

## 1 Parasite hybridization promotes spreading of endosymbiotic viruses

2 Senne Heeren<sup>1,2,3</sup>, Ilse Maes<sup>1</sup>, Mandy Sanders<sup>4</sup>, Lon-Fye Lye<sup>5</sup>, Jorge Arevalo<sup>6</sup>, Alejandro  
3 Llanos-Cuentas<sup>6</sup>, Lineth Garcia<sup>7</sup>, Philippe Lemey<sup>2</sup>, Stephen M Beverley<sup>5</sup>, James A Cotton<sup>4</sup>,  
4 Jean-Claude Dujardin<sup>1,3,\*</sup>, Frederik Van den Broeck<sup>1,2,\*</sup>

5

6 <sup>1</sup> Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

7 <sup>2</sup> Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical  
8 Research, Katholieke Universiteit Leuven, Leuven, Belgium

9 <sup>3</sup> Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium

10 <sup>4</sup> Parasite Genomics Group, Wellcome Sanger Institute, Hinxton, United Kingdom

11 <sup>5</sup> Department of Molecular Microbiology, Washington University School of Medicine, St.  
12 Louis, United States

13 <sup>6</sup> Instituto de Medicina Tropical Alexander von Humboldt, Universidad Peruana Cayetano  
14 Heredia, Lima, Peru

15 <sup>7</sup> Instituto de Investigación Biomédicas e Investigación Social, Universidad Mayor de San  
16 Simon, Cochabamba, Bolivia

17

### 18 \*Correspondence:

19 Frederik Van den Broeck ([fvandenbroeck@gmail.com](mailto:fvandenbroeck@gmail.com))

20 Jean-Claude Dujardin ([icdujardin@itg.be](mailto:icdujardin@itg.be))

21

22

### 23 ABSTRACT

24 Viruses are the most abundant biological entities on Earth and play a significant role in  
25 the evolution of many organisms and ecosystems. In pathogenic protozoa, the presence of  
26 endosymbiotic viruses has been linked to an increased risk of treatment failure and severe  
27 clinical outcome. Here, we studied the molecular epidemiology of the zoonotic disease  
28 cutaneous leishmaniasis in Peru and Bolivia through a joint evolutionary analysis of  
29 *Leishmania braziliensis* parasites and their endosymbiotic *Leishmania* RNA virus. We show  
30 that parasite populations circulate in isolated pockets of suitable habitat and are associated  
31 with single viral lineages that appear in low prevalence. In contrast, groups of hybrid parasites  
32 were geographically and ecologically dispersed, and commonly infected from a pool of  
33 genetically diverse viruses. Our results suggest that parasite hybridization, likely due to  
34 increased human migration and ecological perturbations, increased the frequency of  
35 endosymbiotic interactions known to play a key role in disease severity.

## 36 INTRODUCTION

37 Viruses have the ability to infect virtually any cellular life form on Earth. Particularly  
38 fascinating are RNA viruses that infect simple eukaryotes<sup>1-3</sup>, some of which have important  
39 biological roles such as limiting fungal pathogenicity<sup>4</sup> or increasing protist fecundity<sup>5</sup>. Among  
40 the RNA viruses, the double-stranded RNA (dsRNA) totivirids have evolved significant diversity  
41 and are present in phyla separated by billion years of evolution, with closely related viruses  
42 identified in almost all genera of yeasts, fungi and protozoa studied<sup>6</sup>. Totivirids have no lytic  
43 infectious phase and thus adopted a lifestyle of coexistence favoring long-term symbiotic  
44 persistence, passing from cell to cell mainly through mating and cell division<sup>7</sup>. Because of this  
45 intimate association, it is postulated that these viruses have a mutualistic co-evolutionary  
46 history with their hosts<sup>8</sup>.

47 Totivirids encompass most viral endosymbionts identified in pathogenic protozoa causing  
48 widespread severe illnesses such as trichomoniasis, giardiasis and leishmaniasis<sup>6</sup>. An icon  
49 group of human pathogenic parasites is the genus *Leishmania* (Family Trypanosomatidae),  
50 causing the vector-borne disease leishmaniasis in about 88 countries, mainly in the tropics  
51 and subtropics<sup>9</sup>. Members of the *Leishmania* genus are associated with the *Leishmania* RNA  
52 virus (LRV) (Family *Totiviridae*), forming a tripartite symbiosis with the mammalian or  
53 arthropod host. Phylogenetic studies suggest that the virus was most likely present in the  
54 common ancestor of *Leishmania*, prior to the divergence of these parasites into different  
55 species around the world<sup>8,10</sup>. This is reflected by the two types of viruses (LRV1 and LRV2)  
56 that are carried by members of the subgenera *Viannia* and *Leishmania*, respectively<sup>11</sup>. It was  
57 shown that the dsRNA of LRV1 is recognized by Toll-like receptor 3 (TLR3) which directly  
58 activates a hyperinflammatory response causing increased disease pathology, parasite  
59 numbers and immune response in murine models<sup>12</sup>. In human infections, the presence of  
60 LRV1 has been associated with an increased risk of drug-treatment failures and acute  
61 pathology<sup>13,14</sup>. The virus thus confers survival advantage to *Leishmania* and plays a key role  
62 in the severity of human leishmaniasis<sup>15,16</sup>.

63 Given the epidemiological and biomedical relevance of the *Leishmania*-LRV symbiosis<sup>17</sup>,  
64 there is a clear need to understand the diversity and dissemination of the virus in parasite  
65 populations. To this end, we investigated the joint evolutionary history of *L. braziliensis* (*Lb*)  
66 parasites and LRV1 using whole genome sequencing data. The *Lb* parasite is a zoonotic  
67 pathogen circulating mainly in rodents and other wild mammals (e.g. marsupials) in  
68 Neotropical rainforests<sup>18</sup>. The parasite is the major cause of cutaneous leishmaniasis (CL) in  
69 Central and South America and occasionally develops the disfiguring mucocutaneous disease  
70 where the parasite spreads to mucosal tissue. Our previous work in Peru has shown that LRV1  
71 was present in >25% of sampled *Lb* parasites, and that the virus was significantly associated  
72 with an increased risk of treatment failure<sup>14</sup>. Here, we characterize the co-evolutionary  
73 dynamics between *Lb* and LRV1 from Peru and Bolivia.

74

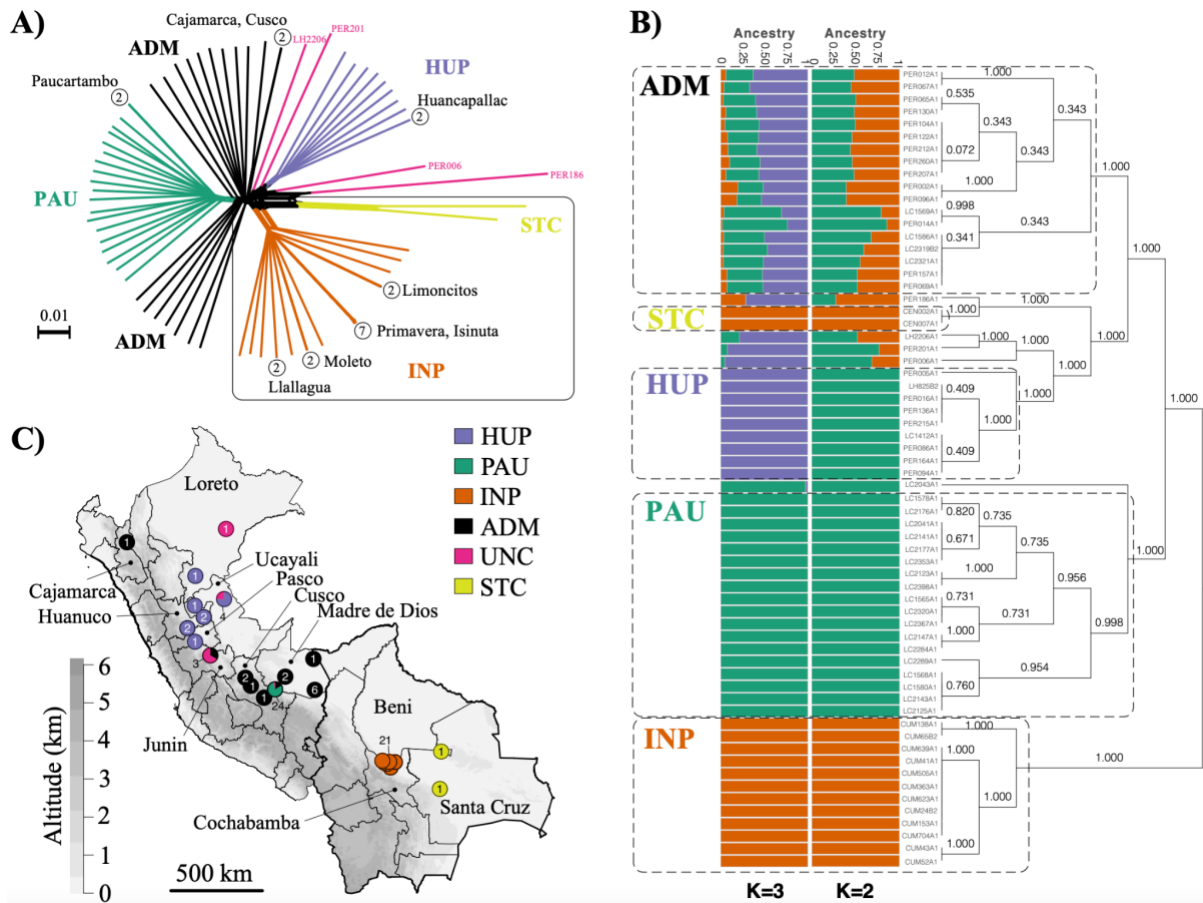
## 75 RESULTS

### 76 Natural genome diversity of *Lb* parasites

77 Our study included 79 *Lb* isolates (Supp. Table 1; Supp. Fig. 1) from Peru (N=55) and  
78 Bolivia (N=24) that were sampled during various studies on the genetics and epidemiology of  
79 leishmaniasis <sup>14,19–23</sup>, and for which the LRV infection status was previously characterized <sup>14</sup>.  
80 Sequence reads of *Lb* were aligned against the *Lb* M2904 reference genome, resulting in a  
81 median coverage of 58x (min=35x, max=121x).

82 Variant discovery was done with GATK HaplotypeCaller to uncover high-quality Single  
83 Nucleotide Polymorphisms (SNPs) and small insertions/deletions (INDELs). The number of  
84 SNPs identified between each genome and the M2904 reference genome was relatively  
85 consistent across the panel, varying between 97,777 SNPs in CEN002 and 109,798 SNPs in  
86 LC2319 (median = 105,096; mean = 106,530). Exceptions were isolates PER231 (126,178),  
87 LC2318 (128,544) and CUM68 (134,882) that showed larger SNP densities and double the  
88 number of heterozygous sites compared to the other isolates (Supp. Fig. 2). When  
89 investigating the genome-wide distribution of normalized allele frequencies at heterozygous  
90 sites (which should be centered around 0.5 in diploid individuals) <sup>24</sup>, we discovered that  
91 isolates CUM68, LC2318 and PER231 showed unbalanced read counts, symptomatic of  
92 tetraploidy <sup>24</sup>, although we could not rule out contamination or a mixed infection (Supp. Fig.  
93 3). Removing these three isolates for downstream analyses, the resulting dataset comprised  
94 a total of 407,070 SNPs and 69,604 bi-allelic INDELs called across 76 *Lb* genomes. The SNP  
95 allele frequency spectrum was dominated by low-frequency variants, with over 66% of SNPs  
96 being at  $\leq 1\%$  MAF. Respectively 41.5% and 5.7% of the SNPs and INDELs were found in the  
97 coding region of the genome, including 437 SNPs and 55 INDELs with a deleterious impact  
98 (e.g. introducing stop codons). Most of these deleterious mutations were rare in our panel of  
99 parasites (83.6% being at  $\leq 1\%$  MAF), and the remaining 80 mutations were not linked to *Lb*  
100 population structure or LRV infection status (Supp. Table 2).

101 Chromosome and gene copy numbers were investigated using normalized median read  
102 depths. This revealed that the majority of chromosomes were diploid (Supp. Fig. 4).  
103 Chromosome 31 was highly polysomic with at least four copies present in all isolates, a  
104 consistent observation in *Leishmania* <sup>25,26</sup>. Chromosomes 10, 26, 32 and 34 were diploid in all  
105 *Lb* isolates (Supp. Fig. 4). Excluding chromosome 31, we found 21 *Lb* isolates that were entirely  
106 diploid and 25 *Lb* isolates with at least 5 polysomic chromosomes, including 6 *Lb* isolates with  
107 at least 10 polysomic chromosomes (Supp. Fig. 4). When investigating variation in copy  
108 numbers for 8,573 coding DNA sequences, we found 201-286 amplifications and 13-33  
109 deletions per *Lb* isolate (Supp. Table. 3), as expected for *Leishmania* and other eukaryotic  
110 genomes <sup>24</sup>. Variation in chromosome and gene copy numbers was not associated with *Lb*  
111 population structure or LRV infection status.



112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

### **Epidemic clones against a background of prevalent recombination in *Lb***

130

131

132

133

134

It has been postulated since 1990 that protozoan parasites may have a predominantly clonal mode of reproduction and that sexual recombination events are rare<sup>27</sup>, although this theory has been the subject of intense debate for more than 30 years<sup>28</sup>. For *Lb*, studies using multilocus microsatellite profiles revealed contradictory results, including moderate degrees of inbreeding in populations from Peru and Bolivia<sup>19,29</sup> and significant levels of recombination

135 in populations from the Brazilian Atlantic Coast<sup>30</sup>. These results may be biased due to the  
136 presence of population substructure (Wahlund effect) or because of the low discriminatory  
137 power of the molecular markers used<sup>28</sup>.

138 Our phylogenetic network based on genome-wide SNPs revealed a star-like topology  
139 whereby the majority of isolates were separated by long branches (Fig. 1A), a pattern  
140 symptomatic of recombination. Indeed, levels of linkage-disequilibrium (LD) were low ( $r^2$   
141 decayed to  $<0.1$  within  $<20\text{bp}$ ) (Supp. Fig. 5) and distributions of per-site inbreeding  
142 coefficients per population were unimodal and centered around zero (Supp. Fig. 6; Supp.  
143 Table 4), after correcting for population structure and spatio-temporal Wahlund effects. Our  
144 genome-scale data thus indicate that the distinct populations in Peru and Bolivia are  
145 approximately in Hardy-Weinberg and linkage equilibrium, suggesting that recombination  
146 may be a prevalent and universal process in this species.

147 While the majority of *Lb* genomes differed at an average 9,866 fixed SNP differences  
148 (median 9,494), we also identified seven small groups of isolates with near-identical genomes  
149 that cluster terminally in the phylogenetic network (Fig. 1A, circles with numbers). Isolates  
150 within each of these groups displayed no or few fixed SNP differences and a relatively small  
151 amount of heterozygous SNP differences (Supp. Table 5), and are thus likely the result of  
152 clonal propagation in the wild. With the exception of a pair of near-identical genomes found  
153 in two distant Departments of Peru (Cajamarca and Cusco), each of the clonal lineages were  
154 geographically confined (Fig. 1A; Supp. Table 1). This is best exemplified by four clonal  
155 lineages that each circulated in a different location in the National Park of Isiboro in Bolivia  
156 (Fig. 1A). One of these lineages was sampled over a period of 12 years during an epidemic  
157 outbreak of leishmaniasis in Cochabamba (Bolivia), suggesting that the observed clonal  
158 population structure is temporally stable (Supp. Table 5).

159 Our data thus suggest that *Lb* population structure follows an epidemic/semi-clonal  
160 model of evolution<sup>31,32</sup>, as proposed for other protozoan parasites<sup>33</sup>. This model assumes  
161 frequent recombination within all members of a given population, but that occasionally a  
162 successful individual increases in frequency to produce an epidemic clone<sup>31,32</sup>.

163

#### 164 **Parasite populations are isolated in pockets of suitable habitat**

165 Divergent *Lb* ecotypes were described in Peru<sup>20,21</sup> and Eastern Brazil<sup>34</sup> that are each  
166 associated with particular ecological niches, suggesting that the environment may play a key  
167 role in *Lb* population structure. In Peru, our previous work revealed the existence of two  
168 Andean and one Amazonian *Lb* lineage<sup>20</sup>. Here, we characterized in more detail the ancestry  
169 of the Amazonian *Lb* parasites.

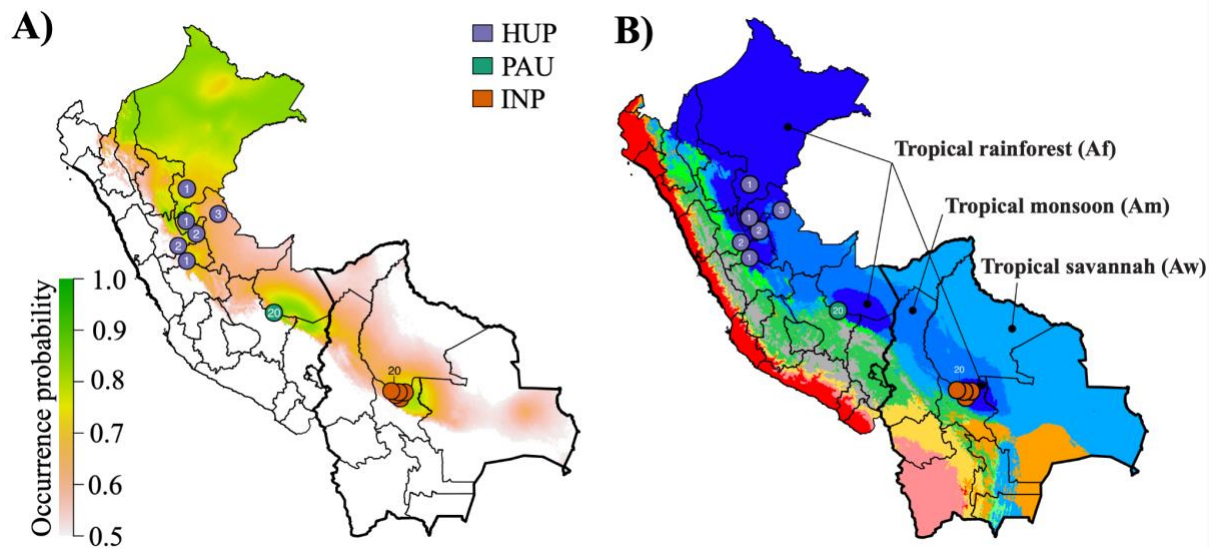
170 We included one isolate from each clonal lineage and removed SNPs showing strong  
171 Linkage Disequilibrium (LD), leaving a total of 176,143 SNPs for investigating population  
172 structure. We identify three distinct ancestry components (PAU, HUP, INP) corresponding to  
173 groups of *Lb* parasites with geographically restricted distributions (Fig. 1B,C; Supp. Fig 7).

174 These three populations showed signatures of spatial and temporal genetic structure: PAU  
175 (N=19) was sampled between 1991 and 1994 in Paucartambo (Southern Peru), INP (N=21)  
176 was sampled between 1994 and 2002 in the Isiboro National Park (Central Bolivia) and HUP  
177 (N=10) was sampled between 1990 and 2003 in Huanuco, Ucayali and Pasco (Central Peru)  
178 (Fig. 1B,C; Supp. Fig 7). Assuming  $K = 2$  populations, the two Peruvian populations PAU and  
179 HUP were grouped as one population (Fig. 1B). While our results indicate that *Lb* may be  
180 genetically clustered at sampling sites, we also identify several groups of *Lb* parasites with  
181 uncertain ancestries (ADM, UNC and STC), one of which contained a total of 19 *Lb* isolates  
182 sampled between 1991 and 2003 across Central and Southern Peru (ADM) (Fig. 1B,C).

