Supplementary Information for

Parasite hybridization promotes spreading of endosymbiotic viruses

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

Frederik Van den Broeck: fvandenbroeck@gmail.com, 0032 3 247 67 94
Jean-Claude Dujardin: jcdujardin@itg.be, 0032 3 247 63 58

SUPPLEMENTARY METHODS

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

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

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

Variation partitioning through RDA analysis was performed on Hellinger transformed SNP data, longitude, latitude and standardized environmental variables, using the 'decostand' and 'rda' functions from the 'vegan' R-package[4], to disentangle the influence of geography (isolation-by-distance) and climate (Isolation-by-environment). Finally, the different variance components were compared based on their adjusted R-squared and each explanatory component was tested for significance using the vegan::anova.cca function. In addition to RDA variation partitioning, we constructed generalized dissimilarity models (GDMs) using the gdm R-package[6], to investigate spatio-environmental patterns of the *Lb* genetic variability in a non-linear way [6,7]. The GDMs constituted a genetic distance matrix (Bray-Curtis dissimilarity of Hellinger transformed SNP genotypes) as response variable and the bioclimatic variables (as selected by the mod-A and mod-M variable selection procedures) as explanatory variables. We accounted for geographic distance effects on the genomic variability by fitting two GDMs per variable selection approach including and excluding the inter-individual geographic distance matrix. Relative variable importance in each GDM was estimated based on the I-spline basis function (i.e., the maximum height of the response curves) along with uncertainty assessment by performing 1000 iterations of each GDM model [8].

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

SUPPLEMENTARY RESULTS

Landscape genomics of *Lb* parasites

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

To investigate what other factors besides geography influenced the population divergence among the *Lb* populations, we investigated the potential impact of the abiotic environment through RDA-based variation partitioning and GDM analysis including geographic distance and 19 bioclimatic variables (Supp. Table 6). Two variable selection approaches were adopted to reduce model overfitting and multicollinearity among bioclimatic variables (Supp. methods). The mod-A variable selection initially resulted in six variables, although they revealed large variance inflation factors (vif) which prompted us to remove variables with a vif > 10 in a stepwise manner, retaining only two variables: 'isothermality' (bio3) and 'Precipitation of the driest month' (bio14). In contrast, the mod-M approach resulted in a final selection of five variables, each with vif < 10: 'isothermality' (bio3), 'Precipitation of the driest month' (bio14), 'precipitation of warmest quarter' (bio18), 'precipitation seasonality' (bio15), 'Annual mean diurnal range' (bio2). (Supp. Table 7).

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

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

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

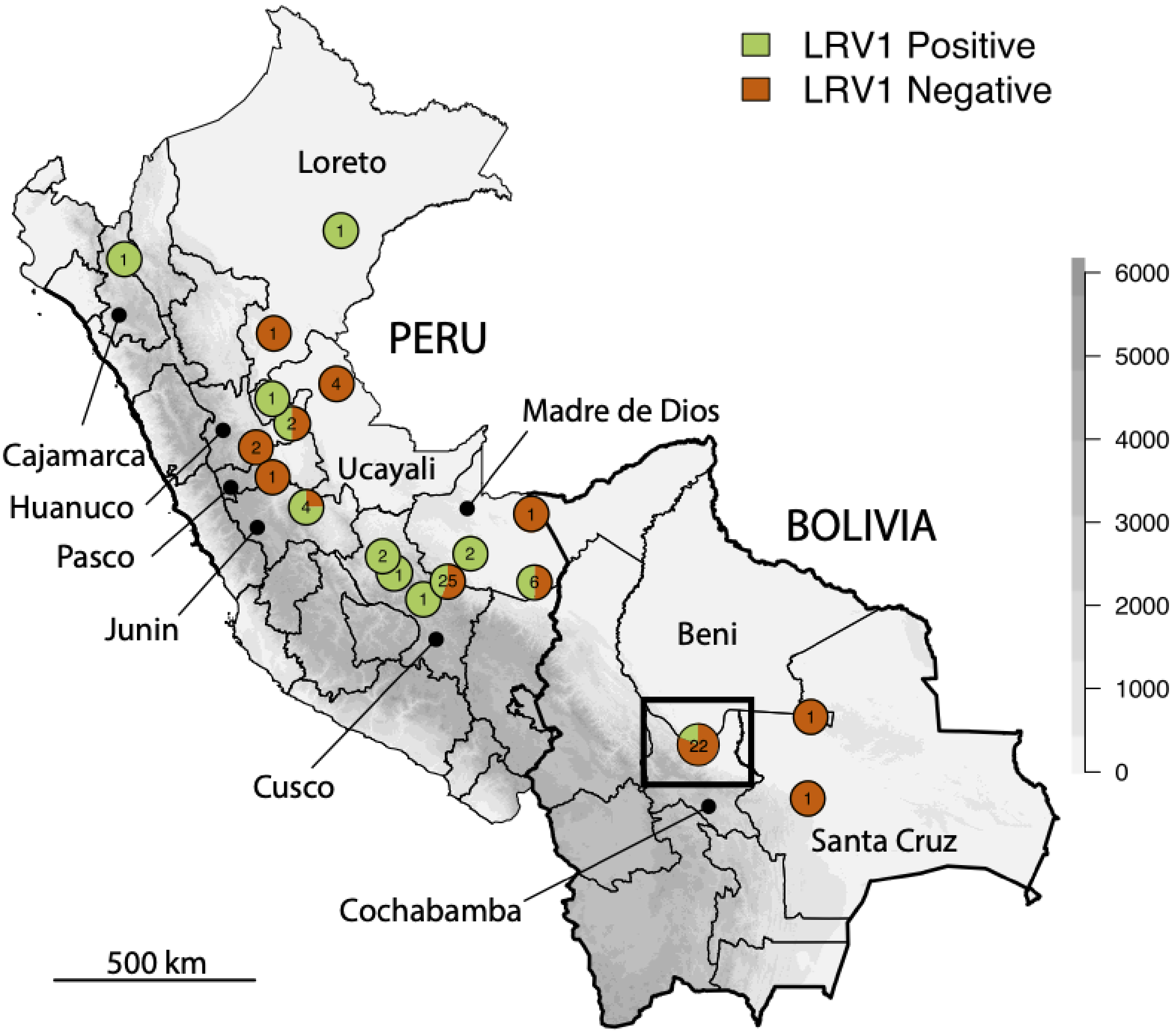
Quality assessment of the LRV1 genome assemblies

