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

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 sandflies. 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. 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.

Nucleotide diversity across the genome is similar in females (π = 0.00177 ± 0.00233 SD female assembly; π = 0.00170 ± 0.00228 SD male assembly) and males from Colombia (π = 0.00171 ± 0.00236 SD female assembly; π = 0.00159 ± 0.00225 SD male assembly). However, concordant with other results, nucleotide diversity was lower on chromosome 1 in female sandflies (π = 0.00138 ± 0.00192 SD female assembly; π = 0.00131 ± 0.00196 SD male assembly) than males (π = 0.00165 ± 0.00239 SD female assembly; π = 0.00160 ± 0.00225 SD male assembly).

	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

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.

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

Transposable elements (TEs) are considered important drivers of sex chromosome evolution[32,33], so we quantified transposable element content in the female and male assemblies. RepeatMasker identified and masked a total of 16.99% (female) and 16.70% (male) of the genome, 15.69% and 15.4% of which were TEs in the female and male assemblies respectively. Most repeats were DNA transposons (12.8% and 12.3% of the genome in females and males respectively) or Retroelements (2.9% and 3.1% of the genome in females and males

respectively). A small proportion of the genome were simple repeats (0.9% and 0.8% of the genome in females and males respectively). The difference in relative proportion of transposable elements between the sexes is largely driven by chromosome 1. This chromosome is not only longer in males (49 Mbp vs 47.7 Mbp), but also has a larger overall proportion of bases masked in males (15.89% vs 13.60%). The male chromosome 1 has a total of 7.8 Mbp of transposable elements masked by our repeat masking pipeline, whereas the female chromosome has 6.499 Mbp of transposable elements masked. The difference, 1.3 Mbp, corresponds to the total difference in size between the assembled male and female chromosome 1 (see Figure S2). In the three autosomal chromosomes, the proportion of transposable elements to assembled chromosome length is <0.5% different between females and males.

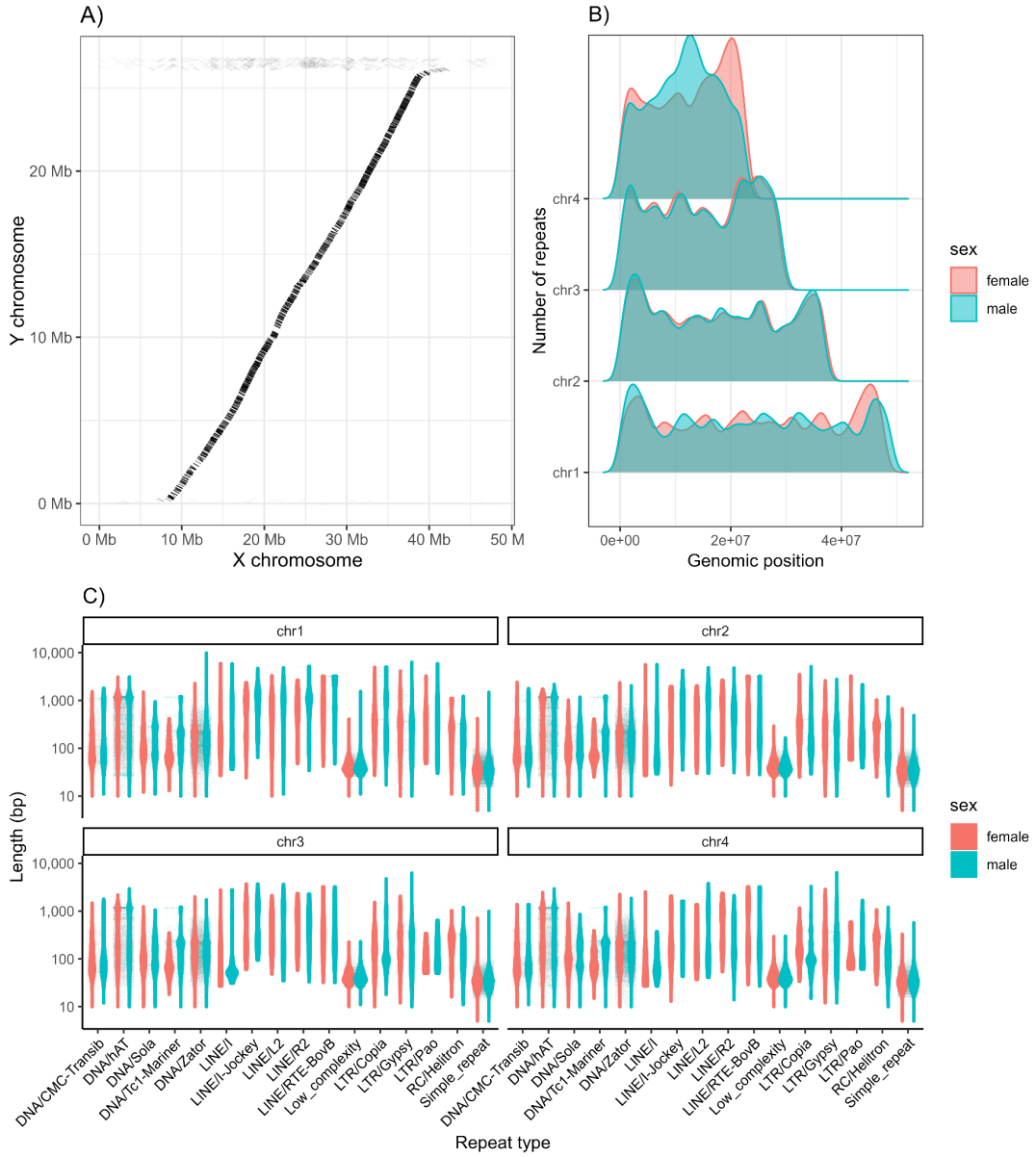


Figure 1. The sex chromosomes of *Lu. longipalpis*. A. Synteny plot of alignments between assembled male (presumptive) X and Y chromosome sequences. **B.** Number of repeats in the

female and male genome assemblies based on genomic position. **C.** Violin plots of TEs in the female and male genome assemblies and their length.

Genome annotation:

We annotated both sexes of our Colombian *Lu. longipalpis.* In total, we annotated 14,559 genes in the male assembly; 10,874 of these predicted genes were in the four chromosomes (4008, 2607, 2371, and 1888 in chromosomes 1-4 respectively). In the female assembly, we annotated 12,029 genes in total; 10,934 of these predicted genes were in the four chromosomes (3922, 2662, 2371, and 1979 in chromosomes 1-4 respectively). The predicted genes are approximately equal to RefSeq predictions in *Lu. longipalpis* from other localities (GCF_024334085.1), which has 12,045 predicted genes. In both sexes of our assemblies, the majority of predicted genes are found in the chromosomes. However, our male assembly has approximately 2,000 more predicted genes than existent assemblies, this is due to the presence of multiple copies of genes between the heteromorphic chromosomes--most of which are represented as unplaced contigs. While females overall had more predicted genes in the regions of the genome that are located in the chromosomes, males had a greater number of predicted genes owing to the predicted genes found in un-scaffolded contigs. Further, males had more predicted genes in chromosome 1, the heteromorphic chromosome. We identified a total of 3,462 genes annotated in the highly differentiated region of chromosome 1 (average Fst > 0.3 between females and males) in the male assembly. Of these, we tentatively assigned identifications to 3,611 genes based on protein evidence from OrthDB version 11 and VectorBase version 67. Among the genes found in this region of high differentiation between females and males within our Colombian population are a number of genes related to courtship (e.g., *takeout*), sex-differentiation (e.g., *sex-lethal, fruitless, doublesex- and mab-3-related transcription factor A2, lethal(2)*), testes-specific genes (e.g., ejaculatory bulb-specific protein 3, ejaculatory bulb-specific protein 3-like, testis-expressed protein 10 homolog, motile sperm domain-containing protein), putatively related to feeding and finding mates (e.g., gustatory receptor genes, gustatory receptor for sugar taste 64f, odorant receptor genes), or related to blood meals (e.g., venom allergen 5).

Next, we studied the extent of differentiation between the *X*-chromosome and the potentially male-restricted chromosome using PCA. We saw evidence of genomic differentiation between female and male genomes in our study. PC1, which corresponds to the sex of the sample, explained ~11.2% of genome-wide differentiation (Fig. 2A). This signal was clearly driven by chromosome 1, which showed a similar pattern in which sex explains ~25.1% of variance (Fig 1B). This signal was entirely absent when we did PCA for the putative "autosomes", chromosomes 2-4, with comparatively very little of the variation explained (PC1: 4%, PC2: 2% Fig 1C) and no correspondence to the sex of the sample. Additionally, we saw a long, extended region of differentiation (i.e., elevated F_{ST}) that occured between female and male sandflies on chromosome 1 (Fig 1D). This region also shows an extended region of decreased nucleotide diversity in female sandflies (Fig 1E), but there was no concomitant decrease in nucleotide diversity within that region in male sandflies (Fig 1F). 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.

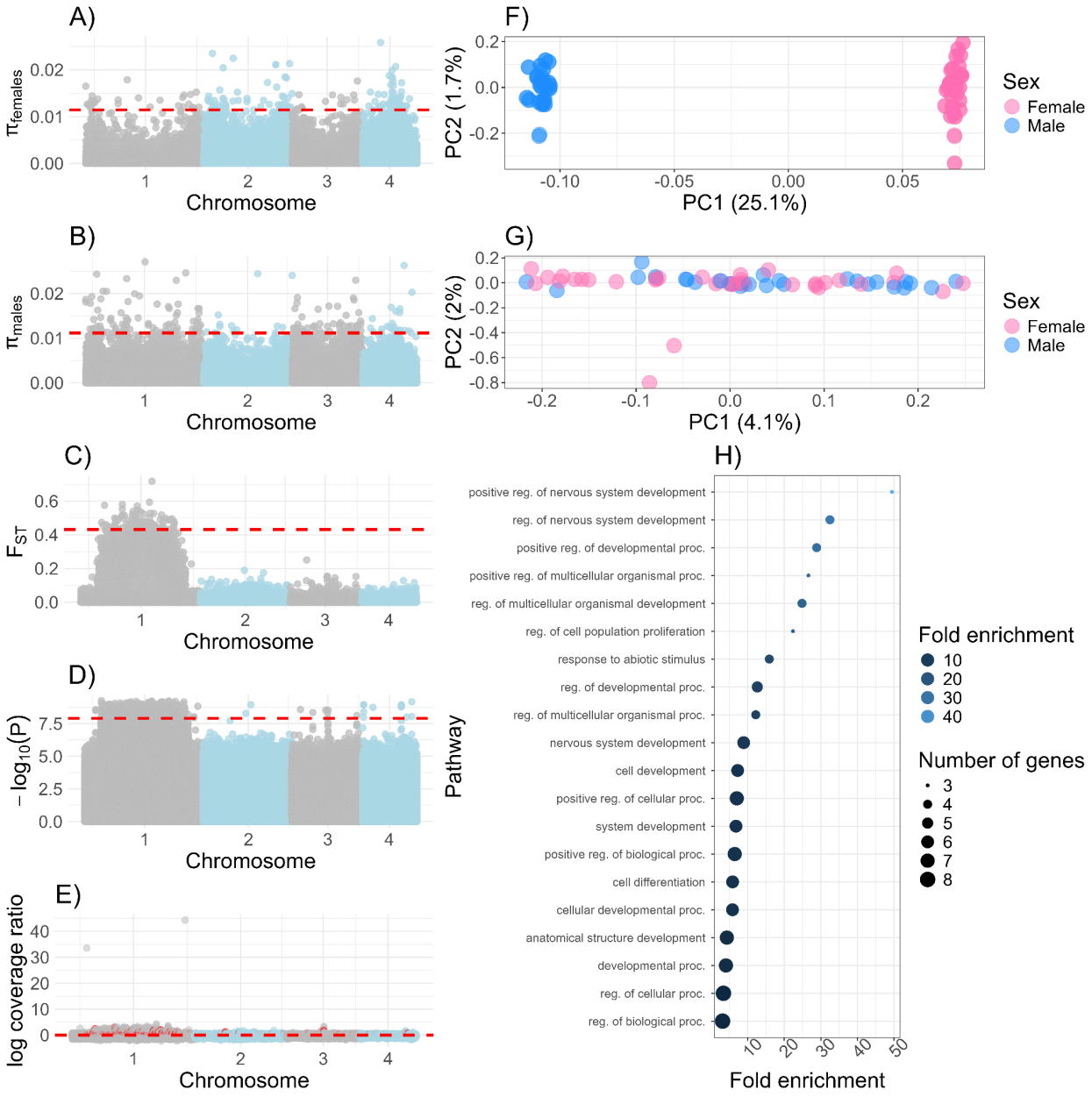


Figure 2. Sex chromosome differentiation in *Lu. longipalpis*. A. π in female sandflies from Colombia. **B.** π in male sandflies from Colombia. **C.** F_{ST} between the sexes. **D.** Manhattan plot of genome-wide association study between the sexes in Colombian sandflies. **E.** Coverage represented as a ratio of average male depth divided by average female depth, calculated as a 2,000bp mean. Two regions in chromosome 1 have been removed that have average log coverage of >30 (with mean positions at 5847001 and 45239001). **F.** PCA of only chromosome

1. **G.** PCA of the autosomes, chromosomes 2-4. **H.** GO biological processes analysis from GWAS results.

Genome wide association studies revealed autosomal genes associated with sex

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

Population genomics across the continent:

Lutzomyia longipalpis has been proposed to be differentiated into distinct populations and even separate species across South and Central America[15,36,37]. We used a two-pronged approach to study the population structure between our Colombian samples and samples of females from several Brazilian populations[31]. Sites are depicted in Figure 3B. Notably, Sobral was split *a priori* by Labbé et al.[31] into Sobral 1S and Sobral 2S based on morphology. First, we used PCA to represent the genetic variation across the continent. The first three principal components (PCs) explain 61% of the genetic variance of South American *Lu. longipalpis*. 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). PC2 and PC3 also capture considerable variance (16.4 and 10.6% respectively; see Supplemental Figure 3). 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. Second, we used phylogenetic trees. An unrooted genealogy based on 100kb windows revealed similar clustering

to that observed in the PCA as clustering was determined by geographical location. Jacobina and Sobral 1S were formed by two distinct and unrelated clusters. We further examined gene concordance between populations as an estimate of gene flow. We created a single “population” tree (akin to a species tree) from our multiple individual sampling of both the whole genome, and of BUSCO genes. When forced to 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 S6). This phylogenetic discordance may occur in instances of recent divergence, incomplete lineage sorting, hybridization, and poor tree inference which may be occuring 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 populations 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.