183 The strong signatures of geographical isolation of the three inferred ancestry  
184 components (PAU, INP and HUP) prompted us to investigate the geographical and/or  
185 environmental variables that impacted the population structure of *Lb* in the region. This was  
186 done through redundancy analysis (RDA) and generalized dissimilarity modeling (GDM),  
187 including geographic distance and 19 bioclimatic variables (Suppl Table 6). Variable selection  
188 using two approaches consistently pinpointed differences in isothermality (bio3) and  
189 precipitation of driest month (bio14) between sampling locations as the environmental  
190 contributors to parasite genetic distance (Suppl Table 7; Suppl. Results). The RDA model  
191 revealed that environmental differences and geographic distance explained one-third (27.3%)  
192 of the genomic variability, of which 10.2% was contributed by environmental differences and  
193 7.5% by geographic distance (Suppl Fig 8A; Suppl. Table 8; Suppl. Results). Similar results were  
194 obtained with the GDM model (Suppl. Results).

195 Environmental niche modeling (ENM) using present and past bioclimatic variables  
196 predicted the putative suitable habitat of the *Lb* populations during the Last Glacial Maximum  
197 (LGM; 21 kya) and the LIG (130 kya). Present-day predicted suitability regions for *Lb* coincided  
198 with tropical rainforests as predicted by the Köppen-Geiger (KG) climate classification, and  
199 shows that the three *Lb* populations are surrounded by the less suitable tropical monsoon  
200 forests and the non-suitable Andean ecoregions (Fig 2; Supp. Fig. 9, 10). Suitability predictions  
201 for the LGM and the LIG periods revealed largely similar ecological niches, suggesting that *Lb*  
202 suitable habitats have been relatively stable over the past 130,000 years (Supp. Fig. 9, 10).





203

204 **Figure 2. Distribution of *Lb* parasite populations and their endosymbiotic LRV1 lineages in Peru and Bolivia. A)**  
205 Present-day predicted suitability regions for *Lb* parasites as revealed with ecological niche modeling. The  
206 continuous-scale legend represents habitat suitability (probability of occurrence). Map shows the distribution of  
207 isolates belonging to the three *Lb* parasite populations HUP, PAU and INP. **B)** Ecological regions in Peru and  
208 Bolivia as per Köppen-Geiger climate classification. Only three ecological regions were labeled for visibility  
209 reasons. The distribution is shown for *Lb* isolates belonging to the three *Lb* parasite populations HUP, PAU and  
210 INP.

211

## 212 **Hybrid parasites are geographically and ecologically dispersed**

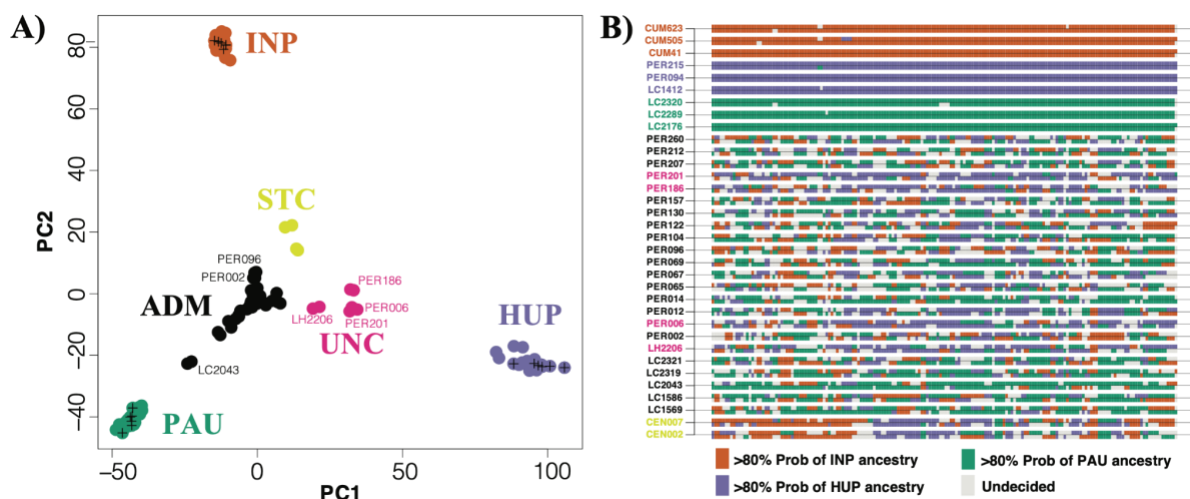
213 In addition to the three *Lb* populations, we also identified three groups of *Lb* parasites  
214 with mixed ancestries: one large group of 19 *Lb* isolates (ADM) sampled between 1991 and  
215 2003 mainly from Southern Peru (Junin, Cusco and Madre de Dios), four *Lb* isolates (UNC)  
216 sampled between 2001 and 2003 from Central/Northern Peru (Junin, Ucayali and Loreto) and  
217 one group of two isolates (STC) sampled in 1984/1985 from the Santa Cruz Department in  
218 Central Bolivia (Fig. 1). The genetic diversity of the largest group of hybrid parasites (ADM)  
219 was much larger compared to the three *Lb* populations: the number of mitochondrial  
220 haplotypes and nuclear segregating sites was 1.4-4 times higher in the ADM group (309,543  
221 SNPs; 8 haplotypes) compared to that of the PAU (189,748 SNPs; 4 haplotypes), INP (185,927  
222 SNPs; 4 haplotypes) and HUP (215,741 SNPs; 2 haplotypes) populations. The observation of a  
223 large number of mitochondrial haplotypes in the ADM group suggests that these parasites are  
224 descendants from multiple independent hybridization events (Supp. Fig. 11).

225 We used PCAdmix to infer the genome-wide ancestry of all admixed individuals and three  
226 control individuals from each source (Fig. 3). Ancestry was assigned to phased haplotype  
227 blocks of 30 SNPs by comparing them to reference panels of PAU, INP and HUP. While the  
228 control samples were assigned 92.9-99.4% to their respective populations, the admixed  
229 individuals showed mixed ancestries between PAU (21.1-73.1%), INP (14.3-36.7%) and HUP  
230 (0.1-54.4%). Plots of local ancestry revealed a complex and heterogeneous pattern of mosaic  
231 ancestry between the three sources (Fig. 3B, Supp. Fig. 12), suggesting that the hybrid

232 parasites experienced many cycles of recombination following the initial admixture event(s).  
233 We next used the three-population statistic  $f_3$  to formally test the potential source  
234 populations for introgressed alleles in the ADM and STC groups. When testing (test; A,B), a  
235 negative result indicates that the test group is an admixed population from A and B. We found  
236 a significantly negative  $f_3$  statistic when ADM was the test group with PAU/HUP ( $f_3 = -0.0013$ ,  
237 Z-score = -18.2) and PAU/INP ( $f_3 = -0.0002$ , Z-score = -2.8) as sources, but not with INP/HUP  
238 ( $f_3 = 0.0013$ , SD=, Z-score = 14.8) as sources. All estimated  $f_3$  statistics were positive when STC  
239 was used as the test group. Hence, the ancestry of STC remains largely unresolved, but may  
240 involve admixture with divergent *Lb* parasites not captured in this study, as indicated by their  
241 distant position in a haplotype network of mitochondrial maxicircle SNPs (Supp. Fig. 11) and  
242 by the comparatively large number of fixed SNP differences between STC and other *Lb* groups  
243 (Supp. Fig. 13).

244 The distribution of the three groups of uncertain ancestries largely pair with the  
245 distribution of the *Lb* populations: the ADM group is mainly found within ecological regions  
246 surrounding the location of the PAU population (Southern Peru), the UNC group is found  
247 within ecological regions surrounding the location of the HUP population (Central/Northern  
248 Peru) and the STC group is found near the INP population (Central Bolivia) (Fig. 1C). In  
249 addition, the median distance between hybrids from the ADM group (310 km), STC group  
250 (203.8 km) and UNC group (349 km) was much larger compared to the median distance  
251 between parasites of each of the populations PAU (0 km), HUP (155 km) and INP (37 km).

252 Altogether, our observations show that hybrid parasites are geographically and  
253 ecologically more widespread compared to the three parasite populations, suggesting that  
254 secondary contacts occurred following migration out of the suitable rainforest habitats.



255  
256 **Figure 3. Mosaic ancestry in hybrid *Lb* parasites.** **A)** PCA-based ancestry assignment of parasites of uncertain  
257 ancestry (ADM, UNC, STC) assuming three *Lb* populations (ADM, INP, HUP). Plus signs indicate three randomly  
258 selected genomes of each source population that were used as controls. **B)** PCA-based local ancestry assignment  
259 to PAU, INP and HUP source populations of the 23 ADM isolates, 2 STC isolates and three randomly selected  
260 isolates from each source population. Ancestry was assigned in windows of 30 SNPs along chromosome 35  
261 (examples for other chromosomes are shown in Supp. Fig. 12).



## 262 **High viral prevalence and lineage diversity within group of hybrid parasites**

263 A total of 31 out of 79 included *Lb* isolates (39%) were LRV1+, as revealed by previous  
264 work<sup>14</sup>. While these originated mainly from Peru (N=27), the geographic distribution of LRV1+  
265 and LRV1- *Lb* isolates was approximately similar (Supp. Fig. 1). Here, we recovered LRV1  
266 genomes for 29/31 LRV1+ *Lb* isolates from Peru and Bolivia following a *de novo* assembly of  
267 total RNA sequencing reads (Supp. Table 9; Supp. Results). The procedure failed for two *Lb*  
268 isolates, either because of difficulties in growing cultures (PER096) or because assembly  
269 yielded a partial LRV1 genome (PER231). Two *Lb* isolates (CUM65 and LC2321) each harbored  
270 two LRV1 genomes (Supp. Results), differing at 999 (for CUM65) and 60 (for LC2321)  
271 nucleotides, bringing the total to 31 viral genomes. While only 0.04% (0.004%-0.1%) of the  
272 RNA sequencing reads aligned against the LRV1 assemblies, median coverages were relatively  
273 high, ranging between 31X and 868X (median = 372X) (Supp. Table 9). Analogous genomic  
274 regions in the assembled sequences were identical to ~1 kb sequences as obtained with  
275 conventional Sanger sequencing, confirming the high quality of our assemblies (Supp.  
276 Results).

277 Sequences were 4,738 - 5,285 bp long, covering the full-length coding sequence of the  
278 virus, and showing an average GC content of 46% (45.4%-46.8%) (Supp. Table 9). All but two  
279 genomes included two open reading frames without internal stop codons, encoding the  
280 Capsid Protein (CP; 2,229bp) and the RNA-dependent RNA Polymerase (RDRP; 2,637bp). Two  
281 isolates (PER014 and PER201) each contained 1 internal stop codon at amino acid positions  
282 875 (TAA) and 882 (TAG), respectively. Sequence identities between these novel LRV1  
283 genomes from Peru and Bolivia, and a previously published LRV1 genome from French Guiana  
284 (YA70; KY750610) ranged between 80% and 81%. Amino acid identity of both genes against  
285 YA70 ranged between 94%-96% for CP, and 85%-87% for RDRP (Supp. Table 9).

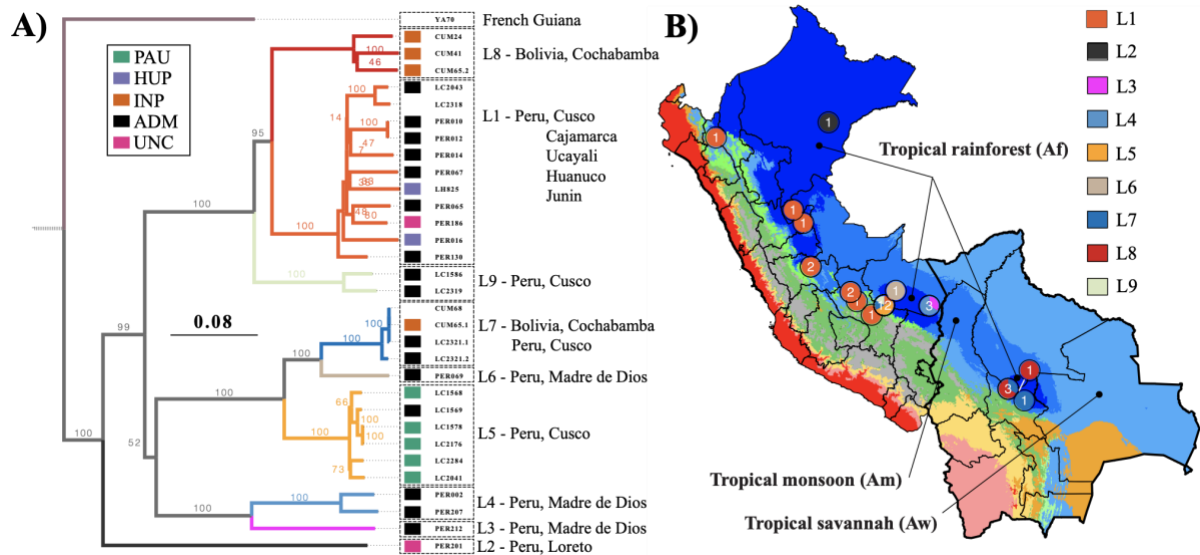
286 We defined a total of nine divergent viral lineages in Peru and Bolivia (L1-9), all supported  
287 by high bootstrap values (Fig. 4) and low pairwise genetic distances (<0.09; Supp. Table 10;  
288 Supp. Table 11). No evidence of recombination was found in our set of LRV1 sequences,  
289 following pairwise homoplasy index (PHI) tests ( $p=0.99$ ). The number of LRV1 lineages  
290 sampled per locality was associated with the number of sampled parasites ( $t=5.72$ ;  $p < 0.001$ )  
291 (Supp. Figure 14). For instance, the most densely sampled location (Paucartambo, Cusco,  
292 Peru) in terms of *Lb* parasites (N = 25) also contained the most viral lineages (Supp. Figure 14),  
293 two of which (L5 and L9) were only found in this location (Fig. 4). Other locations that include  
294 multiple viral lineages are Tambopata (Madre de Dios, Peru) (L3 and L4) and Moleto  
295 (Cochabamba, Bolivia) (L7, L8) (Fig. 4; Supp. Figure 14). These results show that multiple  
296 divergent viral lineages could co-occur within the same geographic region, and that a more  
297 dense sampling of parasites may uncover more viral lineages.

298 The majority of viral lineages (L2-6, L8-9) were found in a single locality (Fig. 4), suggesting  
299 that the geographic spread of most LRV1 lineages is restricted. Two viral lineages were more  
300 widely distributed: one large group of viral strains (L1) was found along the Andes from  
301 Northern Peru to Southern Peru, and one group (L7) was found in Cusco (Peru) and

302 Cochabamba (Bolivia) (Fig. 4). Three viral strains of lineage L7 were virtually identical: viral  
303 sequences of the Bolivian *Lb* parasites CUM68 and CUM65 were identical, and differed by one  
304 nucleotide from a viral sequence of the Peruvian *Lb* strain LC2321. The distal position of the  
305 two Bolivian L7 strains within a clade of Peruvian viral lineages suggests that the L7 viral  
306 lineage was introduced in Bolivia from Southern Peru (Fig. 4). Finally, all early-diverging  
307 lineages (L2-L9) were found in the tropical rainforests (KG-Af) of Peru and Bolivia, except for  
308 the widespread lineage L1 that was found in tropical rainforests (KG-Af), tropical monsoon  
309 (KG-Am), tropical savannah (KG-Aw) and temperate climate (KG-Cwb) (Fig. 4). This suggests  
310 that LRV1 predominantly evolved in the lowland tropical rainforests before spreading to other  
311 ecological regions.

312 When investigating the distribution of viral lineages across the different *Lb* parasite  
313 groups, we found two impactful observations. First, LRV1 prevalence differs between the  
314 different *Lb* groups, with a significantly lower prevalence in the *Lb* populations PAU (30%;  
315 6/20), INP (14.3%; 3/21) and HUP (20%; 2/10) compared to the ADM (78.9%; 15/19) and UNC  
316 (50%; 2/4) groups that included *Lb* parasites of uncertain ancestry. Second, the three *Lb*  
317 populations PAU, INP and HUP (comprising a total of 50 *Lb* isolates) each consisted of strictly  
318 one viral lineage: the two LRV+ *Lb* isolates from HUP comprised the L1 viral lineage, the five  
319 LRV+ *Lb* isolates from PAU comprised the L5 viral lineage and the three LRV+ *Lb* isolates from  
320 INP comprised the L8 viral lineage, with the exception of isolate CUM65 that was coinfecting  
321 with a viral strain from lineage L7 (Fig. 4). These observations strongly suggest that the three  
322 isolated *Lb* populations are each associated with a unique viral lineage. In contrast, the group  
323 of widespread hybrid *Lb* parasites (ADM) comprised almost all viral lineages (L1, L3-L9), four  
324 of which were found exclusively in the ADM *Lb* group (Fig. 4). The extensive viral lineage  
325 diversity within the ADM group suggests that parasite gene flow and hybridization has  
326 promoted the dissemination of viral lineages. This is best exemplified by the widespread L1  
327 viral lineage that is mainly associated with *Lb* parasites from the ADM group, suggesting that  
328 the spread of L1 viruses along the Andes was mediated by parasite hybridization.

329



330  
 331 **Figure 4. Evolutionary history of LRV1 in Peru and Bolivia.** A) Midpoint-rooted maximum likelihood  
 332 phylogenetic tree based on 31 LRV1 genomes of *Lb* from Peru and Bolivia, one LRV1 genome (YA70) of *Lb* from  
 333 French Guiana and 26 LRV1 genomes of *Leishmania guyanensis* from Brazil, Suriname and French Guiana (the  
 334 latter were omitted for visibility reasons). The position of the root is indicated with a dashed line. Clades are  
 335 colored according to the different viral lineages (L1-L9). Colored boxes at the tips of each branch represent the  
 336 population structure of *Lb* (see legend), with colors matching the different groups of parasites as shown in Figure  
 337 1. Note that the tetraploid hybrid parasites LC2318 and CUM68 from Peru and Bolivia, and isolate YA70 from  
 338 French Guiana were omitted from the analyses of parasite population structure; these were thus not assigned  
 339 to any parasite group. B) Ecological regions in Peru and Bolivia as per Köppen-Geiger climate classification. Only  
 340 three ecological regions were labeled for visibility reasons. The distribution is shown for LRV+ *Lb* isolates  
 341 belonging to the nine viral lineages L1-L9.