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

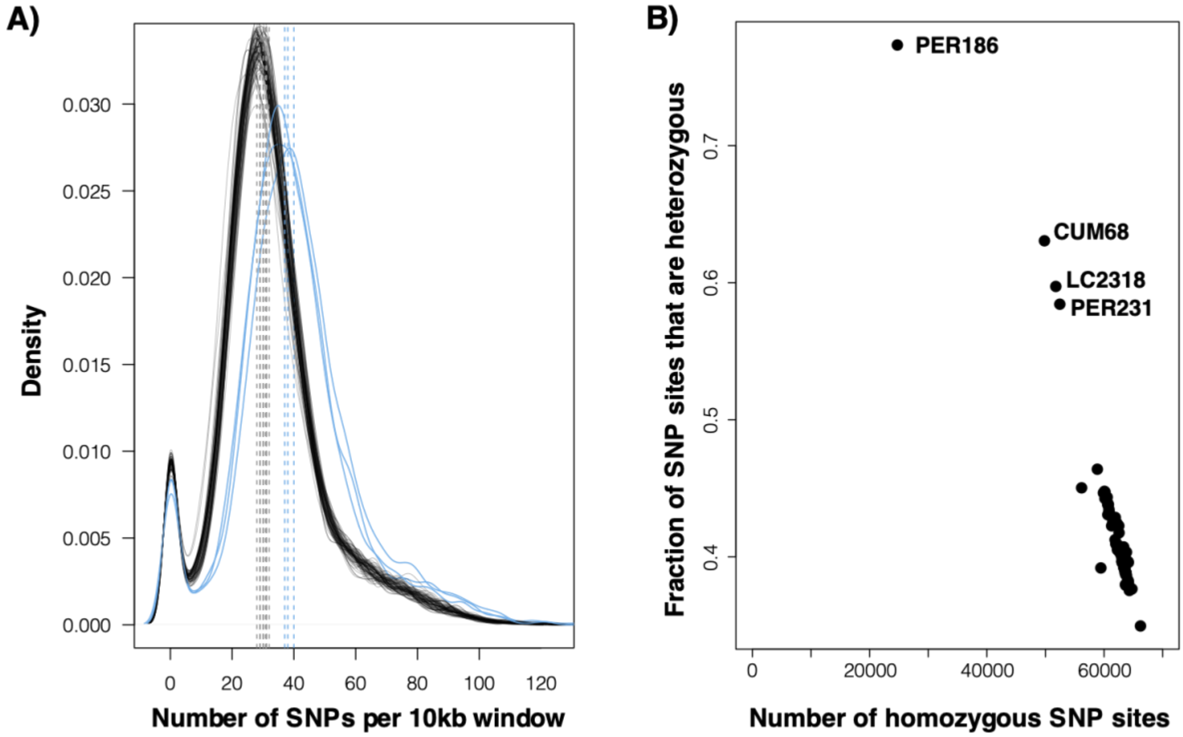
From the average 0.04% of reads that mapped against their respective LRV1 contig (Supp. Table 9), an average of 85.7% (81.3% - 91.5%) of the reads were properly paired (i.e., correctly oriented reads with respect to each other and with proper insert sizes). In addition, most LRV1 strains contained few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

The sequence identity of the assembled genomes was assessed by comparison with analogous ~1kb (1197bp) sequences obtained through conventional Sanger Sequencing. This revealed for 83.8% (26/31) of the genomes a sequence identity of 100% (Supp. Table 9). The remaining five genomes encompassed both genomes of CUM65 and LC2321 (mixed LRV1 infections) and PER212. For CUM65, we observed a sequence identity of 99% (12 mismatches) and 86% (152 mismatches) for LRV1-Lb-CUM65.1 and LRV1-Lb-CUM65.2, respectively. For LC2321, both resolved strains showed a sequence identity of 99% with three and ten nucleotide mismatches in LRV1-Lb-LC2321.1 and LRV1-Lb-LC2321.2, respectively. These mismatches were unique to each strain, which might indicate the Sanger sequence is a chimera of both strains. Finally, the 99% identity of PER212 with its respective partial Sanger sequence showed only one nucleotide mismatch, not corresponding to the identified heterozygous SNP, suggesting a badly called base during the Sanger sequencing.