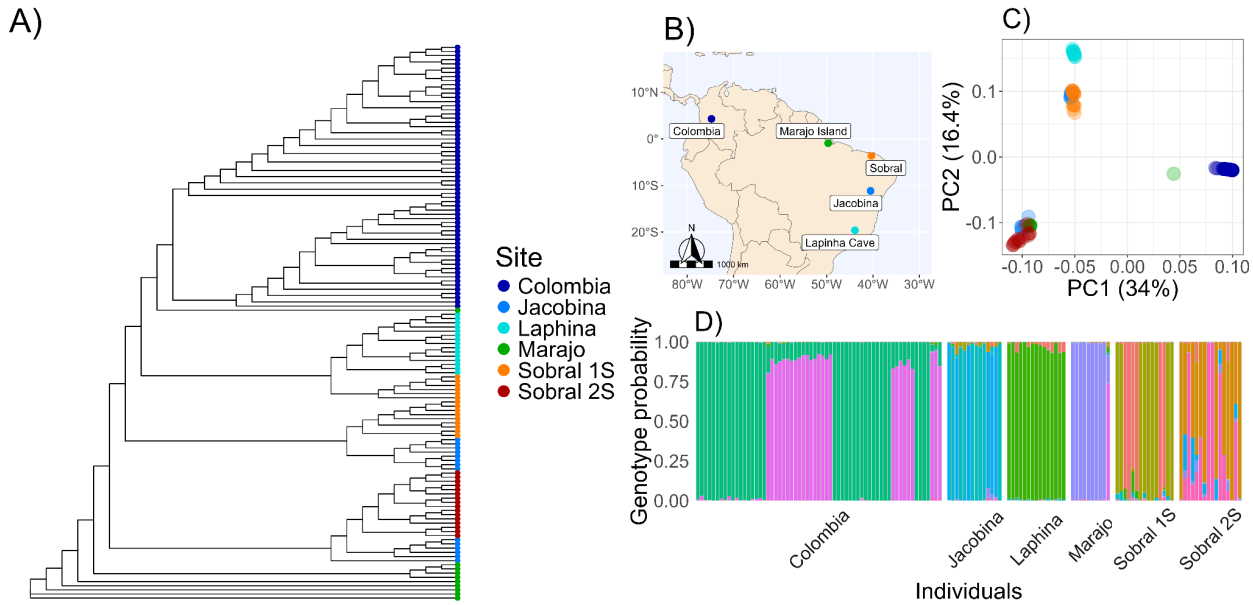


Figure 3. Partition of genetic diversity in *Lu. longipalpis* across South America. A. Phylogeny of sampled *Lutzomyia longipalpis* across South America. **B.** *Sampling localities with genomic data included in population genetics analyses.* ***C.*** *Principal component analysis of Colombian sandflies 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 even nonexistent—gene flow between the populations we identified. We used *PCAngsd*, which identifies the most likely number of clusters, but also estimates admixture proportions from allele

frequencies. Genetic variance was best explained by the existence of 10 clusters. Two of these clusters correspond to male and female *Lu. longipalpis* in Colombia. Colombia was two clusters representing males and females with little contribution to other lineages (a probability of <1% all Brazilian samples, except for the Colombian genotype that corresponds to males, which had a 73.9% probability in one Marajo sample, leading us to believe it was a male sandfly). This analysis (Figure 3D) also separated Colombian and Brazilian sandflies. The genetic variation in Brazilian samples was best explained by the existence of eight groups which correspond largely to the five geographic sites from which they were collected. The only exceptions were Sobral 1S, Sobral 2S, and Jacobina where we detected at least two main clusters per locality. The majority of individuals show limited evidence of admixed ancestry. In Colombia, Laphina, Jacobina, and all but one Marajo sandfly, admixture proportions are <10% per individual. This is particularly notable in the sites with syntopic lineages which could potentially have the chance to interbreed and yet do not appear to (Jacobina, some individuals within Sobral), as evidenced by divergent genotype clusters driving nearly the entirety of individual genetic makeup (e.g., blue vs teal individuals in Jacobina; Figure 3D). However, Sobral showed significant admixture within certain individuals (those of the “2S” lineage, Figure S7). We then proceeded to examine gene flow between populations using fbranch statistics. These analyses indicated high gene flow from Sobral 2S into Jacobina, Jacobina into Sobral 1S and Sobral 1S into Laphina (with Sobral “1S” and “2S” being *a priori* identified as different *Lu. longipalpis* lineages by[31]; Figure S8). Thus, there is either significant gene flow between populations or our consensus population tree does not adequately reflect cryptic diversification and potential syntopic lineages. Indeed, the consensus phylogeny of all individuals indicates the latter, as Sobral 2S individuals cluster with some Jacobina individuals. Overall, these results confirm that the different lineages of *Lu. longipalpis* are divergent, indicates advanced divergence towards speciation, and likely the presence of syntopic and hybridizing lineages.

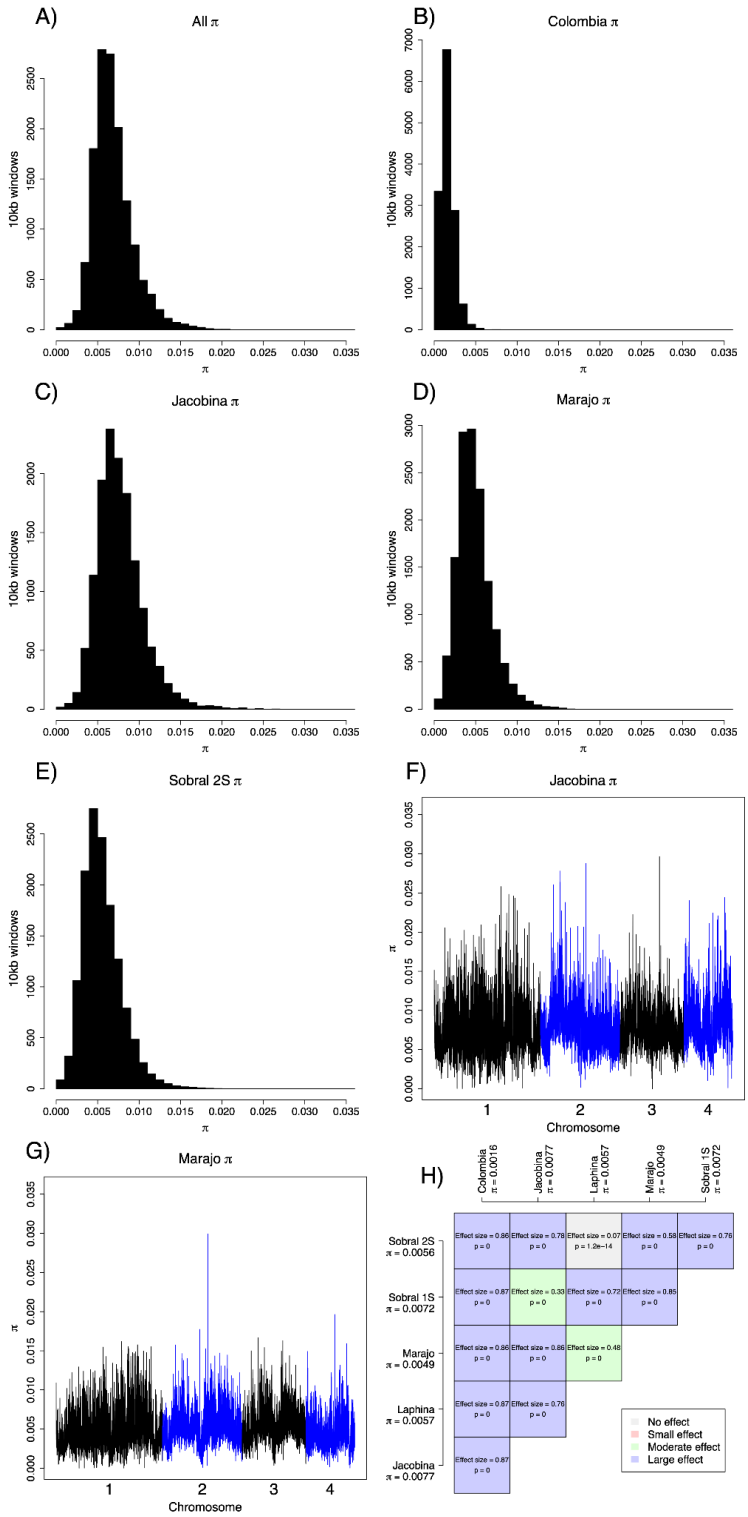


Figure 4. Genetic diversity in different lineages of *Lu. longipalpis*. A. Histograms of π in 10kb window sizes along the genome. **B.** π in Colombia **C**. π in Jacobina **D.** π in Marajo **E.** π in Sobral 2S. **F**. πalong the genome in Jacobina. **G.** π along the genome in Marajo. Figure 1 shows the π in Colombia. Figure S9 shows π along the genome for the remaining four populations. **H.** Pairwise comparisons show differences in π among pairs of populations.

Next, we studied the partitioning of genetic diversity within each *Lu. longipalpis* lineage. First, we quantified the extent of heterozygosity in each lineage. Some of the populations are less genetically diverse than others (Figure 4). Colombia shows the lowest heterozygosity (0.036); Sobral-1S and Marajo showed the highest heterozygosity (~0.1) and were three times more diverse than Colombia. All pairwise comparisons were significant, indicating differences in the level of genetic diversity across *Lu. longipalpis* populations. 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 and Figures S10). The pattern is not driven by a single individual and is a property of the whole population (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 5). 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.

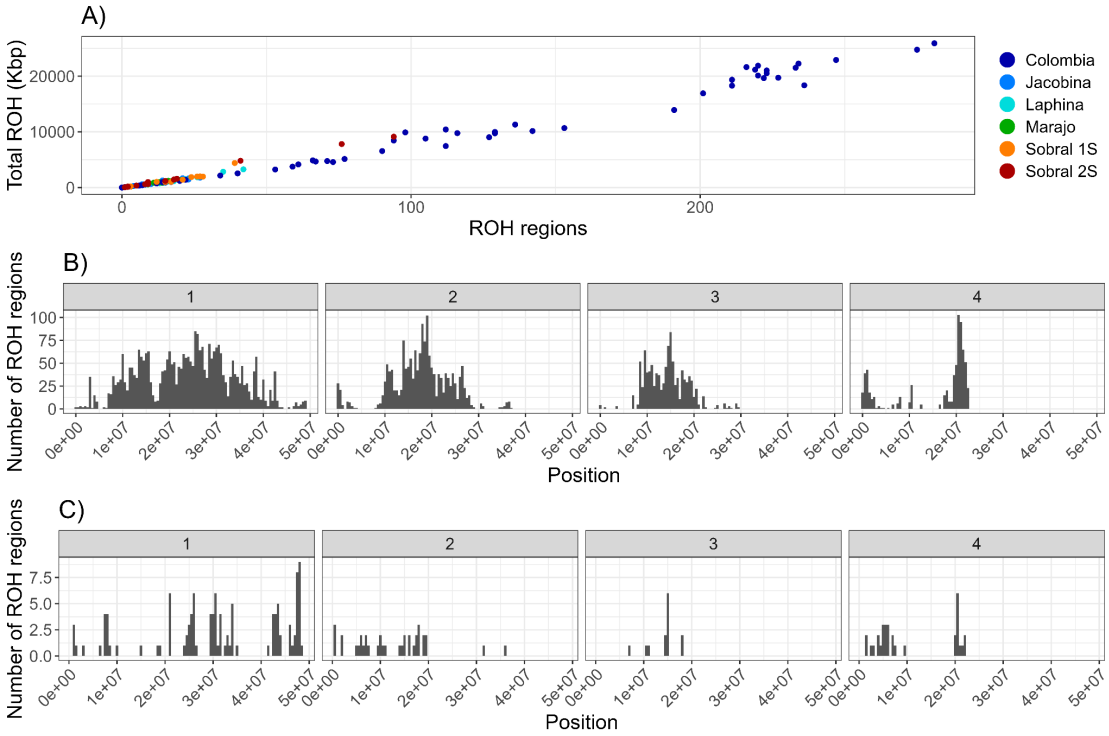


Figure 5. Runs of homozygosity in *Lu. longipalpis*. **A**. Plot depicting the total number of ROH regions over XX size and their total length by populations. **B**. Size distribution across genomic regions for Colombian sandflies (binwidth = 500000). **C**. Size distribution across genomic regions for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

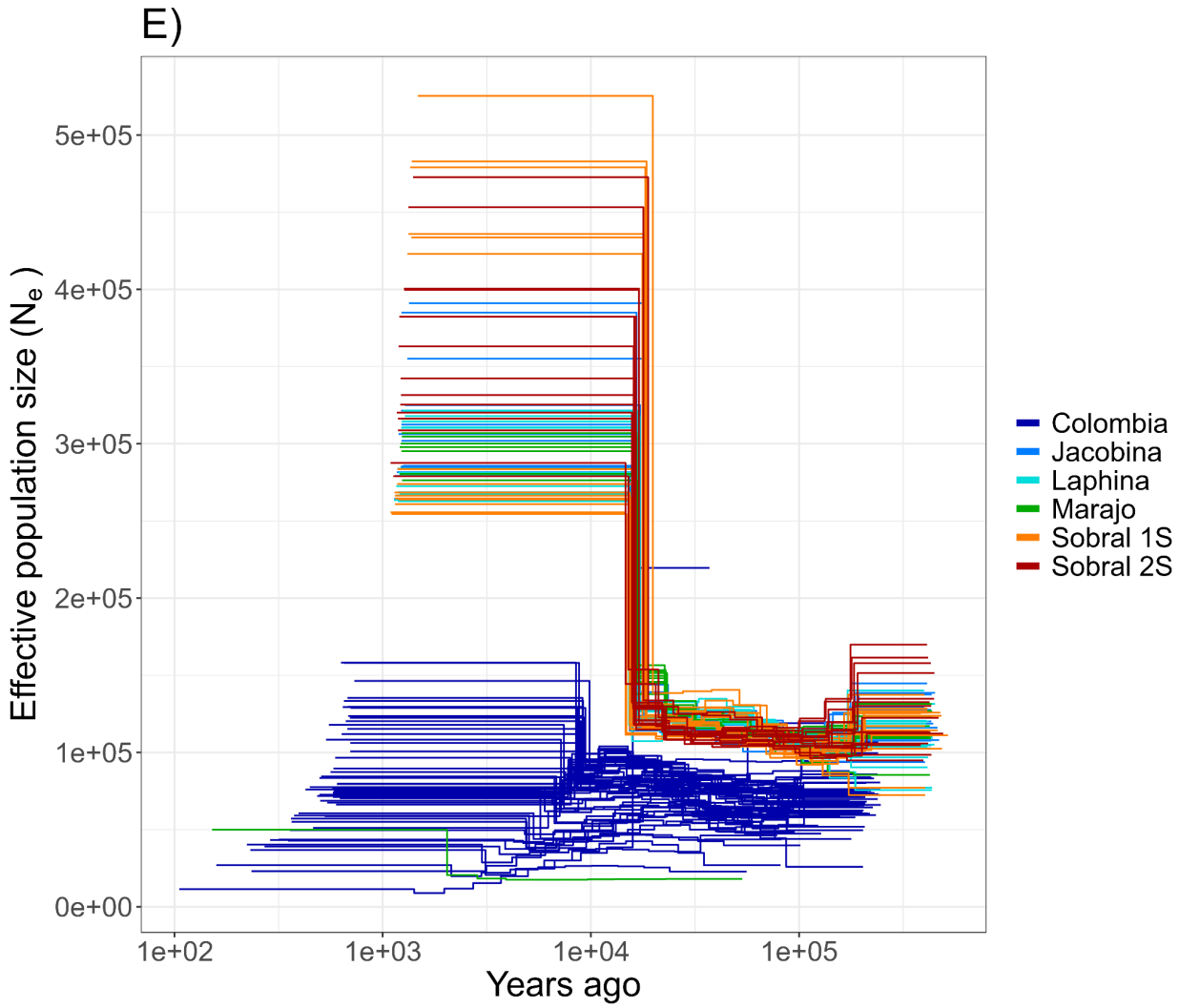


Figure 6. Demographic history of sandflies indicating temporal patterns of predicted effective population size.

