93.9	93.49
Duplicated copies	6.73	1.13	1.13	1.19	1.19
Fragmented copies	0.33	0.46	0.46	0.4	0.33
Incomplete copies	0	0	0	0	0
Missing	2.89	4.29	4.26	5.02	4.99
RepeatMasker (Mbp)					
Percentage of repeats (%)	-	-	15.69	-	15.4
Repeat bases (Mbp)	-	-	23.7	-	24.6
Retroelements	-	-	4.37	-	4.92
DNA transposons	-	-	19.3	-	19.6
Small RNA	-	-	0.39	-	0.0319
Simple repeat	-	-	1.3	-	1.29
Low complexity	-	-	0.47	-	0.5
Total bases masked	-	-	25.69	-	26.67

144

145 *Comparison between the sexes indicates increased transposable element content in males on*
146 *the sex chromosome:*

147 Transposable elements (TEs) are considered important drivers of sex chromosome
148 evolution[32,33], so we quantified transposable element content in the female and male
149 assemblies. RepeatMasker identified and masked a total of 16.99% (female) and 16.70% (male)
150 of the genome, 15.69% and 15.4% of which were TEs in the female and male assemblies
151 respectively (Figure 1B). Most repeats were DNA transposons (12.8% and 12.3% of the
152 genome in females and males respectively) or Retroelements (2.9% and 3.1% of the genome in
153 females and males respectively, Figure 1C). A small proportion of the genome were simple
154 repeats (0.9% and 0.8% of the genome in females and males respectively). The difference in
155 relative proportion of transposable elements between the sexes is largely driven by
156 chromosome 1. This chromosome is not only longer in males (49 Mbp vs 47.7 Mbp), but also
157 has a larger overall proportion of bases masked in males (15.89% vs 13.60%, Figure 1C). The

158 male chromosome 1 has a total of 7.8 Mbp of transposable elements masked by our repeat
159 masking pipeline, whereas the female chromosome has 6.499 Mbp of transposable elements
160 masked. The difference, 1.3 Mbp, corresponds to the total difference in size between the
161 assembled male and female chromosome 1 (see Figure S1). In the three autosomal
162 chromosomes, the proportion of transposable elements to assembled chromosome length is
163 <0.5% different between females and males.



164
165 **Figure 1. The sex chromosomes of *Lu. longipalpis*. A.** Synteny plot of alignments between
166 assembled male (presumptive) X and Y chromosome sequences. **B.** Number of repeats in the
167 female and male genome assemblies based on genomic position. **C.** Violin plots of TEs in the
168 female and male genome assemblies and their length.
169
170
171

172 *Genome annotation:*
173 We annotated both sexes of our Colombian *Lu. longipalpis.* In total, we annotated
174 14,559 genes in the male assembly; 10,874 of these predicted genes were in the four
175 chromosomes (4008, 2607, 2371, and 1888 in chromosomes 1-4 respectively), as opposed to in
176 the contigs. In the female assembly, we annotated 12,029 genes in total; 10,934 of these
177 predicted genes were in the four chromosomes (3,922; 2,662; 2,371; and 1,979 in
178 chromosomes 1-4 respectively). The predicted genes are approximately equal to RefSeq
179 predictions in *Lu. longipalpis* from other localities (GCF_024334085.1), which has 12,045
180 predicted genes. In both sexes of our assemblies, the majority of predicted genes are found in
181 the chromosomes. However, our male assembly has approximately 2,000 more predicted genes
182 than existent assemblies. This is due to the presence of multiple copies of genes between the
183 heteromorphic chromosomes — most of which are represented as unplaced contigs. While
184 females had more predicted genes overall in the regions of the genome that are located in the
185 chromosomes, males had a greater number of predicted genes owing to the predicted genes
186 found in un-scaffolded contigs. Further, males had more predicted genes in chromosome 1, the
187 heteromorphic chromosome. We identified a total of 3,462 genes annotated in the highly
188 differentiated region of chromosome 1 (average F_{ST} > 0.3 between females and males) in the
189 male assembly. Of these, we tentatively assigned identifications to 3,611 genes based on
190 protein evidence from OrthDB version 11 and VectorBase version 67. Among the genes found
191 in this region of high differentiation between females and males within our Colombian population
192 are a number of genes related to courtship (e.g., *takeout*), sex-differentiation (e.g., *sex-lethal,*
193 *fruitless, doublesex- and mab-3-related transcription factor A2, lethal(2)*), testes-specific genes
194 (e.g., ejaculatory bulb-specific protein 3, ejaculatory bulb-specific protein 3-like, testis-expressed
195 protein 10 homolog, motile sperm domain-containing protein), putatively related to feeding and
196 finding mates (e.g., gustatory receptor genes, gustatory receptor for sugar taste 64f, odorant
197 receptor genes), or related to blood meals (e.g., venom allergen 5).
198 Nucleotide diversity across the genome is similar in females (π = 0.00177 ± 0.00233 SD
199 female assembly; π = 0.00170 ± 0.00228 SD male assembly) and males from Colombia (π =
200 0.00171 ± 0.00236 SD female assembly; π = 0.00159 ± 0.00225 SD male assembly; Figures
201 2A-2B). However, concordant with other results, nucleotide diversity was lower on chromosome
202 1 in female sandflies (π = 0.00138 ± 0.00192 SD female assembly; π = 0.00131 ± 0.00196 SD
203 male assembly) than males (π = 0.00165 ± 0.00239 SD female assembly; π = 0.00160 ±
204 0.00225 SD male assembly).
205 Next, we studied the extent of differentiation between the *X*-chromosome and the
206 potentially male-restricted chromosome using PCA. We saw evidence of genomic differentiation
207 between female and male genomes in our study. PC1, which corresponds to the sex of the
208 sample, explained ~11.2% of genome-wide differentiation. This signal was clearly driven by
209 chromosome 1, which showed a similar pattern in which sex explains ~25.1% of variance in the
210 PCA (Figure 2C). This signal was entirely absent when we did PCA for the putative
211 "autosomes", chromosomes 2-4, with comparatively very little of the variation explained (PC1:
212 4%, PC2: 2% Figure 2D) and no correspondence to the sex of the sample. Genealogies based
213 on chromosome 1 showed a similar result (Figure S2). Additionally, we saw a long, extended
214 region of differentiation (i.e., elevated F_{ST}) that occurred between female and male sandflies on
215 chromosome 1 (Figure 2E). This region also shows an extended region of decreased nucleotide

diversity in female sandflies (Figure 2A), but there was no concomitant decrease in nucleotide diversity within that region in male sandflies (Figure 2B). Given this finding, we compared π between sexes, with the specific hypothesis that π was lower on chromosome 1 in females than on autosomes. A linear model indicated that sex (t = 0.744, p = 0.4821) and the interaction of sex*chromosome (t = -0.984, p = 0.4821) were not a significant explanatory variable, but that chromosome was (t = 4.260, p = 0.381), and it was driven by lower π in the sex chromosome.

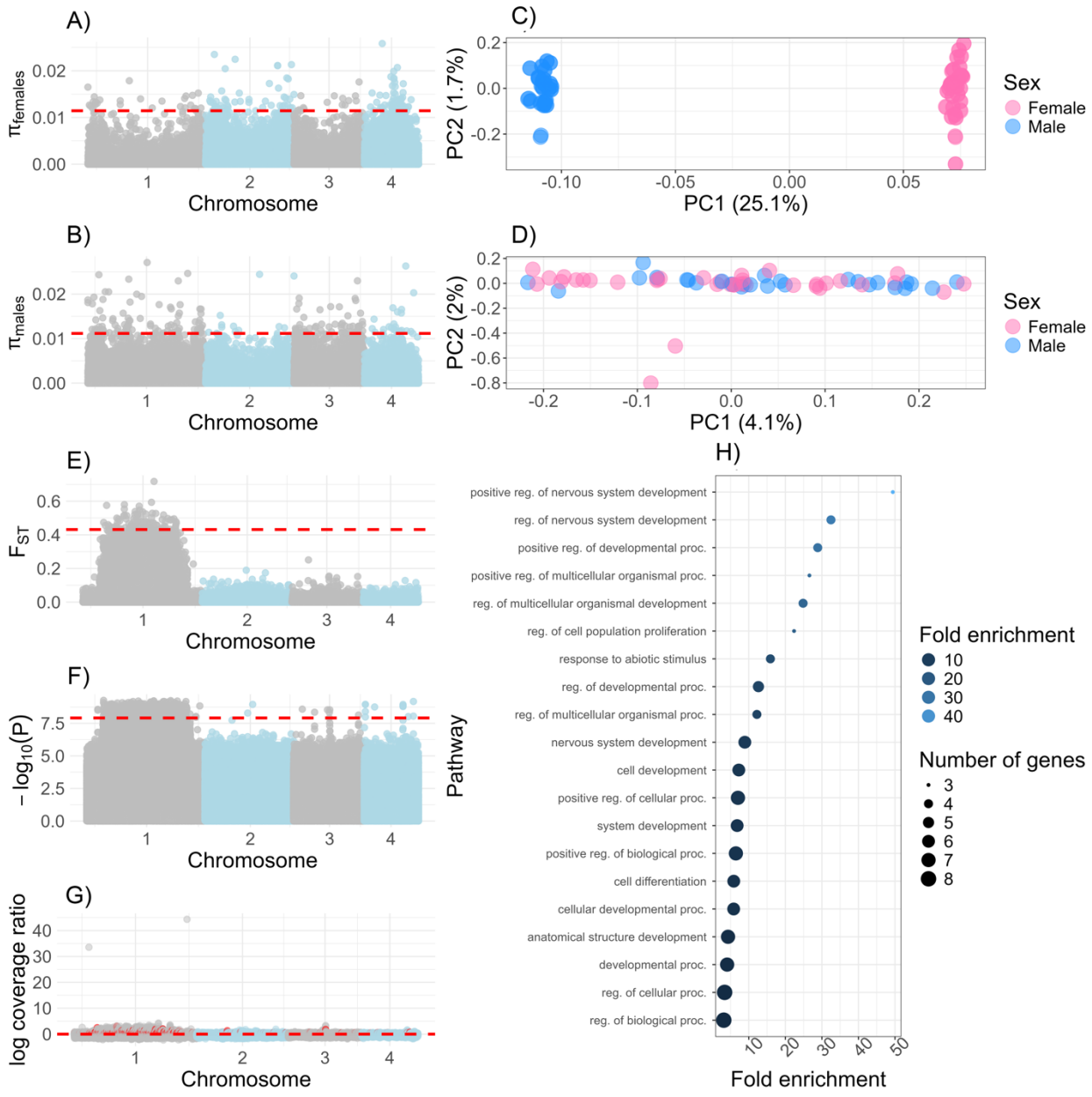


Figure 2. Sex chromosome differentiation in *Lu. longipalpis*. A. π in female sandflies from Colombia. **B.** π in male sandflies from Colombia. **C.** PCA of only chromosome 1. **D.** PCA of the autosomes, chromosomes 2-4. **E.** F_{ST} between the sexes. **F.** Manhattan plot of genome-wide association study between the sexes in Colombian sandflies. **G.** Coverage represented as a

227 ratio of average male depth divided by average female depth, calculated as a 2,000bp mean.
228 Two regions in chromosome 1 have been removed that have average log coverage of >30 (with
229 mean positions at 5847001 and 45239001). **H.** GO biological processes analysis from GWAS
230 results.