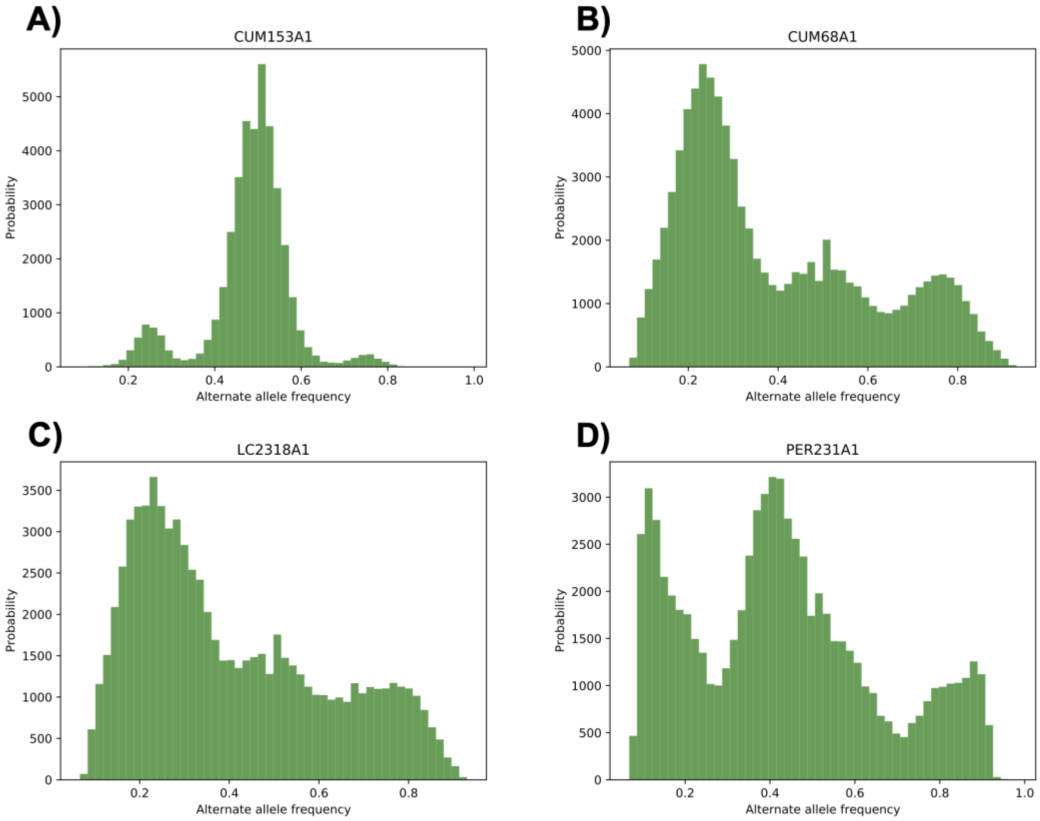
SUPPLEMENTARY FIGURES



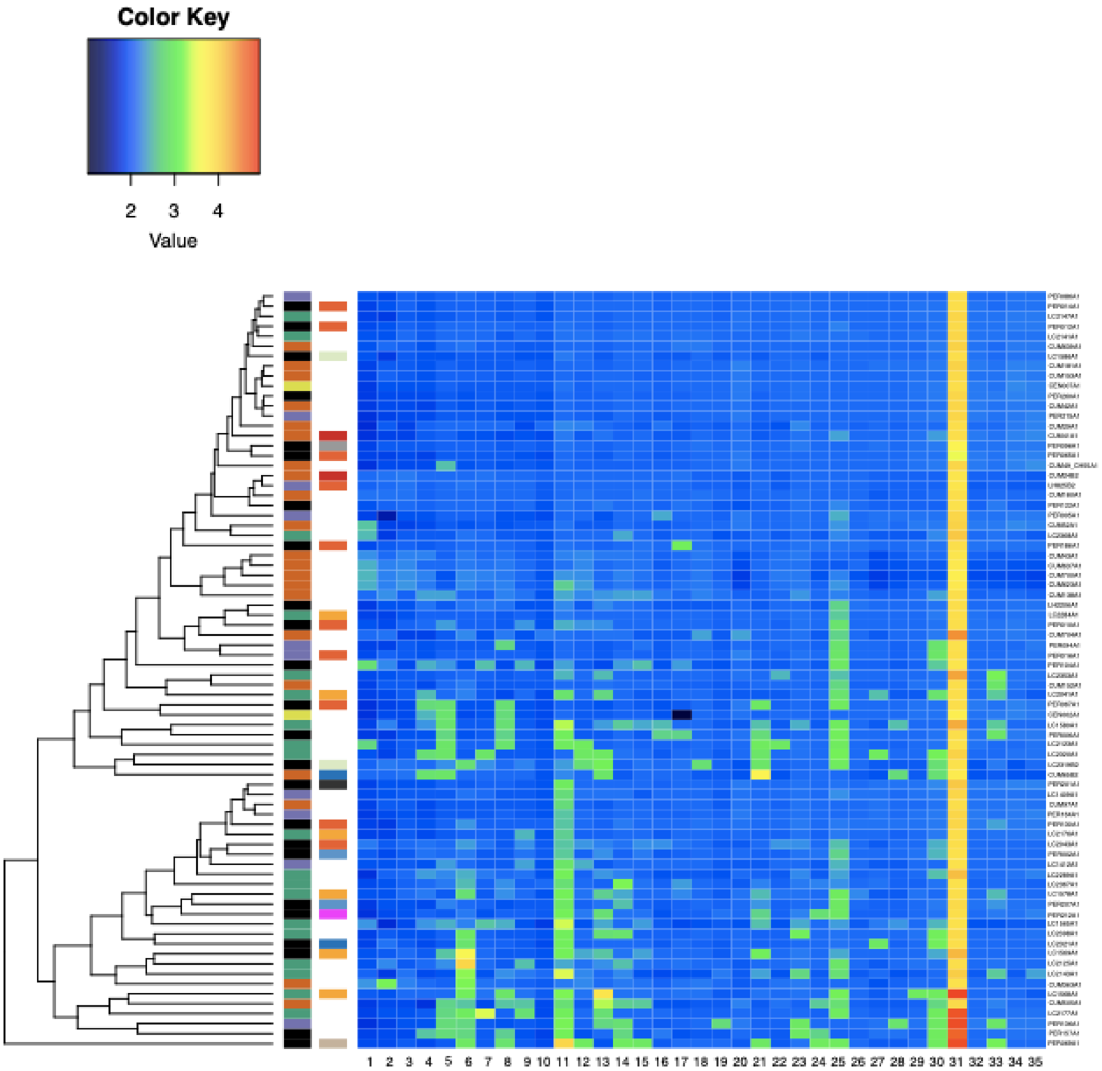
Supplementary Figure 1. Geographic origin of 79 *Lb* isolates from Peru and Bolivia that were included in this study. Rectangular box indicates the location of the Isiboro National Park that extends between the Departments of Cochabamba and Beni. Gray-scale represents altitude in meters.



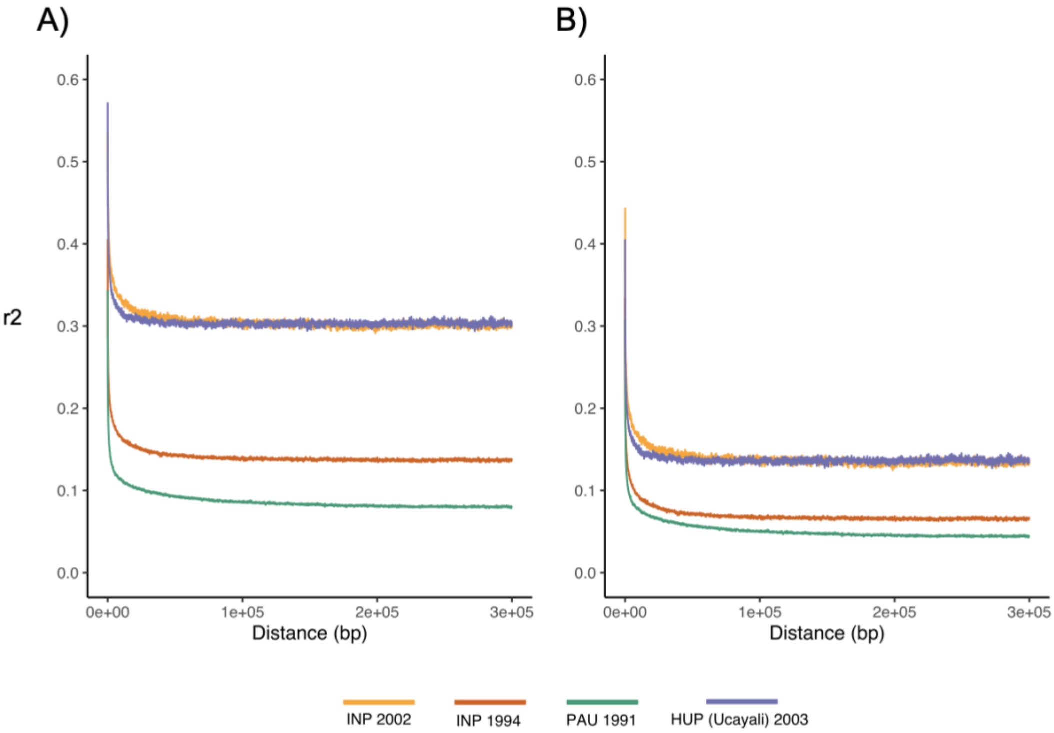
Supplementary Figure 2. A) Kernel density plots of the number of SNPs per 10kb window for each of the 79 *Lb* genomes. The median number of SNPs per 10kb window is indicated with gray vertical dashed lines and ranges between 28 and 32 SNPs for the majority of isolates. Three isolates showed slightly 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.



Supplementary Figure 3. Genome-wide distribution of alternate allele read depth frequencies at heterozygous sites. **(A)** Example of a largely diploid individual (CUM153) with allele frequencies centered around 0.5, which was observed for 77/79 *Lb* genomes included in this study. **(B-C)** Two isolates (CUM68 and LC2318) were symptomatic of tetraploidy, with modes of allele frequencies equal to 0.25, 0.5 and 0.75. **(C)** One isolate (PER231) showed a skewed distribution, suggesting that it may be the result of a mixed infection or contamination.