342

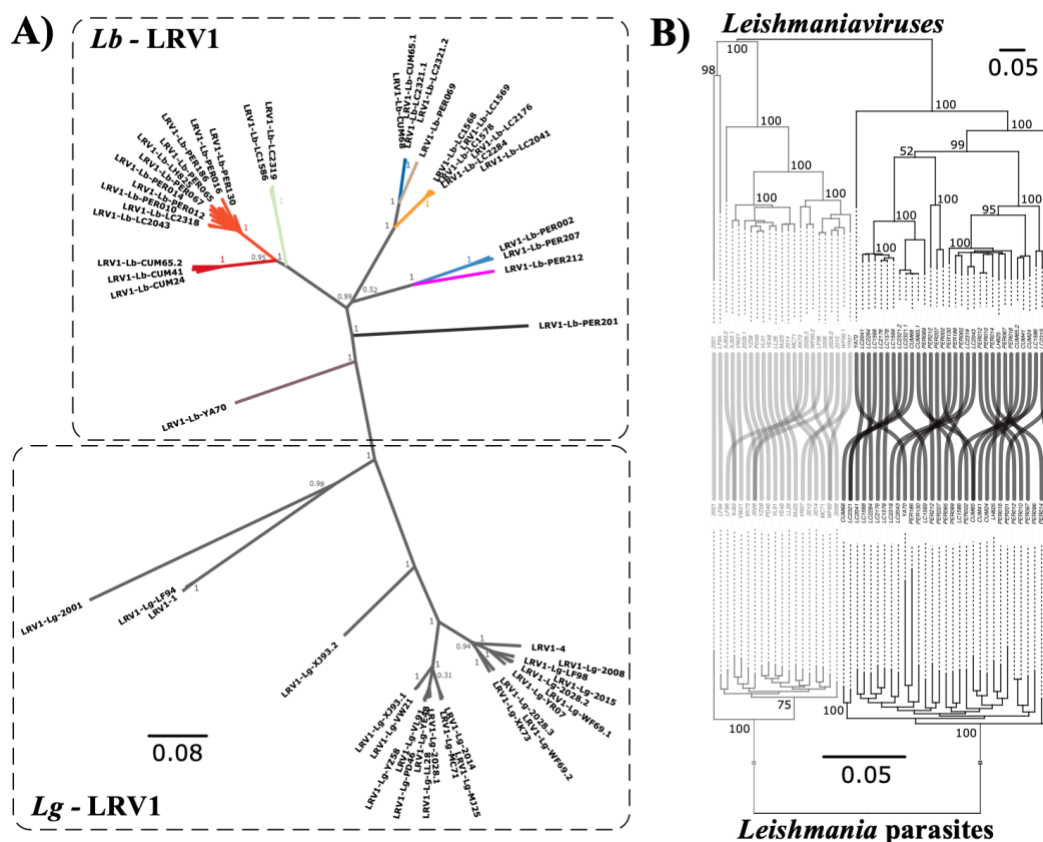
### 343 **Leishmaniviruses co-diverge with *Leishmania* (*Viannia*) host species**

344 Phylogenetic studies indicate that LRV co-diverged with *Leishmania* spp.<sup>35</sup>, though the  
 345 virus is not identified in all *Leishmania* species<sup>11,36</sup> and the presence of viruses in some  
 346 *Leishmania* species may be due to relatively recent horizontal transfer<sup>37-39</sup>. The latter  
 347 observation corroborates experimental evidence that LRV transfer among *Leishmania* species  
 348 may occur via exosomes that are shed by infected *Leishmania* cells<sup>40</sup>. The degree of LRV1 co-  
 349 divergence between *Leishmania Viannia* species has not yet been documented.

350 Viral evolutionary analyses were done using partial (N = 70) and full-length genome (N =  
 351 57) sequences, including publicly available LRV1 genomes of *Lb*, *L. guyanensis* (*Lg*) and *L.*  
 352 *shawii* (*Ls*) from Brazil, French Guiana and Suriname. Despite the allopatric sample, a  
 353 phylogeny based on LRV1 genomes revealed two groups of viral strains: one including *Lb* viral  
 354 strains from Peru, Bolivia and French Guiana, and one including *Lg* viral strains from Brazil,  
 355 French Guiana and Suriname (Fig. 5A). Similar results were obtained based on partial  
 356 sequences of LRV1, whereby viral strains from Brazil clustered either with the *Lb* or *Lg* viral  
 357 strains (Supp. Fig. 15). In addition, nucleotide differences were on average higher between *Lb*  
 358 and *Lg* viral genomes (median = 1,114, min = 1,065, max = 1,208) than between viral genomes  
 359 of *Lb* (median = 908, min = 0, max = 991) and *Lg* (median = 647, min = 24, max = 1,156) (Supp.

360 Fix. 16; Supp. Table 12). These results show that LRV1 consists of divergent viral lineages that  
 361 are grouped by *Leishmania* host species.

362 In order to investigate whether LRV1 co-diverged with *Lb* and *Lg* parasites, we  
 363 reconstructed phylogenies based on *Lb+Lg* viral genomes and their corresponding *Lb+Lg* host  
 364 genotypes. *Leishmania* genotypes were based on SNPs called with GATK Haplotypecaller  
 365 across our set of 79 *Lb* genomes and a set of 20 publicly available *Lg* genomes (see methods).  
 366 Despite the relatively low coverage of *Lg* genomes (median 3x, min 1x, max 4x), genotyping  
 367 recovered a total of 7,571 high-quality bi-allelic SNPs called across 99 *Leishmania* genomes.  
 368 Similar to results obtained for LRV1 (Fig. 5A), a phylogenetic network revealed a clear  
 369 dichotomy between *Lb* and *Lg* parasites (Supp. Fig. 17). Co-phylogenetic analyses revealed a  
 370 split of both viral and parasite genomes at the deepest phylogenetic node, confirming that  
 371 LRV1 viruses cluster with their *Leishmania* host species (Fig. 5B). A consecutive permutation  
 372 test for co-speciation confirmed the topological similarity between both phylogenetic trees  
 373 (RF=70, p-value = 0.001). The extremely low coverage of publicly available sequence data for  
 374 *Lg* precluded us to study intraspecific patterns of co-divergence and host-switching within a  
 375 co-phylogenetic framework.



376  
 377 **Figure 5. Co-divergence between *Leishmania* and LRV1.** **A)** Unrooted maximum likelihood phylogenetic tree  
 378 based on 31 LRV1 genomes of *Lb* from Peru and Bolivia, one LRV1 genome (YA70) of *Lb* from French Guiana and  
 379 26 LRV1 genomes of *Lg* from Brazil, Suriname and French Guiana. LRV1 clades of *Lb* are coloured according to  
 380 the different viral lineages identified in this study (L1-L9) (see Figure 4). **B)** Phylogenetic tangle plot revealing  
 381 clear patterns of virus-parasite co-divergence between *Lb* and *Lg* and their respective monophyletic LRV1 clades.  
 382 The gray and black branches and tree links correspond to *Lg* and *Lb*, respectively.

## 383 DISCUSSION

384 Our main goal was to understand the evolution and dissemination of endosymbiotic  
385 viruses (here LRV1) in an important group of human pathogenic parasites (here *Lb*). Genetic  
386 studies have shown that both *Lb*<sup>20,29,30,34,41</sup> and LRV1<sup>43</sup> populations are structured according  
387 to their geographical origin. Here, we characterized the joint ancestry of the two species in  
388 Peru and Bolivia based on whole genome sequence analysis. Our results show that both *Lb*  
389 and LRV1 are genetically highly heterogeneous and predominantly evolved within the lowland  
390 tropical rainforests. Models of landscape genomics reveal that geographic distance and in  
391 particular environmental differences between sampling locations contributed to partitioning  
392 *Lb* diversity. This indicates that the large diversity and population substructure of *Lb* and LRV1  
393 may have been driven by the extremely diversified ecosystem of the Amazonian rainforest,  
394 including various host–vector communities<sup>44</sup>.

395 The degree of LRV1 co-divergence and host-switching between and within *Leishmania*  
396 species has not yet been documented. We show that LRV1 lineages cluster according to  
397 different *Leishmania* species, indicating that horizontal transfer of LRV1 between parasite  
398 species is rare<sup>43,45</sup>. The agreement between LRV1 and *Leishmania* phylogenies corroborates  
399 the general hypothesis that LRV1 is an ancient virus that has co-evolved with its parasite<sup>8,11,35</sup>.  
400 While we observed a clear pattern of co-divergence at the species level, we did not detect  
401 such a pattern when investigating co-phylogenies of LRV1 and *Lb*. Our data suggest that  
402 recombination in *Lb* may drive prevalent horizontal transmission of LRV1 at the intraspecific  
403 level. Such intraspecific phylogenetic incongruences have also been observed for other  
404 endosymbionts, such as *Wolbachia*<sup>46</sup>.

405 One third of our panel showed signatures of mixed ancestry, supporting a growing body  
406 of genomic evidence for extensive genetic exchange in protozoan parasites in the wild<sup>20,28,47–</sup>  
407<sup>52</sup>. We propose that increased human migration and ecological perturbations over the past  
408 decades, including movement of hemerophile reservoirs/hosts such as dogs and rats, may  
409 have resulted in the displacement of *Lb* parasites out of the tropical rainforests<sup>29,53</sup>. Probably  
410 the most exciting outcome of our study is that hybrid parasites were geographically and  
411 ecologically widely distributed and commonly infected from a pool of genetically diverse  
412 LRV1, in contrast to the *Lb* populations that were confined to tropical rainforests and  
413 infrequently associated with a single LRV1 lineage. Our data thus suggests that parasite  
414 hybridization increased the frequency of *Lb*-LRV1 symbiotic interactions, which play a key role  
415 in the severity of human leishmaniasis<sup>15,16</sup> and which may have profound epidemiological  
416 consequences in the region.

417



418 **METHODS**

419

420 **Nucleic acid extraction and sequencing of DNA parasites and their dsRNA viruses**

421 This study included 79 *Lb* isolates (Supp. Table 1) from Peru (N=55) and Bolivia (N=24)  
422 that were sampled within the context of various studies on the genetics and epidemiology of  
423 leishmaniasis. In Bolivia, the majority of *Lb* isolates (N=21) were sampled between 1994 and  
424 2002 within the context of a CL outbreak in the Isiboro National Park (Department of  
425 Cochabamba). Two *Lb* isolates were sampled between 1984 and 1985 within the Santa Cruz  
426 Department and one is of unknown origin (Supp. Fig. 1). In Peru, half of the *Lb* isolates were  
427 sampled between 1991 and 2003 in the Cusco state (N=29), mainly from the Paucartambo  
428 province (N=25). The remaining 26 *Lb* isolates originated from Madre de Dios (N=9), Ucayali  
429 (N=5), Huanuco (N=4), Junin (N=4), Loreto (N=2), Pasco (N=1) and Cajamarca (N=1) (Supp. Fig.  
430 1).

431 All 79 *Lb* isolates were cultured for 3-4 days on a HOMEM medium with Fetal Bovine  
432 Serum (FBS) at the Antwerp Institute of Tropical Medicine. Parasites were subjected to a small  
433 number of passages (mean =  $18 \pm 5$ ) to reduce potential culture-related biases in parasite  
434 genomic characterization<sup>54</sup>. Parasite cells were centrifuged into dry pellets and their DNA was  
435 extracted using the QIAGEN QIAmp DNA Mini kit following the manufacturer's protocol. At  
436 the Wellcome Sanger Institute, genomic DNA was sheared into 400 to 600 base pair fragments  
437 by focused ultrasonication (Covaris Inc.), and amplification-free Illumina libraries were  
438 prepared<sup>55</sup>. One hundred base pair paired-end reads were generated on the HiSeq 2000, and  
439 150 base pair paired end reads were generated on the HiSeq x10 according to the  
440 manufacturer's standard sequencing protocol.

441 Previous work has shown that 31 out of the 79 *Lb* isolates were positive for LRV1<sup>14</sup>. The  
442 majority of the LRV1-positive *Lb* isolates originated from Peru (N=27) while the remaining four  
443 isolates were sampled in the Isiboro National Park (Cochabamba, Bolivia) (Supp. Fig 1). More  
444 than half of the LRV1-positive isolates in Peru originated from Cusco (N=15) of which 11 were  
445 sampled in the Paucartambo province. LRV1 was also sampled in Madre de Dios (N=5), Junin  
446 (N=3), Cajamarca (N=1), Huanuco (N=1), Loreto (N=1) and Ucayali (N=1) (Supp. Fig 1).

447 The 31 LRV1-positive *Lb* isolates<sup>14</sup> were grown on a HOMEM medium with FBS for 2-3  
448 weeks (mean passage number =  $18 \pm 5$ ) at the Antwerp Institute of Tropical Medicine,  
449 ensuring high parasite yields (ca.  $10^7$ - $10^8$  parasites/ml). Isolation of dsRNA was performed as  
450 previously described<sup>56</sup>. In short, isolation involved a TRIZOL reagent (Invitrogen) RNA  
451 extraction followed by a RNase-free DNase I (NEB.) and a S1 nuclease (Sigma-Aldrich)  
452 treatment along with an additional purification step (Zymoclean). Double-stranded RNA of  
453 approximately 5.2kb. was visualized on 0.8% agarose gel (TAE buffer) stained with ethidium  
454 bromide. Extracts of dsRNA were sequenced at Genewiz (Leipzig, Germany) using the  
455 NovaSeq platform (Illumina) generating 150bp paired end reads.

456

## 457 **Bioinformatics and population genomics of *Lb* parasites**

458 Sequencing reads were mapped against the *Lb* M2904 reference genome using SMALT  
459 v.0.7.4 (<https://www.sanger.ac.uk/tool/smalt-0/>) as previously described<sup>20</sup>. The reference  
460 genome comprises 35 chromosomes (32.73 Mb) and a complete mitochondrial maxicircle  
461 sequence (27.69 kb), and is available on <https://tritrypdb.org/> as  
462 *Lbraziliensis*MHOM/BR/75/M2904\_2019. Genome wide variant calling (SNPs, INDELs) was  
463 done using GATK v.4.0.1.0<sup>57,58</sup>. More specifically, we used GATK HaplotypeCaller for  
464 generating genotype VCF files (gVCF) for each *Lb* isolate. Individual gVCF files were combined  
465 and jointly genotyped using CombineGVCFs and GenotypeGVCFs, respectively. SNPs and  
466 INDELs were separated using SelectVariants. Low quality variants were excluded using GATK  
467 VariantFiltration following GATK's best practices<sup>59</sup> and BCFtools v.1.10.2<sup>60</sup> view and filter.  
468 Specifically, SNPs were excluded when QD < 2, FS > 60.0, SOR > 3.0, MQ < 40.0, MQRankSum  
469 < -12.5, ReadPosRankSum < -8.0, QUAL < 250, format DP < 10, format GQ < 25, or when SNPs  
470 occurred in clusters (ClusterSize=3, clusterWindowSize=10). INDELs were excluded when QD  
471 < 2, FS < 200.0, ReadPosRankSum < -20.0. The final set of SNPs and INDELs were annotated  
472 using the *Lb* M2904 annotation file with SNPEFF v4.5<sup>61</sup>. At heterozygous SNP sites, the  
473 frequencies of the alternate allele read depths<sup>24</sup> were calculated using the vcf2freq.py script  
474 (available at [github.com/FreBio/mytools](https://github.com/FreBio/mytools)).

475 Chromosomal and gene copy number variation were estimated using normalized read  
476 depths. To this end, per-site read depths were calculated with SAMtools depth (-a option).  
477 Haploid copy numbers (HCN) were obtained for each chromosome by dividing the median  
478 chromosomal read depth by the median genome-wide read depth. Somy variation was then  
479 obtained by multiplying HCN by two (assuming diploidy). To obtain gene HCN, the median  
480 read depth per coding DNA sequence (CDS) was divided by the median genome-wide read  
481 depth. The HCN per CDS were summed up per orthologous gene group. Gene copy number  
482 variations were then defined where the z-score was lower than -1 (deletions) or larger than 1  
483 (amplifications).

484 A NeighborNet tree was reconstructed based on uncorrected *p*-distances with SplitsTree  
485 v.4.17.0<sup>62</sup>. The population genomic structure of *Lb* was examined with ADMIXTURE v.1.3.0<sup>63</sup>  
486 and fineSTRUCTURE v.4.1.1<sup>64</sup>. ADMIXTURE was run assuming 1 to 10 populations (*K*),  
487 performing a 5-fold cross-validation procedure, and after removing SNPs with strong LD with  
488 plink v.1.9<sup>65</sup> (--indep-pairwise 50 10 0.3). CHROMOPAINTER analysis (as part of  
489 fineSTRUCTURE) was run to infer the genetic ancestry based on haplotype similarity. To this  
490 end, individual genotypes were computationally phased with BEAGLE v.5.2<sup>66</sup> using default  
491 parameters, after which fineSTRUCTURE was run using 8e06 MCMC iterations with 50% as  
492 burn-in, and 2e06 maximization steps for finding the best state for building the tree<sup>48</sup>. Local  
493 ancestry was assigned with PCAdmix<sup>67</sup> using phased genotype data (i.e. haplotypes) as  
494 obtained with the BEAGLE v5.4<sup>66</sup>. *F*<sub>3</sub>-statistics were calculated with Treemix v1.13<sup>68</sup>. Finally,  
495 Hardy-Weinberg equilibrium (HWE) was examined by calculating per site the inbreeding  
496 coefficient as  $Fis = 1 - Ho/He$ ; with *Ho* representing the observed heterozygosity and *He* the

497 expected heterozygosity. Decay of LD was calculated and visualized using PopLDdecay<sup>69</sup>. To  
498 control for spatio-temporal Wahlund effects, we calculated Fis and LD decay using subsets of  
499 isolates that were isolated close in time (year of isolation) and space (locality), and taking into  
500 account population genomic structure (Supp. Table 4).

501 To investigate the spatio-environmental impact on genetic variation among the three *Lb*  
502 populations we included geographic distances among sampling locations and extracted 19  
503 bioclimatic variables of the WorldClim2 database<sup>70</sup>. We firstly investigated the role of  
504 geography on the genomic differentiation of the *Lb* components through linear and non-linear  
505 regression analyses of distance matrices (Supp. Methods). Secondly, we used redundancy  
506 analysis (RDA)<sup>71</sup> and generalized dissimilarity modeling (GDM)<sup>72</sup> to test the impact of  
507 environmental differences and geographic distance on parasite genetic distance. To account  
508 for model overfitting and multicollinearity, we performed two variable selection approaches  
509 (mod-A, mod-M) (Supp. Methods). Environmental Niche Modeling (ENM) was done based on  
510 present-day and past bioclimatic variables using Maxent, as implemented in the 'dismo' R  
511 package<sup>73,74</sup>. A more detailed description on the landscape genomics analyses (variable  
512 selection, RDA, GDM and ENM) is presented in the Supplementary method section.

513

#### 514 **Bioinformatics and phylogenetic analyses of LRV1**

515 Raw sequencing reads were trimmed and filtered with fastp<sup>75</sup> using the following  
516 settings: a minimum base quality (-q) of 30; the percentage of unqualified bases (-u) set to 10;  
517 per read sliding window trimming based on mean quality scores (-5, front to tail; -3, tail to  
518 front) with a window size (-W) of 1 and mean quality score (-M) of 30; right-cutting reads (--  
519 cut\_right) per 10bp windows (--cut\_right\_window\_size) when mean quality score (--  
520 cut\_right\_mean\_quality) is below 30; only considering reads between 100 (-l) and 150bp (-b).  
521 LRV1 sequences were assembled *de novo* with MEGAHIT<sup>76</sup> and identified using BLASTn<sup>77</sup>  
522 against conventional LRV1 reference genomes LRV1-1<sup>78</sup> and LRV1-4<sup>79</sup> (accession numbers  
523 M92355 and U01899, respectively). In order to check and improve the quality of the  
524 assemblies, trimmed reads were mapped against the LRV1 contigs with SMALT as described  
525 above, with the minimal nucleotide identity (-y) set to 95%. Alignments were examined with  
526 Artemis<sup>80</sup> and used to improve assemblies with Pilon v.1.23<sup>81</sup>. Genome sequences were  
527 aligned using MAFFT v.7.49<sup>82</sup>.

528 For comparative purposes, we included 26 (near-) complete LRV1 genomes and 13 partial  
529 LRV1 sequences of *Lb* (N=8), *Lg* (N=26) and *Lsh* (*L. shawi*) (N=1) from French Guiana, Brazil  
530 and Suriname<sup>43,45,78,79</sup>. Multiple sequence alignments were generated by trimming and re-  
531 aligning i) all available genomes to 5,189 bp sequences (N=57) and ii) all available sequences  
532 to 755 bp sequences (N=70). Maximum likelihood (ML) trees were generated using IQtree  
533 v.1.6.12<sup>83</sup> with 100 bootstrap replicates and implementation of the ModelFinder function<sup>84</sup>  
534 to determine the best substitution model based on the lowest Bayesian Information Criterion  
535 (BIC). The best substitution model identified for the genome alignment was the GTR+F+R4  
536 model, which was also applied on the partial sequence alignment for consistency. Pairwise

537 genetic distance among LRV1 genomes were calculated with the ‘ape’ R-package<sup>85</sup> (model=  
538 ‘raw’). Viral genomes with a genetic distance below 0.09 (i.e. 9% of the sites that are dissimilar  
539 among two genomes) were grouped into distinct viral lineages. Nucleotide diversity and Fst  
540 statistics were calculated within and between viral lineages using the ‘PopGenome’ package  
541 in R<sup>86</sup> while recombination was tested by pairwise homoplasmy index (PHI) tests implemented  
542 in SplitsTree<sup>62,87</sup>.