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 information points to the presence of multiple syntopic lineages. We discuss each of these findings in the next paragraphs.

Our first finding is that females and males differ in their genomic composition. Although recent genomic work has indicated no heteromorphy in the Psychodidae[48], prior work has identified the presence of heteromorphic chromosomes in sandflies[49–51]. Here, we sequenced individuals from both sexes and produced highly continuous assemblies of both sexes. We found that the longest chromosome in *Lu. longipalpis* is heteromorphic and ~1.3 Mbp longer in male sandflies than female sandflies. Our study builds on these previous observations and reveals that there are differences in TE content and overall genomic composition between the *X* and *Y* chromosomes. Further, our analysis of repeats within the genome indicates that this difference in chromosomal length is likely driven by repeat content in male sandflies, which have an extra ~1.3 Mbp of repeats in this chromosome. Our results have implications for the study of sex chromosome evolution in dipterans. Further, because only females are hematophagous, and leishmaniasis is known to have negative impacts on longevity and fecundity of sandflies[52,53], leishmaniasis is likely to be a strong selective force that may impart different selective regimes on female and male genomes, which may be evident either between the sexes or potentially in heteromorphic sex chromosomes.

We identified a large, highly differentiated region between the sexes in a heteromorphic sex chromosome. This region shows a substantial dip in nucleotide diversity in females relative to males and encompasses a suite of genes related to sex determination, including the commonly characterized dipteran sex-determining genes *sex-lethal, fruitless, doublesex- and mab-3-related transcription factor A2, lethal(2)*[54]. Some of these genes, *fru* and *dsx*, have been implicated in sex determination in other phlebotomine species[55]. There are also a variety of testes-specific genes (e.g., *ejaculatory bulb-specific protein 3*, *ejaculatory bulb-specific protein 3-like*, *testis-expressed protein 10 homolog*, *motile sperm domain-containing protein*). One gene in this region, *takeout,* shows a high degree of structural conservation and has a conserved role in courtship across insects as diverse as fruit flies, mosquitos, and moths[56]. Other genes in this region of differentiation may be related to basic ecology and life history such as finding mates, finding prey, and effective hematophagy. Among these are several odorant receptor and gustatory receptor genes. Genes in these families play a key role in finding mates and lekking behavior[57], finding food or blood meals[58–60], and oviposition site choice[61,62]. In other hematophagous lineages, odorant receptor genes are heavily regulated and exhibit sex-specific and satiation-specific signatures of expression[63]. In *Lu. longipalpis,* pheromones play a key role in attracting mates[18,64–66]; these show population-specific signatures[18] and are perceived by odorant receptor genes which also show population specific variation [57]. These may play a role

in the cryptic *Lu. longipalpis* species complex via mate choice, and may be a source of prezygotic isolation. Although these roles of odorant receptor genes are well-known, these genes often have disparate functions among dipterans[67]. In sum, there are a suite of genes in the genomic region that are highly differentiated between the sexes that are related to ecology including finding food, finding mates, and oviposition that are potentially playing important, sex-specific roles.

A complementary result, also related to sex differentiation in *Lutzomyia*, is the finding of autosomal alleles associated with sex. In addition to these regions in chromosome 1, our genome wide association study between female and male sandflies in Colombia yielded significant peaks that were intragenic in three predicted genes, one per chromosome on chromosomes 2-4. We were unable to identify the gene product on chromosome 2 via Uniprot, OrthoDB, VectorBase, or NCBI blast. The predicted gene on chromosome 3 has also been predicted in a previous assembly of *Lu. longipalpis* [31], but remains an uncharacterized protein product. The intragenic SNPs our GWAS identified as associated with sex differences on chromosome 4 is Apoptotic protease-activating factor 1 (*apaf-1*). The *C. elegans* homolog of this gene (*CED-4*) interacts with the *C. elegans* gene *fem-1* which is crucial for male development[34] and *apaf-1* deficient mice exhibit male infertility due to degeneration of spermatogonia[35]. The *Drosophila* homolog of *apaf-1* is the gene Apaf-1–related killer (*ark,* or sometimes *dark* for *Drosophila* Apaf-1–related killer), which is required for sperm formation[68]. The SNPs that we identified were present in intronic regions between predicted exons 3 and 4 in our gene annotation. However, the publicly available RNAseq data that we used for annotation only contained female sandflies, potentially missing key, sex-specific isoforms of this gene. We postulate that genomic differentiation between the sexes in this gene is related to sex-specific differences in gametogenesis, or sex conflict (reviewed in [69]).

Perhaps most notably, our study provides strong evidence of divergence and population structure across the continent. Our analyses combine novel population genomics data for Colombian *Lu. longipalpis* populations with publicly-available whole genome sequence data, which allowed us to carry out the first population genetics study of this critical vector species across South America. Overall divergence between Brazilian and Colombian sandflies was high, supportive of previous work suggesting that *Lu. longipalpis* is a cryptic species complex[15,16,36,37]. In addition, we found evidence that Colombian *Lu. longipalpis* shows lower genetic variability than all of the Brazilian populations. This result manifested in two different facets, lower nucleotide diversity and a substantial increase in ROH tracts. ROH tracts were more common in the *X*-chromosome in all populations but were particularly noticeable in Colombia. It is unclear why Colombian flies exhibit such dramatic patterns of low diversity, particularly because it does not appear to be an isolated population. A potential explanation is reduced vagility in the Colombian lineage. For example if Colombian sandflies roam less during their lifetime than Brazilian flies, they are more likely to experience matings with close relatives and thus show more signs of inbreeding. An additional possibility is that Colombian *Lu. longipalpis* exhibit facultative parthenogenesis, a trait seen in several dipteran species[70], including in *Lu. mamedei*[71]. Both of these possibilities remain highly speculative and a systematic survey of the reproductive biology of the different lineages of *Lu. longipalpis* is needed to understand the patterns we observed. Our demographic analyses concur with the inference of low genetic diversity in Colombia and indicate that all populations had a similar *Ne*

between 16,000 and 14,000 years ago depending on the generation time in the wild; Brazilian populations underwent a large population expansion, Colombian *Lu. longipalpis* did not. This timing is similar to the arrival and expansion of humans in South America[72]. Given that *Lu. longipalpis* is one of the most anthropophilic species, population expansion in Brazilian sandflies may have occurred as a result of human colonization and expansion.

Colombian sandflies (from our single sampling locality) showed structure only between female and male sandflies, whereas multiple Brazilian populations (Jacobina, Sobral) were highly structured in all our analyses. Together, these results indicate the presence of multiple syntopic lineages--a result consistent with the previously hypothesized presence of cryptic species within *Lu. longipalpis*, *sensu lato. Lutzomyia longipalpis* has been proposed to encompass different chemotypes[64], and in some instances the proposals have been more ambitious and have proposed the existence of a species complex which encompasses different species with stark differences in mating behavior (e.g., mating song, pheromones and lekking behavior). Notably, and despite the differentiation we describe for these lineages, we find some evidence of admixture between lineages. Hybridization seems to be common in *Lutzomyia* (reviewed in[73]), which poses the question of how frequently differentiating lineages exchange alleles. Here we provide evidence for the existence of genome-wide admixture within *Lu. longipalpis*. While this introgression does not appear sufficient to erode the genetic differentiation between lineages, even when they are syntopic, this allele sharing might serve as fodder for adaptation[74,75]. The importance of introgression has been previously hypothesized across the *Lutzomyia* genus[76–78], and among *Lu. longipalpis* lineages[79,80]; our results represent a formal genome-wide test of the occurrence of this phenomenon in this genus of disease vectors.

Our manuscript leverages the power of whole genome sequencing to understand the evolutionary processes that govern genetic variation in the critical leishmaniasis vector *Lu. longipalpis*. Our results incorporate a genomic perspective into the study of differentiation within *Lu. longipalpis* and represent a resource to not only understand sex chromosome evolution in disease vectors, but also open the door to understanding the molecular and evolutionary mechanisms of the interactions between the vectors, the *Leishmania* parasites, and the hosts affected by the disease. Now that we have a better snapshot of the extent of genetic variation in the vector, the next frontier is understanding whether the evolutionary interactions between sandfly vectors and the *Leishmania* pathogen can provide the key to mitigating the spread of leishmaniasis and the human health threat it poses.

MATERIALS AND METHODS

Permits and ethics. All sandflies collections were performed following the guidelines of the Colombian decree No 1376. Insects were collected under the permits 001566 and 002662 of the Autoridad de Licencias Ambientales. Sandflies were collected on private property and permission was received from landowners before sampling.

Sample collection. Samples were collected from a semi-forested location outside of Ricaurte, Cundinamarca, Colombia in November 2021 and October 2022**.** Sandflies were collected via mechanical aspiration with a mouth aspirator and placed into vials to temporarily hold the insects until processing. Prior to processing, vials containing the sandflies were put in a cooler

with ice to immobilize the specimens. Vials were emptied into a Petri dish where the sandflies were selected and classified by species under a Leica dissecting scope. To identify sandfly species, we used taxonomic keys[81,82]. Individual sandflies were removed with tweezers and placed into individual 1.5mL tubes either in liquid nitrogen for genome assembly in 2022 or in RNA-stabilizing buffer (RNALater™; Ambion, USA) for whole genome resequencing in 2021. Samples were exported to North Carolina, USA for sequencing.

Genome sequencing. Due to their small size (thorax length in females: mean = 0.052 ± 0.0025 mm; males: mean = 0.0483 ± 0.0027 mm;[83] individuals were pooled prior to genomic DNA extraction for whole genome sequencing (females = 9, males = 10). We used two approaches to obtain reads for the genome assemblies. First, we generated PacBio HiFi reads. We homogenized sandflies with a pestle and incubated them at 55℃ for 30 minutes prior to performing a standard phenol chloroform extraction. Extracted high molecular weight gDNA was sent to the Genomics Core Facility at the Icahn School of Medicine at Mount Sinai, where libraries were prepared using PacBio's ultra-low DNA input HiFi library prep which utilizes PCR amplification of genomic material. Libraries were pooled and sequenced across two SMRT cells on a Pacific Biosciences Sequel IIe to a coverage of 71.8x and 75.7x coverage for females and males.

Second, we used chromatin configuration. We sent 25 female and 25 male *Lu. longipalpis* in ethanol to Arima Genomics for DNA extraction and sequencing. Hi-C data was generated using the Arima HiC kit and the Arima Library Prep Module according to the Arima Genomics manufacturer's protocols. Briefly, Arima High Coverage Hi-C is designed to capture the sequence and 3D conformation of genomes. Chromatin from tissues, cell lines, or blood is first crosslinked to preserve genomic structure, then digested with a restriction enzyme cocktail optimized for uniform coverage. The digested ends are labeled with biotinylated nucleotides during a fill-in step and subsequently ligated to capture spatially proximal DNA fragments. These proximally-ligated DNA fragments are purified, fragmented, and enriched for biotin-labeled sequences.The female sample was sequenced to 189.5 million read pairs and the male sample to 180 million read pairs on an Illumina NovaSeq on 'paired-end' mode.

Genome assemblies. We produced separate genome assemblies for the female and male sandflies. While we initially attempted to use HiFiasm[84], we discovered that haplotype diversity in our pooled data exceeded the allowed diversity in the assembly algorithm. Therefore, we chose to use wtdbg2[85], an assembler with an algorithm that allows for greater diversity and collapses haplotypes into a single, haploid assembly. We used the "wtpoa-cns" command in wtdbg2 to polish the assembly.

We then scaffolded our initial draft genome assemblies with our Hi-C data. Splits were calculated using Juicer version 1.6[86] and reads were mapped to the draft assemblies with BWA version 0.7.17[87]. We then ran Samtools version 1.17 to sort the aligned reads and Picard's "MarkDuplicates" version 2.27.5 to mark duplicates[88]. We indexed our initial contig assembly using Samtools "faidx" prior to running the YaHS (Yet Another Hi-C Scaffolding tool) pipeline[89] with a MAPQ score of 30. We used Juicer "pre" to create a HiC contact map file which we imported into Juicebox Assembly Tools[90] to manually curate each assembly. Curation included fixing one major translocation (female assembly) and splitting a chromosome that had an

incorrect telomere-to-telomere join (male assembly). We then used Juicer "post" to update the manually curated final assemblies for females and males.

After examination of whole genome sequence results, we identified systemic and dramatic differentiation between female and male sequences along the majority of chromosome 1. We were unable to initially phase a ZW or XY assembly using our assembly approaches, likely due to the pooled nature of our data (i.e., because male sandflies were too small for 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.

WGS resequencing. 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.

WGS mapping, variant calling, and diversity summary statistics. Next we measure the extent of genetic diversity and differentiation within *Lu. longipalpis*. We followed the GATK best practices for variant calling[115]. We used BBMap version 38.87 (https://github.com/BioInfoTools/BBMap) to trim adaptors from whole genome resequencing data and then mapped reads to the female reference genome using BWA version 0.7.17[87]. Read groups were added using GATK's AddOrReplaceReadGroups. For samples sequenced across two lanes, aligned data were then merged using the "merge" tool in samtools version 1.14[109]. Aligned data were cleaned with Picard's CleanSam, sorted with SortSam, and duplicates marked with MarkDuplicates (Picard version 2.23.4[88]. Genotypes were then called with GATK HaplotypeCaller, before joint genotyping data with GenotypeGVCFs. All GATK functions were from GATK version 4.1.9.0.

We merged bam files for each sample that were sequenced on multiple lanes using samtools version 1.14 merge function[109]. We generated a Beagle file of genotype probabilities using *angsd* version: 0.921-11-g20b0655, and then calculated principal components using the *angsd* script "pcangsd.py"[116]. We then extracted eigenvalues and calculated the percent of the variance in our data explained by each PC axis, which we then plotted in R version 4.1.1 ("Kick Things") using ggplot2[117]. In addition to running this PCA pipeline with whole genome data, we subsetted the alignment bam file in order to separately examine differentiation between the presumptive sex chromosome (chromosome 1) and the autosomes (chromosomes 2-4) using the same approach.

We examined basic population statistics (nucleotide diversity, π, and population divergence [F_{ST}, Dxy]) using pixy[118], which accounts for invariant sites. We calculated pairwise differences in nucleotide diversity and Weir-Cockerham F_{ST} values[119]. We were particularly interested in potential differentiation between female and male sandflies, given that only females are hematophagous and thus only females are vectors of leishmaniasis. Per pixy's instructions, we first created a vcf file of invariant sites using the "-all-sites" flag in GATK's GenotypeGVCFs. We then filtered the corresponding gvcf file to remove sites with a high proportion of missing data (--max-missing 0.9), low coverage (--min-meanDP 3), and high overall depth (--max-meanDP 40) using vcftools 0.1.15[120]. We then used pixy to calculate F_{ST} and Dxy across 1,000bp windows between female and male sandflies. Additionally, we subset our gvcf file into a female-only and a male-only gvcf which we used to calculate nucleotide diversity (π) across 1,000bp windows in female and male sandflies. We additionally calculated Tajima's D in 1,000 bp windows using VCFtools.