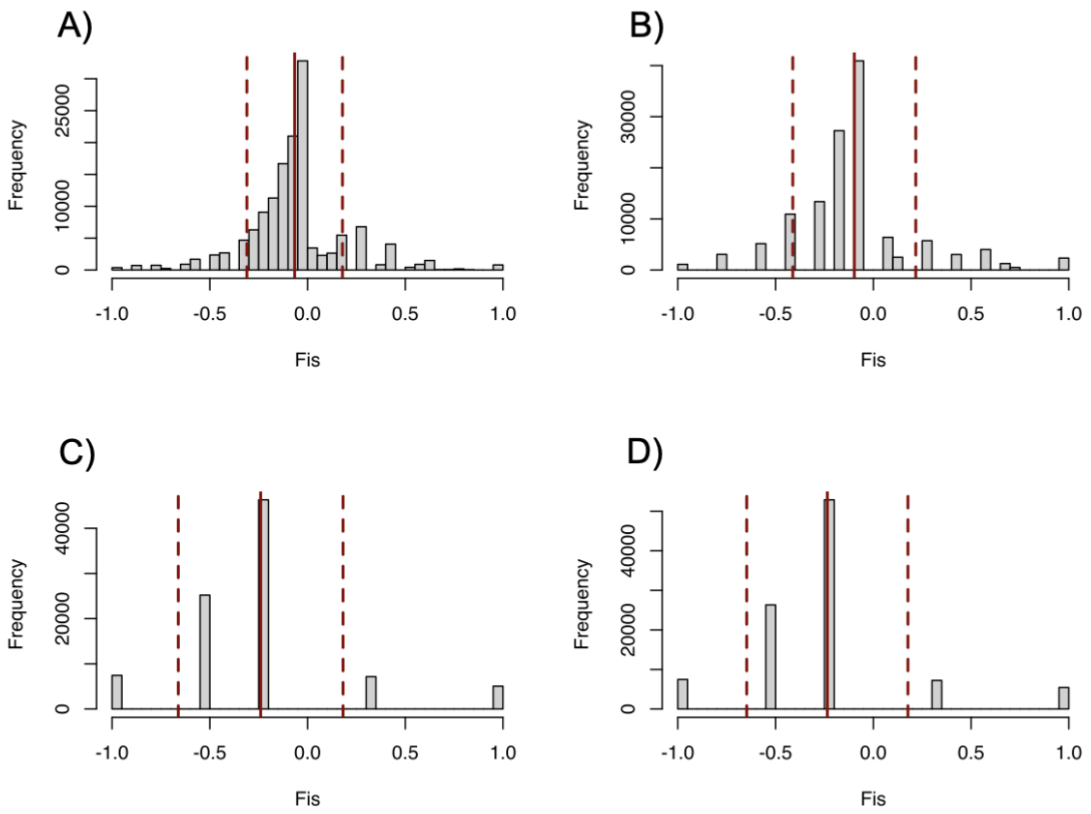
Supplementary Figure 4. Variation in chromosome copy numbers in a panel of 76 *Lb* genomes. Isolates are clustered according to similarity in somy estimation with aneuploid individuals at the bottom of the heatmap and overall diploid individuals at the top of the heatmap. Coloured boxes on the left of 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).



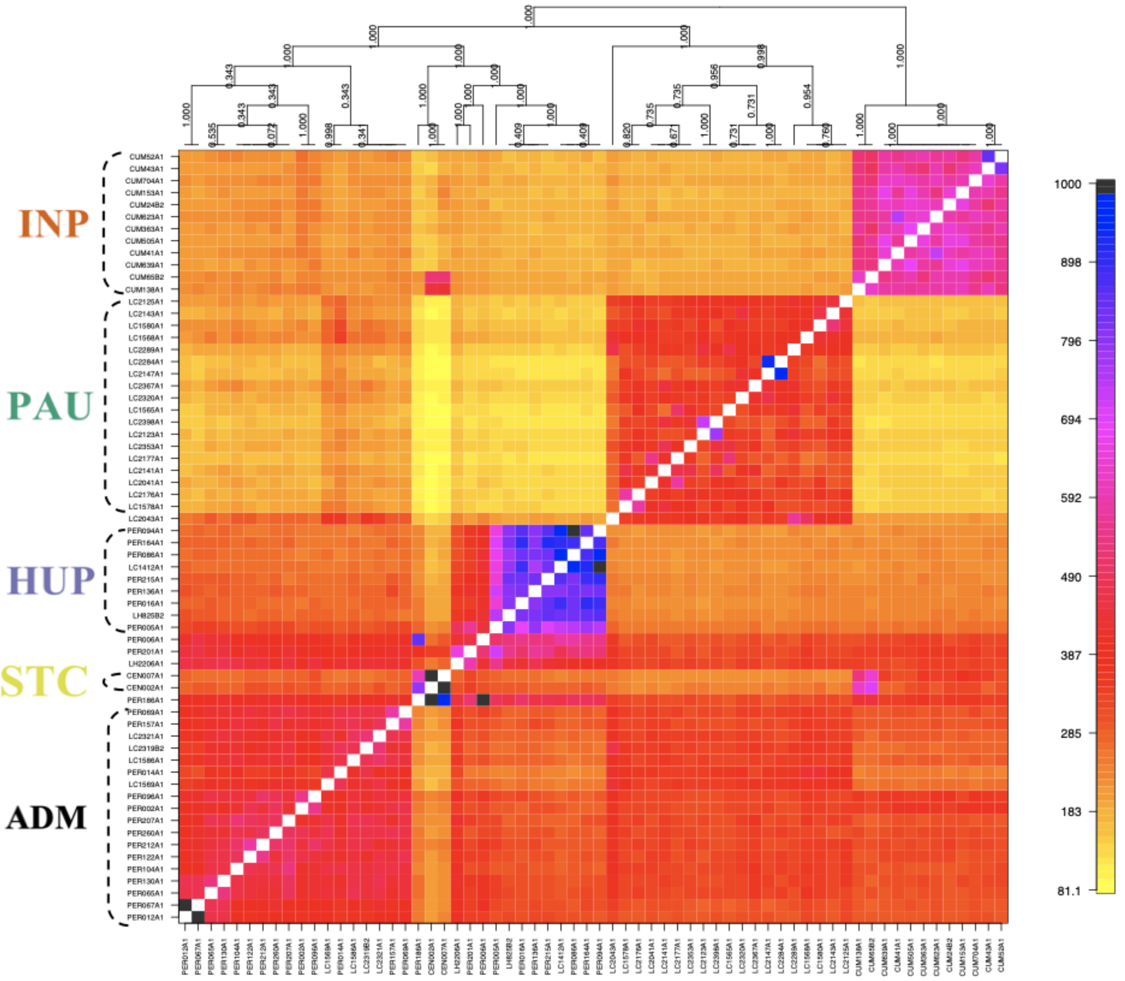
213

214 **Supplementary Figure 5.** Linkage Disequilibrium decay plots after correction for population structure and spatio-temporal Wahlund effects. **A)** Uncorrected for sample size. **B)** Corrected for sample size (r2 - 1/(2n)) [16]**.**