543

#### 544 **Co-phylogenetic analysis of LRV1 and *Leishmania***

545 Co-phylogenetic analyses were done at both the species (between LRV1 infecting *Lb* and  
546 *Lg*) and at the population (LRV1 infecting *Lb* from Peru and Bolivia) level. These analyses were  
547 performed using the phytools R package<sup>88</sup>. To assess the co-evolutionary history of LRV1 with  
548 *Lb* and *Lg*, we added 24 LRV1 genomes and 19 *Lg* and 1 *Lb* SNP genotypes from French Guiana  
549 (N=19) and Brazil (N=1)<sup>45</sup> (*Lg* reads accession: PRAJNA371487; LRV1 genome accessions:  
550 KY750607 to KY750630). The phylogenetic trees were tested for topological similarity  
551 calculating the Robinson-Foulds distance<sup>89</sup> between both trees with comparison against a  
552 null distribution of 1,000 permuted un-correlated trees. For LRV1 we reconstructed a ML tree  
553 as described above, including the 23 LRV1 genomes of *Lg* and 32 LRV1 genomes of *Lb*. For  
554 *Leishmania*, sequence reads of *Lg* were mapped against the M2904 reference genome and  
555 GATK Haplotypecaller was run as described above. We then performed joint genotyping on a  
556 dataset including the 19 *Lg* genomes and 80 *Lb* genomes. SNPs were filtered following similar  
557 criteria as described above (QD<2, FS>200.0, SOR>3.0, MQ<40.0, MQRankSum<-12.5,  
558 ReadPosRankSum<-8.0, QUAL<250, info DP < 10, ClusterSize=3 and ClusterWindowSize=10).  
559 The *Leishmania* ML tree was generated using IQtree based on 7,571 jointly called bi-allelic  
560 SNPs with 100 bootstrap replicates and GTR+F+R5 as best performing substitution model<sup>83,84</sup>.  
561 For the co-phylogenetic reconciliation at the intraspecific level, we focused on phylogenies  
562 generated for our dataset of LRV1 (GTR+F+R4 ML tree generated with IQtree) and *Lb*  
563 (bifurcating tree generated by fineSTRUCTURE ) from Peru and Bolivia.

564

565 **DATA AVAILABILITY**

566 Genomic sequence reads of the 79 sequenced *Lb* genomes are available on the European  
567 Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/home>) under accession number  
568 PRJEB4442. The assembled sequences of the 31 LRV1 genomes are available on GenBank  
569 (<https://www.ncbi.nlm.nih.gov/genbank/>) under accession numbers OQ673070-OQ673100.

570

571 **FUNDING**

572 This work received financial support from the European Commission (Contracts TS2-  
573 CT90-0315 and TS3-CT92-0129) and Directie-Generaal Ontwikkelingssamenwerking en  
574 Humanitaire Hulp (DGD) (Belgian cooperation). FVDB and SH acknowledge support from the  
575 Research Foundation Flanders (Grants 1226120N and G092921N). LFL and SMB acknowledge  
576 support from the National Institute of Health (5R01AI130222 and 2R01AI029646).

577

578 **REFERENCES**

- 579 1. Wang, A. L. & Wang, C. C. VIRUSES OF THE PROTOZOA. *Annu. Rev. Microbiol.* **45**, 251–  
580 263 (1991).
- 581 2. Banik, G., Stark, D., Rashid, H. & Ellis, J. Recent Advances in Molecular Biology of  
582 Parasitic Viruses. *Infectious Disorders - Drug Targets* **14**, 155–167 (2015).
- 583 3. Ghabrial, S. A., Castón, J. R., Jiang, D., Nibert, M. L. & Suzuki, N. 50-plus years of fungal  
584 viruses. *Virology* **479-480**, 356–368 (2015).
- 585 4. Chen, B., Geletka, L. M. & Nuss, D. L. Using Chimeric Hypoviruses To Fine-Tune the  
586 Interaction between a Pathogenic Fungus and Its Plant Host. *J. Virol.* **74**, 7562–7567  
587 (2000).
- 588 5. Jenkins, M. C. *et al.* Fecundity of *Cryptosporidium parvum* is correlated with  
589 intracellular levels of the viral symbiont CPV. *Int. J. Parasitol.* **38**, 1051–1055 (2008).
- 590 6. Bruenn, J. A. Viruses of Fungi and Protozoans: Is Everyone Sick? *Viral Ecology* 297–317  
591 (2000).
- 592 7. Hillman, B. I. & Cohen, A. B. Totiviruses (Totiviridae). in *Encyclopedia of Virology (Fourth*



- 593 *edition*) (eds. Bamford, D. H. & Zuckerman, M.) 648–657 (Academic Press, 2021).
- 594 8. Widmer, G. & Dooley, S. Phylogenetic analysis of Leishmania RNA virus and Leishmania  
595 suggests ancient virus-parasite association. *Nucleic Acids Res.* **23**, 2300–2304 (1995).
- 596 9. Torres-Guerrero, E., Quintanilla-Cedillo, M. R., Ruiz-Esmenjaud, J. & Arenas, R.  
597 Leishmaniasis: a review. *F1000Res.* **6**, 750 (2017).
- 598 10. Scheffter, S. M., Ro, Y. T., Chung, I. K. & Patterson, J. L. The complete sequence of  
599 Leishmania RNA virus LRV2-1, a virus of an Old World parasite strain. *Virology* **212**, 84–  
600 90 (1995).
- 601 11. Cantanhêde, L. M. *et al.* The Maze Pathway of Coevolution: A Critical Review over the  
602 Leishmania and Its Endosymbiotic History. *Genes* **12**, (2021).
- 603 12. Ives, A. *et al.* Leishmania RNA virus controls the severity of mucocutaneous  
604 leishmaniasis. *Science* **331**, 775–778 (2011).
- 605 13. Bourreau, E. *et al.* Presence of Leishmania RNA Virus 1 in Leishmania guyanensis  
606 Increases the Risk of First-Line Treatment Failure and Symptomatic Relapse. *J. Infect.*  
607 *Dis.* **213**, 105–111 (2016).
- 608 14. Aduai, V. *et al.* Association of the Endobiont Double-Stranded RNA Virus LRV1 With  
609 Treatment Failure for Human Leishmaniasis Caused by Leishmania braziliensis in Peru  
610 and Bolivia. *J. Infect. Dis.* **213**, 112–121 (2016).
- 611 15. Hartley, M.-A., Ronet, C., Zangger, H., Beverley, S. M. & Fasel, N. Leishmania RNA virus:  
612 when the host pays the toll. *Front. Cell. Infect. Microbiol.* **2**, 99 (2012).
- 613 16. Brettmann, E. A. *et al.* Tilting the balance between RNA interference and replication  
614 eradicates Leishmania RNA virus 1 and mitigates the inflammatory response. *Proc.*  
615 *Natl. Acad. Sci. U. S. A.* **113**, 11998–12005 (2016).
- 616 17. Lafleur, A. & Olivier, M. Viral endosymbiotic infection of protozoan parasites: How it

- 617 influences the development of cutaneous leishmaniasis. *PLoS Pathog.* **18**, e1010910  
618 (2022).
- 619 18. Roque, A. L. R. & Jansen, A. M. Wild and synanthropic reservoirs of Leishmania species  
620 in the Americas. *Int. J. Parasitol. Parasites Wildl.* **3**, 251–262 (2014).
- 621 19. Rougeron, V. *et al.* Extreme inbreeding in Leishmania braziliensis. *Proc. Natl. Acad. Sci.*  
622 *U. S. A.* **106**, 10224–10229 (2009).
- 623 20. Van den Broeck, F. *et al.* Ecological divergence and hybridization of Neotropical  
624 Leishmania parasites. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25159–25168 (2020).
- 625 21. Odiwuor, S. *et al.* Evolution of the Leishmania braziliensis species complex from  
626 amplified fragment length polymorphisms, and clinical implications. *Infect. Genet. Evol.*  
627 **12**, 1994–2002 (2012).
- 628 22. Arevalo, J. *et al.* Influence of Leishmania (Viannia) species on the response to  
629 antimonial treatment in patients with American tegumentary leishmaniasis. *J. Infect.*  
630 *Dis.* **195**, 1846–1851 (2007).
- 631 23. Adai, V. *et al.* Multilocus genotyping reveals a polyphyletic pattern among naturally  
632 antimony-resistant Leishmania braziliensis isolates from Peru. *Infect. Genet. Evol.* **11**,  
633 1873–1880 (2011).
- 634 24. Rogers, M. B. *et al.* Chromosome and gene copy number variation allow major  
635 structural change between species and strains of Leishmania. *Genome Res.* **21**, 2129–  
636 2142 (2011).
- 637 25. Imamura, H. *et al.* Evolutionary genomics of epidemic visceral leishmaniasis in the  
638 Indian subcontinent. *Elife* **5**, (2016).
- 639 26. Franssen, S. U. *et al.* Global genome diversity of the Leishmania donovani complex. *Elife*  
640 **9**, (2020).

- 641 27. Tibayrenc, M. & Ayala, F. J. Reproductive clonality of pathogens: a perspective on  
642 pathogenic viruses, bacteria, fungi, and parasitic protozoa. *Proc. Natl. Acad. Sci. U. S. A.*  
643 **109**, E3305–13 (2012).
- 644 28. Ramírez, J. D. & Llewellyn, M. S. Reproductive clonality in protozoan pathogens-truth or  
645 artefact? *Mol. Ecol.* **23**, 4195–4202 (2014).
- 646 29. De los Santos, M. B., Ramírez, I. M., Rodríguez, J. E., Beerli, P. & Valdivia, H. O. Genetic  
647 diversity and population structure of *Leishmania (Viannia) braziliensis* in the Peruvian  
648 jungle. *PLoS Negl. Trop. Dis.* **16**, e0010374 (2022).
- 649 30. Kuhls, K. *et al.* Population structure and evidence for both clonality and recombination  
650 among Brazilian strains of the subgenus *Leishmania (Viannia)*. *PLoS Negl. Trop. Dis.* **7**,  
651 e2490 (2013).
- 652 31. Maiden, M. C. J. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60**, 561–  
653 588 (2006).
- 654 32. Smith, J. M., Smith, N. H., O’Rourke, M. & Spratt, B. G. How clonal are bacteria?  
655 *Proceedings of the National Academy of Sciences* **90**, 4384–4388 (1993).
- 656 33. Van den Broeck, F., Tavernier, L. J. M., Vermeiren, L., Dujardin, J.-C. & Van Den Abbeele,  
657 J. Mitonuclear genomics challenges the theory of clonality in *Trypanosoma congolense*:  
658 Reply to Tibayrenc and Ayala. *Molecular ecology* vol. 27 3425–3431 (2018).
- 659 34. S L Figueiredo de Sá, B., Rezende, A. M., Melo Neto, O. P. de, Brito, M. E. F. de &  
660 Brandão Filho, S. P. Identification of divergent *Leishmania (Viannia) braziliensis*  
661 ecotypes derived from a geographically restricted area through whole genome analysis.  
662 *PLoS Negl. Trop. Dis.* **13**, e0007382 (2019).
- 663 35. Zangger, H. *et al.* *Leishmania aethiopica* field isolates bearing an endosymbiotic  
664 dsRNA virus induce pro-inflammatory cytokine response. *PLoS Negl. Trop. Dis.* **8**, e2836

- 665 (2014).
- 666 36. Shita, E. Y. *et al.* Prevalence of Leishmania RNA virus in Leishmania parasites in patients  
667 with tegumentary leishmaniasis: A systematic review and meta-analysis. *PLoS Negl.*  
668 *Trop. Dis.* **16**, e0010427 (2022).
- 669 37. Hajjarian, H. *et al.* Detection and molecular identification of leishmania RNA virus (LRV)  
670 in Iranian Leishmania species. *Arch. Virol.* **161**, 3385–3390 (2016).
- 671 38. Saberi, R. *et al.* Presence and diversity of Leishmania RNA virus in an old zoonotic  
672 cutaneous leishmaniasis focus, northeastern Iran: haplotype and phylogenetic based  
673 approach. *Int. J. Infect. Dis.* **101**, 6–13 (2020).
- 674 39. Nalçacı, M. *et al.* Detection of Leishmania RNA virus 2 in Leishmania species from  
675 Turkey. *Trans. R. Soc. Trop. Med. Hyg.* **113**, 410–417 (2019).
- 676 40. Atayde, V. D. *et al.* Exploitation of the Leishmania exosomal pathway by Leishmania  
677 RNA virus 1. *Nat Microbiol* **4**, 714–723 (2019).
- 678 41. Patino, L. H., Muñoz, M., Cruz-Saavedra, L., Muskus, C. & Ramírez, J. D. Genomic  
679 Diversification, Structural Plasticity, and Hybridization in Leishmania (Viannia)  
680 braziliensis. *Front. Cell. Infect. Microbiol.* **10**, 582192 (2020).
- 681 42. Oddone, R. *et al.* Development of a multilocus microsatellite typing approach for  
682 discriminating strains of Leishmania (Viannia) species. *J. Clin. Microbiol.* **47**, 2818–2825  
683 (2009).
- 684 43. Cantanhêde, L. M. *et al.* New insights into the genetic diversity of Leishmania RNA Virus  
685 1 and its species-specific relationship with Leishmania parasites. *PLoS One* **13**,  
686 e0198727 (2018).
- 687 44. Rotureau, B. *et al.* Diversity and ecology of sand flies (Diptera: Psychodidae:  
688 Phlebotominae) in coastal French Guiana. *Am. J. Trop. Med. Hyg.* **75**, 62–69 (2006).

- 689 45. Tirera, S. *et al.* Unraveling the genetic diversity and phylogeny of Leishmania RNA virus  
690 1 strains of infected Leishmania isolates circulating in French Guiana. *PLoS Negl. Trop.*  
691 *Dis.* **11**, e0005764 (2017).
- 692 46. Scholz, M. *et al.* Large scale genome reconstructions illuminate Wolbachia evolution.  
693 *Nat. Commun.* **11**, 5235 (2020).
- 694 47. Rogers, M. B. *et al.* Genomic confirmation of hybridisation and recent inbreeding in a  
695 vector-isolated Leishmania population. *PLoS Genet.* **10**, e1004092 (2014).
- 696 48. Tihon, E., Imamura, H., Dujardin, J.-C., Van Den Abbeele, J. & Van den Broeck, F.  
697 Discovery and genomic analyses of hybridization between divergent lineages of  
698 Trypanosoma congolense, causative agent of Animal African Trypanosomiasis. *Mol.*  
699 *Ecol.* **26**, 6524–6538 (2017).
- 700 49. Schwabl, P. *et al.* Meiotic sex in Chagas disease parasite Trypanosoma cruzi. *Nat.*  
701 *Commun.* **10**, (2019).
- 702 50. Glans, H. *et al.* High genome plasticity and frequent genetic exchange in Leishmania  
703 tropica isolates from Afghanistan, Iran and Syria. *PLoS Negl. Trop. Dis.* **15**, e0010110  
704 (2021).
- 705 51. Van den Broeck, F. *et al.* Emerging hybrid uncovers cryptic Leishmania lineage diversity  
706 in the Neotropics. *under review* (2022).
- 707 52. Inbar, E. *et al.* Whole genome sequencing of experimental hybrids supports meiosis-like  
708 sexual recombination in leishmania. *PLoS Genet.* **15**, 1–28 (2019).
- 709 53. Marzochi, M. C. de A. *et al.* Anthropogenic Dispersal of Leishmania (Viannia) braziliensis  
710 in the Americas: A Plausible Hypothesis. *Frontiers in Tropical Diseases* **2**, 21 (2021).
- 711 54. Domagalska, M. A. *et al.* Genomes of Leishmania parasites directly sequenced from  
712 patients with visceral leishmaniasis in the Indian subcontinent. *PLoS Negl. Trop. Dis.* **13**,



- 713 e0007900 (2019).
- 714 55. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates  
715 improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291–295  
716 (2009).
- 717 56. Kleschenko, Y. *et al.* Molecular Characterization of Leishmania RNA virus 2 in  
718 Leishmaniamajor from Uzbekistan. *Genes* **10**, (2019).
- 719 57. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing  
720 next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- 721 58. Van der Auwera, G. A. & OConnor, B. D. *Genomics in the Cloud Using Docker, GATK, and*  
722 *WDL in Terra (First edition)*. (O’Reilly Media, 2020).
- 723 59. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the  
724 Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**,  
725 11.10.1–11.10.33 (2013).
- 726 60. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
- 727 61. Cingolani, P. *et al.* A program for annotating and predicting the effects of single  
728 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*  
729 strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
- 730 62. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies.  
731 *Mol. Biol. Evol.* **23**, 254–267 (2006).
- 732 63. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in  
733 unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- 734 64. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure  
735 using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
- 736 65. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based

- 737 linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 738 66. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-  
739 scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
- 740 67. Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along  
741 each chromosome in individuals with admixed ancestry from two or more populations.  
742 *Hum. Biol.* **84**, 343–364 (2012).
- 743 68. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-  
744 wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
- 745 69. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective  
746 tool for linkage disequilibrium decay analysis based on variant call format files.  
747 *Bioinformatics* **35**, 1786–1788 (2019).
- 748 70. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for  
749 global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
- 750 71. Oksanen, J. *et al.* *Vegan Community Ecology Package Version 2.5-7.* (2020).
- 751 72. Fitzpatrick, M., Mokany, K., Manion, G., Nieto-Lugilde, D. & Ferrier, S. *gdm: Generalized*  
752 *Dissimilarity Modeling.* (2022).
- 753 73. Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J. & Hijmans, M. R. J. Package ‘dismo’.  
754 *Circles* **9**, 1–68 (2017).
- 755 74. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species  
756 geographic distributions. *Ecol. Modell.* **190**, 231–259 (2006).
- 757 75. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.  
758 *Bioinformatics* **34**, i884–i890 (2018).
- 759 76. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node  
760 solution for large and complex metagenomics assembly via succinct de Bruijn graph.

- 761 *Bioinformatics* **31**, 1674–1676 (2015).
- 762 77. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment  
763 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- 764 78. Widmer, G., Comeau, A. M., Furlong, D. B., Wirth, D. F. & Patterson, J. L.  
765 Characterization of a RNA virus from the parasite *Leishmania*. *Proc. Natl. Acad. Sci. U. S.*  
766 *A.* **86**, 5979–5982 (1989).
- 767 79. Scheffter, S., Widmer, G. & Patterson, J. L. Complete sequence of *Leishmania* RNA virus  
768 1-4 and identification of conserved sequences. *Virology* **199**, 479–483 (1994).
- 769 80. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an  
770 integrated platform for visualization and analysis of high-throughput sequence-based  
771 experimental data. *Bioinformatics* **28**, 464–469 (2012).
- 772 81. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant  
773 detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
- 774 82. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:  
775 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- 776 83. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and  
777 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol.*  
778 *Biol. Evol.* **32**, 268–274 (2015).
- 779 84. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.  
780 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*  
781 **14**, 587–589 (2017).
- 782 85. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R  
783 language. *Bioinformatics* **20**, 289–290 (2004).
- 784 86. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an

- 785            efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**,  
786            1929–1936 (2014).
- 787    87. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting  
788            the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
- 789    88. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other  
790            things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
- 791    89. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–  
792            147 (1981).
- 793

## 1 **Supplementary Information for**

2

3

## 4 **Parasite hybridization promotes spreading of endosymbiotic viruses**

5 Senne Heeren, Ilse Maes, Mandy Sanders, Lon-Fye Lye, Jorge Arevalo, Alejandro Llanos-  
6 Cuentas, Lineth Garcia, Philippe Lemey, Stephen M Beverley, James A Cotton, Jean-Claude  
7 Dujardin, Frederik Van den Broeck

8

9 **Frederik Van den Broeck: [fvandenbroeck@gmail.com](mailto:fvandenbroeck@gmail.com), 0032 3 247 67 94**

10 **Jean-Claude Dujardin: [jcdujardin@itg.be](mailto:jcdujardin@itg.be), 0032 3 247 63 58**

11

12

13

## 14 **SUPPLEMENTARY METHODS**

15

### 16 **Landscape genomic analyses of *Leishmania braziliensis* (*Lb*) parasites**

17 To investigate the spatio-environmental impact on genetic variation among the three *Lb*  
18 populations, we extracted the 19 bioclimatic variables of the WorldClim2 database <sup>1</sup>. All 19  
19 variables were extracted per locality from 1 km spatial resolution raster maps after all layers  
20 were transformed to the same extent, resolution, and coordinate reference system (WGS  
21 1984). Geographic distances among sampling points were calculated as great-circle distances  
22 using geodist R-package (measure= 'haversine' <sup>2</sup>).