We detected runs of homozygosity (ROH) using PLINK 1.90b6.21[121] with the parameters: --allow-extra-chr and --homozyg-kb 5. This command estimates the run length and identifies contiguous segments of homozygosity greater than 5 kbp by assessing the observed genotype data against expected patterns of homozygosity. ROH was detected separately for

each of Colombia and the five Brazil populations. With a 5 kbp minimum size, we identified a minimum of 100 SNPs per region and on average there were 189-276 SNPs per region. We repeated the same analyses using different parameter values for the minimum length of ROH in kbp (5, 10, 25, 50, 100), and results were similar despite fewer ROH regions with higher kb thresholds.

Phylogenetic trees. We used trees to address two questions. First, to examine the genealogy of the largest chromosome relative to the presumptive autosomes in placement of females and males in relation to each other, and second, to examine continent-wide evolutionary relationships. We used the python script "vcf2phylp.py"[122] to split our vcf files into 100,000 bp regions and convert them into the phylip format.

To examine the genealogy of the largest chromosome relative to the presumptive autosomes, we created two separate trees using 1) only the largest chromosome or 2) the presumptive autosomes in just Colombian sandflies. For both trees, we made a phylogenetic tree for each window using IQ-TREE[123] version 2.2.2.7 with DNA specified ("-st DNA") and 1000 bootstraps ("-bb 1000"). We merged the resulting consensus tree files into a single tree and used ASTRAL[124] version 5.7.8 to create a single genome-wide consensus tree. We additionally created trees solely from chromosome 1 and the autosomes.

To examine evolutionary relationships across the continent, we made a single genome-wide tree using the above methods. The approach for this question is identical to the one described immediately above, with the exception that we created a single tree including all chromosomes. Further, we examined gene concordance between the genome wide tree and conserved, core dipteran genes from the BUSCO diptera_odb10 database[113]. To do this, we used ASTRAL[124] version 5.7.8 to create a single genome-wide consensus tree that is constrained to the populations sampled. We additionally extracted all the complete, single copy dipteran BUSCO genes assembled in our male assembly, using the python script "vcf2phylp.py"[122] to convert the vcf of these genomic regions into the phylip format. We created single gene trees as per the windows above using IQ-TREE, and created a single consensus population-level tree using ASTRAL as we did for our whole genome tree. We then used IQ-TREE[123] version 2.2.2.7 to calculate gene concordance between each gene tree and the overall consensus tree. Finally, we examined gene flow and the potential for introgression using fbranch statistics in Dsuite [125]. Briefly, we used our consensus population tree and our variant call file as input to examine fbranch statistics using "Dtrios" to examine all trios and "Fbranch" to examine the output fbranch statistics. Given the clear divergence between Brazilian and Colombian sandflies, we used Colombia as the outgroup in our analysis and focused on Brazilian sandflies.

Population genomics of geographically disparate Lu. longipalpis *populations. Lutzomyia longipalpis* has a broad geographic range from Argentina through Mexico[2,14]. Further, multilocus sequence typing and phenotypic surveys have revealed strong population structure which might correspond to cryptic speciation[17,18] and divergences in courtship song[20], morphology,[21], and pheromones[18]. Recent work sequenced Brazilian *Lu. longipalpis* from four localities in Brazil[31]. Here we examined basic population genomic metrics between Colombian and Brazilian *Lu.*

longipalpis. We were particularly interested in differentiation between populations of putative *Lu. longipalpis,* and as a result focus on metrics such as F_{ST} and Dxy between populations.

Out of the 68 Colombian samples, 63 sandflies (37 females and 26 males) passed our quality metrics for population level analyses (>50% of reads mapping to our *Lu. longipalpis* female assembly and > 5X average coverage). On average, 91.4% of female reads and 84.4% of male reads mapped to the female assembly. Average coverage was 16.8x and 13.2x for females and males respectively. When mapped to the male assembly, 84.3% of female reads and 79.6% of male reads mapped and we had 14.6x and 11.7x coverage for females and males respectively.

Additionally, we downloaded whole genome resequence data from[31] using the SRA toolkit[107]. Table S1 lists the accession numbers of the previously published reads used in this study. We then followed the same protocol that we used for our Colombian samples, the GATK best-practices for variant calling [115]. We used BBMap version 38.87 (https://github.com/BioInfoTools/BBMap) to trim adaptors from whole genome resequencing data and then mapped reads to the female reference genome using BWA version 0.7.17[87]. Read groups were added using GATK's AddOrReplaceReadGroups. For samples sequenced across two lanes, aligned data were then merged using the "merge" tool in samtools version 1.14[109]. Aligned data were cleaned with Picard's CleanSam, sorted with SortSam, and duplicates marked with MarkDuplicates (Picard version 2.23.4[88]. Genotypes were then called with GATK HaplotypeCaller, before joint genotyping data with GenotypeGVCFs. All GATK functions were from GATK version 4.1.9.0.

For samples across the continent of South America, we examined summary population statistics (π, F_{st}, and Dxy) as we did for samples within Colombia (see above). We filtered our gvcf file to remove sites with a high proportion of missing data (--max-missing 0.9), low coverage (--min-meanDP 3), and high overall depth (--max-meanDP 40) using vcftools 0.1.15[120]. We then used pixy to calculate F_{ST} and Dxy across 1,000bp windows between populations. Additionally, we subset our gvcf file into a female-only and a male-only gvcf which we used to calculate nucleotide diversity (π) across 1,000bp windows in female and male sandflies. We additionally calculated Tajima's D in 1,000 bp windows using VCFtools.

In order to examine population structure across the continent, we used angsd[116] to examine the data with a PCA, structure, and *ADMIXTURE* based approaches. We used a similar approach to the within-Colombia examinations. We generated a Beagle file of genotype probabilities using *angsd* version: 0.921-11-g20b0655, and then calculated principal components and genotype probabilities using the *angsd* script "pcangsd.py"[116]. This program calculates ancestry based on individual allele frequency and uses a similar approach to delta K in that it permutes iteratively across a range of ancestry possibilities. We then extracted eigenvalues and calculated the percent variance in our data explained and plotted both this PCA and a STRUCTURE plot in R version 4.1.1 ("Kick Things") using ggplot2[117].

To quantify if sex was responsible for differences in genome size, we used a linear model (function "lm" in R) with average π per window in the sexes as the outcome and sex, chromosome, and the interaction of sex and chromosome as fixed effects.

We estimated demographic histories using MSMC2[126] with one thread and ten time segments of different lengths using the parameters '-t 1 -p 1-6+8-1+1-4'. MSMC2 was run separately for individual samples since we had unphased genotypes. The MSCM2 input files

were generated using msmc-tools[127]. Separate input files were generated for each of the four major chromosomes using 'generate_multihetsep.py' and then concatenated together into separate genomic input files for each individual sample. 'generate_multihetsep.py' requires two input files: a mask of sites with coverage for that individual and a genomic mappability mask. The covered site masks were created in three steps run separately for each chromosome and each individual sample using Samtools 1.21[128], BCFtools[128] and msmc-tools. The genomic mappability mask was created using genmap[129]. MSMC2 results were plotted using R [130] using a mutation rate of μ=1.25e-8 and an estimated generation time of 49 days[40] with time in years calculated as 'time_boundary / μ / gen'. The effective population size was calculated as (1 / λ) / (2 * μ) where lambda is the coalescence rate returned by MSMC2.

Genome wide association study for sex determination. Finally, we ran a genome-wide association study (GWAS) for 133 samples from both Colombia and Brazil using sex as the phenotype (113 females, 20 males) to identify genetic regions associated with sex determination. First, we created the required plink input files from our vcf by filtering for just the four chromosomes and using a minor allele frequency of 0.05. To run the GWAS, we used plink 2.0[121] using a generalized linear model framework with sex as a binary variable[131], and we filtered out SNPs with 20% or more missing data (--geno 0.2) and a minor allele frequency of 0.05 (--maf 0.05). Results are presented from the continent wide analysis, but results are the same using just the Colombian samples yielded qualitatively identical results.

Data availability:
All raw sequence data as well as genome assemblies are available on NCBI under BioProject PRJNA1110158. Code can be found at https://github.com/AdamStuckert/Lu_longipalpis; and will be deposited in Zenodo upon acceptance.

Author contributions:
AMMS: Conceived project, collected samples, conducted wet lab work, analyzed data, wrote manuscript
DT: Analyzed data, wrote manuscript
ERRD: Conducted wet lab work, wrote manuscript
DP: Conducted wet lab work, wrote manuscript
MP: Conducted wet lab work, wrote manuscript
CM: Conceived project
JAR: Wrote manuscript
EB: Conceived project, collected samples
JRW: Conducted wet lab work, wrote manuscript
RV: Conceived project, collected samples, identified sandflies, wrote manuscript
DRM: Conceived project, collected samples, analyzed data, wrote manuscript

Competing interests:
The authors declare they have no competing interests.

Acknowledgments:

The research presented here was supported by the National Institute of General Medical 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. A. 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**, e585 (2010).

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

(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 in Colombia. *J. Med. Entomol.* **30**, 427–435 (1993).

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

39. Ceballos, F. C., Hazelhurst, S. & Ramsay, M. Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. *BMC Genomics* **19**, 106 (2018).

40. Killick-Kendrick, R., Leaney, A. J. & Ready, P. D. The establishment, maintenance and productivity of a laboratory colony of Lutzomyia longipalpis (Diptera: Psychodidae). *J. Med. Entomol.* **13**, 429–440 (1977).

41. Killick-Kendrick, R. Phlebotomine vectors of the leishmaniases: a review. *Med. Vet. Entomol.* **4**, 1–24 (1990).

42. Killick-Kendrick, R. The biology and control of phlebotomine sand flies. *Clin. Dermatol.* **17**, 279–289 (1999).

43. Ready, P. D. Biology of phlebotomine sand flies as vectors of disease agents. *Annu. Rev. Entomol.* **58**, 227–250 (2013).

44. Salomon, O. D. Lutzomyia longipalpis, Gone with the Wind and Other Variables. *Neotrop. Entomol.* **50**, 161–171 (2021).

45. Sousa-Paula, L. C. de, Otranto, D. & Dantas-Torres, F. Lutzomyia longipalpis (Sand Fly). *Trends Parasitol.* **36**, 796–797 (2020).

46. Rêgo, F. & Soares, R. Lutzomyia longipalpis: an update on this sand fly vector. *An. Acad. Bras. Cienc.* **93**, e20200254 (2021).

47. Huang, M. *et al.* Improved high quality sand fly assemblies enabled by ultra low input long read sequencing. *Sci. Data* **11**, 918 (2024).

48. Vicoso, B. & Bachtrog, D. Numerous transitions of sex chromosomes in Diptera. *PLoS Biol.* **13**, e1002078 (2015).

49. Kreutzer, R. D., Modi, G. B., Tesh, R. B. & Young, D. G. Brain cell karyotypes of six species of New and Old World sand flies (Diptera: Psychodidae). *J. Med. Entomol.* **24**, 609–612 (1987).

50. Jiménez, M. E., Bello, F. J., Ferro, C. & Cárdenas, E. Brain cell karyotype of the

phlebotomine sand fly Lutzomyia shannoni (Dyar) (Diptera: Psychodidae). *Mem. Inst. Oswaldo Cruz* **96**, 379–380 (2001).

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

61. Dougherty, M. & Hamilton, G. Dodecanoic Acid Is the Oviposition Pheromone ofLutzomyia longipalpis. *J. Chem. Ecol.* **23**, 2657–2671 (1997).

62. Dougherty, M. J., Guerin, P. M. & Ward, R. D. Identification of oviposition attractants for the sandfly Lutzomyia longipalpis (Diptera: Psychodidae) in volatiles of faeces from vertebrates. *Physiol. Entomol.* **20**, 23–32 (1995).

63. Ni, M. *et al.* Screening for odorant receptor genes expressed in Aedes aegypti involved in host-seeking, blood-feeding and oviposition behaviors. *Parasit. Vectors* **15**, 71 (2022).

64. Spiegel, C. N. *et al.* The *Lutzomyia longipalpis* complex: a brief natural history of aggregation-sex pheromone communication. *Parasit. Vectors* **9**, 580 (2016).

65. Hamilton, J. G. C., Maingon, R. D. C., Alexander, B., Ward, R. D. & Brazil, R. P. Analysis of the sex pheromone extract of individual male Lutzomyia longipalpis sandflies from six regions in Brazil. *Med. Vet. Entomol.* **19**, 480–488 (2005).

66. Watts, P. C. *et al.* Male sex pheromones and the phylogeographic structure of the Lutzomyia longipalpis species complex (Diptera: Psychodidae) from Brazil and Venezuela. *Am. J. Trop. Med. Hyg.* **73**, 734–743 (2005).

67. Maguire, S. E., Afify, A., Goff, L. A. & Potter, C. J. Odorant-receptor-mediated regulation of chemosensory gene expression in the malaria mosquito Anopheles gambiae. *Cell Rep.* **38**, 110494 (2022).

68. Huh, J. R. *et al.* Multiple apoptotic caspase cascades are required in nonapoptotic roles for Drosophila spermatid individualization. *PLoS Biol.* **2**, E15 (2004).

69. Kasimatis, K. R., Nelson, T. C. & Phillips, P. C. Genomic signatures of sexual conflict. *J. Hered.* **108**, 780–790 (2017).

70. Sperling, A. L. & Glover, D. M. Parthenogenesis in dipterans: a genetic perspective. *Proc. Biol. Sci.* **290**, 20230261 (2023).

71. Brazil, R. P. & Oliveira, S. M. Parthenogenesis in the sandfly Lutzomyia mamedei (Diptera: Psychodidae). *Med. Vet. Entomol.* **13**, 463–464 (1999).

72. Goldberg, A., Mychajliw, A. M. & Hadly, E. A. Post-invasion demography of prehistoric humans in South America. *Nature* **532**, 232–235 (2016).

73. Dos Reis, Y. V. & Alevi, K. C. C. Hybridization in Phlebotominae (Diptera: Psychodidae): A mini-review. *Infect. Genet. Evol.* **86**, 104593 (2020).

74. Hedrick, P. W. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* **22**, 4606–4618 (2013).

75. Arnold, M. L. & Martin, N. H. Adaptation by introgression. *J. Biol.* **8**, 82 (2009).

76. Mazzoni, C. J. *et al.* Multilocus analysis of introgression between two sand fly vectors of leishmaniasis. *BMC Evol. Biol.* **8**, 141 (2008).

77. Mazzoni, C. J., Souza, N. A., Andrade-Coelho, C., Kyriacou, C. P. & Peixoto, A. A. Molecular polymorphism, differentiation and introgression in the period gene between Lutzomyia intermedia and Lutzomyia whitmani. *BMC Evol. Biol.* **6**, 85 (2006).

78. Testa, J. M., Montoya-Lerma, J., Cadena, H., Oviedo, M. & Ready, P. D. Molecular identification of vectors of Leishmania in Colombia: mitochondrial introgression in the Lutzomyia townsendi series. *Acta Trop.* **84**, 205–218 (2002).