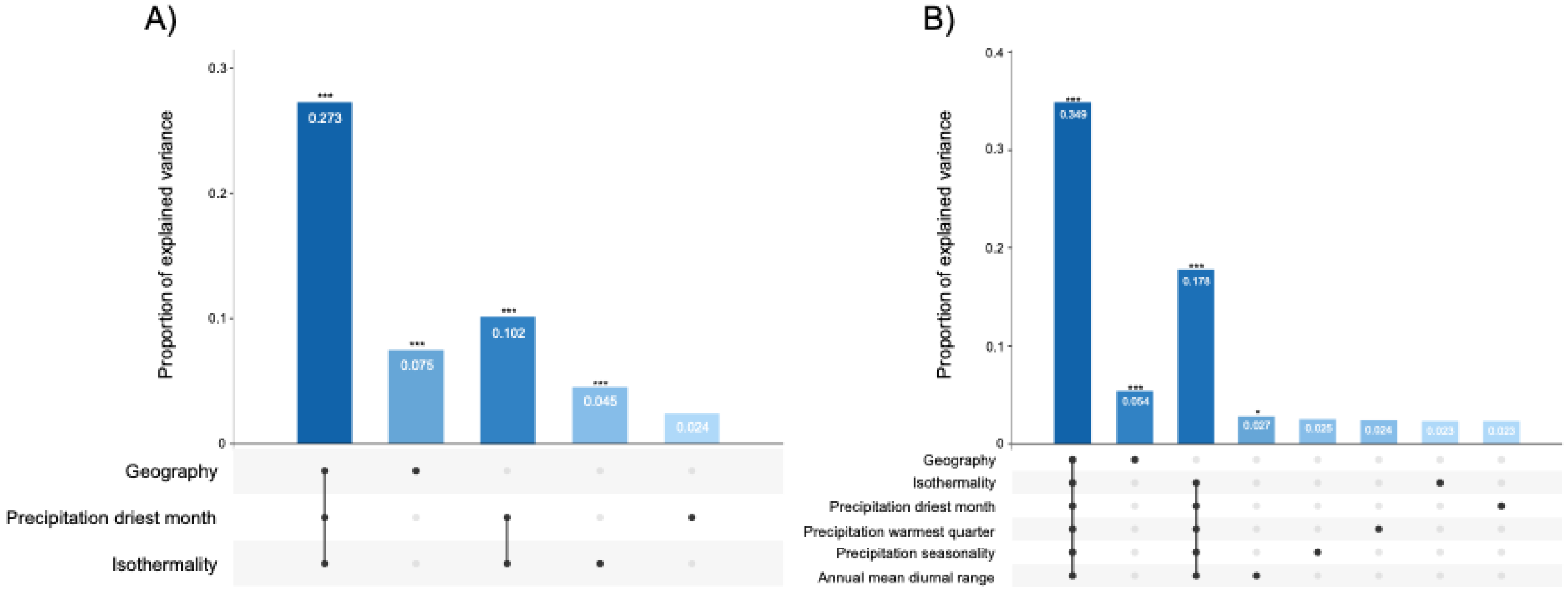
217



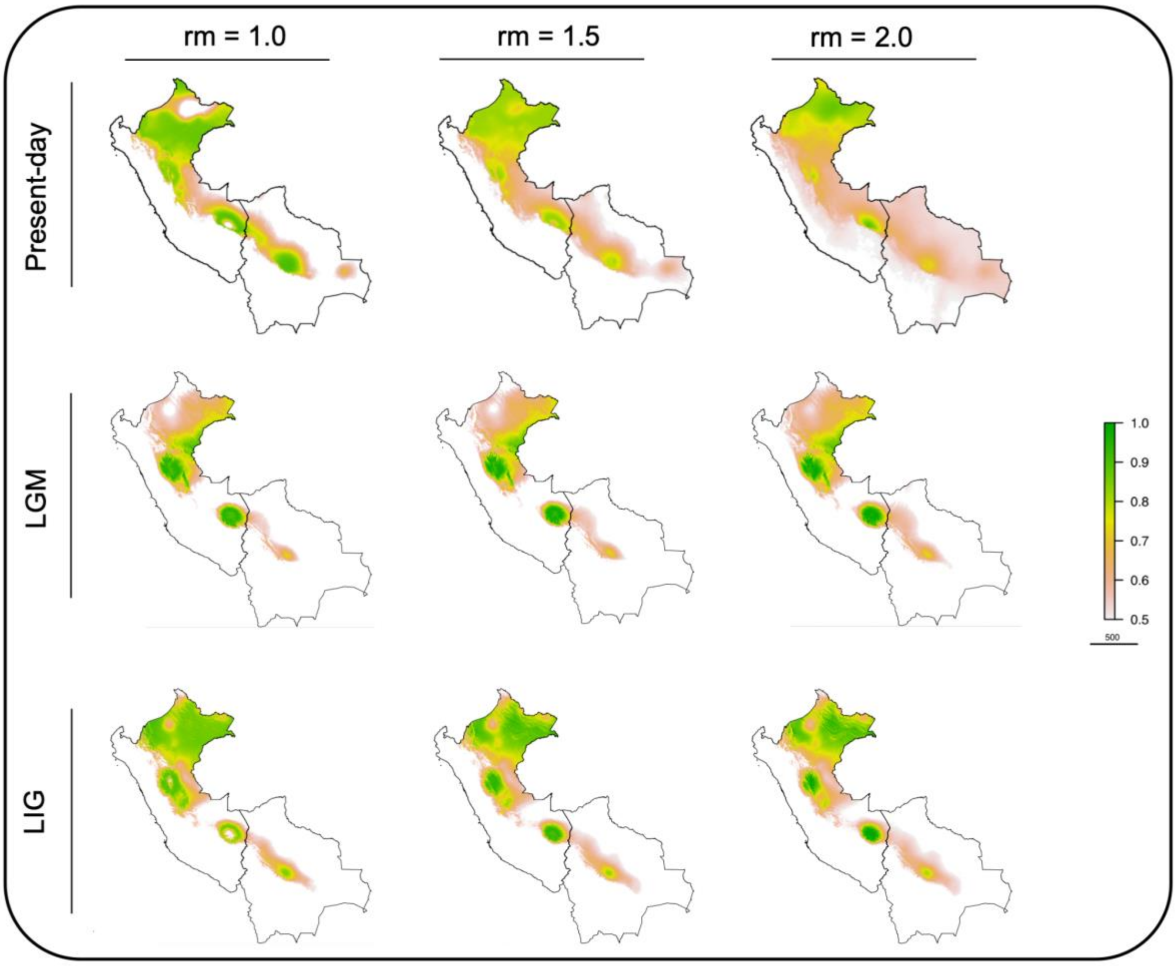
Supplementary Figure 6. Fis distributions after correction for population structure and spatio-temporal Wahlund effects. **A)** Individuals from PAU sampled in 1991 (N=14). **B)** Individuals from INP sampled in 1994 (N=7). **C)** Individuals from INP sampled in 2002 (N=3). **D)** Individuals from HUP (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.