231

232 *Genome wide association studies revealed autosomal genes associated with sex*

233 Next, we did a GWAS to identify genetic regions associated with sex determination. We
234 identified 2,641 genic SNPs, and 3,165 intergenic SNPs associated with sex, the vast majority
235 of which were located in the sex chromosome. The highly differentiated region between sexes in
236 chromosome 1, the same that shows differentiation by our other analyses, showed a strong
237 association with sex (Figure 2F). Nonetheless, there were autosomal markers strongly
238 associated with sex. Sixty-seven SNPs were also associated with significant differences in
239 genomic coverage between the sexes. Eight genic, autosomal SNPs mapped to three predicted
240 genes in the three autosomal chromosomes, one of which has no detectable ortholog in any
241 other dipteran (g5391). A second gene (g7272) showed strong similarity to *mdg3*
242 retrotransposons in *D. melanogaster* (FBgn0043882). This gene also showed a difference in
243 genomic coverage between the sexes in our short read sequencing dataset, with a
244 log2(male:female) coverage ~0.78. The rest of markers showed a uniform distribution of
245 coverage around 1 (Figure 2G). The third and final gene with genic SNPs associated with sex
246 was Apoptotic protease-activating factor 1 (*apaf-1*), which plays a role in sex-specific cell
247 apoptosis and male infertility[34,35]. Twenty-four intergenic SNPs were in the autosomes, and they
248 are in close proximity to seven genes and three transposable elements. Table S2 lists their
249 name, location, association *p*-value, and known orthologs in other dipterans. No gene involved
250 in primary sex determination showed a significant association. Of note, we saw a significant
251 enrichment for genes involved in regulation of nervous system development (Figure 2H,
252 GO:0051962: 49.4X; GO:0051960: 32.5X, FDR ≤ 9.6×10^{-4}). Table S3 lists other GO categories
253 with their associated enrichment values.

254

255 *Population genomics across the continent:*
256 *Lutzomyia longipalpis* has been proposed to be differentiated into distinct populations
257 and even separate species across South and Central America[15,36,37]. We used a two-pronged
258 approach to study the population structure between our Colombian samples and samples of
259 females from several Brazilian populations[31]. Notably, Sobral was split *a priori* by Labbé et al.[31]
260 into Sobral 1S and Sobral 2S. First, we used phylogenetic trees. An unrooted genealogy based
261 on 100kb windows revealed a clustering determined by geographical location consistent with
262 previous studies[31]. Jacobina and Sobral 1S were each formed by two distinct and unrelated
263 clusters (Figure 3A-3B). We further examined gene concordance between populations as an
264 estimate of gene flow. We created a single “population” tree (akin to a species tree) from our
265 multiple individual sampling of both the whole genome, and of BUSCO genes. When forced to
266 recapitulate the populations, we found poor gene concordance support for several branches,

indicating poor alignment between individual gene trees and our consensus population tree (Figure S3). This phylogenetic discordance may occur in instances of recent divergence, incomplete lineage sorting, hybridization, and poor tree inference which may be occurring in this population tree[38]. This system may be an amalgamation of all of these factors. In summary, the two analyses were consistent and suggested significant differentiation between lineages from different locales, particularly between Colombian and Brazilian samples. Further, we observed evidence for differentiation between samples from the same locale in Jacobina, Sobral 1S, and Sobral 2S (Figure 3A) which indicated the existence of differentiated and syntopic lineages. Second, we used PCA to represent the genetic variation across the continent (Figure 3B). The first three principal components (PCs) explain 61% of the genetic variance of South American *Lu. longipalpis* (Figure S4). PC1 accounted for ~34% of genome-wide variation, and appears to reflect differentiation between Colombia and Brazil, and between some Brazilian lineages (i.e., Lapinha cave; Figure 3C). PC2 and PC3 also capture considerable variance (16.4 and 10.6% respectively; see Figure S5). PC2 separated Marajo and Sobral 2S from Ricaurte, Sobral 1S, Laphina with Jacobina samples split between the two groups. Sobral 2S, Marajo, and Ricaurte are similar in PC3, but Jacobina and Sobral 1S are divided into two separate clusters from each locality.

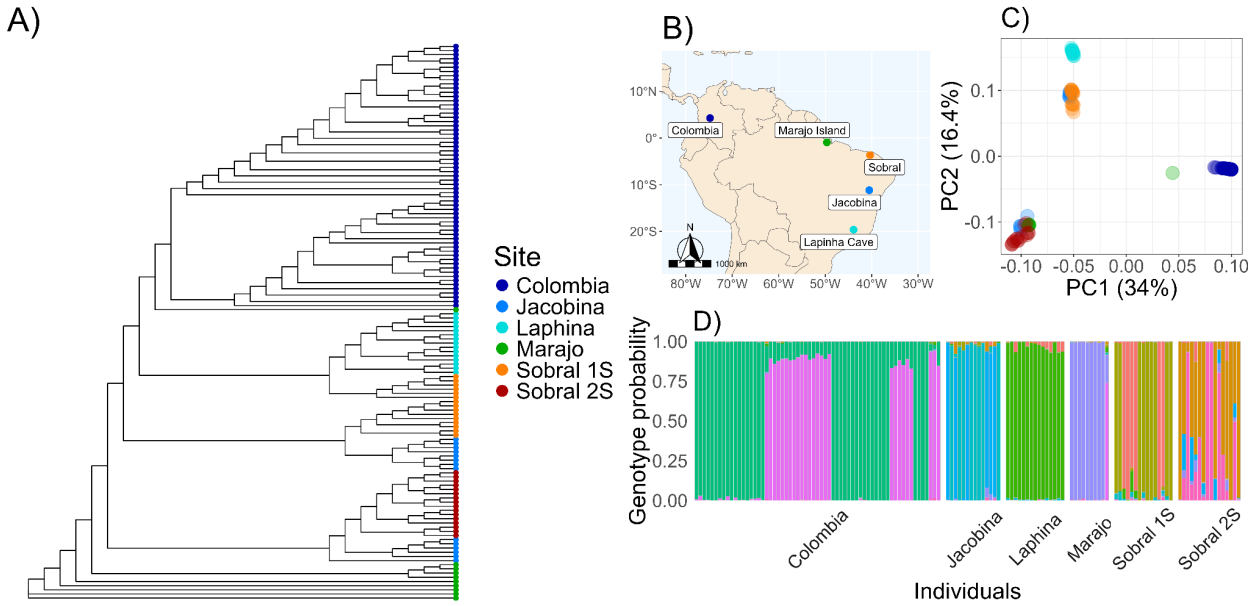


Figure 3. Partition of genetic diversity in *Lu. longipalpis* across South America. A. Phylogeny of sampled *Lutzomyia longipalpis* across South America. **B.** *Sampling localities with genomic data included in population genetics analyses.* ***C.*** *Principal component analysis of Colombian sandflies from Colombian (this study) and Brazilian sandflies.* ***D.*** *Structure plot of genotype probabilities across these same samples. Colombian flies divide into two genotypes driven by sex. All Brazilian sandflies are females, while our Colombian samples also contain males. Panels A and C have the same legend for sampling locality, D has a different color schema due to the much larger number of genotypes predicted. Brazilian populations are from Labbe et al.[31].*

Next, we measured the extent of gene flow between these genetic clusters. Given the strong evidence of reproductive isolation between lineages[21,37], we hypothesized reduced—or

297 even nonexistent—gene flow between the populations we identified. We used *PCAngsd*, which
298 identifies the most likely number of clusters, but also estimates admixture proportions from allele
299 frequencies. Genetic variance was best explained by the existence of 10 clusters. Two of these
300 clusters correspond to male and female *Lu. longipalpis* in Colombia, with little contribution to
301 other lineages (a probability of <1% all Brazilian samples, except for the Colombian genotype
302 that corresponds to males, which had a 73.9% probability in one Marajo sample, leading us to
303 believe it was a male sandfly). This analysis (Figure 3D) also separated Colombian and
304 Brazilian sandflies. The genetic variation in Brazilian samples was best explained by the
305 existence of eight groups which correspond largely to the five geographic sites from which they
306 were collected. The only exceptions were Sobral 1S, Sobral 2S, and Jacobina where we
307 detected at least two main clusters per locality. The majority of individuals show limited
308 evidence of admixed ancestry. In Colombia, Laphina, Jacobina, and all but one Marajo sandfly,
309 admixture proportions are <10% per individual. This is particularly notable in the sites with
310 syntopic lineages which could potentially have the chance to interbreed and yet do not appear
311 to (Jacobina, some individuals within Sobral), as evidenced by divergent genotype clusters
312 driving nearly the entirety of individual genetic makeup (e.g., blue vs teal individuals in Jacobina;
313 Figure 3D). However, Sobral showed significant admixture within certain individuals (those of
314 the "2S" lineage, Figure S6). We then proceeded to examine gene flow between populations
315 using fbranch statistics. These analyses indicated high gene flow from Sobral 2S into Jacobina,
316 Jacobina into Sobral 1S and Sobral 1S into Laphina (with Sobral "1S" and "2S" being *a priori*
317 identified as different *Lu. longipalpis* lineages by[31]; Figure S7). Thus, there is either significant
318 gene flow between populations or our consensus population tree does not adequately reflect
319 cryptic diversification and potential syntopic lineages. Indeed, the consensus phylogeny of all
320 individuals indicates the latter, as Sobral 2S individuals cluster with some Jacobina individuals.
321 Overall, these results confirm that the different lineages of *Lu. longipalpis* are divergent,
322 indicates advanced divergence towards speciation, and likely the presence of syntopic and
323 hybridizing lineages.
324
325
326