79. Casaril, A. E. *et al.* Macrogeographic genetic structure of Lutzomyia longipalpis complex populations using Next Generation Sequencing. *PLoS One* **14**, e0223277 (2019).

80. Araki, A. S. *et al.* Multilocus analysis of divergence and introgression in sympatric and allopatric sibling species of the Lutzomyia longipalpis complex in Brazil. *PLoS Negl. Trop. Dis.* **7**, e2495 (2013).

81. Galati, E. A. B. Phlebotominae (Diptera, Psychodidae): Classification, Morphology and Terminology of Adults and Identification of American Taxa. in *Brazilian Sand Flies : Biology, Taxonomy, Medical Importance and Control* (eds. Rangel, E. F. & Shaw, J. J.) 9–212 (Springer International Publishing, Cham, 2018).

82. Galati, E. A. B., Galvis-Ovallos, F., Lawyer, P., Léger, N. & Depaquit, J. An illustrated guide

for characters and terminology used in descriptions of Phlebotominae (Diptera, Psychodidae). *Parasite* **24**, 26 (2017).

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

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).

106. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, (2023).

107. SRA Toolkit Development Team. *SRA Toolkit*. (2020).

108. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

109. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

110. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency with Extrinsic Data. *bioRxiv* (2024) doi:10.1101/2023.01.13.524024.

111. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–9 (2006).

112. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics* **65**, e57 (2019).

113. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

114. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).

115. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).

116. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 1–13 (2014).

117. Wickham, H. ggplot2. *WIREs Comp Stat* **3**, 180–185 (2011).

118. Korunes, K. L. & Samuk, K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* **21**, 1359–1368 (2021).

119. Weir, B. S. & Cockerham, C. C. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution* **38**, 1358–1370 (1984).

120. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

121. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

122. Comeault, A. A., Kautt, A. F. & Matute, D. R. Genomic signatures of admixture and selection are shared among populations of Zaprionus indianus across the western hemisphere. *Mol. Ecol.* **30**, 6193–6210 (2021).

123. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

124. Rabiee, M., Sayyari, E. & Mirarab, S. Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenet. Evol.* **130**, 286–296 (2019).

125. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* **21**, 584–595 (2021).

126. Wang, K., Mathieson, I., O'Connell, J. & Schiffels, S. Tracking human population structure through time from whole genome sequences. *PLoS Genet.* **16**, e1008552 (2020).

127. Schiffels, S. *Msmc-Tools: Tools and Utilities for Msmc and msmc2*. (Github).

128. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

129. Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics* **36**, 3687–3692 (2020).

130. R Core Team. *R: A Language and Environment for Statistical Computing*. (2024).

131. Hill, A. *et al.* Stepwise distributed open innovation contests for software development: Acceleration of genome-wide association analysis. *Gigascience* **6**, 1–10 (2017).

SUPPLEMENTARY TABLES

Table S1. Sample information for short read genomic data used in this study. This includes the collections from Colombia and a list of accession numbers of the previously published reads used in this study.

Run	Site	Sex
LU_LONG_M_131_CAATTAAC-CGAGATAT_S34	Ricaurte, Colombia	Male
JB_LM30-lutzomyia_male_42_021921_TCCTGAGC-AAGGCTAT_S79	Ricaurte, Colombia	Male
JB_LM11-lutzomyia_male_23_021921_CGTACTAG-CCTAGAGT_S58	Ricaurte, Colombia	Male
JB_LF33-lutzomyia_female_100_021921_CGAGGCTG-GAGCCTTA_S26	Ricaurte, Colombia	Female
JB_LF13-lutzomyia_female_80_021921_TAGGCATG-CCTAGAGT_S4	Ricaurte, Colombia	Female
JB_LF4-lutzomyia_female_71_021921_GGACTCCT-TTCTAGCT_S44	Ricaurte, Colombia	Female
JB_LM3-lutzomyia_male_15_021921_TAAGGCGA-CCTAGAGT_S80	Ricaurte, Colombia	Male
LU_LONG_F_P12_GGACTTGG-CGCAGACG_S20	Ricaurte, Colombia	Female
JB_LM1-lutzomyia_male_13_021921_TAAGGCGA-TCGACTAG_S67	Ricaurte, Colombia	Male
JB_LF52-lutzomyia_female_119_021921_GCTCATGA-TTCTAGCT_S47	Ricaurte, Colombia	Female
JB_LM24-lutzomyia_male_36_021921_AGGCAGAA-TTATGCGA_S72	Ricaurte, Colombia	Male
JB_LM22-lutzomyia_male_34_021921_AGGCAGAA-AAGGCTAT_S70	Ricaurte, Colombia	Male
JB_LM19-lutzomyia_male_31_021921_AGGCAGAA-CCTAGAGT_S66	Ricaurte, Colombia	Male
JB_LM5-lutzomyia_male_17_021921_TAAGGCGA-CTATTAAG_S82	Ricaurte, Colombia	Male
JB_LF45-lutzomyia_female_112_021921_GTAGAGGA-CCTAGAGT_S39	Ricaurte, Colombia	Female

JB_LF51-lutzomyia_female_118_021921_GCTCATGA-TCGACTAG_S46	Ricaurte, Colombia	Female
JB_LM18-lutzomyia_male_30_021921_AGGCAGAA-TTCTAGCT_S65	Ricaurte, Colombia	Male
JB_LM2-lutzomyia_male_14_021921_TAAGGCGA-TTCTAGCT_S78	Ricaurte, Colombia	Male
LU_LONG_F_138_CATAGAGT-TGCCACCA_S5	Ricaurte, Colombia	Female
JB_LF48-lutzomyia_female_115_021921_GTAGAGGA-AAGGCTAT_S42	Ricaurte, Colombia	Female
LU_LONG_F_147_CTCACCAA-CTAGGCAA_S14	Ricaurte, Colombia	Female
JB_LF22-lutzomyia_female_89_021921_CTCTCTAC-GCGTAAGA_S14	Ricaurte, Colombia	Female
JB_LM7-lutzomyia_male_19_021921_TAAGGCGA-GAGCCTTA_S84	Ricaurte, Colombia	Male
LU_LONG_M_127_GCCACAGG-CATGCCAT_S30	Ricaurte, Colombia	Male
JB_LF7-lutzomyia_female_74_021921_GGACTCCT-CTATTAAG_S54	Ricaurte, Colombia	Female
LU_LONG_F_P19_AGTTCAGG-CCAACAGA_S26	Ricaurte, Colombia	Female
JB_LF18-lutzomyia_female_85_021921_TAGGCATG-TTATGCGA_S9	Ricaurte, Colombia	Female
LU_LONG_F_154_GTATGTTC-AACAGGAA_S18	Ricaurte, Colombia	Female
LU_LONG_M_128_ATTGTGAA-TGCATTGC_S31	Ricaurte, Colombia	Male
LU_LONG_F_135_GACGTCTT-GGTTCACC_S3	Ricaurte, Colombia	Female
JB_LM10-lutzomyia_male_22_021921_CGTACTAG-TTCTAGCT_S57	Ricaurte, Colombia	Male
LU_LONG_F_150_CGCTATGT-TCCGACAC_S17	Ricaurte, Colombia	Female
LU_LONG_F_P11_TTATAACC-GATATCGA_S19	Ricaurte, Colombia	Female

JB_LF11-lutzomyia_female_78_021921_TAGGCATG-TCGACTAG_S2	Ricaurte, Colombia	Female
LU_LONG_F_146_GAACCGCG-TGACCTTA_S13	Ricaurte, Colombia	Female
JB_LF40-lutzomyia_female_107_021921_AAGAGGCA-AAGGCTAT_S34	Ricaurte, Colombia	Female
LU_LONG_F_P18_TGGATCGA-GTGCGATA_S25	Ricaurte, Colombia	Female
LU_LONG_F_137_TGCGAGAC-CATTGTTG_S4	Ricaurte, Colombia	Female
JB_LF25-lutzomyia_female_92_021921_CTCTCTAC-GAGCCTTA_S17	Ricaurte, Colombia	Female
LU_LONG_F_143_CTGCTTCC-GATAGATC_S10	Ricaurte, Colombia	Female
JB_LF47-lutzomyia_female_114_021921_GTAGAGGA-CTATTAAG_S41	Ricaurte, Colombia	Female
JB_LM17-lutzomyia_male_29_021921_AGGCAGAA-TCGACTAG_S64	Ricaurte, Colombia	Male
LU_LONG_F_P16_GCTTGTCA-GAACATAC_S23	Ricaurte, Colombia	Female
JB_LF43-lutzomyia_female_110_021921_GTAGAGGA-TCGACTAG_S37	Ricaurte, Colombia	Female
LU_LONG_F_149_TATCGCAC-CTTAGTGT_S16	Ricaurte, Colombia	Female
JB_LF5-lutzomyia_female_72_021921_GGACTCCT-CCTAGAGT_S52	Ricaurte, Colombia	Female
LU_LONG_F_144_TCATCCTT-AGCGAGCT_S11	Ricaurte, Colombia	Female
JB_LM25-lutzomyia_male_37_021921_TCCTGAGC-TCGACTAG_S73	Ricaurte, Colombia	Male
LU_LONG_F_142_GGTCACGA-GTATTATG_S9	Ricaurte, Colombia	Female
LU_LONG_F_145_AGGTTATA-CAGTTCCG_S12	Ricaurte, Colombia	Female
LU_LONG_F_P20_GACCTGAA-TTGGTGAG_S27	Ricaurte, Colombia	Female

JB_LM16-lutzomyia_male_28_021921_CGTACTAG-TTATGCGA_S63	Ricaurte, Colombia	Male
JB_LM14-lutzomyia_male_26_021921_CGTACTAG-AAGGCTAT_S61	Ricaurte, Colombia	Male
LU_LONG_F_140_GTGAATAT-TCTCATTC_S7	Ricaurte, Colombia	Female
LU_LONG_F_148_TCTGTTGG-TCGAATGG_S15	Ricaurte, Colombia	Female
LU_LONG_F_141_AACTGTAG-ACGCCGCA_S8	Ricaurte, Colombia	Female
LU_LONG_F_P14_AAGTCCAA-TATGAGTA_S21	Ricaurte, Colombia	Female
JB_LM13-lutzomyia_male_25_021921_CGTACTAG-CTATTAAG_S60	Ricaurte, Colombia	Male
LU_LONG_F_P17_CAAGCTAG-ACATAGCG_S24	Ricaurte, Colombia	Female
LU_LONG_F_139_ACAGGCGC-CTCTGCCT_S6	Ricaurte, Colombia	Female
LU_LONG_F_P15_ATCCACTG-AGGTGCGT_S22	Ricaurte, Colombia	Female
JB_LF26-lutzomyia_female_93_021921_CTCTCTAC-TTATGCGA_S18	Ricaurte, Colombia	Female
JB_LF15-lutzomyia_female_82_021921_TAGGCATG-CTATTAAG_S6	Ricaurte, Colombia	Female
SRR3109093	Jacobina, Brazil	Female
SRR3109094	Jacobina, Brazil	Female
SRR3109095	Jacobina, Brazil	Female
SRR3109096	Jacobina, Brazil	Female
SRR3109097	Jacobina, Brazil	Female
SRR3109098	Jacobina, Brazil	Female
SRR3109099	Jacobina, Brazil	Female
SRR3109100	Jacobina, Brazil	Female
SRR3109101	Jacobina, Brazil	Female
SRR3109102	Jacobina, Brazil	Female
SRR3109103	Jacobina, Brazil	Female

SRR3109104	Jacobina, Brazil	Female
SRR3109105	Jacobina, Brazil	Female
SRR3109106	Jacobina, Brazil	Female
SRR3109107	Laphina, Brazil	Female
SRR3109108	Laphina, Brazil	Female
SRR3109109	Laphina, Brazil	Female
SRR3109110	Laphina, Brazil	Female
SRR3109111	Laphina, Brazil	Female
SRR3109112	Laphina, Brazil	Female
SRR3109113	Laphina, Brazil	Female
SRR3109114	Laphina, Brazil	Female
SRR3109115	Laphina, Brazil	Female
SRR3109116	Laphina, Brazil	Female
SRR3109117	Laphina, Brazil	Female
SRR3109118	Laphina, Brazil	Female
SRR3109119	Laphina, Brazil	Female
SRR3109120	Laphina, Brazil	Female
SRR3109121	Laphina, Brazil	Female
SRR3109122	Marajo, Brazil	Female
SRR3109123	Marajo, Brazil	Female
SRR3109124	Marajo, Brazil	Female
SRR3109125	Marajo, Brazil	Female
SRR3109126	Marajo, Brazil	Female
SRR3109127	Marajo, Brazil	Female
SRR3109128	Marajo, Brazil	Female
SRR3109129	Marajo, Brazil	Female
SRR3109130	Marajo, Brazil	Female
SRR3109131	Marajo, Brazil	Female
SRR3109132	Sobral 1S, Brazil	Female
SRR3109133	Sobral 1S, Brazil	Female

SRR3109134	Sobral 1S, Brazil	Female
SRR3109135	Sobral 1S, Brazil	Female
SRR3109136	Sobral 1S, Brazil	Female
SRR3109137	Sobral 1S, Brazil	Female
SRR3109138	Sobral 1S, Brazil	Female
SRR3109139	Sobral 1S, Brazil	Female
SRR3109140	Sobral 1S, Brazil	Female
SRR3109141	Sobral 1S, Brazil	Female
SRR3109142	Sobral 1S, Brazil	Female
SRR3109143	Sobral 1S, Brazil	Female
SRR3109144	Sobral 1S, Brazil	Female
SRR3109145	Sobral 1S, Brazil	Female
SRR3109146	Sobral 1S, Brazil	Female
SRR3109148	Sobral 2S, Brazil	Female
SRR3109149	Sobral 2S, Brazil	Female
SRR3109150	Sobral 2S, Brazil	Female
SRR3109151	Sobral 2S, Brazil	Female
SRR3109152	Sobral 2S, Brazil	Female

SRR3109153	Sobral 2S, Brazil	Female
SRR3109154	Sobral 2S, Brazil	Female
SRR3109155	Sobral 2S, Brazil	Female
SRR3109156	Sobral 2S, Brazil	Female
SRR3109157	Sobral 2S, Brazil	Female
SRR3109158	Sobral 2S, Brazil	Female
SRR3109159	Sobral 2S, Brazil	Female
SRR3109160	Sobral 2S, Brazil	Female
SRR3109161	Sobral 2S, Brazil	Female
SRR3109162	Sobral 2S, Brazil	Female
SRR3109163	Sobral 2S, Brazil	Female

Table S2. Significant GWAS hits. The table shows the name, location, association p-value, and known orthologs in other dipterans.

chrom	pos	p	gene_type	gene_id	gene_distance	Flybase_gene_name	gene_name
2	11287131	1.82E-08	Downstream	g4778	2049		
2	11287131	1.82E-08	Upstream	g4777	23544		
2	17800961	5.17E-09	Downstream	g5201	165		uncharacterized protein
2	17800961	5.17E-09	Upstream	g5200	35253		eye-specific diacylglycerol kinase-like