23 The impact of geographic distance on the genetic differentiation of *Lb* populations was  
24 assessed through Mantel tests between i) interdeme geographic distance and the Weir-  
25 Cockerham's  $F_{ST}$ <sup>3</sup>; ii) the geographic distance and Bray-Curtis genetic dissimilarity among  
26 individuals. The environmental and geographic influence on the ancestral genomic population  
27 structure of *Lb* was disentangled through redundancy analysis (RDA) and generalized  
28 dissimilarity modeling (GDM) including geographic distance and a selection of bioclimatic  
29 variables to reduce model overfitting and multicollinearity. Variable selection was performed  
30 adopting two approaches: i) mod-A: an RDA-based forward selection procedure using  
31 'ordiR2step' function of the vegan R-package <sup>4,5</sup>; ii) mod-M: a manual variable selection  
32 procedure, selecting variables if their added contribution increased the adjusted R-squared  
33 and if both the overall RDA model and individual variable were significant (p-val < 0.05).

34 Variation partitioning through RDA analysis was performed on Hellinger transformed SNP  
35 data, longitude, latitude and standardized environmental variables, using the 'decostand' and  
36 'rda' functions from the 'vegan' R-package<sup>4</sup>, to disentangle the influence of geography  
37 (isolation-by-distance) and climate (Isolation-by-environment). Finally, the different variance  
38 components were compared based on their adjusted R-squared and each explanatory

39 component was tested for significance using the `vegan::anova.cca` function. In addition to  
40 RDA variation partitioning, we constructed generalized dissimilarity models (GDMs) using the  
41 `gdm` R-package<sup>6</sup>, to investigate spatio-environmental patterns of the *Lb* genetic variability in  
42 a non-linear way<sup>6,7</sup>. The GDMs constituted a genetic distance matrix (Bray-Curtis dissimilarity  
43 of Hellinger transformed SNP genotypes) as response variable and the bioclimatic variables  
44 (as selected by the mod-A and mod-M variable selection procedures) as explanatory variables.  
45 We accounted for geographic distance effects on the genomic variability by fitting two GDMs  
46 per variable selection approach including and excluding the inter-individual geographic  
47 distance matrix. Relative variable importance in each GDM was estimated based on the I-  
48 spline basis function (i.e., the maximum height of the response curves) along with uncertainty  
49 assessment by performing 1000 iterations of each GDM model<sup>8</sup>.

50         Based on the most important environmental variables influencing the *Lb* population  
51 genomic structure, we attempted to estimate and map patches of suitable habitat for *Lb*  
52 within our study region based on present-day<sup>1</sup>, LGM<sup>9,10</sup> and LIG<sup>10,11</sup> bioclimatic data.  
53 Ecological niche models (ENMs) were constructed using Maxent, as implemented in the `dismo`  
54 R-package<sup>12,13</sup> for both environmental variable selection methods with the following  
55 parameters: linear, quadratic, product, threshold, hinge, 10 cross-validation replications and  
56 regularization multiplier (rm) set to 1, 1.5 and 2. Habitat suitability maps were constructed by  
57 averaging the predictions from all 10 replicates on present-day, LGM or LIG environmental  
58 data. A jackknife procedure was included to measure relative variable contribution and  
59 importance.

60  
61

## 62 **SUPPLEMENTARY RESULTS**

63

### 64 **Landscape genomics of *Lb* parasites**

65         Upon the strong signatures of geographical isolation of the three ancestral *Lb*  
66 components (Fig 2A) and the lack of association between the inter-population (great-circle)  
67 geographic distance and the Weir & Cockerham's  $F_{ST}$  (Supp Fig. 18; Supp Table 13), we  
68 investigated the differential influence of geography and environmental variables on the *Lb*  
69 population structure in the region. When addressing the inter-individual association of  
70 geographic distance with genetic distance (Bray-Curtis dissimilarity of SNP genotypes) we  
71 picked up a pattern resembling case-IV isolation-by-distance (i.e., increasing genetic distance  
72 with geographic distance up to a certain point after which the relationship weakens down;  
73 Supp. Fig. 19; Supp. Table 14). This revealed that isolation-by-distance mainly plays a role  
74 within populations over distances up to ca. 500km, while IBD diminishes on an inter-  
75 population level when geographic distances become too great (> 500km).

76         To investigate what other factors besides geography influenced the population  
77 divergence among the *Lb* populations, we investigated the potential impact of the abiotic  
78 environment through RDA-based variation partitioning and GDM analysis including



79 geographic distance and 19 bioclimatic variables (Supp. Table 6). Two variable selection  
80 approaches were adopted to reduce model overfitting and multicollinearity among  
81 bioclimatic variables (Supp. methods). The mod-A variable selection initially resulted in six  
82 variables, although they revealed large variance inflation factors (vif) which prompted us to  
83 remove variables with a vif > 10 in a stepwise manner, retaining only two variables:  
84 'isothermality' (bio3) and 'Precipitation of the driest month' (bio14). In contrast, the mod-M  
85 approach resulted in a final selection of five variables, each with vif < 10: 'isothermality' (bio3),  
86 'Precipitation of the driest month' (bio14), 'precipitation of warmest quarter' (bio18),  
87 'precipitation seasonality' (bio15), 'Annual mean diurnal range' (bio2). (Supp. Table 7).

88 Variation partitioning of the automated variable selection model revealed that about  
89 one-third (27.3%) of the total genomic variability could be explained by the environment  
90 (bio3, bio14) and geography together, of which both components contributed 10.2% and  
91 7.5%, respectively (Supp. Fig 8A; Supp. Table 8). In addition, the remaining 9.6% of the  
92 explained genomic variability indicates a strong confounding effect among the environment  
93 and geography components with the RDA model, meaning that about one-third of the  
94 explainable genomic variation cannot be attributed to one specific explanatory component  
95 (Supp. Table 8). In parallel, generalized dissimilarity models (GDM), including the same  
96 variables, revealed similar patterns in explaining the genomic variability by the environment  
97 (bio3, bio14) relative to geography (Supp. Fig 20). Here, GDMs could explain 55.86%  
98 (excluding geography) to 65.38% (including geography) of the genomic deviance (null  
99 deviance).

100 In contrast, variation partitioning of the RDA-model based on the manual variable  
101 selection approach (bio2, bio3, bio14, bio15, bio18) revealed a stronger environmental  
102 contribution in explaining the genomic variation in *Lb*. From the full RDA-model, explaining  
103 34.9% of the total genomic variability, about 51% could be explained by the entire  
104 environmental component whereas only 5.4% could be explained by geography (Supp. Fig 8B;  
105 Supp. Table 8). Additional GDM, explaining 59.98% (excluding geography) to 65.04%  
106 (including geography) of the genomic deviance, gave consistently similar results to the RDA  
107 model (Supp. Fig 20).

108 In accordance with the different variation partitioning models (RDA and GDM),  
109 revealing the key role in explaining the *Lb* population divergence, present-day habitat  
110 suitability predictions showed that regions of suitable (abiotic) habitat for *Lb* coincided with  
111 tropical rainforests, as predicted by the Koppen-Geiger climate classification, where the three  
112 ancestral *Lb* populations were surrounded by less suitable tropical monsoon forests (Fig 2B,C).  
113 Additional suitability predictions using Last Glacial Maximum (LGM) and Last Interglacial (LIG)  
114 periods revealed similar regions of suitable habitat, suggesting the suitable regions for  
115 ancestral *Lb* populations have been relatively stable over the past 120,000 years (Supp. Fig 9,  
116 10).

117

118

## 119 **Quality assessment of the LRV1 genome assemblies**

120 The LRV1 genomes included in this study were generated through dsRNA extraction  
121 from 31 LRV1-positive *L. braziliensis* isolates that were re-cultured following total RNA  
122 sequencing, *de novo* assembly and LRV1 contig extraction by a local BLAST search (see  
123 methods). This procedure failed for two *Lb* isolates, either due to difficulties during culturing  
124 (PER096) or because the assembly yielded a partial LRV1 genome (PER231). Two *Lb* isolates  
125 (CUM65 and LC2321) each harbored two LRV1 genomes, differing at 999 (for CUM65) and 60  
126 (for LC2321) nucleotides, bringing the total to 31 viral genomes. The assembly quality of the  
127 genomes was examined by investigating the coverage of mapped (paired) reads, SNP counts  
128 after mapping, comparison with analogous genomic regions to ~1kb sequences obtained  
129 through conventional Sanger sequencing and by read-based computational assembly  
130 improvement using Pilon <sup>14</sup>.

131 From the average 0.04% of reads that mapped against their respective LRV1 contig  
132 (Supp. Table 9), an average of 85.7% (81.3% - 91.5%) of the reads were properly paired (i.e.,  
133 correctly oriented reads with respect to each other and with proper insert sizes). In addition,  
134 most LRV1 strains contained few SNPs (zero to three) of which all were heterozygous except  
135 one. PER130 showed one homozygous SNP, which was located at the 5' end (position 6) of  
136 the assembled genome which was trimmed off in downstream analyses. However, in LRV-Lb-  
137 LC2321 we encountered 51 heterozygous SNPs (data not shown), suggesting the possibility of  
138 a mixed viral infection (i.e. two LRV1 strains present in one parasite isolate of which one is  
139 much lower in abundance than the other). To examine this in more detail, we extracted the  
140 reads mapping to the potential chimeric LRV1 genome and re-assembled the reads using a  
141 recently developed strain-resolving *de novo* assembler <sup>15</sup>, developed to extract various viral  
142 strains from mixed infection samples. This resulted in two LRV1 contigs: LC2321.1 (5,260bp)  
143 and LC2321.2 (4,738bp), and considerably dropped the number of heterozygous SNPs found  
144 in both strains. LC2321.1 showed nine heterozygous SNP of which eight were located on a  
145 non-resolved part of the LC2321.2 genome. LC2321.2 on the other hand did not reveal any  
146 SNPs. Furthermore, the remaining quality statistics of both strain resolved genomes from  
147 LC2321 were similar to the other assemblies (Supp. Table 9).

148 The sequence identity of the assembled genomes was assessed by comparison with  
149 analogous ~1kb (1197bp) sequences obtained through conventional Sanger Sequencing. This  
150 revealed for 83.8% (26/31) of the genomes a sequence identity of 100% (Supp. Table 9). The  
151 remaining five genomes encompassed both genomes of CUM65 and LC2321 (mixed LRV1  
152 infections) and PER212. For CUM65, we observed a sequence identity of 99% (12 mismatches)  
153 and 86% (152 mismatches) for LRV1-Lb-CUM65.1 and LRV1-Lb-CUM65.2, respectively. For  
154 LC2321, both resolved strains showed a sequence identity of 99% with three and ten  
155 nucleotide mismatches in LRV1-Lb-LC2321.1 and LRV1-Lb-LC2321.2, respectively. These  
156 mismatches were unique to each strain, which might indicate the Sanger sequence is a  
157 chimera of both strains. Finally, the 99% identity of PER212 with its respective partial Sanger

158 sequence showed only one nucleotide mismatch, not corresponding to the identified  
159 heterozygous SNP, suggesting a badly called base during the Sanger sequencing.

160

161

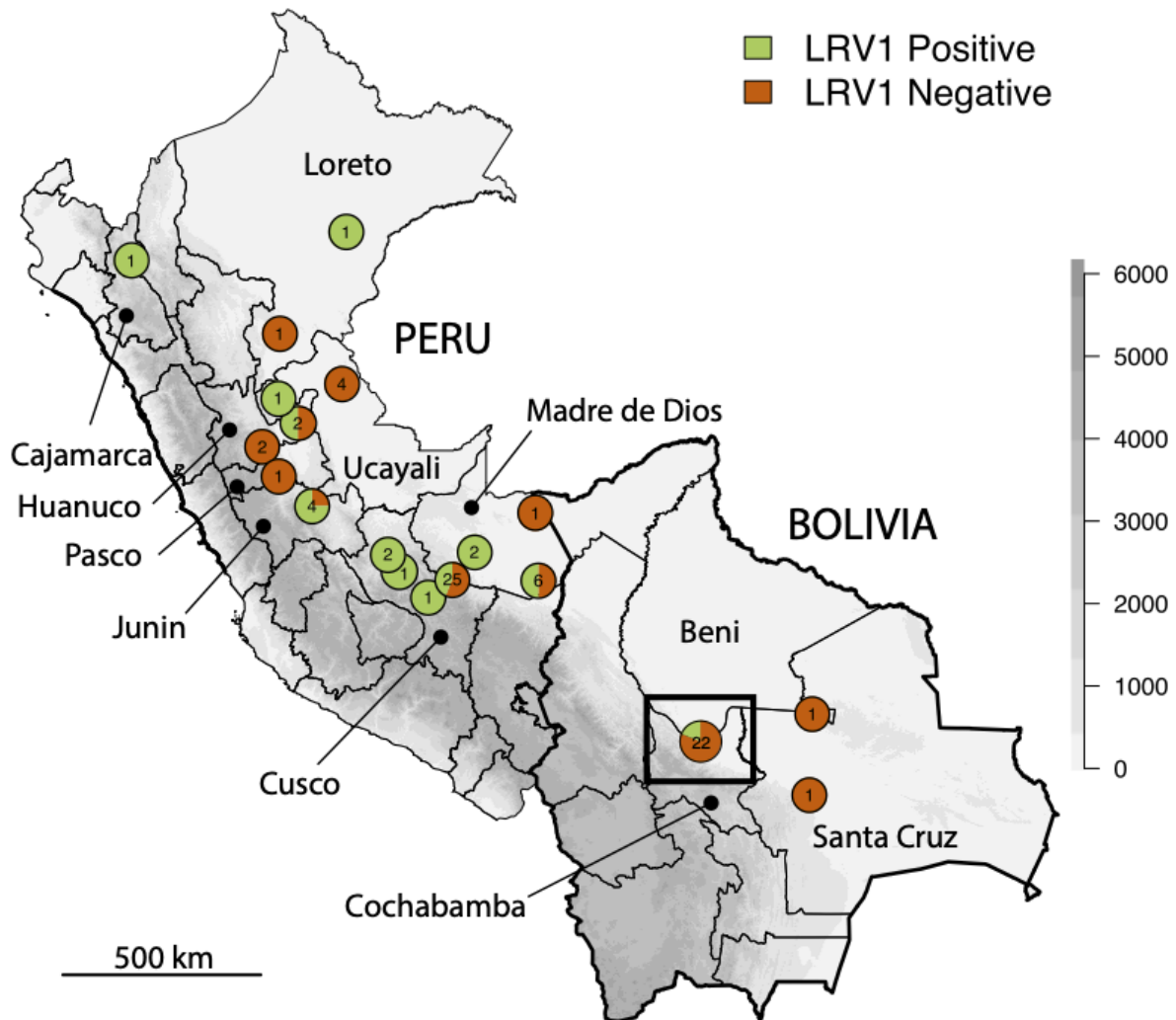
162

163

164

165 SUPPLEMENTARY FIGURES

166



167

168 **Supplementary Figure 1.** Geographic origin of 79 *Lb* isolates from Peru and Bolivia that were  
169 included in this study. Rectangular box indicates the location of the Isiboro National Park that  
170 extends between the Departments of Cochabamba and Beni. Gray-scale represents altitude  
171 in meters.

172

173

174

175

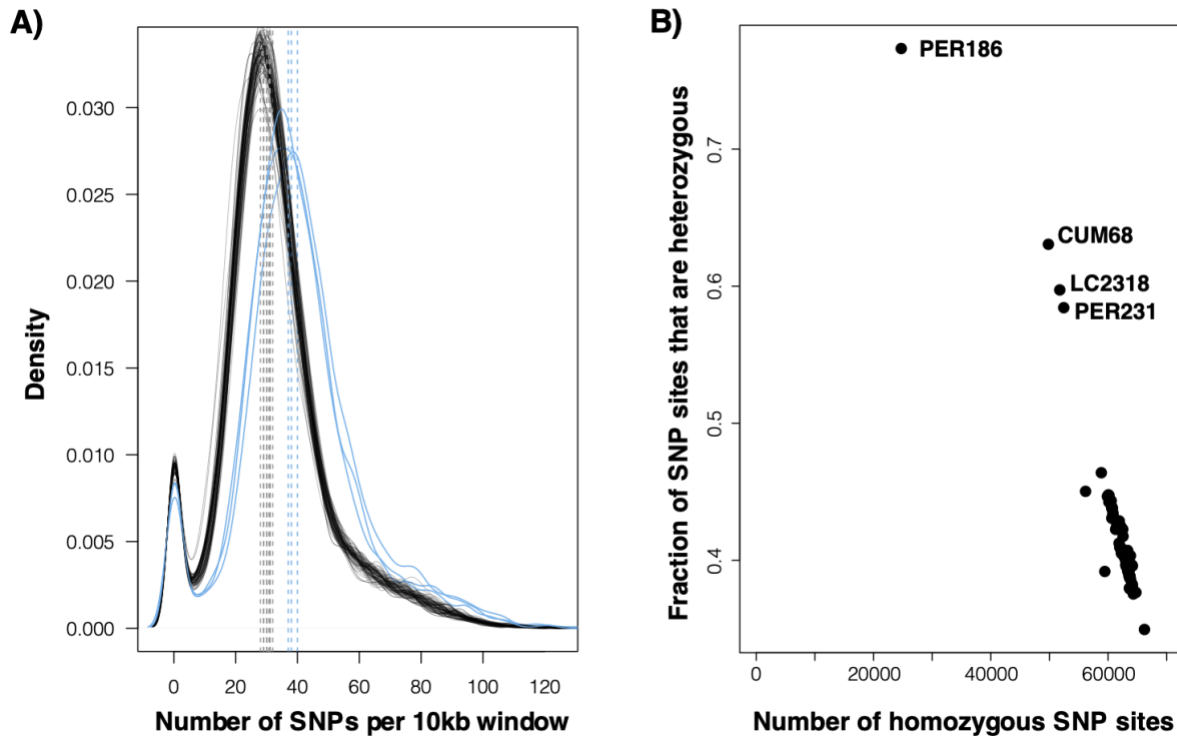
176

177

178

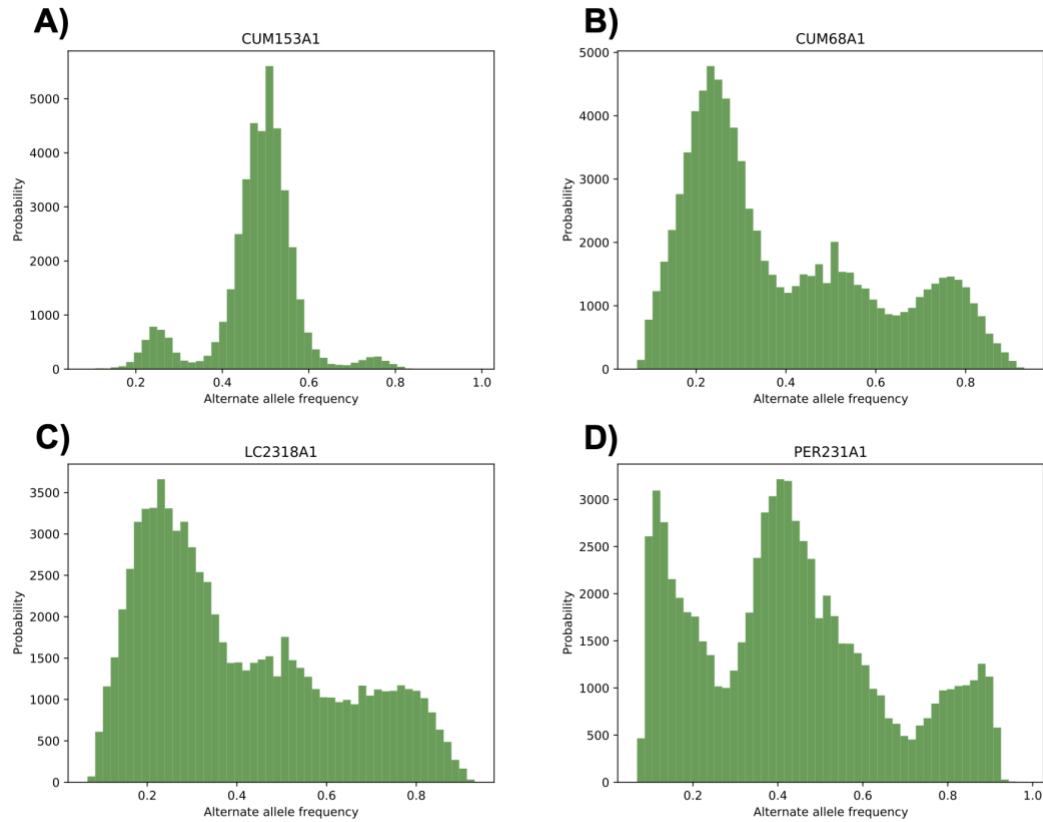
179

180



181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192

**Supplementary Figure 2. A)** Kernel density plots of the number of SNPs per 10kb window for each of the 79 *Lb* genomes. The median number of SNPs per 10kb window is indicated with gray vertical dashed lines and ranges between 28 and 32 SNPs for the majority of isolates. Three isolates showed slightly larger SNP densities (indicated with blue lines in the plot): 37 SNPs in PER231, 38 SNPs in LC2318 and 40 SNPs in CUM68. **B)** Fraction of SNP sites that are heterozygous versus the number of homozygous SNP sites in each of the 79 *Lb* genomes.