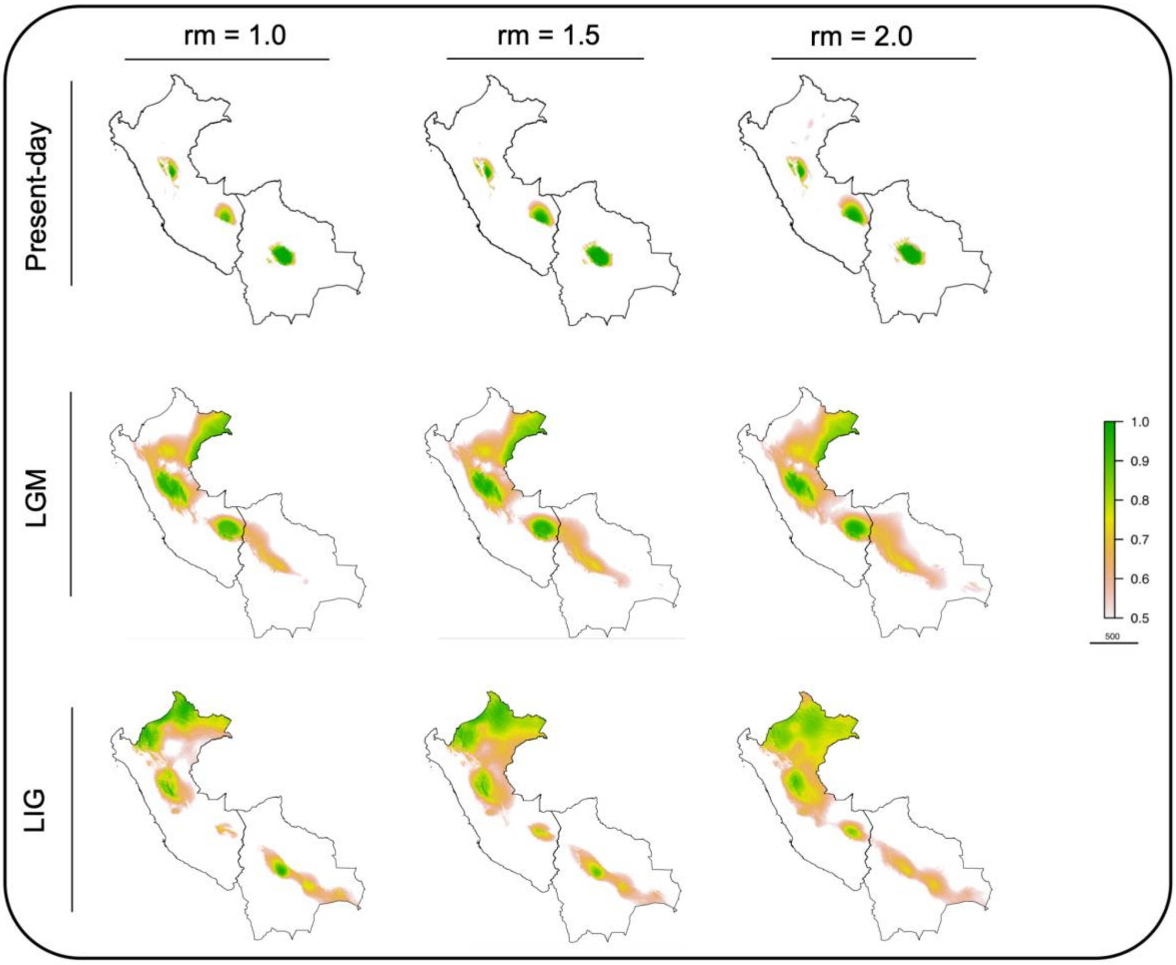
Supplementary Figure 7. Co-ancestry matrix, as inferred by fineSTRUCTURE, depicting the pairwise number of received (rows) and donated (columns) haplotype segments between two parasite genomes. Color key on the right shows the amount of shared haplotype segments and is capped on 1000 for visibility reasons. 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).



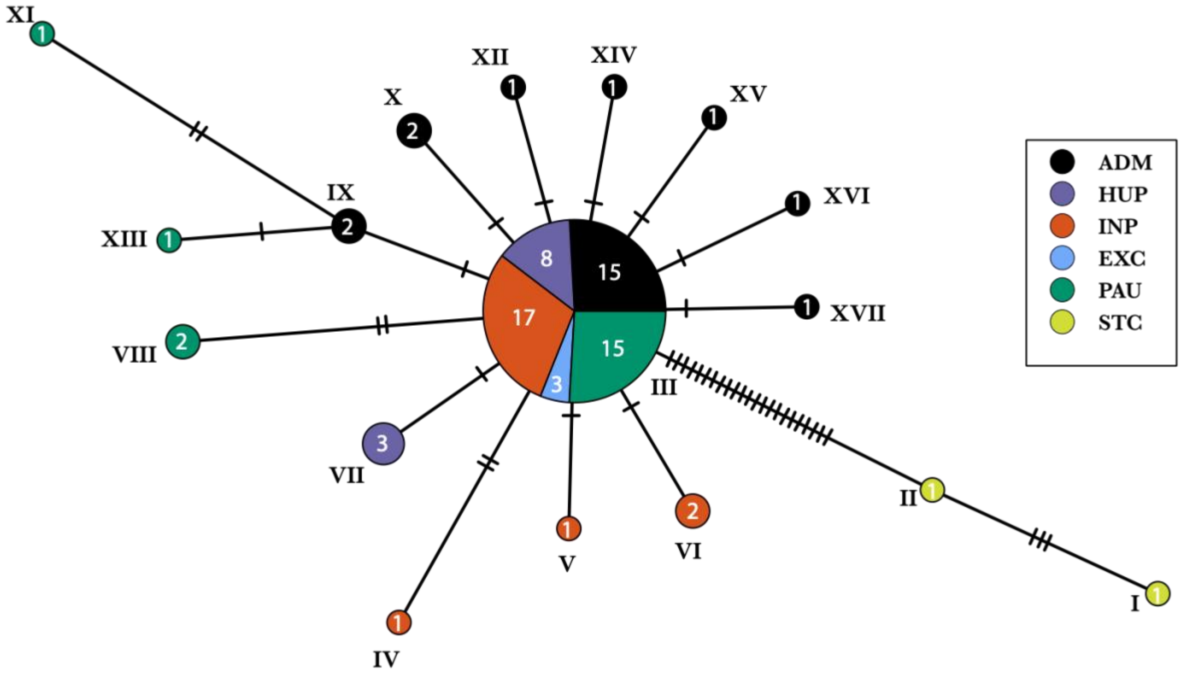
Supplementary Figure 8: Partial RDA models showing the influence of the environment (bioclimatic variables) and geography on the genomic variability between the three inferred populations in Peru and Bolivia. A) Partial RDA model including geography, isothermality (bio3) and precipitation of driest month (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).