2	19878643	1.12E-09	Genic	g5391			
3	3650443	2.59E-09	Downstream	g6882	36835	CG17870	14-3-3 protein zeta
3	3650443	2.59E-09	Upstream	g6881	1646		
3	8753610	8.95E-08	Genic	g7272		CAA65152.1	mdg3 element ORF
3	8756972	4.43E-09	Genic	g7272		CAA65152.1	mdg3 element ORF
3	8757227	5.06E-08	Genic	g7272		CAA65152.1	mdg3 element ORF
3	14121068	2.91E-09	Downstream	g7711	14311		
3	14121068	2.91E-09	Upstream	g7710	14987		
3	14909442	1.22E-08	Downstream	g7789	4284		Orf2
3	14909442	1.22E-08	Upstream	g7788	6504	CG15715	zinc finger protein 706
3	14909452	4.62E-09	Downstream	g7789	4274		Orf2
3	14909452	4.62E-09	Upstream	g7788	6514	CG15715	zinc finger protein 706
3	15059348	4.94E-08	Downstream	g7799	3787		integrator complex subunit 12 #(low query coverage/confidence)
3	15059348	4.94E-08	Upstream	g7798	8033		uncharacterized protein K02A2.6-like
3	15061297	2.94E-08	Downstream	g7799	1838		integrator complex subunit 12 #(low query coverage/confidence)
3	15061297	2.94E-08	Upstream	g7798	9982		uncharacterized protein K02A2.6-like
3	15062377	2.89E-09	Downstream	g7799	758		integrator complex subunit 12 #(low query coverage/confidence)

3	15062377	2.89E-09	Upstream	g7798	11062		uncharacterized protein K02A2.6-like
3	27024829	7.38E-09	Downstream	g8767	7639	AAD37797	pleiotrophin isoform
3	27024829	7.38E-09	Upstream	g8766	15256		
3	27024841	3.28E-08	Downstream	g8767	7627	AAD37797	pleiotrophin isoform
3	27024841	3.28E-08	Upstream	g8766	15268		
3	27024846	4.65E-08	Downstream	g8767	7622	AAD37797	pleiotrophin isoform
3	27024846	4.65E-08	Upstream	g8766	15273		
3	27024848	4.27E-08	Downstream	g8767	7620	AAD37797	pleiotrophin isoform
3	27024848	4.27E-08	Upstream	g8766	15275		
4	89619	6.71E-08	Downstream	g8996	39230		na
4	89619	6.71E-08	Upstream	g8995	10970		LIM/homeobox protein Lhx2 #(low confidence)
4	477670	1.12E-09	Genic	g9046		CG6829	uncharacterized protein
4	477673	1.12E-09	Genic	g9046		CG6829	uncharacterized protein
4	477798	1.03E-08	Genic	g9046		CG6829	uncharacterized protein
4	477808	3.89E-09	Genic	g9046		CG6829	uncharacterized protein
4	4318823	1.76E-09	Downstream	g9424	853	CG2095	exocyst complex component 4
4	4318823	1.76E-09	Upstream	g9423	355	AAL13990.1	heat shock factor protein
4	16129633	1.33E-09	Downstream	g10362	4748	CG8333	enhancer of split mgamma protein

4	16129633	1.33E-09	Upstream	g10361	18627		
4	16129745	1.13E-09	Downstream	g10362	4636	CG8333	enhancer of split mgamma protein
4	16129745	1.13E-09	Upstream	g10361	18739		
4	16129961	7.69E-08	Downstream	g10362	4420	CG8333	enhancer of split mgamma protein
4	16129961	7.69E-08	Upstream	g10361	18955		
4	16130036	1.61E-09	Downstream	g10362	4345	CG8333	enhancer of split mgamma protein
4	16130036	1.61E-09	Upstream	g10361	19030		
4	16130047	1.61E-09	Downstream	g10362	4334	CG8333	enhancer of split mgamma protein
4	16130047	1.61E-09	Upstream	g10361	19041		
4	16131132	1.82E-08	Downstream	g10362	3249	CG8333	enhancer of split mgamma protein
4	16131132	1.82E-08	Upstream	g10361	20126		
4	18327554	9.57E-09	Downstream	g10493	670	CG6975	tuberin
4	18327554	9.57E-09	Upstream	g10491	349	CG11660	serine/threonine-protein kinase RIO1
4	20445947	8.85E-09	Downstream	g10649	17732		
4	20445947	8.85E-09	Upstream	g10648	5814		uncharacterized protein
4	20445956	6.66E-10	Downstream	g10649	17723		
4	20445956	6.66E-10	Upstream	g10648	5823		uncharacterized protein

Table S3. GO categories showing enrichment in the GWAS hits.

Enrichment	FDR	nGenes	Genes	Enrichment	GO ID

		Pathway	Fold	Pathways	
1.9E-03	3	106	49.4	Positive reg. of nervous system development	GO:0051962
9.6E-04	4	215	32.5	Reg. of nervous system development	GO:0051960
9.6E-04	4	242	28.9	Positive reg. of developmental proc.	GO:0051094
5.4E-03	3	197	26.6	Positive reg. of multicellular organismal proc.	GO:0051240
1.2E-03	4	281	24.9	Reg. of multicellular organismal development	GO:2000026
7.6E-03	3	234	22.4	Reg. of cell population proliferation	GO:0042127
3.3E-03	4	439	15.9	Response to abiotic stimulus	GO:0009628
1.4E-03	5	691	12.6	Reg. of developmental proc.	GO:0050793
5.9E-03	4	572	12.2	Reg. of multicellular organismal proc.	GO:0051239
1.1E-03	6	1176	8.9	Nervous system development	GO:0007399
2.0E-03	6	1443	7.3	Cell development	GO:0048468
9.6E-04	7	1739	7	Positive reg. of cellular proc.	GO:0048522
2.7E-03	6	1539	6.8	System development	GO:0048731
9.6E-04	7	1892	6.5	Positive reg. of biological proc.	GO:0048518
4.2E-03	6	1782	5.9	Cell differentiation	GO:0030154
4.2E-03	6	1790	5.9	Cellular developmental proc.	GO:0048869
4.2E-03	7	2831	4.3	Anatomical structure development	GO:0048856
5.3E-03	7	3005	4.1	Developmental proc.	GO:0032502
3.3E-03	8	4133	3.4	Reg. of cellular proc.	GO:0050794

SUPPLEMENTARY FIGURES

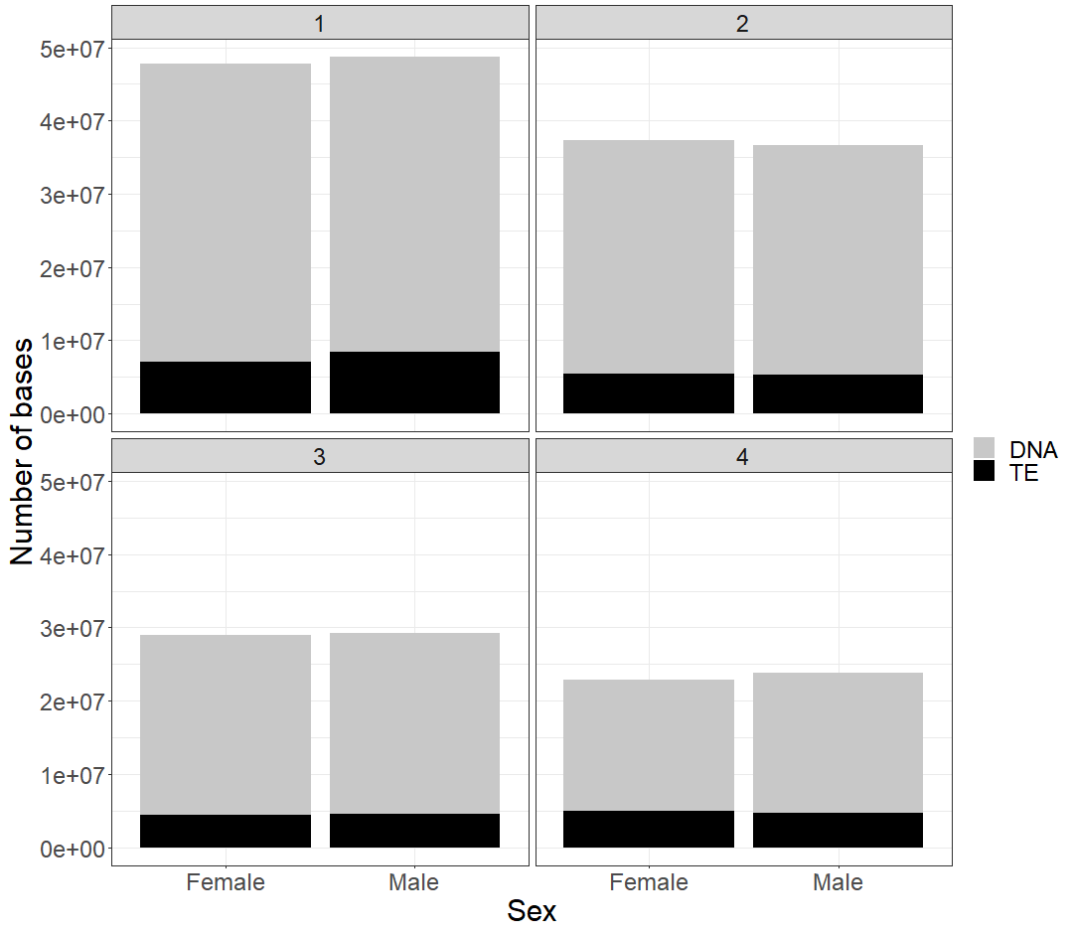


Figure S1. Overall number of bases classified as transposable elements (TE) and not classified as TE (DNA in figure legend). Facets represent the assembled chromosomes. Cumulative stacked bars represent total nucleotide bases per chromosome by sex.

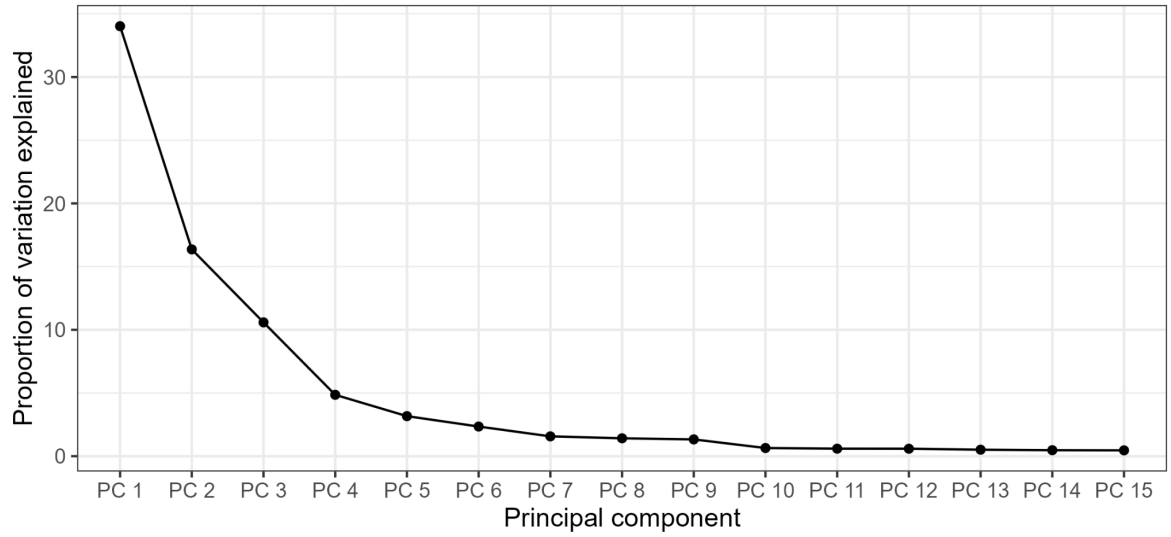


Figure S2. Scree plot of genomic variation in both Colombian and Brazilian sandflies explained. The proportion of variation explained is less than 3 after principal component 5.

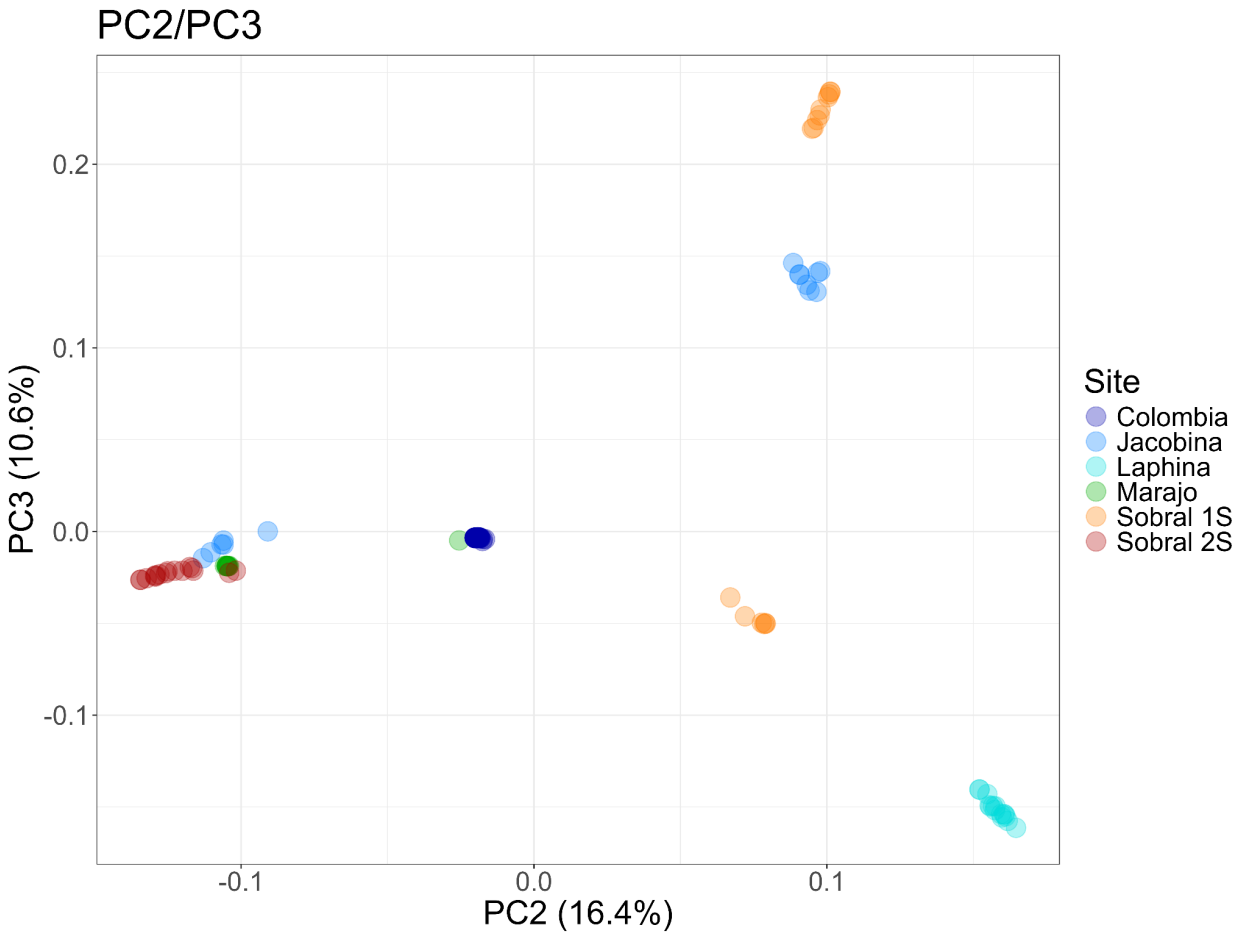
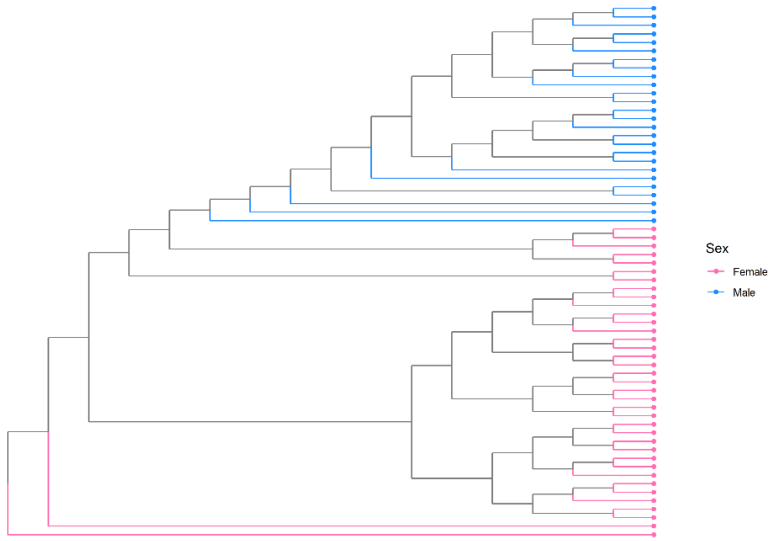
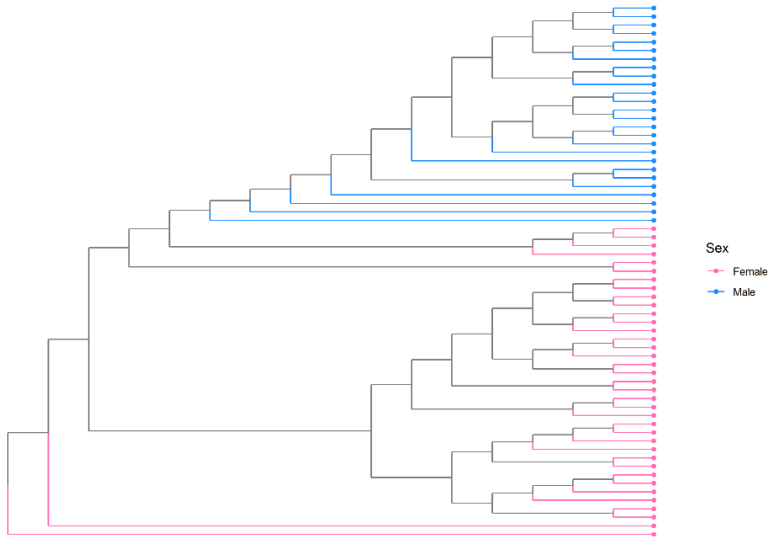