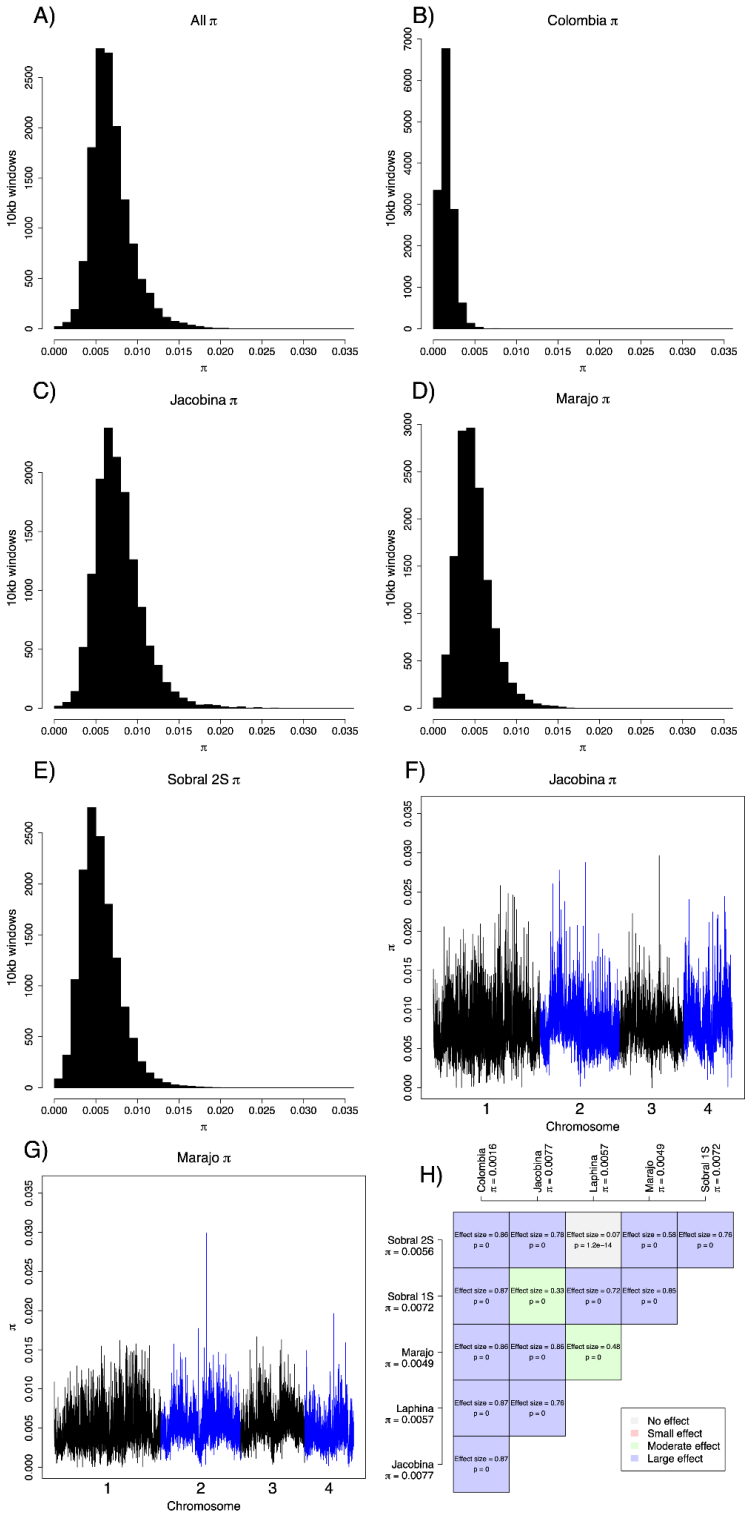


Figure 4. Genetic diversity in different lineages of *Lu. longipalpis*. A. Histograms of π in 10kb window sizes along the genome. **B.** π in Colombia **C**. π in Jacobina **D.** π in Marajo **E.** π in Sobral 2S. **F**. π along the genome in Jacobina. **G.** π along the genome in Marajo. Figure 2A and 2B show the π in Colombia. Figure S8 shows π along the genome for the remaining four populations. **H.** Pairwise comparisons show differences in π among pairs of populations.

The heterozygosity in the whole *L. longipalpis* clade was 0.06 (Figure 4A), but given the population structure described above, we studied the partitioning of genetic diversity within each *Lu. longipalpis* lineage. First, we quantified the extent of heterozygosity in each lineage (Figure 4B-E). Some of the populations are less genetically diverse than others. Colombia shows the lowest heterozygosity (0.0036); Sobral-1S and Marajo showed the highest heterozygosity (~0.01) and were three times more diverse than Colombia. In general, the trends of heterozygosity along the genome were similar across populations (Figures 2A, 2B, 4F, 4G, Figure S8). All pairwise comparisons were significant (Figure 4H), indicating differences in the level of genetic diversity across *Lu. longipalpis* lineages. Second, we looked for runs of homozygosity (ROH), which are identical haplotypes inherited from each parent and thus are a measure of inbreeding[39]. The total length of ROH was correlated with the number of tracks in an individual (Figure 5A). Colombian *Lu. longipalpis* had more runs of homozygosity (ROHs) than any of the Brazilian populations (Figure 5B, C, Figure S9). The pattern is not driven by a single individual and is a property of the whole population (Figure S10) and extends beyond the mean level of linkage disequilibrium (Figure S11). Collectively, these results indicate that the genetic diversity, both at the level of nucleotide and haplotype, of Colombian *Lu. longipalpis* is lower than that of Brazilian populations. Based on this, we hypothesized that the demographic trajectory of *Lu. longipalpis* in Colombia was different from that of the populations in Brazil and thus we examined the trajectory of the effective population size over time of the different lineages of *Lu. longipalpis*. We found that all Brazilian populations experienced an effective population size increase on average around 122,410 generations ago (Figure 6). Captive sandflies have a generation of 6-7 weeks[40], which would suggest that the population increase occurred between 14,086 and 16,433 years ago (for 6 week or 7 week generations, respectively). While we found the same trend in all Brazilian populations, the historic effective population size of the Colombian population shows no evidence of dramatic change. As we hypothesized, the Colombian and Brazilian populations of *Lu. longipalpis* have undergone different demographic trajectories which may help explain their current levels of genetic diversity.

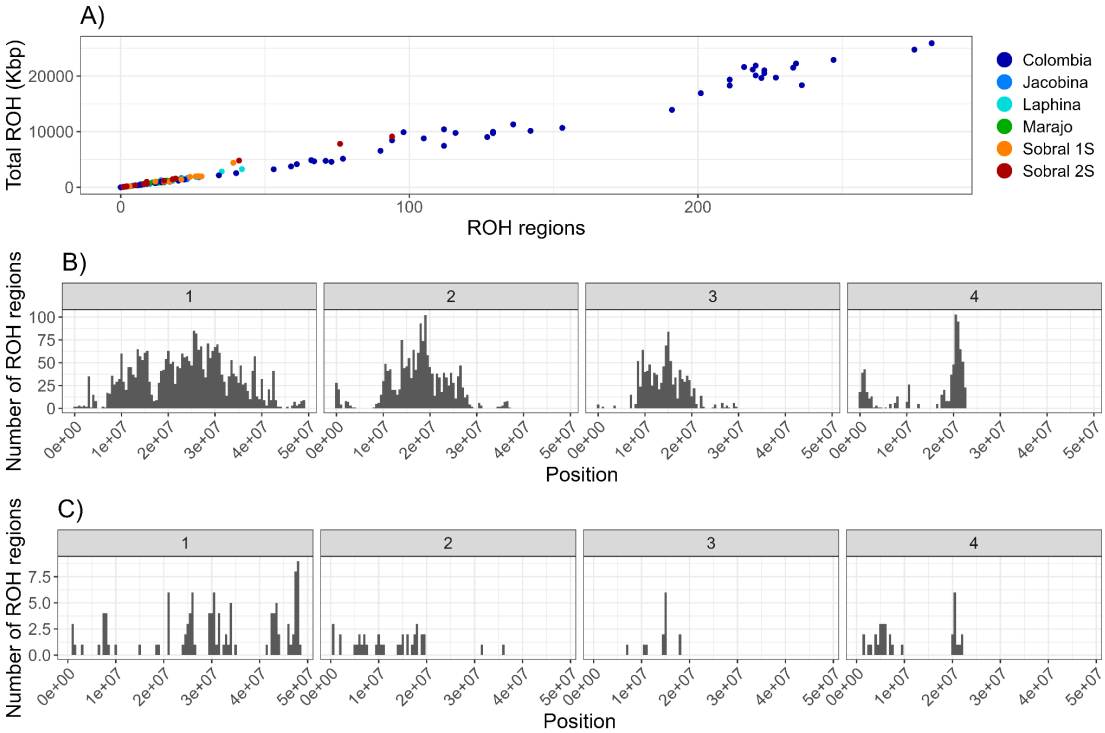


Figure 5. Runs of homozygosity in *Lu. longipalpis*. **A**. Plot depicting the total number of ROH regions over 50kb size and their total length by populations. **B**. Size distribution across genomic regions for Colombian sandflies (binwidth = 500000). **C**. Size distribution across genomic regions for sandflies from Marajo, Brazil (binwidth = 500000). Figure S9 shows the ROH distribution per chromosome for other populations. Figure S10 shows ROH histograms by individual for each population.

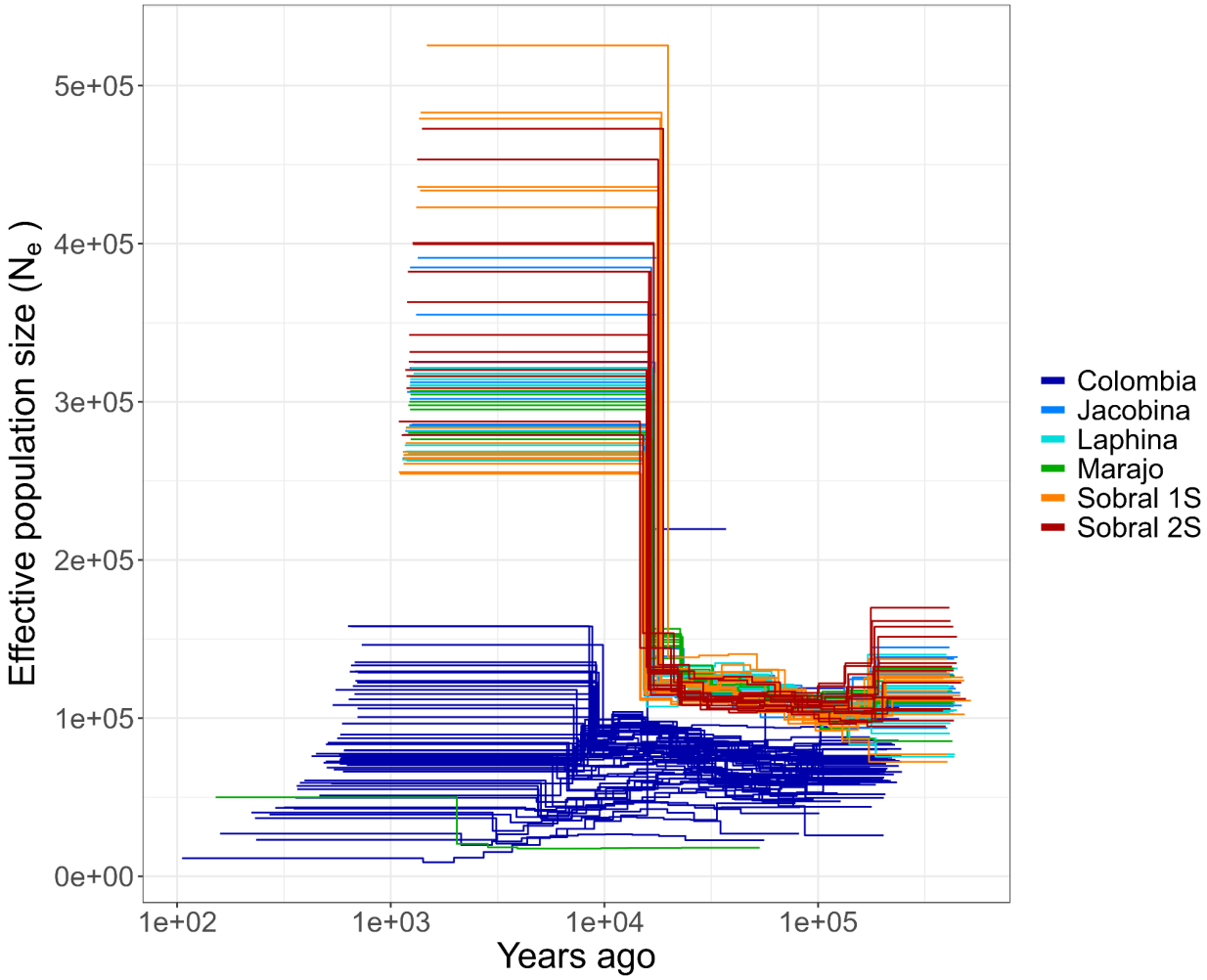


Figure 6. Demographic history of sandflies indicating temporal patterns of predicted effective population size. We assumed a generation time of six weeks.