193

194

195 **Supplementary Figure 3.** Genome-wide distribution of alternate allele read depth

196 frequencies at heterozygous sites. **(A)** Example of a largely diploid individual (CUM153) with

197 allele frequencies centered around 0.5, which was observed for 77/79 *Lb* genomes included

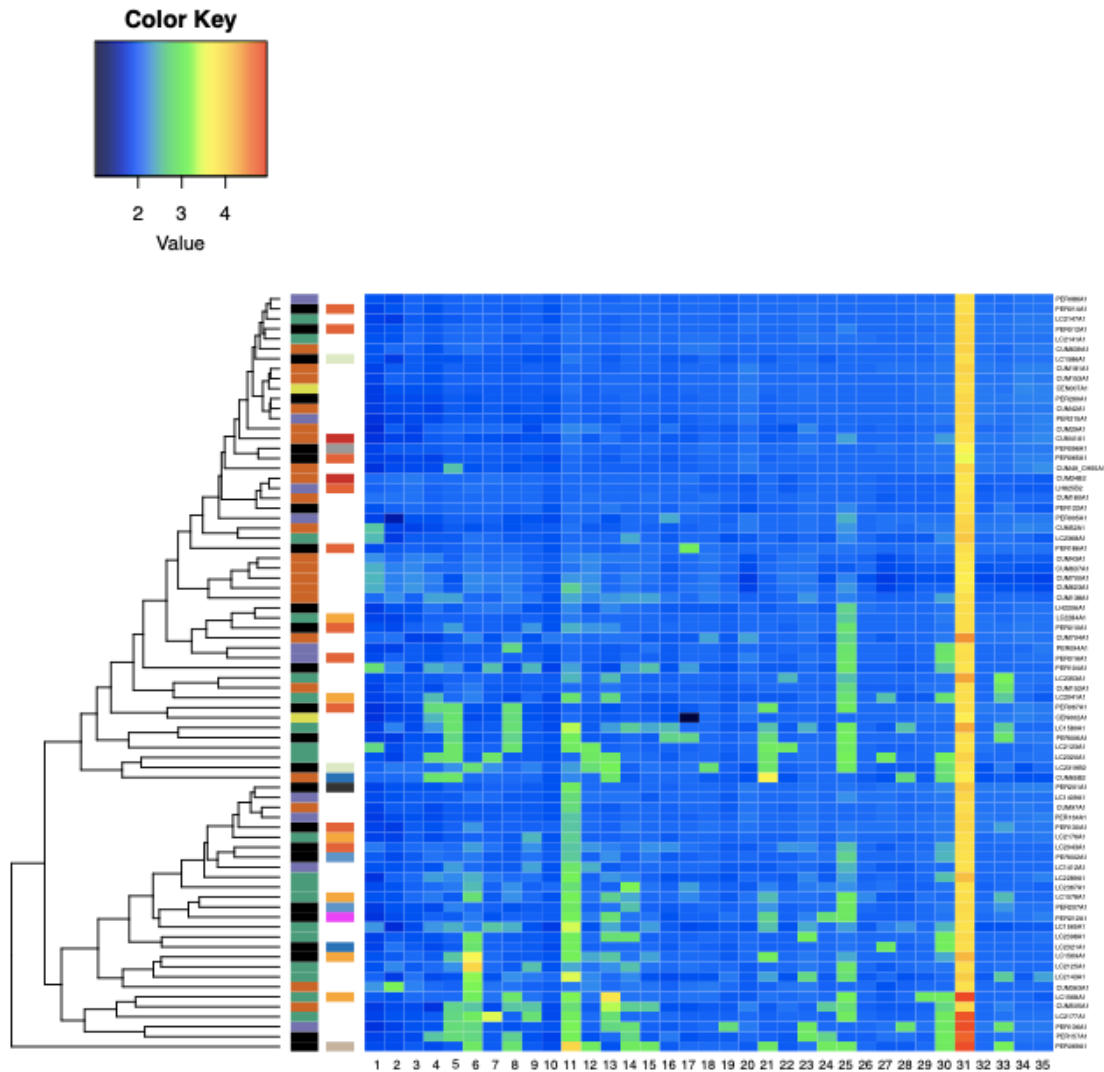
198 in this study. **(B-C)** Two isolates (CUM68 and LC2318) were symptomatic of tetraploidy, with

199 modes of allele frequencies equal to 0.25, 0.5 and 0.75. **(C)** One isolate (PER231) showed a

200 skewed distribution, suggesting that it may be the result of a mixed infection or

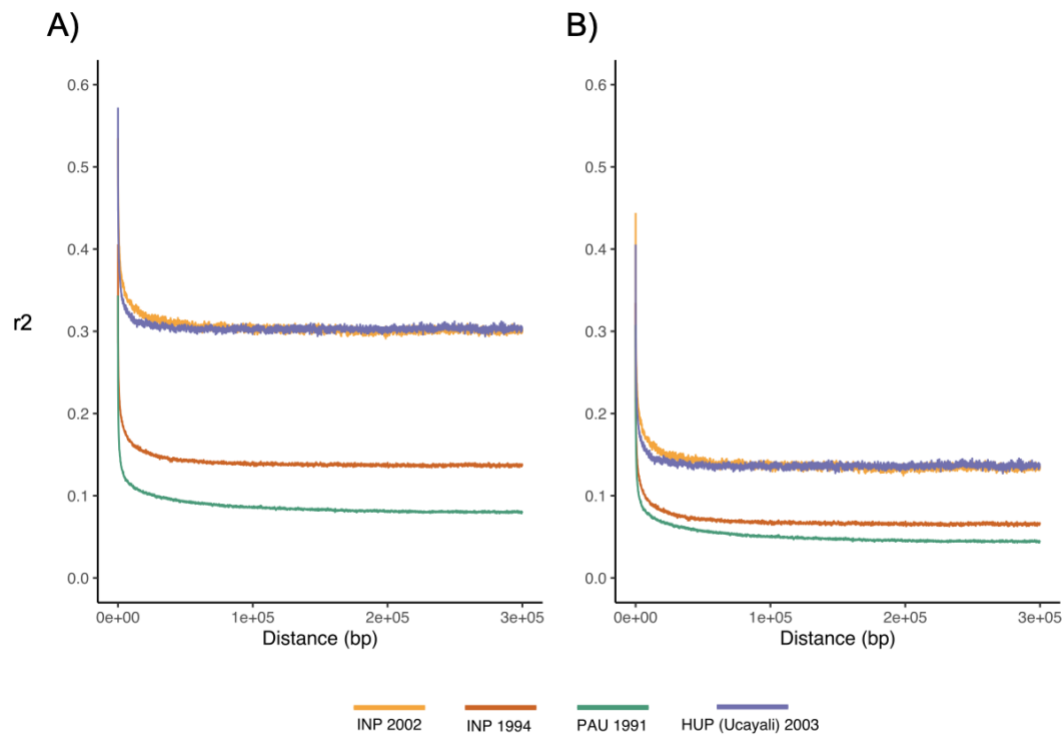
contamination.





201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212

**Supplementary Figure 4.** Variation in chromosome copy numbers in a panel of 76 *Lb* genomes. Isolates are clustered according to similarity in some estimation with aneuploid individuals at the bottom of the heatmap and overall diploid individuals at the top of the heatmap. Coloured boxes on the left of the heatmap represent: *Left* - the inferred *Lb* populations PAU (green), INP (orange), HUP (purple), STC (yellow-green) and ADM (black); *Right* - the identified LRV1 lineages encountered in each *Lb* isolate. I (orange), II (dark gray), III (pink), IV (steelblue), V (yellow), VI (beige), VII (dark blue), VIII (red), IX (light green), NA (light gray).



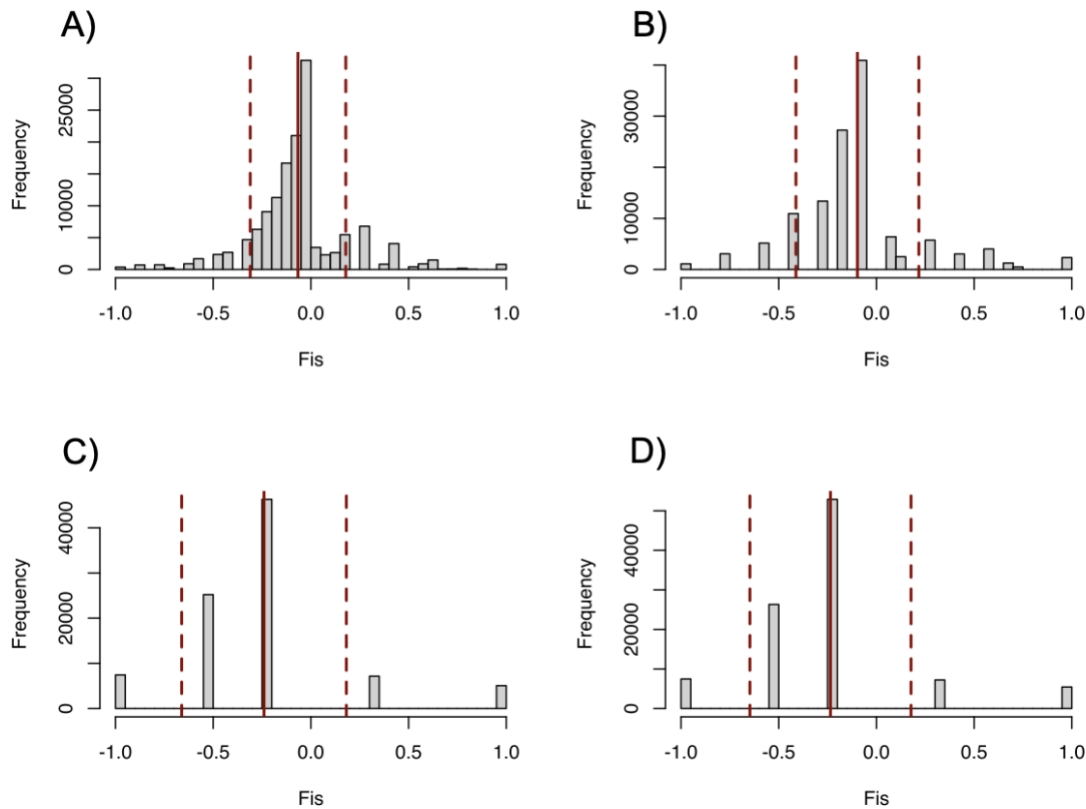
213

214 **Supplementary Figure 5.** Linkage Disequilibrium decay plots after correction for population

215 structure and spatio-temporal Wahlund effects. **A)** Uncorrected for sample size. **B)** Corrected

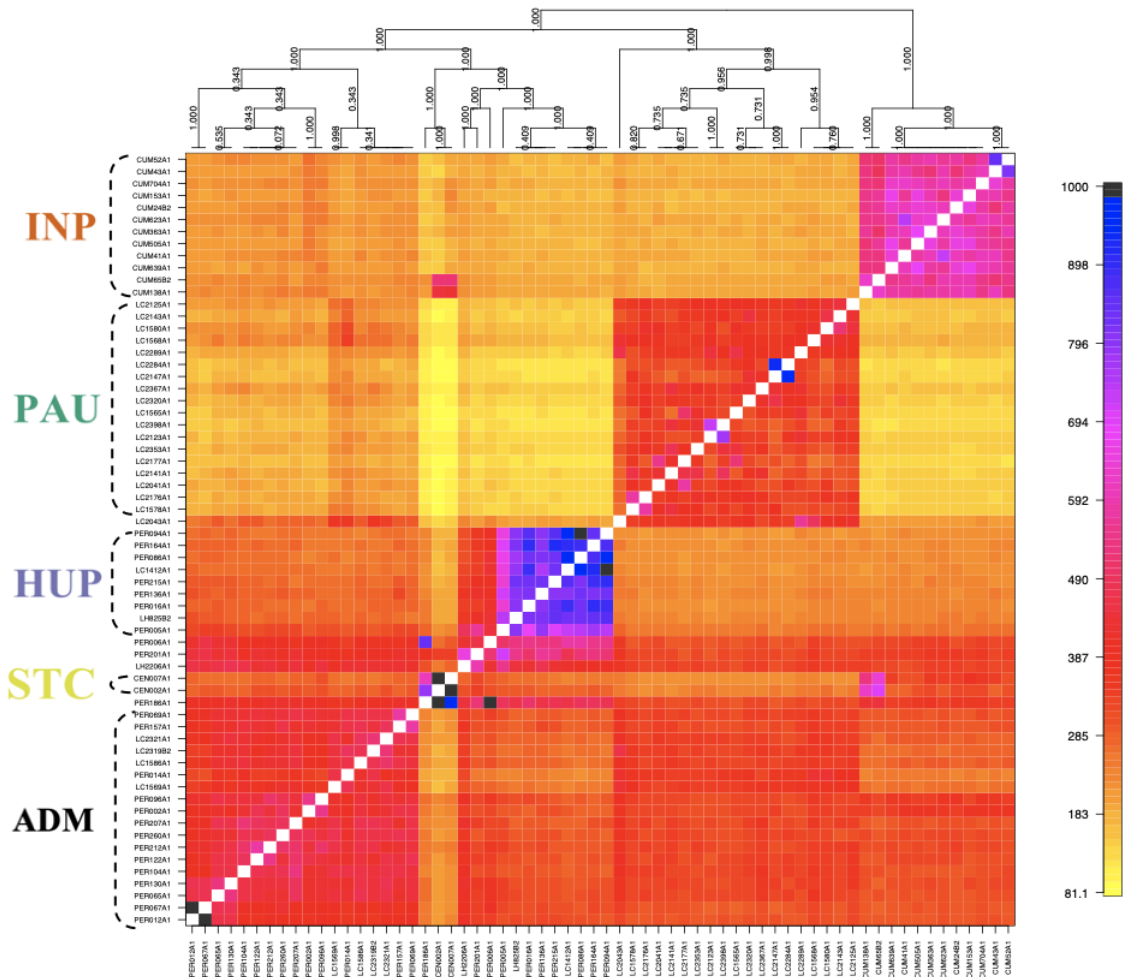
216 for sample size  $(r2 - 1/(2n))^{16}$ .

217



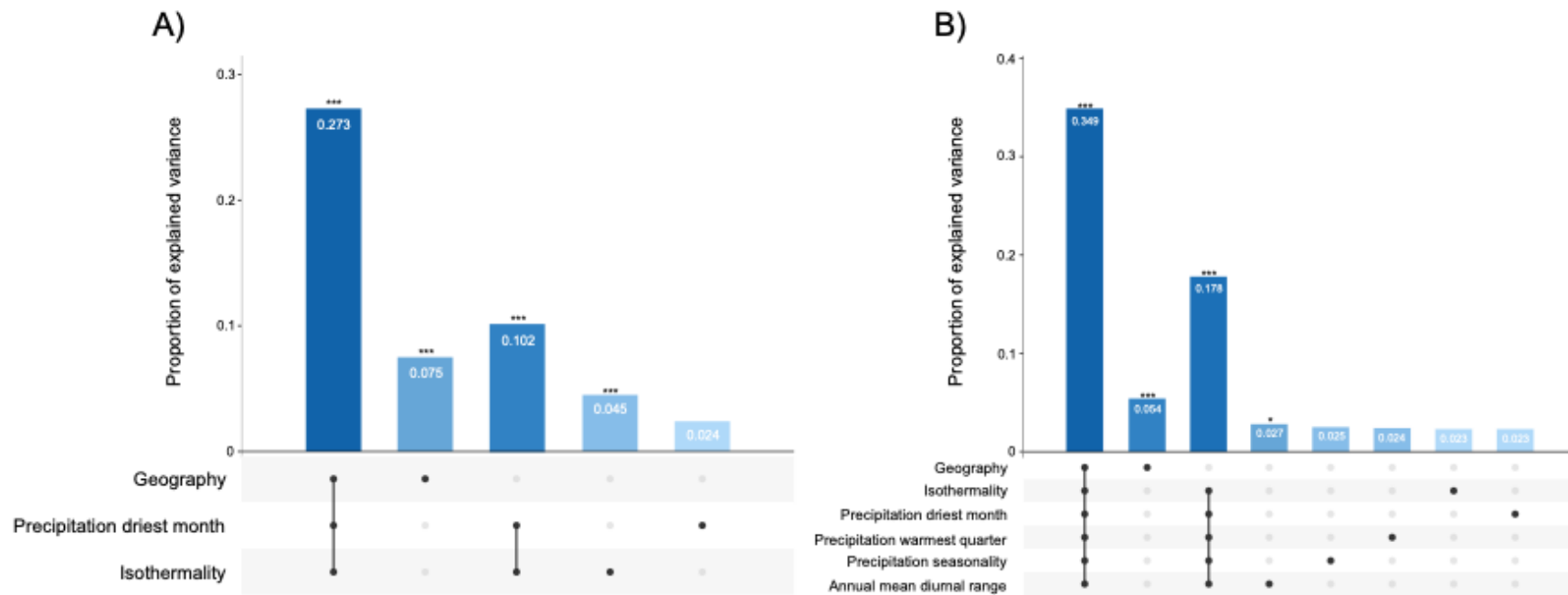
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230

**Supplementary Figure 6.** Fis distributions after correction for population structure and spatio-temporal Wahlund effects. **A)** Individuals from PAU sampled in 1991 (N=14). **B)** Individuals from INP sampled in 1994 (N=7). **C)** Individuals from INP sampled in 2002 (N=3). **D)** Individuals from HUP (Ucayali) sampled in 2003 (N=3). Solid red lines depict the mean Fis value for each population. Dashed red lines represent the mean's standard deviation.