Figure S3. Principal components 2 and 3 from whole genome resequence data.

Whole genome



Chr1



Autosomes

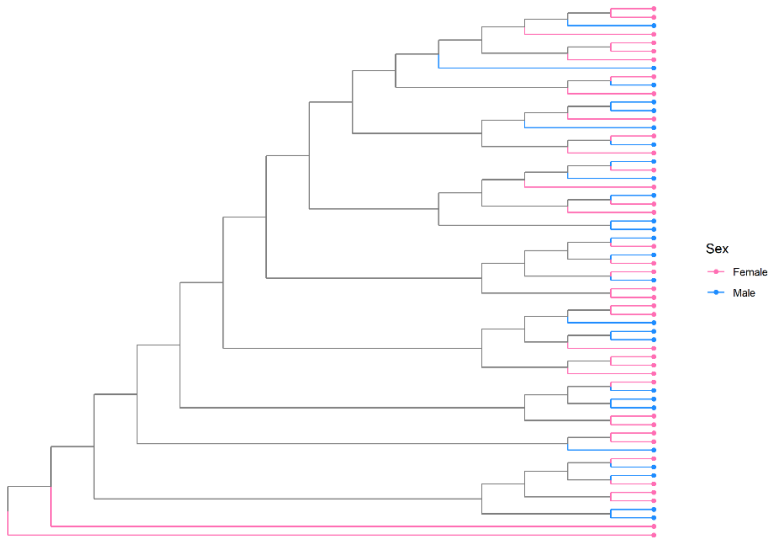


Figure S4. Phylogenetic trees of inter-relatedness between individuals across the whole genome, chromosome 1, and autosomes.

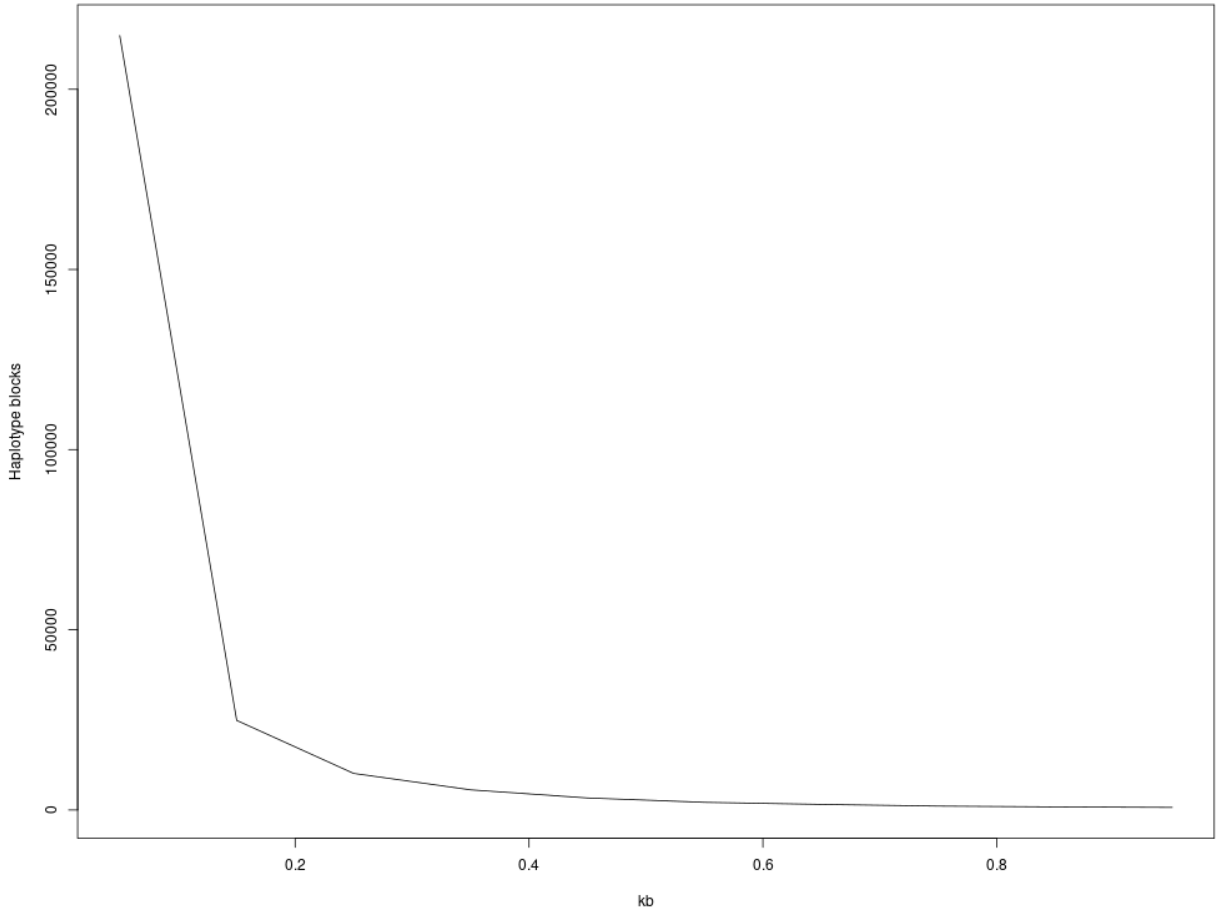


Figure S5. Linkage disequilibrium decays rapidly in *Lutzomyia longipalpis* with most haplotype blocks < 200bp.

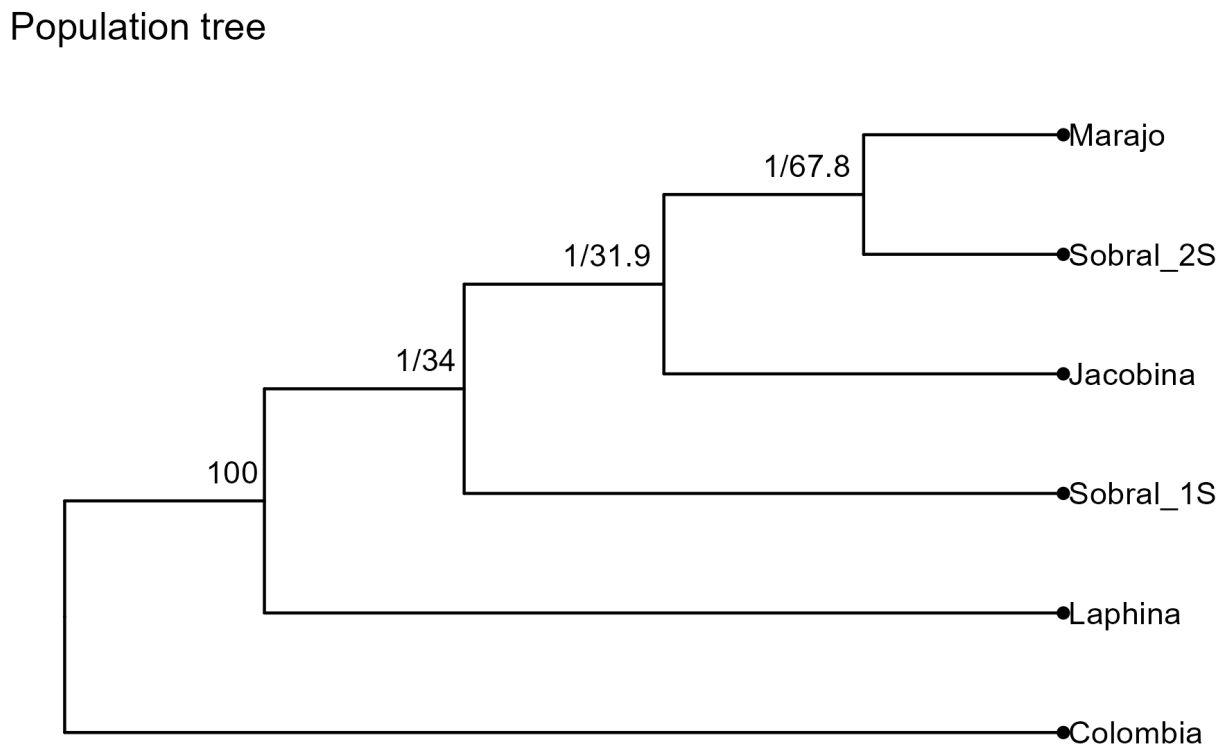


Figure S6. Consensus population tree from ASTRAL. The second values at the branches indicate gene concordance values from IQTREE2.

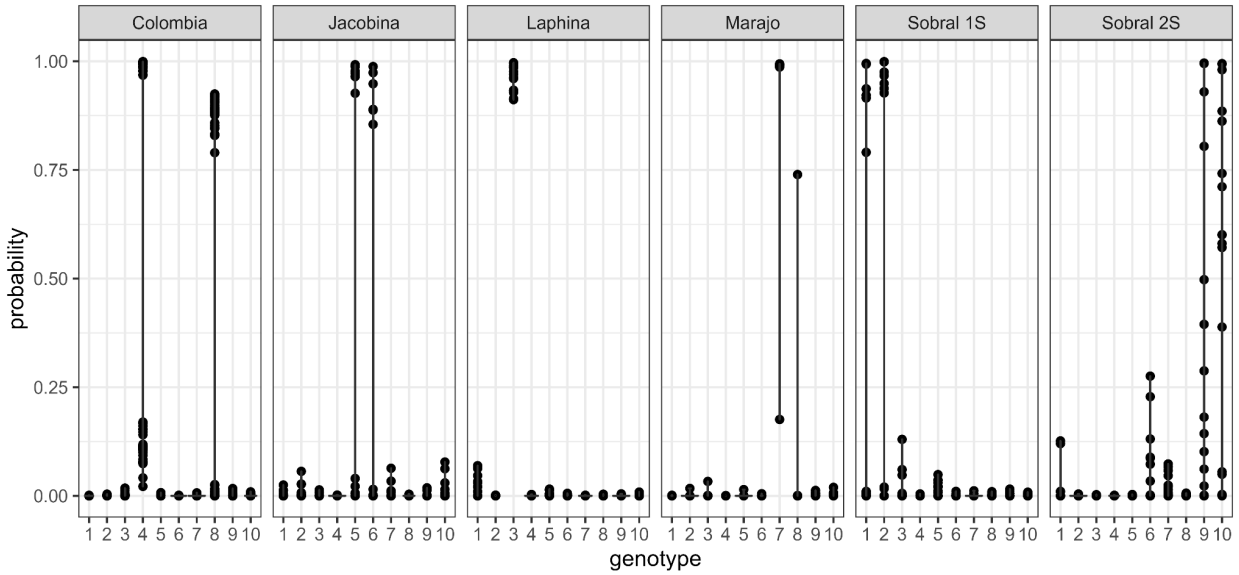


Figure S7. Admixture proportions by genotype for each individual sandfly.