DISCUSSION

Phlebotomine sandflies are disease vectors for several human and animal pathogens, and as a result, pose a serious human health threat[41–43]. One of the more notable diseases they transmit is the neglected tropical disease leishmaniasis, and in the Neotropics, one of the most common and important disease vectors of leishmaniasis is *Lutzomyia longipalpis*[44,45]. This sandfly species is widespread across the South American continent[2,14], poorly characterized[15,16], and is thought to be increasingly associated with anthropogenic and urbanized habitats[27]. Despite the clear human health risk posed by these vector species, little is known about their basic biology, hampering control and mitigation efforts[46]. Although there are some genomic resources[31,47], these are limited to Brazilian populations which represent geographically distant populations within a broader cryptic species complex of *Lu. longipalpis*. Here, we present highly continuous genome assemblies for both female and male sandflies from Colombia, which will be a critical resource for future work. We identified signatures of differentiation between the sexes and high divergence across the continent. Further, we discovered that the demographic histories of the different lineages of *Lu. longipalpis sensu lato* are complex, and that genetic

389 information points to the presence of multiple syntopic lineages. We discuss each of these
390 findings in the next paragraphs.
391 Our first finding is that females and males differ in their genomic composition. Although
392 recent genomic work has indicated no heteromorphy in the Psychodidae[48], prior work identified
393 the presence of heteromorphic chromosomes in sandflies[49–51]. Here, we sequenced individuals
394 from both sexes and produced highly continuous assemblies of both sexes. We found that the
395 longest chromosome in *Lu. longipalpis* is heteromorphic and ~1.3 Mbp longer in male sandflies
396 than in female sandflies. Our study builds on these previous observations and reveals that there
397 are differences in TE content and overall genomic composition between the *X* and *Y*
398 chromosomes. Further, our analysis of repeats within the genome indicates that this difference
399 in chromosomal length is likely driven by repeat content in male sandflies, which have an extra
400 ~1.3 Mbp of repeats in this chromosome. Our results have implications for the study of sex
401 chromosome evolution in dipterans. Further, because only females are hematophagous, and
402 leishmaniasis is known to have negative impacts on longevity and fecundity of sandflies[52,53],
403 leishmaniasis is likely to be a strong selective force that may impart different selective regimes
404 on female and male genomes, which may be evident either between the sexes or potentially in
405 heteromorphic sex chromosomes.
406 We identified a large, highly differentiated region between the sexes in a heteromorphic
407 sex chromosome. This region shows a substantial dip in nucleotide diversity in females relative
408 to males and encompasses a suite of genes related to sex determination, including the
409 commonly characterized dipteran sex-determining genes *sex-lethal, fruitless* (*fru*)*, doublesex-*
410 *and mab-3-related transcription factor A2* (*dsx*)*, lethal(2)*[54]. Some of these genes, *fru* and *dsx*,
411 have been implicated in sex determination in other phlebotomine species[55]. There are also a
412 variety of testes-specific genes (e.g., *ejaculatory bulb-specific protein 3*, *ejaculatory bulb-*
413 *specific protein 3-like*, *testis-expressed protein 10 homolog*, *motile sperm domain-containing*
414 *protein*). One gene in this region, *takeout,* shows a high degree of structural conservation and
415 has a conserved role in courtship across insects as diverse as fruit flies, mosquitos, and
416 moths[56]. Other genes in this region of differentiation may be related to basic ecology and life
417 history such as finding mates, finding prey, and effective hematophagy. Among these are
418 several odorant receptor and gustatory receptor genes. Genes in these families play a key role
419 in finding mates and lekking behavior[57], finding food or blood meals[58–60], and oviposition site
420 choice[61,62]. In other hematophagous lineages, odorant receptor genes are heavily regulated and
421 exhibit sex-specific and satiation-specific signatures of expression[63]. In *Lu. longipalpis,*
422 pheromones play a key role in attracting mates[18,64–66]; these show population-specific
423 signatures[18] and are perceived by odorant receptor genes which also show population specific
424 variation [57]. These may play a role in the cryptic *Lu. longipalpis* species complex via mate
425 choice, and may be a source of prezygotic isolation. Although these roles of odorant receptor
426 genes are well-known, these genes often have disparate functions among dipterans[67]. In sum,
427 there are a suite of genes in the genomic region that are highly differentiated between the sexes
428 that are related to ecology including finding food, finding mates, and oviposition that are
429 potentially playing important, sex-specific roles.
430 A complementary result, also related to sex differentiation in *Lutzomyia*, is the finding of
431 autosomal alleles associated with sex. In addition to these regions in chromosome 1, our
432 genome wide association study between female and male sandflies in Colombia yielded

433 significant peaks that were intragenic in three predicted genes, one per chromosome on
434 chromosomes 2-4. We were unable to identify the gene product on chromosome 2 via Uniprot,
435 OrthoDB, VectorBase, or NCBI blast. The predicted gene on chromosome 3 has also been
436 predicted in a previous assembly of *Lu. longipalpis* [31], but remains an uncharacterized protein
437 product. The intragenic SNPs our GWAS identified as associated with sex differences on
438 chromosome 4 is Apoptotic protease-activating factor 1 (*apaf-1*). The *C. elegans* homolog of
439 this gene (*CED-4*) interacts with the *C. elegans* gene *fem-1* which is crucial for male
440 development[34] and *apaf-1* deficient mice exhibit male infertility due to degeneration of
441 spermatogonia[35]. The *Drosophila* homolog of *apaf-1* is the gene Apaf-1–related killer (*ark,* or
442 sometimes *dark* for *Drosophila* Apaf-1–related killer), which is required for sperm formation[68].
443 The SNPs that we identified were present in intronic regions between predicted exons 3 and 4
444 in our gene annotation. However, the publicly available RNAseq data that we used for
445 annotation only contained female sandflies, potentially missing key, sex-specific isoforms of this
446 gene. We postulate that genomic differentiation between the sexes in this gene is related to sex-
447 specific differences in gametogenesis, or sex conflict (reviewed in [69]).
448 Perhaps most notably, our study provides strong evidence of divergence and population
449 structure across the continent. Our analyses combine novel population genomics data for
450 Colombian *Lu. longipalpis* populations with publicly-available whole genome sequence data,
451 which allowed us to carry out the first population genetics study of this critical vector species
452 across South America. Overall divergence between Brazilian and Colombian sandflies was
453 high, supportive of previous work suggesting that *Lu. longipalpis* is a cryptic species
454 complex[15,16,36,37]. In addition, we found evidence that Colombian *Lu. longipalpis* shows lower
455 genetic variability than all of the Brazilian populations. This result manifested in two different
456 facets, lower nucleotide diversity and a substantial increase in ROH tracts. ROH tracts were
457 more common in the *X*-chromosome in all populations but were particularly noticeable in
458 Colombia. It is unclear why Colombian flies exhibit such dramatic patterns of low diversity,
459 particularly because it does not appear to be an isolated population. A potential explanation is
460 reduced vagility in the Colombian lineage. For example if Colombian sandflies roam less during
461 their lifetime than Brazilian flies, they are more likely to experience matings with close relatives
462 and thus show more signs of inbreeding. An additional possibility is that Colombian *Lu.*
463 *longipalpis* exhibit facultative parthenogenesis, a trait seen in several dipteran species[70],
464 including in *Lu. mamedei*[71]. Both of these possibilities remain highly speculative and a
465 systematic survey of the reproductive biology of the different lineages of *Lu. longipalpis* is
466 needed to understand the patterns we observed. Our demographic analyses concur with the
467 inference of low genetic diversity in Colombia and indicate that all populations had a similar *Ne*
468 between 16,000 and 14,000 years ago depending on the generation time in the wild; Brazilian
469 populations underwent a large population expansion, Colombian *Lu. longipalpis* did not. This
470 timing is similar to the arrival and expansion of humans in South America[72]. Given that *Lu.*
471 *longipalpis* is one of the most anthropophilic species, population expansion in Brazilian sandflies
472 may have occurred as a result of human colonization and expansion.
473
474 Colombian sandflies (from our single sampling locality) showed structure only between
475 female and male sandflies, whereas multiple Brazilian populations (Jacobina, Sobral) were
476 highly structured in all our analyses. Together, these results indicate the presence of multiple

477 syntopic lineages—a result consistent with the previously hypothesized presence of cryptic
478 species within *Lu. longipalpis*, *sensu lato. Lutzomyia longipalpis* has been proposed to
479 encompass different chemotypes[64], and in some instances the proposals have been more
480 ambitious and have proposed the existence of a species complex which encompasses different
481 species with stark differences in mating behavior (e.g., mating song, pheromones and lekking
482 behavior). Notably, and despite the differentiation we describe for these lineages, we find some
483 evidence of admixture between lineages. Hybridization seems to be common in *Lutzomyia*
484 (reviewed in[73]), which poses the question of how frequently differentiating lineages exchange
485 alleles. Here we provide evidence for the existence of genome-wide admixture within *Lu.*
486 *longipalpis*. While this introgression does not appear sufficient to erode the genetic
487 differentiation between lineages, even when they are syntopic, this allele sharing might serve as
488 fodder for adaptation[74,75]. The importance of introgression has been previously hypothesized
489 across the *Lutzomyia* genus[76–78], and among *Lu. longipalpis* lineages[79,80]; our results represent
490 a formal genome-wide test of the occurrence of this phenomenon in this genus of disease
491 vectors.
492 Our manuscript leverages the power of whole genome sequencing to understand the
493 evolutionary processes that govern genetic variation in the critical leishmaniasis vector *Lu.*
494 *longipalpis*. Our results incorporate a genomic perspective into the study of differentiation within
495 *Lu. longipalpis* and represent a resource to not only understand sex chromosome evolution in
496 disease vectors, but also open the door to understanding the molecular and evolutionary
497 mechanisms of the interactions between the vectors, the *Leishmania* parasites, and the hosts
498 affected by the disease. Now that we have a better snapshot of the extent of genetic variation in
499 the vector, the next frontier is understanding whether the evolutionary interactions between
500 sandfly vectors and the *Leishmania* pathogen can provide the key to mitigating the spread of
501 leishmaniasis and the human health threat it poses.
502
503 **MATERIALS AND METHODS**
504 *Permits and ethics.* All sandflies collections were performed following the guidelines of the
505 Colombian decree No 001566 and 002662. Insects were collected under the permits 001566
506 and 002662 of the Autoridad de Licencias Ambientales. Sandflies were collected on private
507 property and permission was received from landowners before sampling.
508
509 *Sample collection.* Samples were collected from a semi-forested location outside of Ricaurte,
510 Cundinamarca, Colombia in November 2021 and October 2022**.** Sandflies were collected via
511 mechanical aspiration with a mouth aspirator and placed into vials to temporarily hold the
512 insects until processing. Prior to processing, vials containing the sandflies were put in a cooler
513 with ice to immobilize the specimens. Vials were emptied into a Petri dish where the sandflies
514 were selected and classified by species under a Leica dissecting scope. To identify sandfly
515 species, we used taxonomic keys[81,82]. Individual sandflies were removed with tweezers and
516 placed into individual 1.5mL tubes either in liquid nitrogen for genome assembly in 2022 or in
517 RNA-stabilizing buffer (RNALater™; Ambion, USA) for whole genome resequencing in 2021.
518 Samples were exported to North Carolina, USA for sequencing.
519