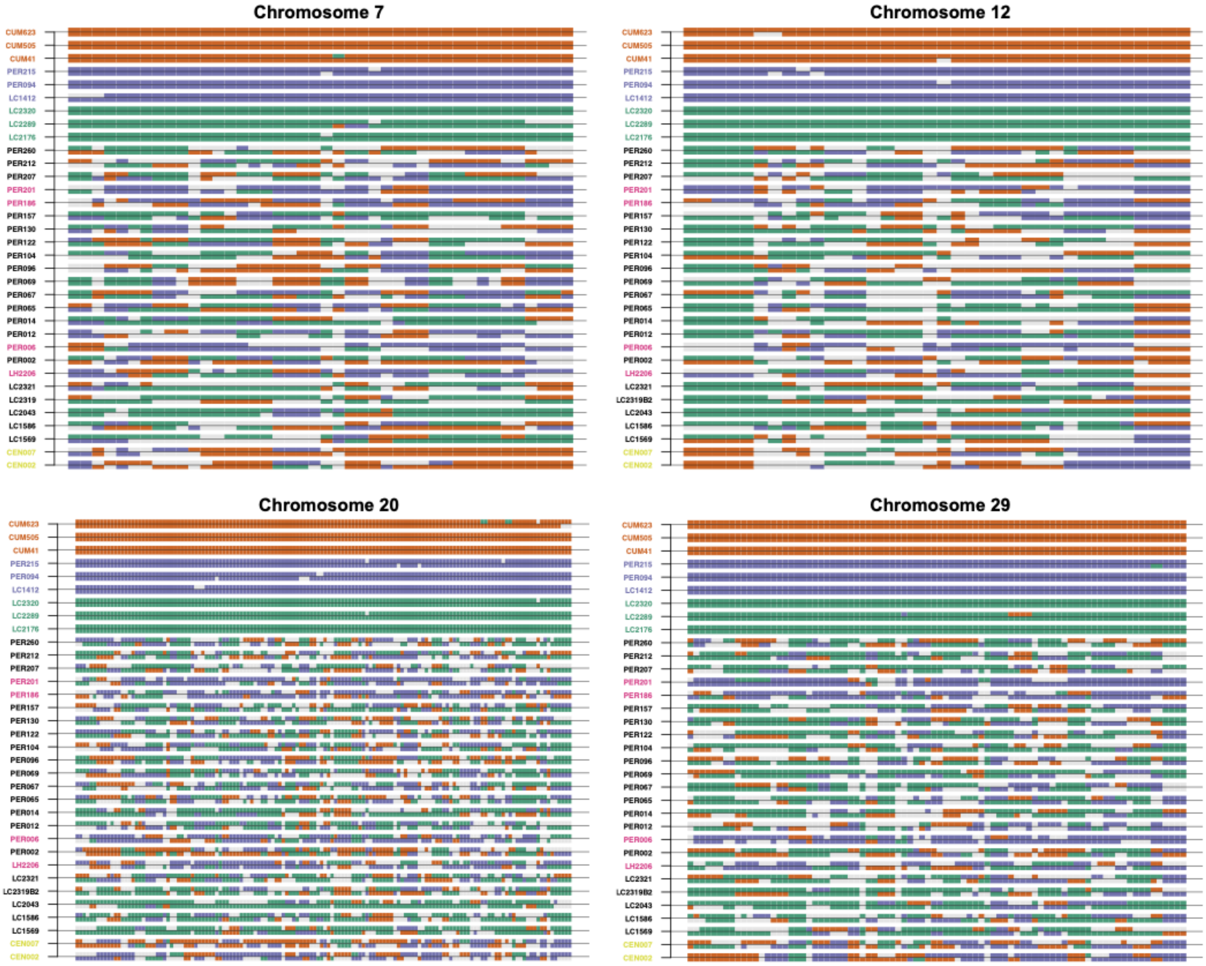
Supplementary Figure 9. Ecological niche models based on Present-day, LGM and LIG data of isothermality (bio3) and Precipitation of the driest month (bio14) (mod-A variable selection approach) with different regularization values (rm = 1, 1.5 and 2). The continuous-scale legend represents habitat suitability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



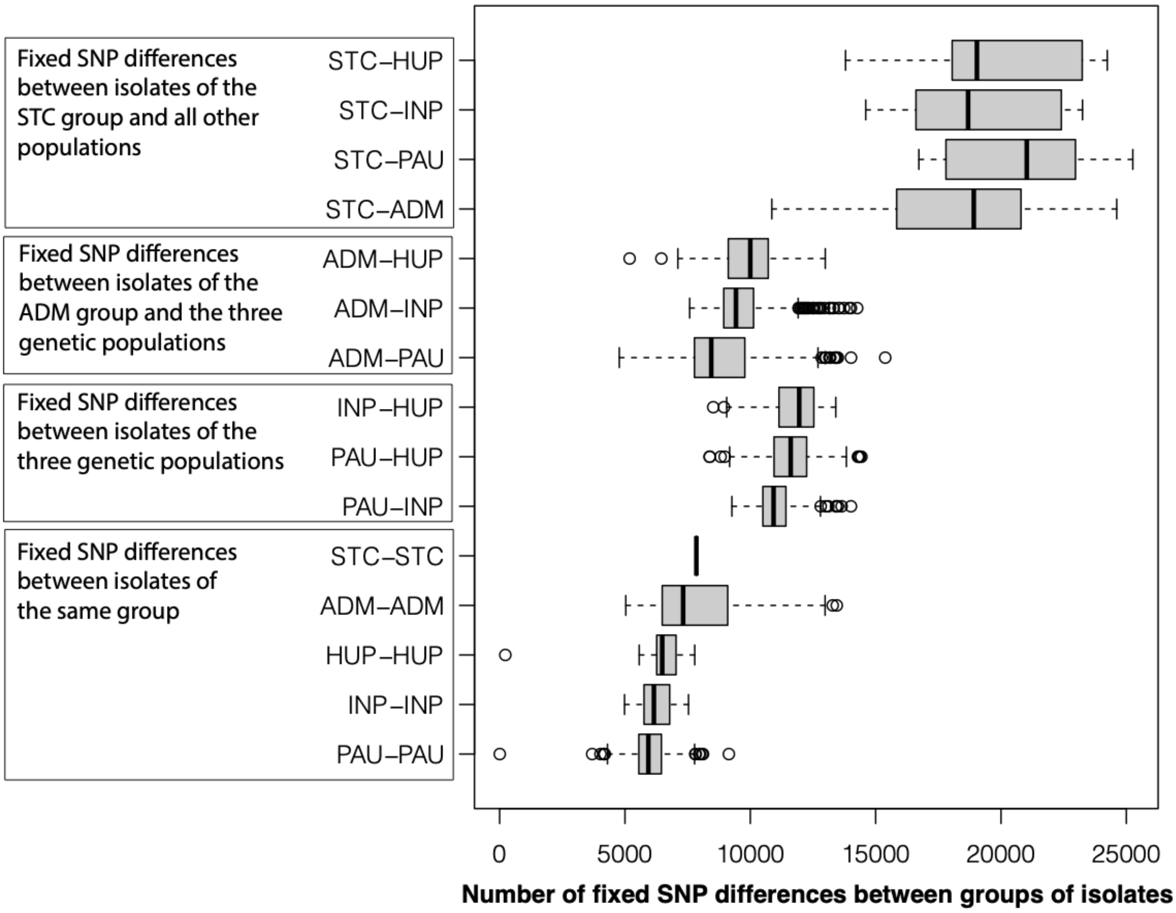
Supplementary Figure 10. Ecological niche models based on Present-day, LGM and LIG data of isothermality (bio3), Precipitation of the driest month (bio14), precipitation of the warmest quarter (bio18), precipitation seasonality (bio15) and annual mean diurnal range (bio2) (mod-M variable selection approach)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