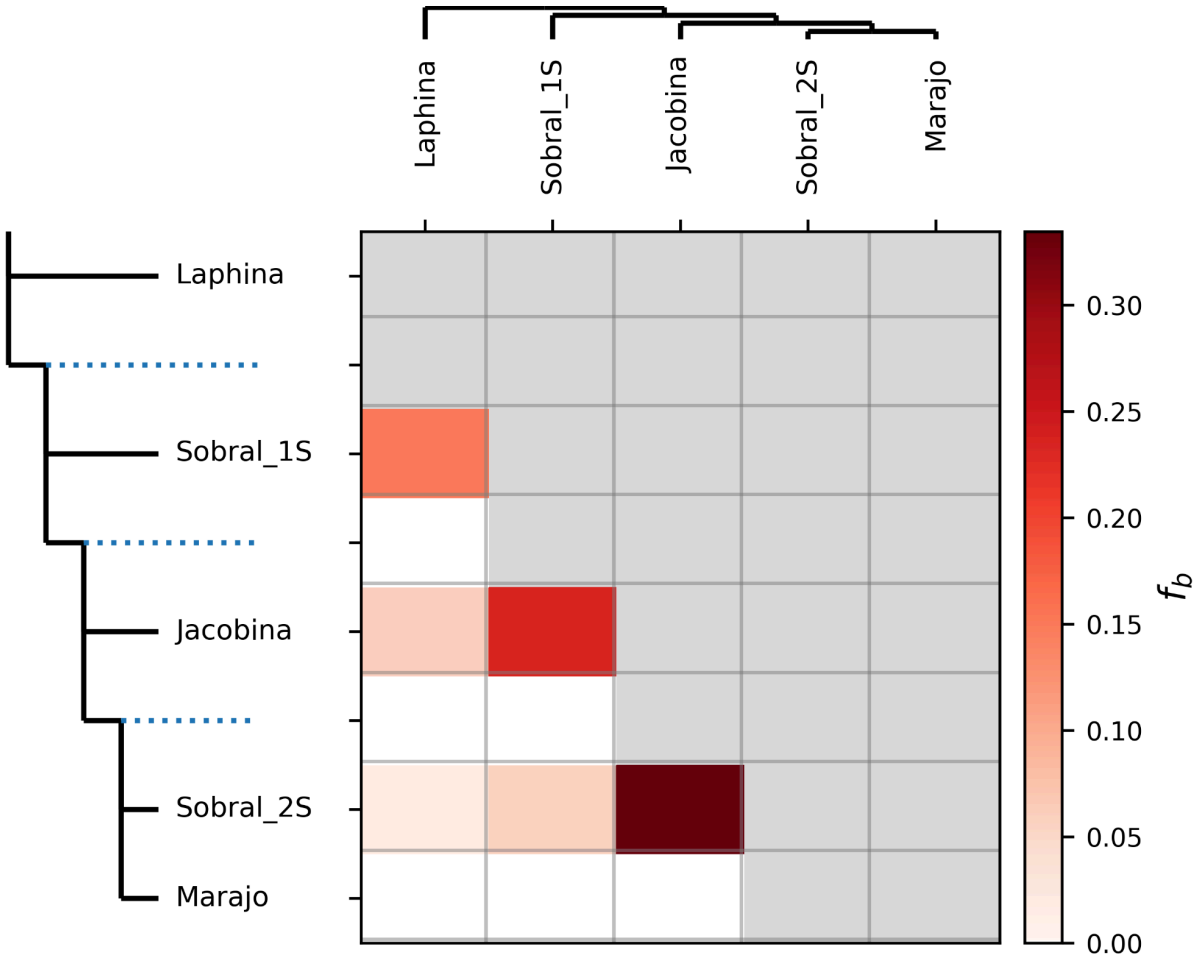


Figure S8. Fbranch statistics and gene flow inferred by Dsuite.

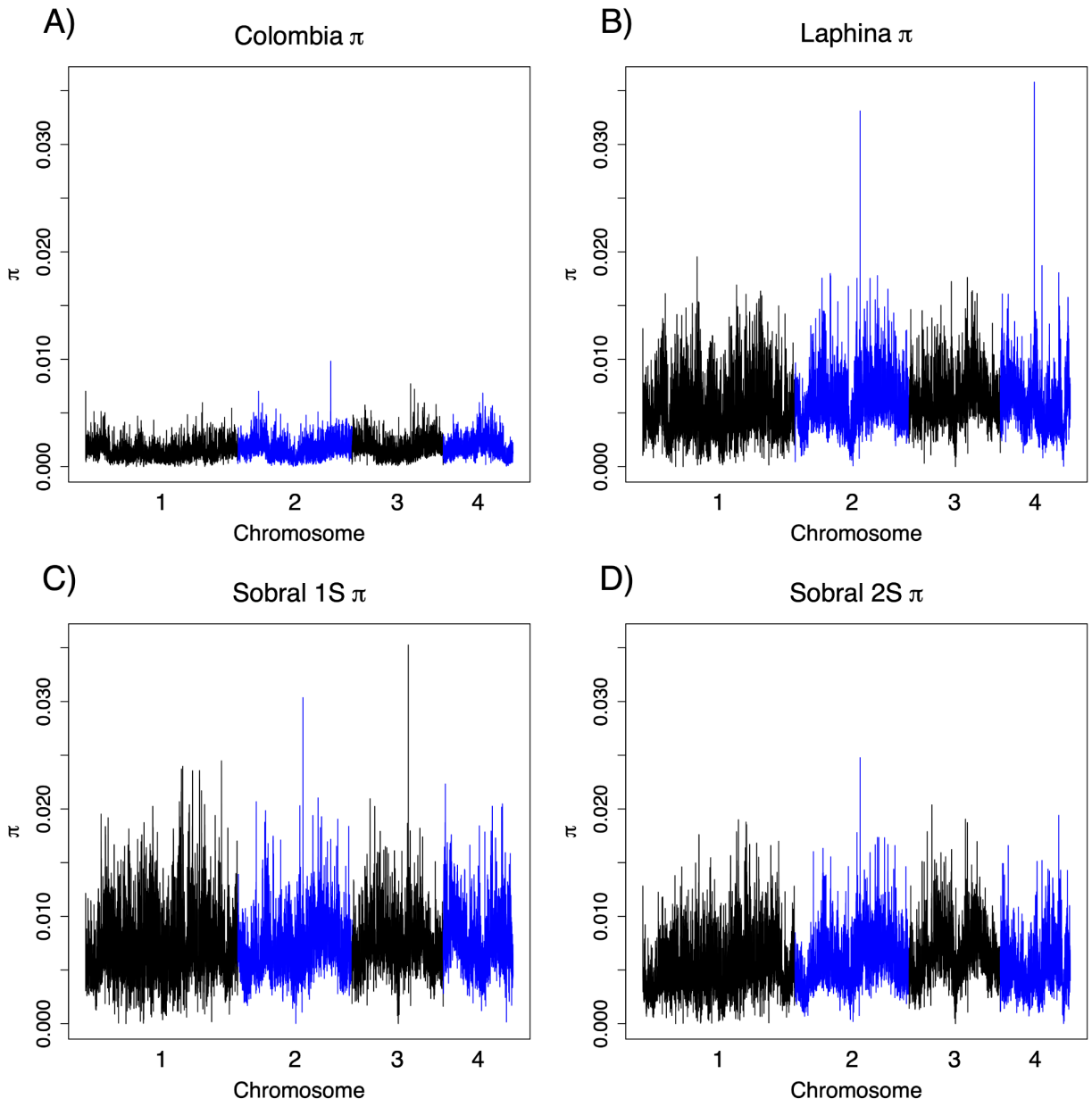


Figure S9. π in 10kb windows. **A.** π along the genome in Colombia. **B.** π along the genome in Laphina. **C.** π along the genome in Sobral 1S. **D.** π along the genome in Sobral 2S.

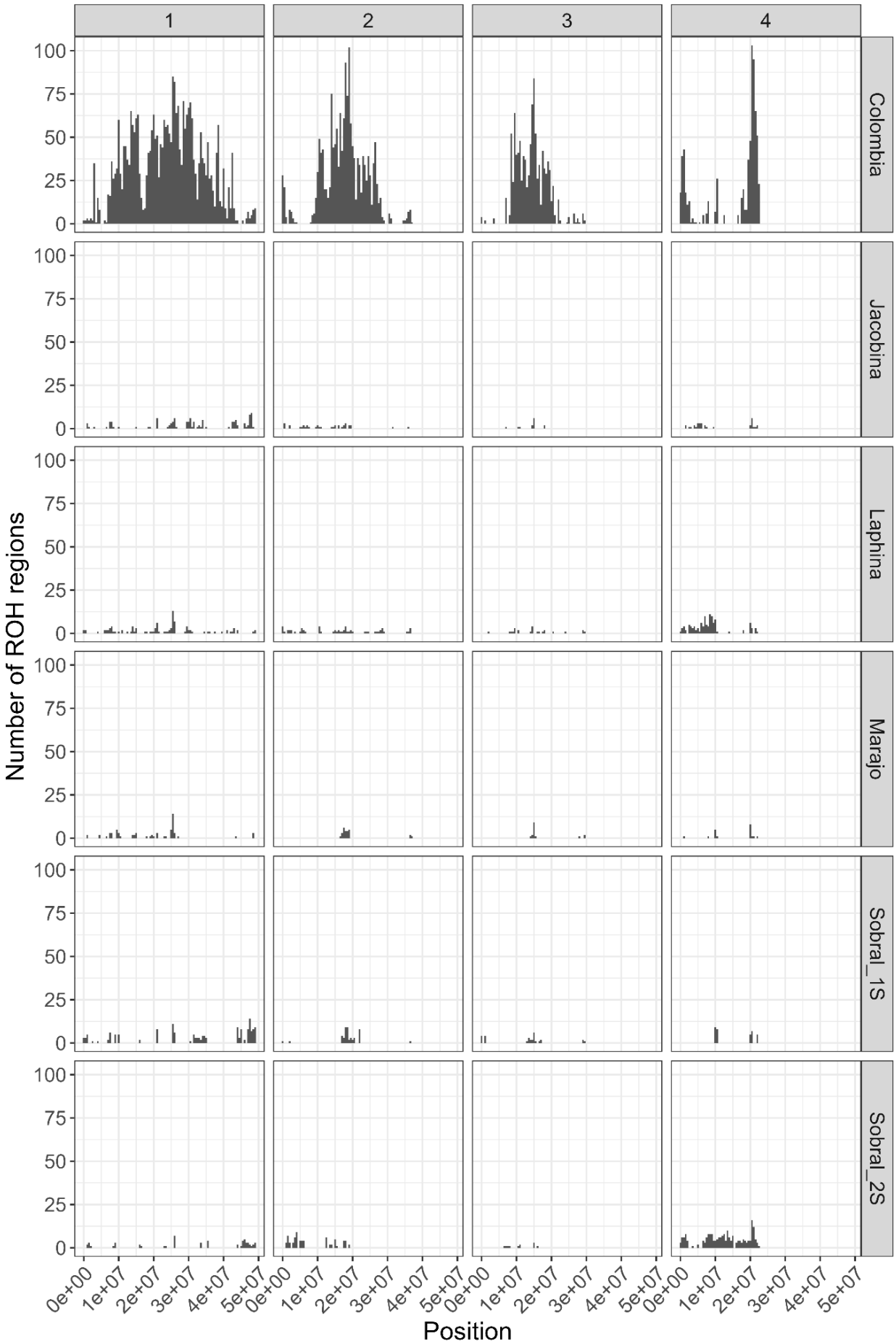

1
2
3
4
Colombia
Jacobina
Laphina
Marajo
Sobral_1S
Sobral_2S
Number of ROH regions
Position
0
25
50
75
100
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07

Figure S10. Runs of Homozygosity in genomic regions from each population.

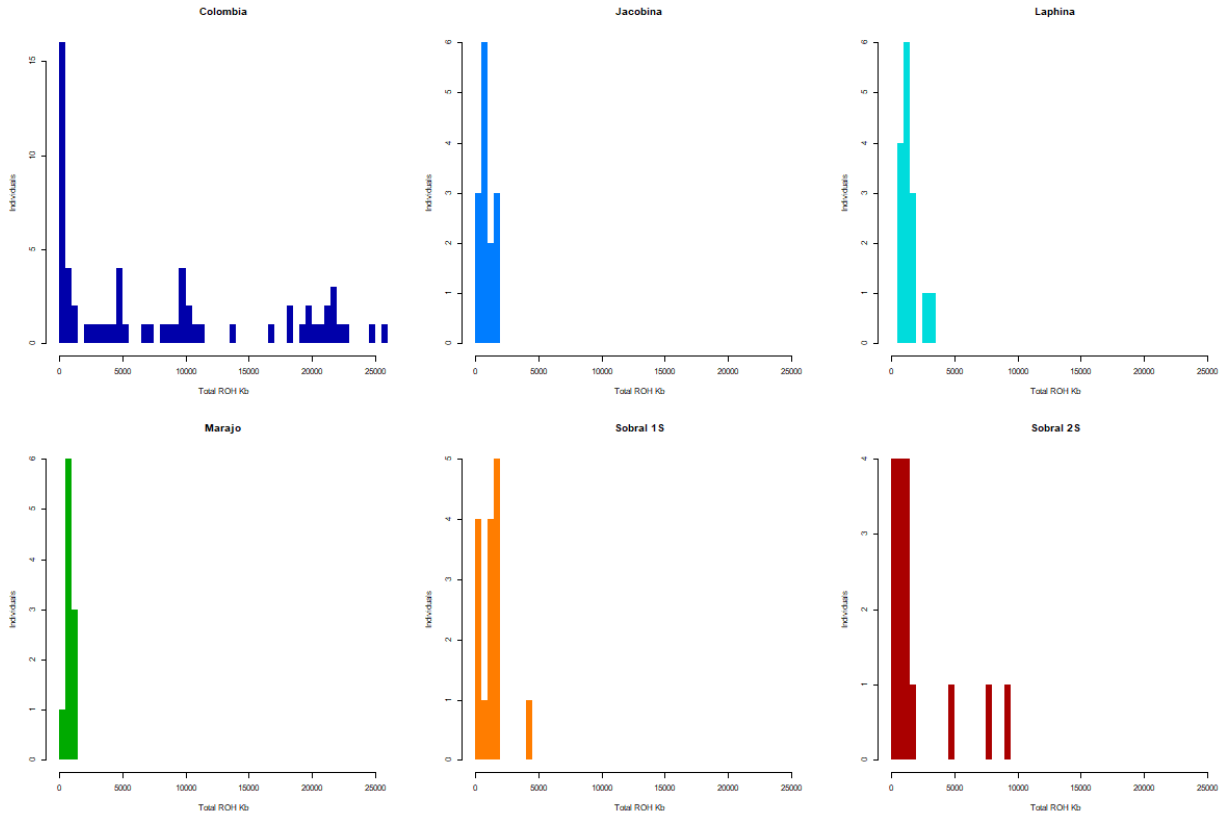


Figure S11. Histograms of ROH distributions by individual, clearly showing that the data from Colombia is not driven by a single individual.

DNA Extraction Protocol

Materials Needed:

- Zymo Genomic DNA Clean & Concentrator Kit
- Homogenization buffer (HB)
- 30mM Tris HCL (pH 8.0)
- 10mM EDTA
- 0:1M NaCl
- 0.5% Triton X-100
- SDS 10%
- RNase A
- Proteinase K

Note: This protocol is adapted from Dr. Jeremy Wang's high molecular weight extraction protocol for nanopore sequencing. Be sure that all solutions from the Zymo kit have the appropriate volume of 100% ethanol added.

1. Add 500 µL of HB to a 1:5mL tube with 20-40 frozen flies and then homogenize the flies using a clean pestle until you can no longer recognize any fly parts.
2. Add 125 µL of SDS and 10 µL of Proteinase K then vortex for 15 seconds.
3. Incubate at 55C for 1 hour (inverting every 10 minutes) and after 30 minutes add 10 µL of RNase A.
4. 4. Add twice the volume of ChIP DNA Binding Buffer to each volume of DNA sample and mix thoroughly.
5. Pipette the mixture into a Zymo-Spin Column in a collection tube, centrifuge at 16,000 x g for 30 seconds, and discard the flow-through.
6. Add 200 µL of DNA Wash Buffer to the Zymo-Spin Column, centrifuge at 16,000 x g for 1 minute, and discard the flow-through.
7. Repeat step 6.
8. Transfer the Zymo-Spin Column to a 1:5mL tube and to elute the DNA add 32 µL of DNA Elution Buffer directly onto the column matrix and centrifuge at 16,000 x g for 30 seconds.
9. Store the eluted DNA in the cold room until quantification.