520 *Genome sequencing.* Due to their small size (thorax length in females: mean = 0.052 ± 0.0025
521 mm; males: mean = 0.0483 ± 0.0027 mm;[83] individuals were pooled prior to genomic DNA
522 extraction for whole genome sequencing (females = 9, males = 10). We used two approaches to
523 obtain reads for the genome assemblies. First, we generated PacBio HiFi reads. We
524 homogenized sandflies with a pestle and incubated them at 55℃ for 30 minutes prior to
525 performing a standard phenol chloroform extraction. Extracted high molecular weight gDNA was
526 sent to the Genomics Core Facility at the Icahn School of Medicine at Mount Sinai, where
527 libraries were prepared using PacBio's ultra-low DNA input HiFi library prep which utilizes PCR
528 amplification of genomic material. Libraries were pooled and sequenced across two SMRT cells
529 on a Pacific Biosciences Sequel IIe to a coverage of 71.8x and 75.7x coverage for females and
530 males.
531 Second, we used chromatin configuration. We sent 25 female and 25 male *Lu.*
532 *longipalpis* in ethanol to Arima Genomics for DNA extraction and sequencing. Hi-C data was
533 generated using the Arima HiC kit and the Arima Library Prep Module according to the Arima
534 Genomics manufacturer's protocols. Briefly, Arima High Coverage Hi-C is designed to capture
535 the sequence and 3D conformation of genomes. Chromatin from tissues, cell lines, or blood is
536 first crosslinked to preserve genomic structure, then digested with a restriction enzyme cocktail
537 optimized for uniform coverage. The digested ends are labeled with biotinylated nucleotides
538 during a fill-in step and subsequently ligated to capture spatially proximal DNA fragments.
539 These proximally-ligated DNA fragments are purified, fragmented, and enriched for biotin-
540 labeled sequences.The female sample was sequenced to 189.5 million read pairs and the male
541 sample to 180 million read pairs on an Illumina NovaSeq on 'paired-end' mode.
542
543 *Genome assemblies.* We produced separate genome assemblies for the female and male
544 sandflies. While we initially attempted to use HiFiasm[84], we discovered that haplotype diversity
545 in our pooled data exceeded the allowed diversity in the assembly algorithm. Therefore, we
546 chose to use wtdbg2[85], an assembler with an algorithm that allows for greater diversity and
547 collapses haplotypes into a single, haploid assembly. We used the "wtpoa-cns" command in
548 wtdbg2 to polish the assembly.
549 We then scaffolded our initial draft genome assemblies with our Hi-C data. Splits were
550 calculated using Juicer version 1.6[86] and reads were mapped to the draft assemblies with BWA
551 version 0.7.17[87]. We then ran Samtools version 1.17 to sort the aligned reads and Picard's
552 "MarkDuplicates" version 2.27.5 to mark duplicates[88]. We indexed our initial contig assembly
553 using Samtools "faidx" prior to running the YaHS (Yet Another Hi-C Scaffolding tool) pipeline[89]
554 with a MAPQ score of 30. We used Juicer "pre" to create a HiC contact map file which we
555 imported into Juicebox Assembly Tools[90] to manually curate each assembly. Curation included
556 fixing one major translocation (female assembly) and splitting a chromosome that had an
557 incorrect telomere-to-telomere join (male assembly). We then used Juicer "post" to update the
558 manually curated final assemblies for females and males.
559 After examination of whole genome sequence results, we identified systemic and
560 dramatic differentiation between female and male sequences along the majority of chromosome
561 1. We were unable to initially phase a ZW or XY assembly using our assembly approaches,
562 likely due to the pooled nature of our data (i.e., because male sandflies were too small for
563 effective single-fly sequencing, and we were forced to pool multiple individuals). As a result, we

took a secondary approach to assemble what we predicted to be a Y chromosome in males using the software SRY[91] to identify reads that are likely to be from sex-limited chromosomes. We used female and male Illumina whole genome sequence reads to produce sex-specific kmers. We then mapped these kmers to the raw male-only PacBio HiFi reads. We then extracted those reads that were highlighted by presumptive sex-limited kmers from our Illumina data and used wtdbg2[85] to assemble a Y chromosome from those HiFi reads. We then used RagTag[92] to scaffold the male Y chromosome using the male X chromosome (chromosome 1) as our backbone. We aligned this to the male "X" chromosome using minimap2[93] assuming ~0.1% sequence divergence ("-cx asm5") and visualized the alignment using "*pafr*" in R [94].

After producing our scaffolded assemblies, we identified repeat elements in the assemblies. We used Repeat Modeler version 2.0.4[95] to model repeat elements within each assembly. Most of these repeat elements were unidentified by Repeat Modeler, so we ran Transposon Classifier RFS[96] to classify these repeat regions. We then created a dipteran and *Lu. longipalpis* specific database of repeat elements. We combined the curated DFAM database version 3.4[97] with the RepBase version 25.05 within RepeatMasker version 4.1.2-p1[98]. We used the function "famdb.py" to extract all dipteran sequences from these databases, then added our sex-specific sandfly repeat databases. We then identified and soft-masked repeats throughout the genome using RepeatMasker version 4.1.2-p1[98] and specified our curated vertebrate and *Lu. longipalpis* repeat library as the library of repeats.

We then annotated assemblies using Braker version 3.0.8[99–102]. We used a combination of protein evidence and RNA seq evidence to annotate putative protein coding regions of the genome. We used the OrthoDB v11 database[103] and the VectorBase 67 database[104,105] as input protein data for annotation. Protein sequences were aligned to the genomes using miniprot[106]. We used the SRA toolkit[107] to download publicly-available RNA seq data from *Lu. longipalpis* to annotate the genome. We downloaded samples with accession numbers SRR535756-SRR535778. RNA seq reads were mapped to our genomes using HISAT2[108] and sorted/manipulated using samtools[109]. We followed this with Braker3, utilizing the GeneMark-ETP pipeline[110] to predict protein coding genes from RNAseq evidence. This was then compared to the protein evidence, OrthoDB and VectorBase in this case, to find homologous proteins that were then mapped back to the genome with ProtHint[101]. Augustus[111,112] then predicted a second set of genes. We additionally used the diptera_odb10 database[113] in Augustus to predict potentially missing protein products. We then used TSEBRA[114] to merge gene predictions from Augustus and the GeneMark-ETP pipelines.

<u>*WGS resequencing.*</u> We resequenced a total of 63 *Lu. longipalpis* from Ricaurte, Cundinamarca, Colombia. Of these, 37 sandflies were female and 26 were males sample information in Table S1). Briefly, we conducted a homogenization and incubation similar to that of the WGS gDNA extraction. Following the incubation, we followed a modified version of the instructions on the Zymo Genomic DNA Clean and Concentrator kit, and quantified eluted DNA using a Qubit. Our full extraction protocol can be found in the supplemental materials. Library prep and sequencing were conducted at the University of North Carolina-Chapel Hill's High-Throughput Sequencing Facility. Libraries were prepped using a ThruPLEX method for low DNA concentrations (CHiP-seq). Following library preparation, pooled paired-end 150bp reads were sequenced on an Illumina NovaSeq 6000 SP.

608
609 *WGS mapping, variant calling, and diversity summary statistics.* Next we measure the extent of
610 genetic diversity and differentiation within *Lu. longipalpis*. We followed the GATK best practices
611 for variant calling[115]. We used BBMap version 38.87 (https://github.com/BioInfoTools/BBMap) to
612 trim adapters from whole genome resequencing data and then mapped reads to the female
613 reference genome using BWA version 0.7.17[87]. Read groups were added using GATK's
614 AddOrReplaceReadGroups. For samples sequenced across two lanes, aligned data were then
615 merged using the "merge" tool in samtools version 1.14[109]. Aligned data were cleaned with
616 Picard's CleanSam, sorted with SortSam, and duplicates marked with MarkDuplicates (Picard
617 version 2.23.4[88]. Genotypes were then called with GATK HaplotypeCaller, before joint
618 genotyping data with GenotypeGVCFs. All GATK functions were from GATK version 4.1.9.0.
619 We merged bam files for each sample that were sequenced on multiple lanes using
620 samtools version 1.14 merge function[109]. We generated a Beagle file of genotype probabilities
621 using *angsd* version: 0.921-11-g20b0655, and then calculated principal components using the
622 *angsd* script "pcangsd.py"[116]. We then extracted eigenvalues and calculated the percent of the
623 variance in our data explained by each PC axis, which we then plotted in R version 4.1.1 ("Kick
624 Things") using ggplot2[117]. In addition to running this PCA pipeline with whole genome data, we
625 subsetted the alignment bam file in order to separately examine differentiation between the
626 presumptive sex chromosome (chromosome 1) and the autosomes (chromosomes 2-4) using
627 the same approach.
628 We examined basic population statistics (nucleotide diversity, π, and population
629 divergence [F_{ST}, Dxy]) using pixy[118], which accounts for invariant sites. We calculated pairwise
630 differences in nucleotide diversity and Weir-Cockerham F_{ST} values[119]. We were particularly
631 interested in potential differentiation between female and male sandflies, given that only females
632 are hematophagous and thus only females are vectors of leishmaniasis. Per pixy's instructions,
633 we first created a vcf file of invariant sites using the "-all-sites" flag in GATK's GenotypeGVCFs.
634 We then filtered the corresponding gvcf file to remove sites with a high proportion of missing
635 data (--max-missing 0.9), low coverage (--min-meanDP 3), and high overall depth (--max-
636 meanDP 40) using vcftools 0.1.15[120]. We then used pixy to calculate F_{ST} and Dxy across
637 1,000bp windows between female and male sandflies. The window size was chosen five the
638 extent of LD in *Lu. longipalpis*.) . Additionally, we subset our gvcf file into a female-only and a
639 male-only gvcf which we used to calculate nucleotide diversity (π) across 1,000bp windows in
640 female and male sandflies. We used PLINK 1.90b6.21[121] to calculate linkage disequilibrium. We
641 additionally calculated Tajima's D in 1,000 bp windows using VCFtools.
642 We detected runs of homozygosity (ROH) using PLINK 1.90b6.21[121] with the
643 parameters: --allow-extra-chr and --homozyg-kb 5. This command estimates the run length and
644 identifies contiguous segments of homozygosity greater than 5 kbp by assessing the observed
645 genotype data against expected patterns of homozygosity. ROH was detected separately for
646 each of Colombia and the five Brazil populations. With a 5 kbp minimum size, we identified a
647 minimum of 100 SNPs per region and on average there were 189-276 SNPs per region. We
648 repeated the same analyses using different parameter values for the minimum length of ROH in
649 kbp (5, 10, 25, 50, 100), and results were similar despite fewer ROH regions with higher kb
650 thresholds.
651