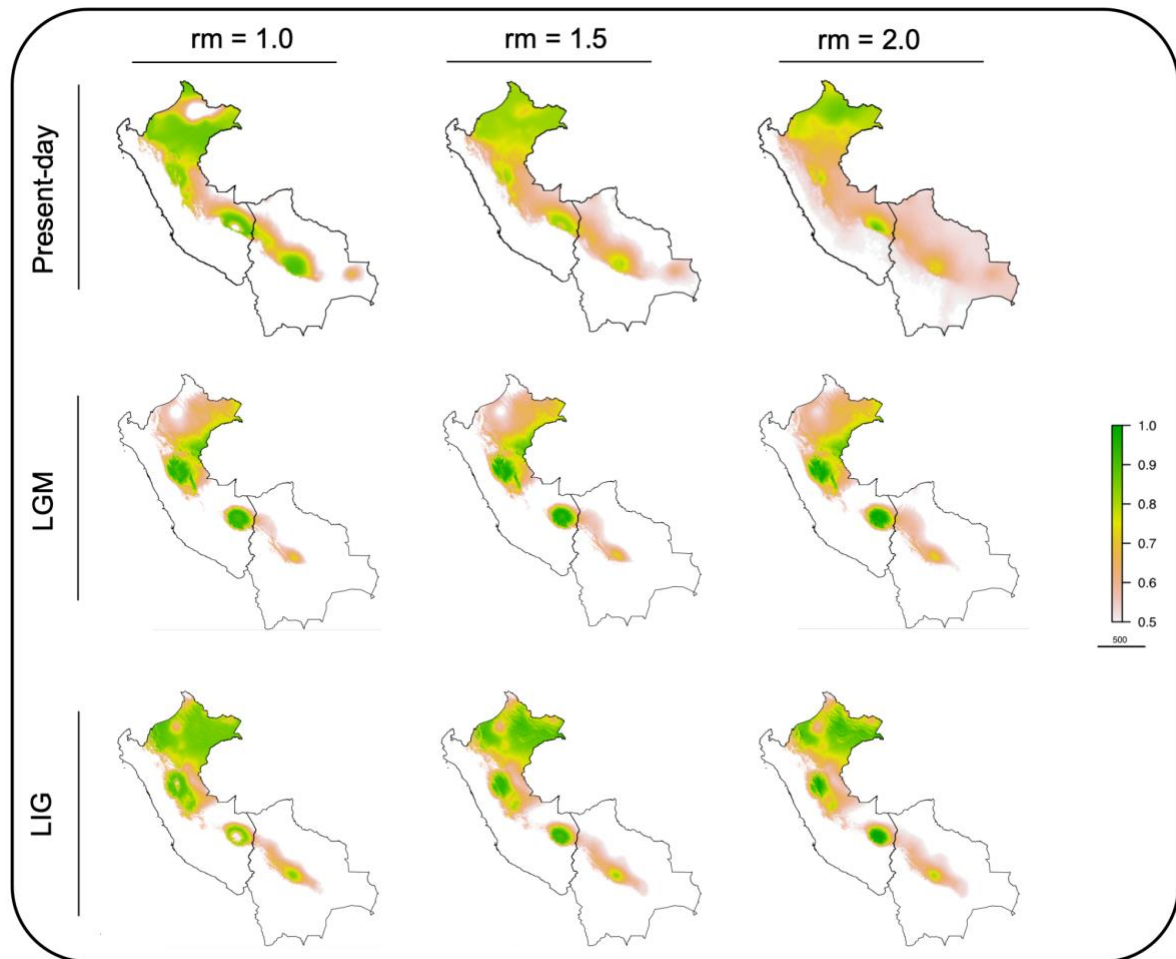


231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242

**Supplementary Figure 7.** Co-ancestry matrix, as inferred by fineSTRUCTURE, depicting the pairwise number of received (rows) and donated (columns) haplotype segments between two parasite genomes. Color key on the right shows the amount of shared haplotype segments and is capped on 1000 for visibility reasons. Individuals are ordered according to the fineSTRCUTURE clustering outcome (above matrix) and dashed accolades indicate the three main parasite groups (INP, HUP, PAU) and the two main groups of admixed parasites (ADM, STC); the remainder of the parasites were of uncertain ancestry (UNC).

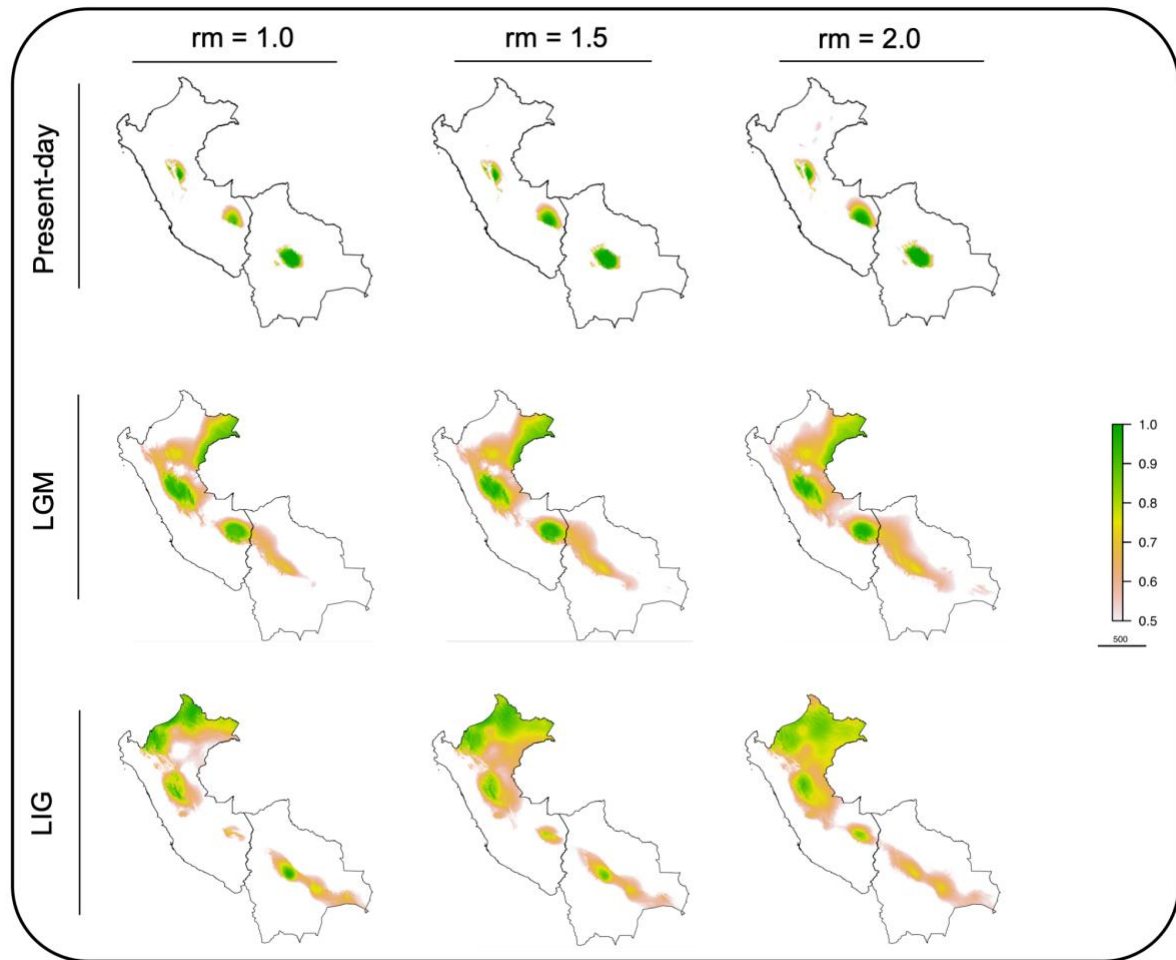


**Supplementary Figure 8:** Partial RDA models showing the influence of the environment (bioclimatic variables) and geography on the genomic variability between the three inferred populations in Peru and Bolivia. A) Partial RDA model including geography, isothermality (bio3) and precipitation of driest month (bio14). Bioclimatic variables were selected based on the automated variable selection approach (suppl. Table 6; see methods). B) Partial RDA model including geography, Isothermality (bio3), precipitation driest month (bio14), precipitation warmest quarter (bio18), precipitation seasonality (bio15) and annual mean diurnal range (bio2). Bioclimatic variables were selected based on the manual variable selection approach (suppl. 6; see methods).

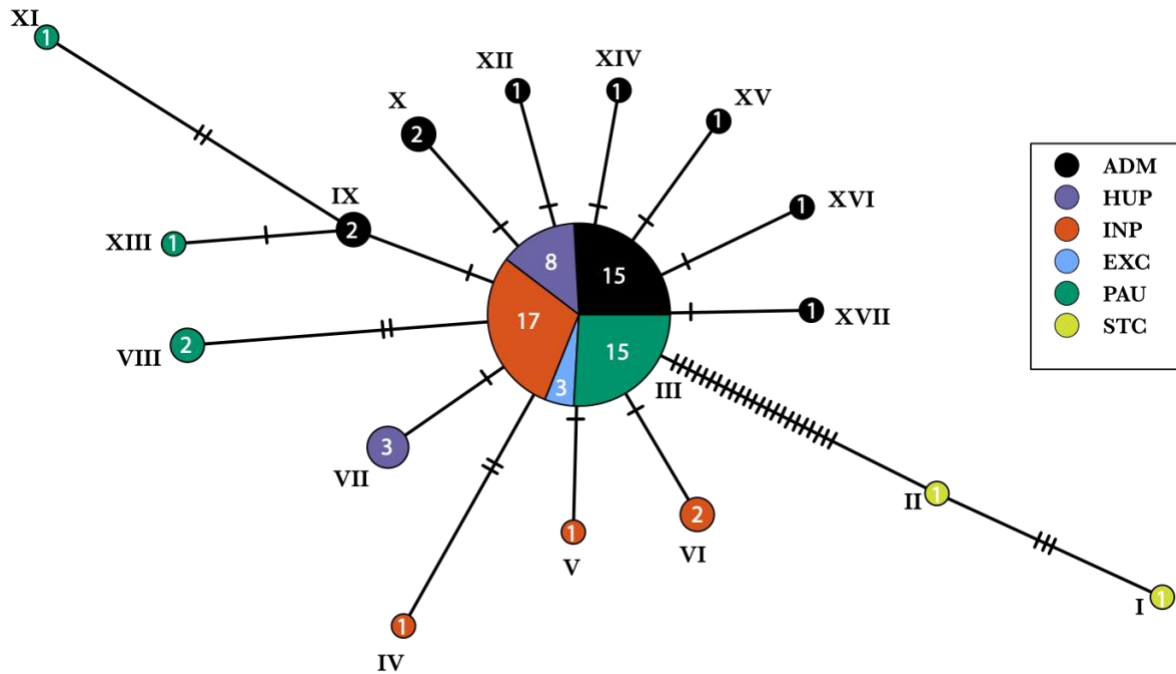


**Supplementary Figure 9.** Ecological niche models based on Present-day, LGM and LIG data of isothermality (bio3) and Precipitation of the driest month (bio14) (mod-A variable selection approach) with different regularization values ( $rm = 1, 1.5$  and  $2$ ). The continuous-scale legend represents habitat suitability (probability of occurrence).

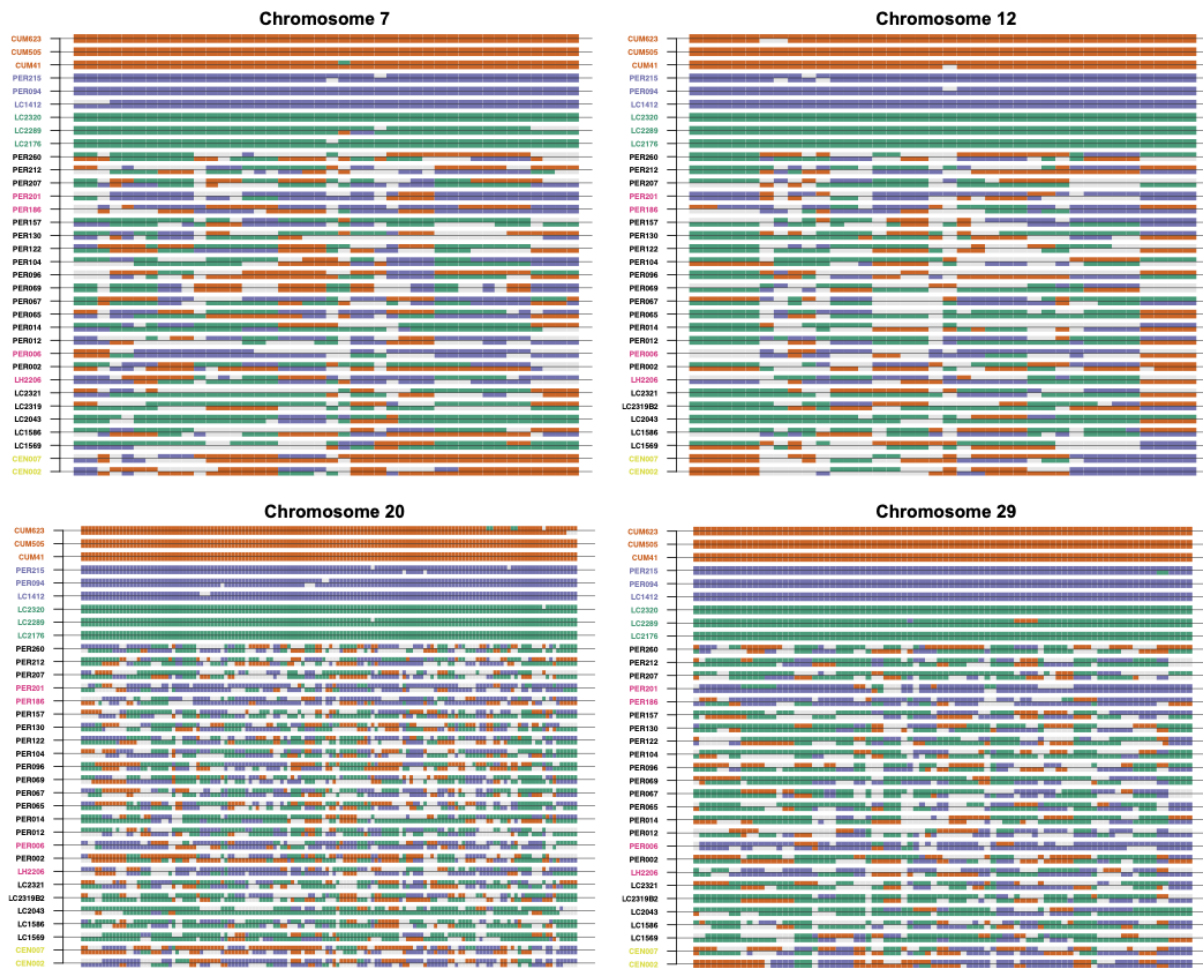




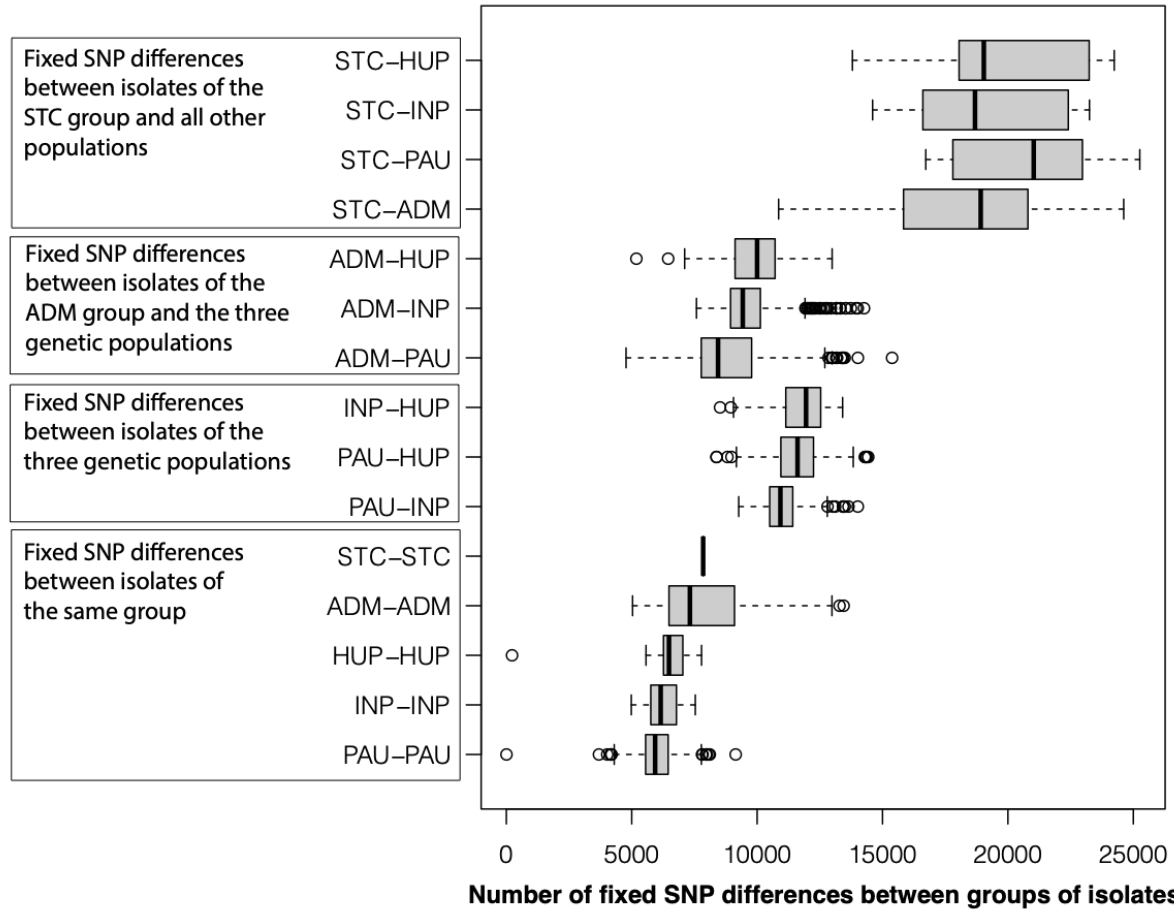
**Supplementary Figure 10.** Ecological niche models based on Present-day, LGM and LIG data of isothermality (bio3), Precipitation of the driest month (bio14), precipitation of the warmest quarter (bio18), precipitation seasonality (bio15) and annual mean diurnal range (bio2) (mod-M variable selection approach) with different regularization values ( $rm = 1, 1.5$  and  $2$ ). The continuous-scale legend represents habitat suitability (probability of occurrence).