652 *Phylogenetic trees.* We used trees to address two questions. First, to examine the genealogy of
653 the largest chromosome relative to the presumptive autosomes in placement of females and
654 males in relation to each other, and second, to examine continent-wide evolutionary
655 relationships. We used the python script "vcf2phylp.py"[122] to split our vcf files into 100,000 bp
656 regions and convert them into the phylip format.
657 To examine the genealogy of the largest chromosome relative to the presumptive
658 autosomes, we created two separate trees using 1) only the largest chromosome or 2) the
659 presumptive autosomes in just Colombian sandflies. For both trees, we made a phylogenetic
660 tree for each window using IQ-TREE[123] version 2.2.2.7 with DNA specified ("-st DNA") and 1000
661 bootstraps ("-bb 1000"). We merged the resulting consensus tree files into a single tree and
662 used ASTRAL[124] version 5.7.8 to create a single genome-wide consensus tree. We additionally
663 created trees solely from chromosome 1 and the autosomes.
664 To examine evolutionary relationships across the continent, we made a single genome-
665 wide tree using the above methods. The approach for this question is identical to the one
666 described immediately above, with the exception that we created a single tree including all
667 chromosomes. Further, we examined gene concordance between the genome wide tree and
668 conserved, core dipteran genes from the BUSCO diptera_odb10 database[113]. To do this, we
669 used ASTRAL[124] version 5.7.8 to create a single genome-wide consensus tree that is
670 constrained to the populations sampled. We additionally extracted all the complete, single copy
671 dipteran BUSCO genes assembled in our male assembly, using the python script
672 "vcf2phylp.py"[122] to convert the vcf of these genomic regions into the phylip format. We created
673 single gene trees as per the windows above using IQ-TREE, and created a single consensus
674 population-level tree using ASTRAL as we did for our whole genome tree. We then used IQ-
675 TREE[123] version 2.2.2.7 to calculate gene concordance between each gene tree and the overall
676 consensus tree. Finally, we examined gene flow and the potential for introgression using
677 fbranch statistics in Dsuite [125]. Briefly, we used our consensus population tree and our variant
678 call file as input to examine fbranch statistics using "Dtrios" to examine all trios and "Fbranch" to
679 examine the output fbranch statistics. Given the clear divergence between Brazilian and
680 Colombian sandflies, we used Colombia as the outgroup in our analysis and focused on
681 Brazilian sandflies.
682
683 *Population genomics of geographically disparate* Lu. longipalpis *populations. Lutzomyia*
684 *longipalpis* has a broad geographic range from Argentina through Mexico[2,14]. Further, multilocus
685 sequence typing and phenotypic surveys have revealed strong population structure which might
686 correspond to cryptic speciation[17,18] and divergences in courtship song[20], morphology,[21], and
687 pheromones[18]. Recent work sequenced Brazilian *Lu. longipalpis* from four localities in Brazil[31].
688 Here we examined basic population genomic metrics between Colombian and Brazilian *Lu.*
689 *longipalpis.* We were particularly interested in differentiation between populations of putative *Lu.*
690 *longipalpis,* and as a result focus on metrics such as F_{ST} and Dxy between populations.
691 Out of the 68 Colombian samples, 63 sandflies (37 females and 26 males) passed our
692 quality metrics for population level analyses (>50% of reads mapping to our *Lu. longipalpis*
693 female assembly and > 5X average coverage). On average, 91.4% of female reads and 84.4%
694 of male reads mapped to the female assembly. Average coverage was 16.8x and 13.2x for
695 females and males respectively. When mapped to the male assembly, 84.3% of female reads

696 and 79.6% of male reads mapped and we had 14.6x and 11.7x coverage for females and males
697 respectively.
698 Additionally, we downloaded whole genome resequence data from[31] using the SRA
699 toolkit[107]. Table S1 lists the accession numbers of the previously published reads used in this
700 study. We then followed the same protocol that we used for our Colombian samples, the GATK
701 best-practices for variant calling [115]. We used BBMap version 38.87
702 (https://github.com/BioInfoTools/BBMap) to trim adaptors from whole genome resequencing
703 data and then mapped reads to the female reference genome using BWA version 0.7.17[87].
704 Read groups were added using GATK's AddOrReplaceReadGroups. For samples sequenced
705 across two lanes, aligned data were then merged using the "merge" tool in samtools version
706 1.14[109]. Aligned data were cleaned with Picard's CleanSam, sorted with SortSam, and
707 duplicates marked with MarkDuplicates (Picard version 2.23.4[88]. Genotypes were then called
708 with GATK HaplotypeCaller, before joint genotyping data with GenotypeGVCFs. All GATK
709 functions were from GATK version 4.1.9.0.
710 For samples across the continent of South America, we examined summary population
711 statistics (π, F_{st}, and Dxy) as we did for samples within Colombia (see above). We filtered our
712 gvcf file to remove sites with a high proportion of missing data (--max-missing 0.9), low
713 coverage (--min-meanDP 3), and high overall depth (--max-meanDP 40) using vcftools
714 0.1.15[120]. We then used pixy to calculate F_{ST} and Dxy across 1,000bp windows between
715 populations. Additionally, we subset our gvcf file into a female-only and a male-only gvcf which
716 we used to calculate nucleotide diversity (π) across 1,000bp windows in female and male
717 sandflies. We additionally calculated Tajima's D in 1,000 bp windows using VCFtools.
718 In order to examine population structure across the continent, we used angsd[116] to
719 examine the data with a PCA, structure, and *ADMIXTURE* based approaches. We used a
720 similar approach to the within-Colombia examinations. We generated a Beagle file of genotype
721 probabilities using *angsd* version: 0.921-11-g20b0655, and then calculated principal
722 components and genotype probabilities using the *angsd* script "pcangsd.py"[116]. This program
723 calculates ancestry based on individual allele frequency and uses a similar approach to delta K
724 in that it permutes iteratively across a range of ancestry possibilities. We then extracted
725 eigenvalues and calculated the percent variance in our data explained and plotted both this
726 PCA and a STRUCTURE plot in R version 4.1.1 ("Kick Things") using ggplot2[117].
727 To quantify if sex was responsible for differences in genome size, we used a linear
728 model (function "lm" in R) with average π per window in the sexes as the outcome and sex,
729 chromosome, and the interaction of sex and chromosome as fixed effects.
730 We estimated demographic histories using MSMC2[126] with one thread and ten time
731 segments of different lengths using the parameters '-t 1 -p 1-6+8-1+1-4'. MSMC2 was run
732 separately for individual samples since we had unphased genotypes. The MSCM2 input files
733 were generated using msmc-tools[127]. Separate input files were generated for each of the four
734 major chromosomes using 'generate_multihetsep.py' and then concatenated together into
735 separate genomic input files for each individual sample. 'generate_multihetsep.py' requires two
736 input files: a mask of sites with coverage for that individual and a genomic mappability mask.
737 The covered site masks were created in three steps run separately for each chromosome and
738 each individual sample using Samtools 1.21[128], BCFtools[128] and msmc-tools. The genomic
739 mappability mask was created using genmap[129]. MSMC2 results were plotted using R [130] using

740 a mutation rate of μ=1.25e-8 and an estimated generation time of 49 days[40] with time in years
741 calculated as 'time_boundary / μ / gen'. The effective population size was calculated as (1 / λ) /
742 (2 * μ) where lambda is the coalescence rate returned by MSMC2.
743
744 *Genome wide association study for sex determination.* Finally, we ran a genome-wide
745 association study (GWAS) for 133 samples from both Colombia and Brazil using sex as the
746 phenotype (113 females, 20 males) to identify genetic regions associated with sex
747 determination. First, we created the required plink input files from our vcf by filtering for just the
748 four chromosomes and using a minor allele frequency of 0.05. To run the GWAS, we used plink
749 2.0[121] using a generalized linear model framework with sex as a binary variable[131], and we
750 filtered out SNPs with 20% or more missing data (--geno 0.2) and a minor allele frequency of
751 0.05 (--maf 0.05). Results are presented from the continent wide analysis, but results are the
752 same using just the Colombian samples yielded qualitatively identical results.
753
754 **Data availability:**
755 All raw sequence data as well as genome assemblies are available on NCBI under BioProject
756 PRJNA1110158. Code can be found at https://github.com/AdamStuckert/Lu_longipalpis; and
757 will be deposited in Zenodo upon acceptance.
758
759 **Author contributions:**
760 AMMS: Conceived project, collected samples, conducted wet lab work, analyzed data, wrote
761 manuscript
762 DT: Analyzed data, wrote manuscript
763 ERRD: Conducted wet lab work, wrote manuscript
764 DP: Conducted wet lab work, wrote manuscript
765 MP: Conducted wet lab work, wrote manuscript
766 CM: Conceived project
767 JAR: Wrote manuscript
768 EB: Conceived project, collected samples
769 JRW: Conducted wet lab work, wrote manuscript
770 RV: Conceived project, collected samples, identified sandflies, wrote manuscript
771 DRM: Conceived project, collected samples, analyzed data, wrote manuscript
772
773 **Competing interests:**
774 The authors declare they have no competing interests.
775
776 **Acknowledgments:**
777 The research presented here was supported by the National Institute of General Medical
778 Sciences under award R35GM148244.

References:

1. Maroli, M., Feliciangeli, M. D., Bichaud, L., Charrel, R. N. & Gradoni, L. Phlebotomine sandflies and the spreading of leishmaniases and other diseases of public health concern. *Med. Vet. Entomol.* **27**, 123–147 (2013).
2. Young, D. G. & Duncan, M. M. Guide to the identification and geographic distribution of *Lutzomyia* sand flies in Mexico, the West Indies, Central and South America (Diptera: Psychodidae). *Memoirs of the American Entomological Institute* **54**, 1–887 (1994).
3. World Health Organization. Control of the leishmaniases. *World Health Organ. Tech. Rep. Ser.* xii–xiii, 1–186, back cover (2010).
4. Alvar, J. *et al.* Leishmaniasis Worldwide and Global Estimates of Its Incidence. *PLoS One* **7**, e35671 (2012).
5. Karimkhani, C. *et al.* Global burden of cutaneous leishmaniasis: a cross-sectional analysis from the Global Burden of Disease Study 2013. *Lancet Infect. Dis.* **16**, 584–591 (2016).
6. Bern, C., Maguire, J. H. & Alvar, J. Complexities of assessing the disease burden attributable to leishmaniasis. *PLoS Negl. Trop. Dis.* **2**, e313 (2008).
7. McIlwee, B. E., Weis, S. E. & Hosler, G. A. Incidence of Endemic Human Cutaneous Leishmaniasis in the United States. *JAMA Dermatol.* **154**, 1032–1039 (2018).
8. Curtin, J. M. & Aronson, N. E. Leishmaniasis in the United States: Emerging issues in a region of low endemicity. *Microorganisms* **9**, 578 (2021).
9. Gin, T. E., Lashnits, E., Wilson, J. M., Breitschwerdt, E. B. & Qurollo, B. Demographics and travel history of imported and autochthonous cases of leishmaniosis in dogs in the United States and Canada, 2006 to 2019. *J. Vet. Intern. Med.* **35**, 954–964 (2021).
10. González, C. *et al.* Climate change and risk of leishmaniasis in North America: predictions from ecological niche models of vector and reservoir species. *PLoS Negl. Trop. Dis.* **4**,