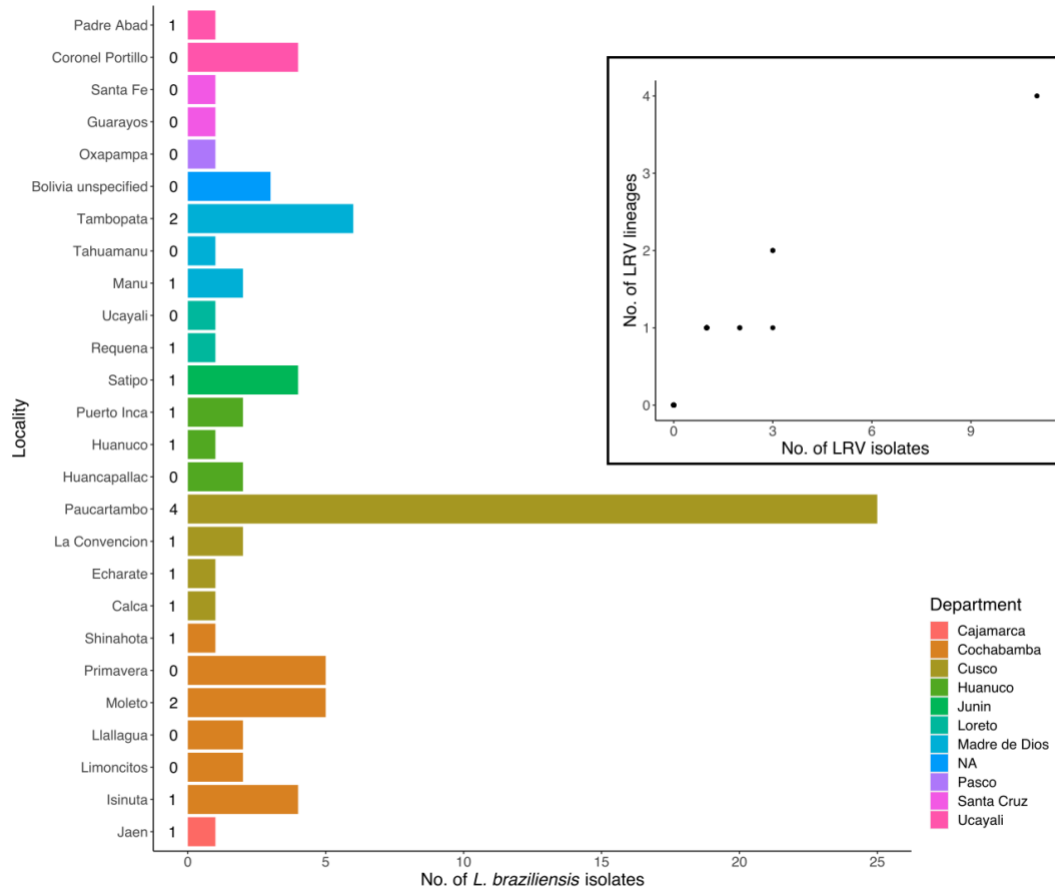
**Supplementary Figure 11.** TCS Haplotype network with PopART based on 53 high-quality SNPs identified within the coding region of the haploid mitochondrial maxicircle. A total of 17 haplotypes (shown with Roman numbers) were identified within our set of 80 *Lb* isolates from Peru and Bolivia. The size of the circles represent the number of sequences that represent a given haplotype; this number is also written in white within each circle. Colors indicate the five groups of parasites as identified with ADMIXTURE and fineSTRUCTURE using genome-wide SNPs; the EXC group indicates the three isolates that were excluded from population structure analyses. The dominant haplotype III is represented by 58 maxicircle sequences (72.5% of the 80 included sequences) and is found in four of the five groups. Black bars represent the number of mutations between two haplotypes.



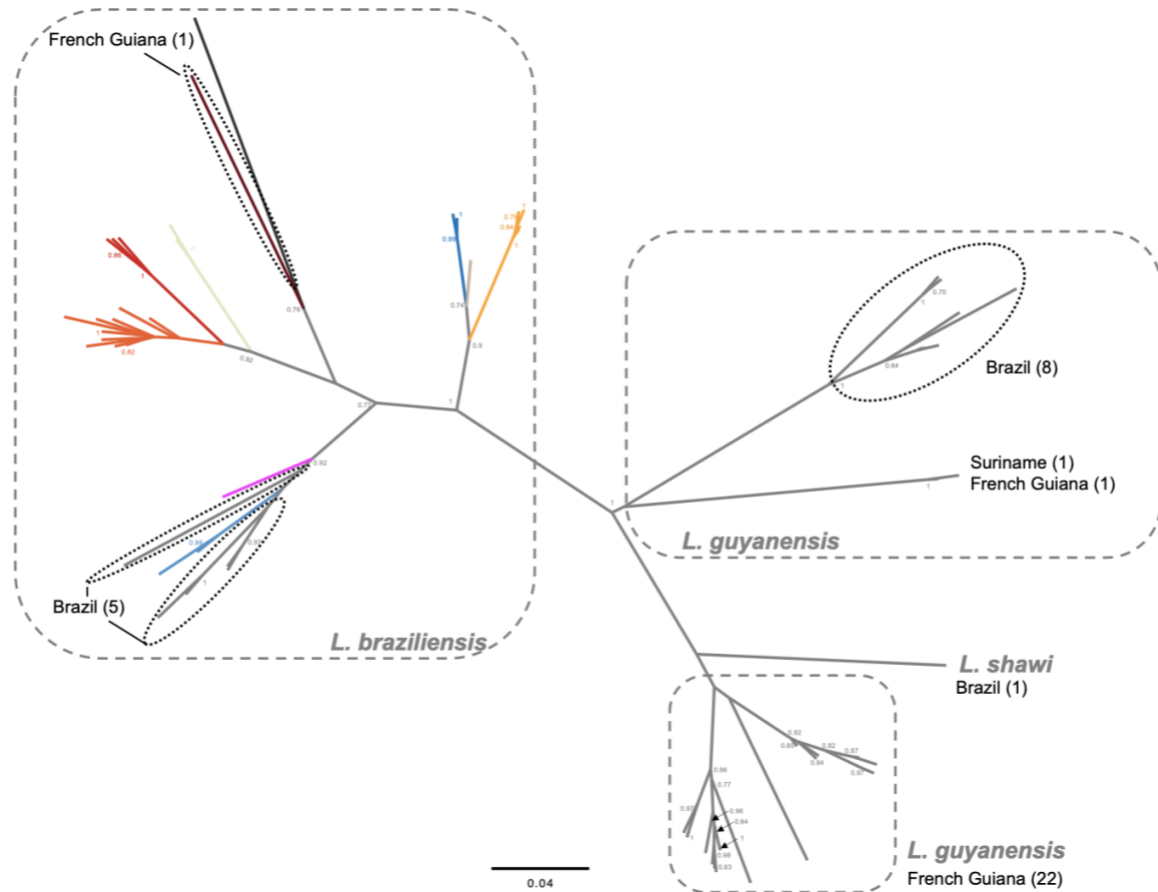
**Supplementary Figure 12.** PCAdmix local ancestry assignment to PAU, INP and HUP source populations of the 19 ADM isolates, 4 UNC isolates, 2 STC isolates and three randomly selected isolates from each source population. Ancestry was assigned in windows of 30 SNPs along each chromosome (here only chromosomes 7, 12, 20 and 29 are shown).



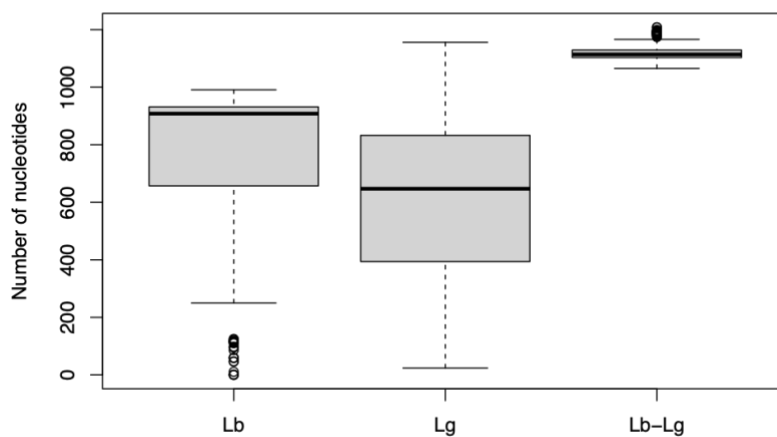
**Supplementary Figure 13.** Number of fixed SNP differences between *Lb* isolates of the same group or between *Lb* isolates of different groups.



**Supplementary Figure 14.** Barplot shows the number of *Lb* isolates per sampling locality. Number on the right of each bar shows the number of *Lb* isolates that were positive for LRV1. Bars are coloured according to the Department. Inset reveals the number of LRV1 lineages versus the number of LRV1 isolates that were recovered in a given locality.

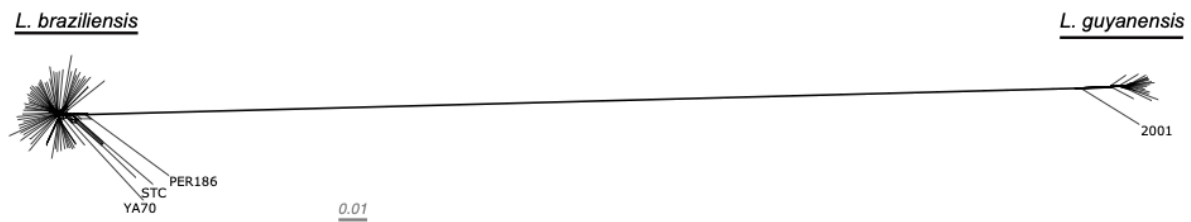


**Supplementary Figure 15.** Maximum likelihood tree based on partial LRV1 sequences (756bp) from *L. braziliensis*, *L. guyanensis* and *L. shawi* originating from Peru, Bolivia, Brazil, French Guiana and Suriname<sup>17–20</sup>. Colored lineages in *L. braziliensis* correspond to the LRV1 lineages described in this study (Fig. 3).

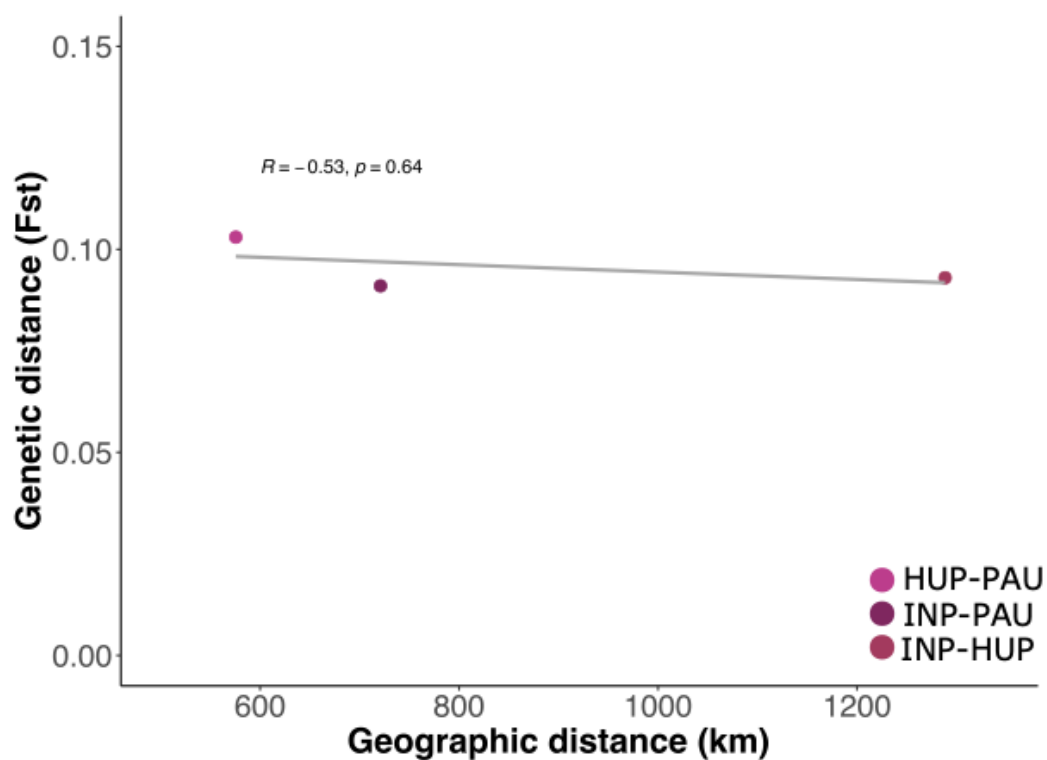


**Supplementary Figure 16.** Nucleotide differences - on average higher between *Lb* and *Lg* viral genomes than within *Lb* and *Lg*.

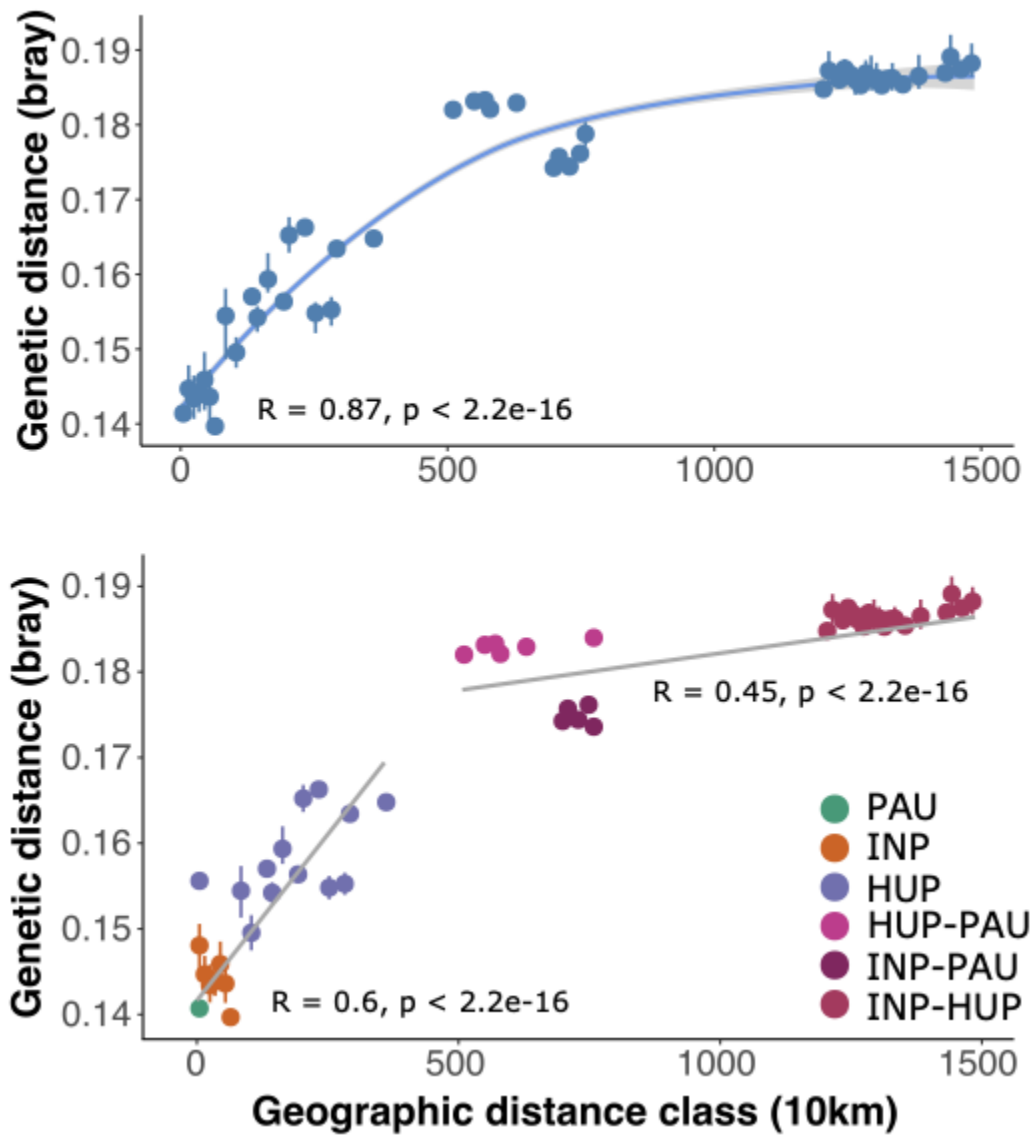




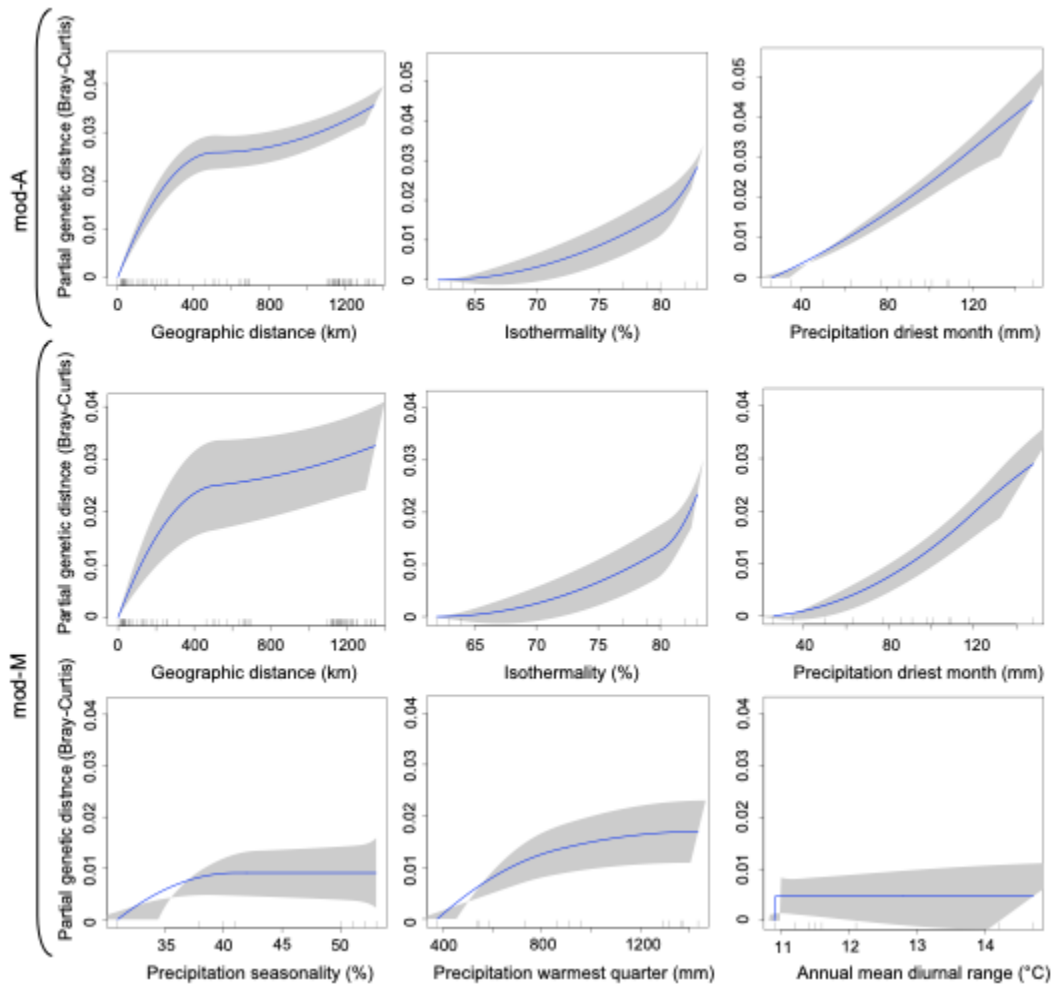
**Supplementary Figure 17.** A phylogenetic network, inferred with SPLITSTREE, based on uncorrected p-distances between 77 *Lb* and 19 *Lg* isolates typed at 7,571 bi-allelic SNPs.



**Supplementary Figure 18.** Linear regression of the interdeme great-circle distance (km) and the Weir & Cockerham's  $F_{st}$  of the three ancestral *L. braziliensis* components.



**Supplementary Figure 19.** Regressions of the inter-individual great-circle distance (km) and pairwise genetic distance (Bray-Curtis dissimilarity of SNP genotypes) revealing a case-IV IBD pattern. a) Loess regression for all individuals with a loess value of 1.2. b) Linear regression of intra- and inter-population genetic distance vs. geographic distance, separately.



**Supplementary Figure 20.** I-spline response curves for each variable included in the different generalized dissimilarity models (mod-A & mod-M). The maximum curve height represents the amount of genetic variability the variable explains (i.e. the variable importance). The curves' slope indicates the degree of the explained genomic dissimilarity along the spatial or environmental gradient, meaning a steeper slope represents greater dissimilarity between two points while a shallower slope suggests less variability among the two points.