805 e585 (2010).
806 11. Clarke, C. F., Bradley, K. K., Wright, J. H. & Glowicz, J. Case report: Emergence of
807 autochthonous cutaneous leishmaniasis in northeastern Texas and southeastern
808 Oklahoma. *Am. J. Trop. Med. Hyg.* **88**, 157–161 (2013).
809 12. Montenegro Quiñonez, C. A., Runge-Ranzinger, S., Rahman, K. M. & Horstick, O.
810 Effectiveness of vector control methods for the control of cutaneous and visceral
811 leishmaniasis: A meta-review. *PLoS Negl. Trop. Dis.* **15**, e0009309 (2021).
812 13. Wilson, A. L. *et al.* The importance of vector control for the control and elimination of vector-
813 borne diseases. *PLoS Negl. Trop. Dis.* **14**, e0007831 (2020).
814 14. Soares, R. P. P. & Turco, S. J. *Lutzomyia longipalpis* (Diptera: Psychodidae:
815 Phlebotominae): a review. *An. Acad. Bras. Cienc.* **75**, 301–330 (2003).
816 15. Bauzer, L. G. S. R., Souza, N. A., Maingon, R. D. C. & Peixoto, A. A. *Lutzomyia longipalpis*
817 in Brazil: a complex or a single species? A mini-review. *Mem. Inst. Oswaldo Cruz* **102**, 1–
818 12 (2007).
819 16. Maingon, R. D. C., Ward, R. D., Hamilton, J. G. C., Bauzer, L. G. S. R. & Peixoto, A. A. The
820 *Lutzomyia longipalpis* species complex: does population sub-structure matter to
821 Leishmania transmission? *Trends Parasitol.* **24**, 12–17 (2008).
822 17. Soto, S. I., Lehmann, T., Rowton, E. D., Vélez B, I. D. & Porter, C. H. Speciation and
823 population structure in the morphospecies *Lutzomyia longipalpis* (Lutz & Neiva) as derived
824 from the mitochondrial *ND4* gene. *Mol. Phylogenet. Evol.* **18**, 84–93 (2001).
825 18. Maingon, R. D. C. *et al.* Genetic identification of two sibling species of *Lutzomyia*
826 *longipalpis* (Diptera: Psychodidae) that produce distinct male sex pheromones in Sobral,
827 Ceará State, Brazil. *Mol. Ecol.* **12**, 1879–1894 (2003).
828 19. Gutierrez, M. A. C. *et al.* DNA barcoding of *Lutzomyia longipalpis* species complex (Diptera:
829 Psychodidae), suggests the existence of 8 candidate species. *Acta Trop.* **221**, 105983
830 (2021).

20. Souza, N. A. *et al.* Analysis of the copulatory courtship songs of *Lutzomyia longipalpis* in six populations from Brazil. *J. Med. Entomol.* **41**, 906–913 (2004).
21. Ward RD, Ribeiro AL, Ready PD, Murtagh A. Reproductive isolation between different forms of *Lutzomyia longipalpis* (Lutz & Neiva),(Diptera: Psychodidae), the vector of *Leishmania donovani chagasi* Cunha & Chagas and its significance to kala-azar distribution in South America. *Memórias do Instituto Oswaldo Cruz* **78**, 269–280 (1983).
22. Mangabeira, O. Sobre a sistemática e biologia dos *Phlebotomus* do Ceará. *Rev. Bras. Malariol. Doencas Trop.* **21**, 3–26 (1969).
23. Lainson, R. & Rangel, E. F. *Lutzomyia longipalpis* and the eco-epidemiology of American visceral leishmaniasis, with particular reference to Brazil: a review. *Mem. Inst. Oswaldo Cruz* **100**, 811–827 (2005).
24. Ferro, C., Pardo, R., Torres, M. & Morrison, A. C. Larval microhabitats of *Lutzomyia longipalpis* (Diptera: Psychodidae) in an endemic focus of visceral leishmaniasis in Colombia. *J. Med. Entomol.* **34**, 719–728 (1997).
25. Ferro, C. *et al.* Spatial Distribution of Sand Fly Vectors and Eco-Epidemiology of Cutaneous Leishmaniasis Transmission in Colombia. *PLoS One* **10**, e0139391 (2015).
26. Costa, P. L. *et al.* Ecology of Lutzomyia longipalpis in an area of visceral leishmaniasis transmission in north-eastern Brazil. *Acta Trop.* **126**, 99–102 (2013).
27. Salomón, O. D., Feliciangeli, M. D., Quintana, M. G., Afonso, M. M. dos S. & Rangel, E. F. *Lutzomyia longipalpis* urbanisation and control. *Mem. Inst. Oswaldo Cruz* **110**, 831–846 (2015).
28. Berrozpe, P. E. *et al.* Spatiotemporal dynamics of *Lutzomyia longipalpis* and macro-habitat characterization using satellite images in a leishmaniasis-endemic city in Argentina. *Med. Vet. Entomol.* **33**, 89–98 (2019).
29. Morrison, A. C., Ferro, C., Morales, A., Tesh, R. B. & Wilson, M. L. Dispersal of the sand fly Lutzomyia longipalpis (Diptera: Psychodidae) at an endemic focus of visceral leishmaniasis

857 in Colombia. *J. Med. Entomol.* **30**, 427–435 (1993).
858 30. Afonso, M. M. D. S., Duarte, R., Miranda, J. C., Caranha, L. & Rangel, E. F. Studies on the
859 Feeding Habits of *Lutzomyia (Lutzomyia) longipalpis* (Lutz & Neiva, 1912) (Diptera:
860 Psychodidae: Phlebotominae) Populations from Endemic Areas of American Visceral
861 Leishmaniasis in Northeastern Brazil. *J. Trop. Med.* **2012**, 858657 (2012).
862 31. Labbé, F. *et al.* Genomic analysis of two phlebotomine sand fly vectors of *Leishmania* from
863 the New and Old World. *PLoS Negl. Trop. Dis.* **17**, e0010862 (2023).
864 32. Srikulnath, K., Ahmad, S. F., Singchat, W. & Panthum, T. Do Ty3/Gypsy transposable
865 elements play preferential roles in sex chromosome differentiation? *Life (Basel)* **12**, 522
866 (2022).
867 33. Zhou, Q. *et al.* The epigenome of evolving *Drosophila* neo-sex chromosomes: dosage
868 compensation and heterochromatin formation. *PLoS Biol.* **11**, e1001711 (2013).
869 34. Chan, S. L., Yee, K. S., Tan, K. M. & Yu, V. C. The *Caenorhabditis elegans* sex
870 determination protein FEM-1 is a CED-3 substrate that associates with CED-4 and
871 mediates apoptosis in mammalian cells. *J. Biol. Chem.* **275**, 17925–17928 (2000).
872 35. Honarpour, N. *et al.* Adult Apaf-1-deficient mice exhibit male infertility. *Dev. Biol.* **218**, 248–
873 258 (2000).
874 36. Uribe, S. The status of the *Lutzomyia longipalpis* species complex and possible implications
875 for Leishmania transmission. *Mem. Inst. Oswaldo Cruz* **94**, 729–734 (1999).
876 37. Souza, N. A., Brazil, R. P. & Araki, A. S. The current status of the *Lutzomyia longipalpis*
877 (Diptera: Psychodidae: Phlebotominae) species complex. *Mem. Inst. Oswaldo Cruz* **112**,
878 161–174 (2017).
879 38. Morales-Briones, D. F. *et al.* Disentangling sources of gene tree discordance in
880 phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae s.L. *Syst. Biol.*
881 **70**, 219–235 (2021).
882 39. Ceballos, F. C., Hazelhurst, S. & Ramsay, M. Assessing runs of Homozygosity: a

883 comparison of SNP Array and whole genome sequence low coverage data. *BMC*
884 *Genomics* **19**, 106 (2018).
885 40. Killick-Kendrick, R., Leaney, A. J. & Ready, P. D. The establishment, maintenance and
886 productivity of a laboratory colony of *Lutzomyia longipalpis* (Diptera: Psychodidae). *J. Med.*
887 *Entomol.* **13**, 429–440 (1977).
888 41. Killick-Kendrick, R. Phlebotomine vectors of the leishmaniases: a review. *Med. Vet.*
889 *Entomol.* **4**, 1–24 (1990).
890 42. Killick-Kendrick, R. The biology and control of phlebotomine sand flies. *Clin. Dermatol.* **17**,
891 279–289 (1999).
892 43. Ready, P. D. Biology of phlebotomine sand flies as vectors of disease agents. *Annu. Rev.*
893 *Entomol.* **58**, 227–250 (2013).
894 44. Salomon, O. D. *Lutzomyia longipalpis*, Gone with the Wind and Other Variables. *Neotrop.*
895 *Entomol.* **50**, 161–171 (2021).
896 45. Sousa-Paula, L. C. de, Otranto, D. & Dantas-Torres, F. *Lutzomyia longipalpis* (Sand Fly).
897 *Trends Parasitol.* **36**, 796–797 (2020).
898 46. Rêgo, F. & Soares, R. Lutzomyia longipalpis: an update on this sand fly vector. *An. Acad.*
899 *Bras. Cienc.* **93**, e20200254 (2021).
900 47. Huang, M. *et al.* Improved high quality sand fly assemblies enabled by ultra low input long
901 read sequencing. *Sci. Data* **11**, 918 (2024).
902 48. Vicoso, B. & Bachtrog, D. Numerous transitions of sex chromosomes in Diptera. *PLoS Biol.*
903 **13**, e1002078 (2015).
904 49. Kreutzer, R. D., Modi, G. B., Tesh, R. B. & Young, D. G. Brain cell karyotypes of six species
905 of New and Old World sand flies (Diptera: Psychodidae). *J. Med. Entomol.* **24**, 609–612
906 (1987).
907 50. Jiménez, M. E., Bello, F. J., Ferro, C. & Cárdenas, E. Brain cell karyotype of the
908 phlebotomine sand fly *Lutzomyia shannoni* (Dyar) (Diptera: Psychodidae). *Mem. Inst.*

909 *Oswaldo Cruz* **96**, 379–380 (2001).
910 51. Vigoder, F. M., Araripe, L. O. & Carvalho, A. B. Identification of the sex chromosome
911 system in a sand fly species, *Lutzomyia longipalpis* s.l. *G3* **11**, (2021).
912 52. El Sawaf, B. M., El Sattar, S. A., Shehata, M. G., Lane, R. P. & Morsy, T. A. Reduced
913 longevity and fecundity in Leishmania-infected sand flies. *Am. J. Trop. Med. Hyg.* **51**, 767–
914 770 (1994).
915 53. Agrela, I. F. & Feliciangeli, M. D. Effect of *Leishmania* spp infection on the survival, life
916 expectancy, fecundity and fertility of Lutzomyia longipalpiss.l. and Lutzomyia
917 pseudolongipalpis. *Mem. Inst. Oswaldo Cruz* **110**, 611–617 (2015).
918 54. Ashmore, J. S., Slippers, B., Duong, T. A. & Dittrich-Schröder, G. Understanding the
919 genetics of sex determination in insects and its relevance to genetic pest management.
920 *Insect Mol. Biol.* (2024) doi:10.1111/imb.12982.
921 55. Petrella, V. *et al.* Identification of sex determination genes and their evolution in
922 Phlebotominae sand flies (Diptera, Nematocera). *BMC Genomics* **20**, 522 (2019).
923 56. Saurabh, S., Vanaphan, N., Wen, W. & Dauwalder, B. High functional conservation of
924 takeout family members in a courtship model system. *PLoS One* **13**, e0204615 (2018).
925 57. Hickner, P. V. *et al.* Molecular signatures of sexual communication in the phlebotomine
926 sand flies. *PLoS Negl. Trop. Dis.* **14**, e0008967 (2020).
927 58. Takken, W. & Knols, B. G. Odor-mediated behavior of Afrotropical malaria mosquitoes.
928 *Annu. Rev. Entomol.* **44**, 131–157 (1999).
929 59. Foster, W. A. Mosquito sugar feeding and reproductive energetics. *Annu. Rev. Entomol.* **40**,
930 443–474 (1995).
931 60. Valenzuela, J. G., Francischetti, I. M. B., Pham, V. M., Garfield, M. K. & Ribeiro, J. M. C.
932 Exploring the salivary gland transcriptome and proteome of the *Anopheles stephensi*
933 mosquito. *Insect Biochem. Mol. Biol.* **33**, 717–732 (2003).
934 61. Dougherty, M. & Hamilton, G. Dodecanoic Acid Is the Oviposition Pheromone of *Lutzomyia*

935 *longipalpis*. *J. Chem. Ecol.* **23**, 2657–2671 (1997).
936 62. Dougherty, M. J., Guerin, P. M. & Ward, R. D. Identification of oviposition attractants for the
937 sandfly *Lutzomyia longipalpis* (Diptera: Psychodidae) in volatiles of faeces from
938 vertebrates. *Physiol. Entomol.* **20**, 23–32 (1995).
939 63. Ni, M. *et al.* Screening for odorant receptor genes expressed in *Aedes aegypti* involved in
940 host-seeking, blood-feeding and oviposition behaviors. *Parasit. Vectors* **15**, 71 (2022).
941 64. Spiegel, C. N. *et al.* The *Lutzomyia longipalpis* complex: a brief natural history of
942 aggregation-sex pheromone communication. *Parasit. Vectors* **9**, 580 (2016).
943 65. Hamilton, J. G. C., Maingon, R. D. C., Alexander, B., Ward, R. D. & Brazil, R. P. Analysis of
944 the sex pheromone extract of individual male *Lutzomyia longipalpis* sandflies from six
945 regions in Brazil. *Med. Vet. Entomol.* **19**, 480–488 (2005).
946 66. Watts, P. C. *et al.* Male sex pheromones and the phylogeographic structure of the
947 *Lutzomyia longipalpis* species complex (Diptera: Psychodidae) from Brazil and Venezuela.
948 *Am. J. Trop. Med. Hyg.* **73**, 734–743 (2005).
949 67. Maguire, S. E., Afify, A., Goff, L. A. & Potter, C. J. Odorant-receptor-mediated regulation of
950 chemosensory gene expression in the malaria mosquito *Anopheles gambiae*. *Cell Rep.* **38**,
951 110494 (2022).
952 68. Huh, J. R. *et al.* Multiple apoptotic caspase cascades are required in nonapoptotic roles for
953 *Drosophila* spermatid individualization. *PLoS Biol.* **2**, E15 (2004).
954 69. Kasimatis, K. R., Nelson, T. C. & Phillips, P. C. Genomic signatures of sexual conflict. *J.*
955 *Hered.* **108**, 780–790 (2017).
956 70. Sperling, A. L. & Glover, D. M. Parthenogenesis in dipterans: a genetic perspective. *Proc.*
957 *Biol. Sci.* **290**, 20230261 (2023).
958 71. Brazil, R. P. & Oliveira, S. M. Parthenogenesis in the sandfly *Lutzomyia mamedei* (Diptera:
959 Psychodidae). *Med. Vet. Entomol.* **13**, 463–464 (1999).
960 72. Goldberg, A., Mychajliw, A. M. & Hadly, E. A. Post-invasion demography of prehistoric

961 humans in South America. *Nature* **532**, 232–235 (2016).
962 73. Dos Reis, Y. V. & Alevi, K. C. C. Hybridization in Phlebotominae (Diptera: Psychodidae): A
963 mini-review. *Infect. Genet. Evol.* **86**, 104593 (2020).
964 74. Hedrick, P. W. Adaptive introgression in animals: examples and comparison to new
965 mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* **22**, 4606–4618
966 (2013).
967 75. Arnold, M. L. & Martin, N. H. Adaptation by introgression. *J. Biol.* **8**, 82 (2009).
968 76. Mazzoni, C. J. *et al.* Multilocus analysis of introgression between two sand fly vectors of
969 leishmaniasis. *BMC Evol. Biol.* **8**, 141 (2008).
970 77. Mazzoni, C. J., Souza, N. A., Andrade-Coelho, C., Kyriacou, C. P. & Peixoto, A. A.
971 Molecular polymorphism, differentiation and introgression in the *period* gene between
972 *Lutzomyia intermedia* and *Lutzomyia whitmani*. *BMC Evol. Biol.* **6**, 85 (2006).
973 78. Testa, J. M., Montoya-Lerma, J., Cadena, H., Oviedo, M. & Ready, P. D. Molecular
974 identification of vectors of Leishmania in Colombia: mitochondrial introgression in the
975 *Lutzomyia townsendi* series. *Acta Trop.* **84**, 205–218 (2002).
976 79. Casaril, A. E. *et al.* Macrogeographic genetic structure of *Lutzomyia longipalpis* complex
977 populations using Next Generation Sequencing. *PLoS One* **14**, e0223277 (2019).
978 80. Araki, A. S. *et al.* Multilocus analysis of divergence and introgression in sympatric and
979 allopatric sibling species of the *Lutzomyia longipalpis* complex in Brazil. *PLoS Negl. Trop.*
980 *Dis.* **7**, e2495 (2013).
981 81. Galati, E. A. B. Phlebotominae (Diptera, Psychodidae): Classification, Morphology and
982 Terminology of Adults and Identification of American Taxa. in *Brazilian Sand Flies : Biology,*
983 *Taxonomy, Medical Importance and Control* (eds. Rangel, E. F. & Shaw, J. J.) 9–212
984 (Springer International Publishing, Cham, 2018).
985 82. Galati, E. A. B., Galvis-Ovallos, F., Lawyer, P., Léger, N. & Depaquit, J. An illustrated guide
986 for characters and terminology used in descriptions of Phlebotominae (Diptera,

987 Psychodidae). *Parasite* **24**, 26 (2017).
988 83. Vivero-Gomez, R. *et al.* Humidity and temperature preference in two Neotropical species of
989 sand flies. *Parasit. Vectors* **17**, 246 (2024).
990 84. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
991 assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
992 85. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* (2019)
993 doi:10.1038/s41592-019-0669-3.
994 86. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-
995 C Experiments. *Cell Syst* **3**, 95–98 (2016).
996 87. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
997 transform. *Bioinformatics* (2009).
998 88. Picard Toolkit. *Broad Institute, GitHub Repository* (2019).
999 89. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *bioRxiv*
1000 2022.06.09.495093 (2022) doi:10.1101/2022.06.09.495093.
1001 90. Dudchenko, O. *et al.* The Juicebox Assembly Tools module facilitates de novo assembly of
1002 mammalian genomes with chromosome-length scaffolds for under $1000.
1003 doi:10.1101/254797.
1004 91. Wang, X.-B., Liu, Q.-Y., Li, A.-L. & Ruan, J. SRY: An effective method for sorting long reads
1005 of sex-limited chromosome. *bioRxiv* 2020.05.25.115592 (2020)
1006 doi:10.1101/2020.05.25.115592.
1007 92. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato
1008 system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
1009 93. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
1010 3100 (2018).
1011 94. Winter, D. *Pafr: Read, Manipulate and Visualize 'Pairwise mApping Format' Data*. (2020).
1012 95. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable

element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).

96. Riehl, K., Riccio, C., Miska, E. A. & Hemberg, M. TransposonUltimate: software for transposon classification, annotation and detection. *bioRxiv* 2021.04.30.442214 (2021) doi:10.1101/2021.04.30.442214.

97. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).

98. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. (2013-2015).

99. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv* (2023) doi:10.1101/2023.06.10.544449.

100. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).

101. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, lqaa108 (2021).

102. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation with BRAKER. *Methods Mol. Biol.* **1962**, 65–95 (2019).

103. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).

104. Butler, R. The design and development of Vectorbase: A bioinformatic resource center for invertebrate vectors of human pathogens. (2010).

105. Giraldo-Calderón, G. I. *et al.* VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **43**, D707–D713 (2014).

1039 106. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, (2023).
1040 107. SRA Toolkit Development Team. *SRA Toolkit*. (2020).
1041 108. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome
1042 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915
1043 (2019).
1044 109. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1045 2079 (2009).
1046 110. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP: Automatic Gene Finding in
1047 Eukaryotic Genomes in Consistency with Extrinsic Data. *bioRxiv* (2024)
1048 doi:10.1101/2023.01.13.524024.
1049 111. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids*
1050 *Res.* **34**, W435–9 (2006).
1051 112. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr.*
1052 *Protoc. Bioinformatics* **65**, e57 (2019).
1053 113. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
1054 BUSCO: Assessing genome assembly and annotation completeness with single-copy
1055 orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
1056 114. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector
1057 for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
1058 115. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome
1059 Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33
1060 (2013).
1061 116. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation
1062 Sequencing Data. *BMC Bioinformatics* **15**, 1–13 (2014).
1063 117. Wickham, H. ggplot2. *WIREs Comp Stat* **3**, 180–185 (2011).
1064 118. Korunes, K. L. & Samuk, K. pixy: Unbiased estimation of nucleotide diversity and

1065 divergence in the presence of missing data. *Mol. Ecol. Resour.* **21**, 1359–1368 (2021).
1066 119. Weir, B. S. & Cockerham, C. C. Estimating f-statistics for the analysis of population
1067 structure. *Evolution* **38**, 1358–1370 (1984).
1068 120. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
1069 (2011).
1070 121. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
1071 datasets. *Gigascience* **4**, 7 (2015).
1072 122. Comeault, A. A., Kautt, A. F. & Matute, D. R. Genomic signatures of admixture and
1073 selection are shared among populations of *Zaprionus indianus* across the western
1074 hemisphere. *Mol. Ecol.* **30**, 6193–6210 (2021).
1075 123. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective
1076 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**,
1077 268–274 (2015).
1078 124. Rabiee, M., Sayyari, E. & Mirarab, S. Multi-allele species reconstruction using ASTRAL.
1079 *Mol. Phylogenet. Evol.* **130**, 286–296 (2019).
1080 125. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - Fast D-statistics and related admixture
1081 evidence from VCF files. *Mol. Ecol. Resour.* **21**, 584–595 (2021).
1082 126. Wang, K., Mathieson, I., O'Connell, J. & Schiffels, S. Tracking human population structure
1083 through time from whole genome sequences. *PLoS Genet.* **16**, e1008552 (2020).
1084 127. Schiffels, S. *Msmc-Tools: Tools and Utilities for Msmc and msmc2*. (Github).
1085 128. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008
1086 (2021).
1087 129. Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: ultra-fast computation
1088 of genome mappability. *Bioinformatics* **36**, 3687–3692 (2020).
1089 130. R Core Team. *R: A Language and Environment for Statistical Computing*. (2024).
1090 131. Hill, A. *et al.* Stepwise distributed open innovation contests for software development:

1091 Acceleration of genome-wide association analysis. *Gigascience* **6**, 1–10 (2017).