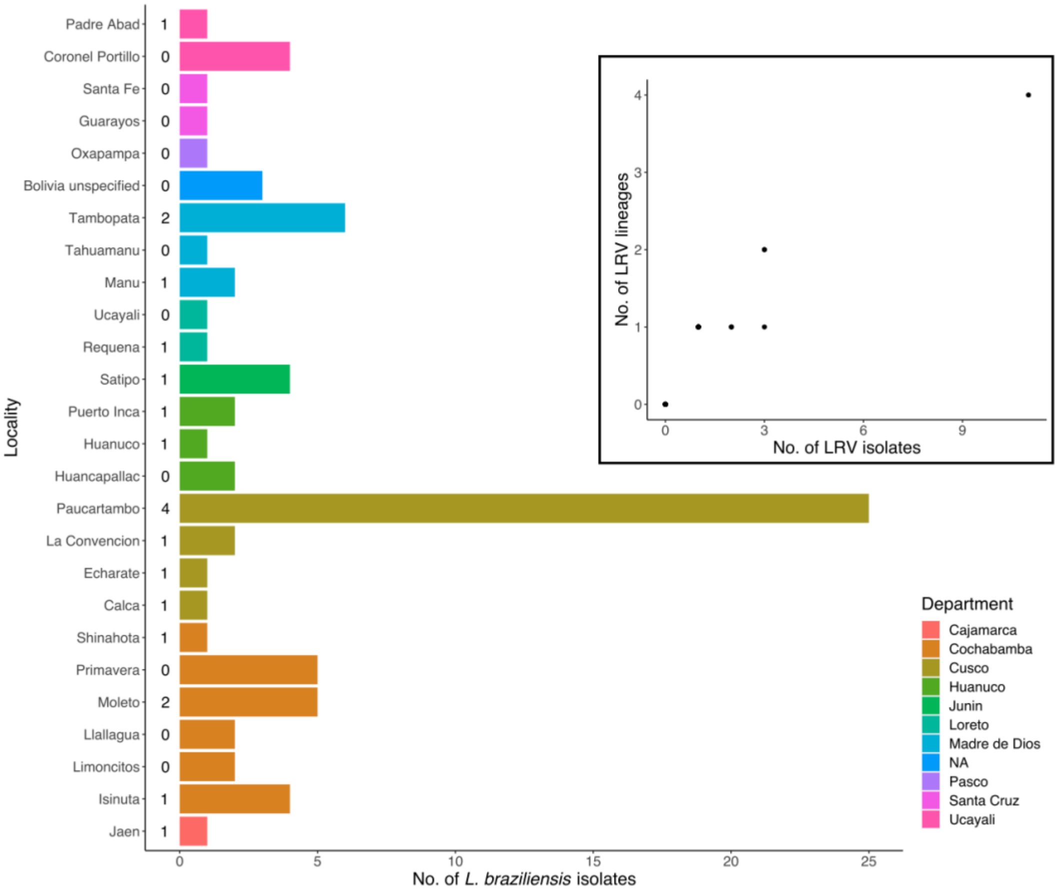
Supplementary Figure 11. TCS Haplotype network with PopART based on 53 high-quality SNPs identified within the coding region of the haploid mitochondrial maxicircle. A total of 17 haplotypes (shown with Roman numbers) were identified within our set of 80 *Lb* isolates from Peru and Bolivia. The size 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.



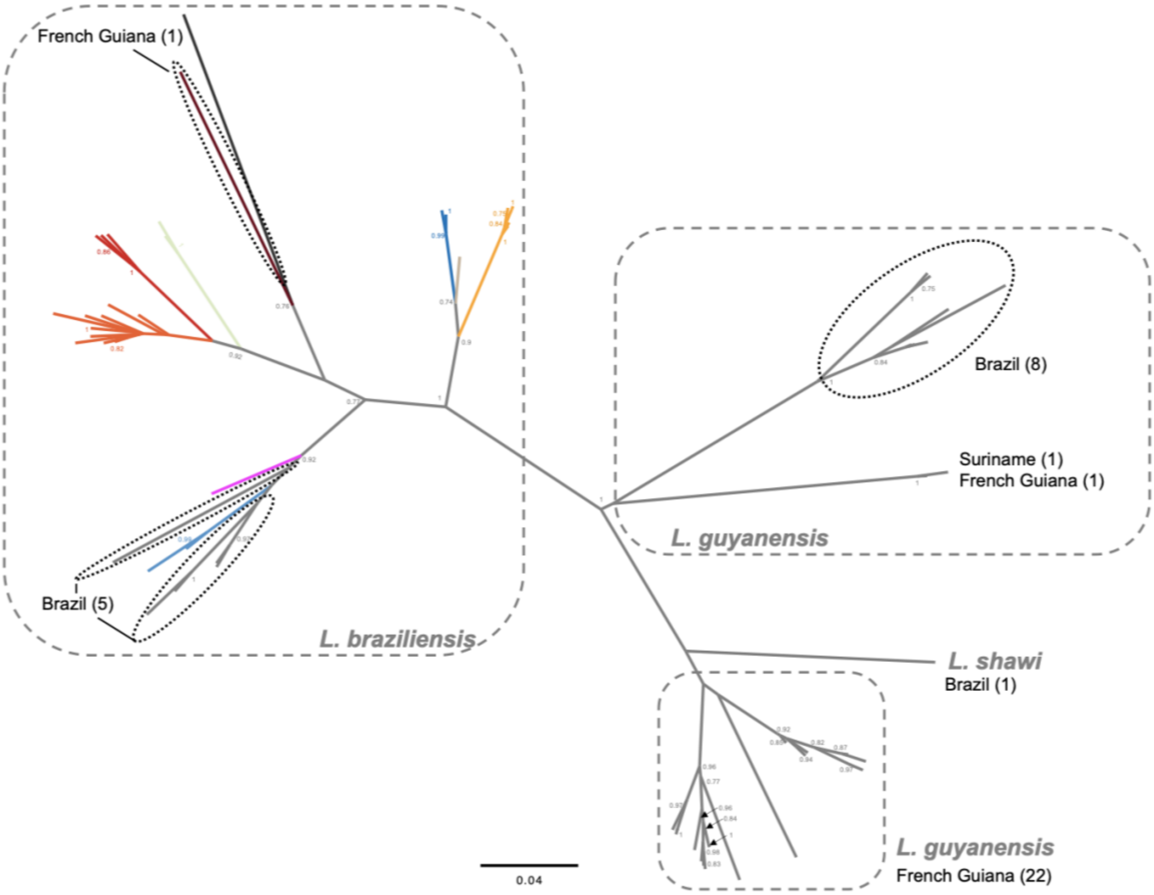
Supplementary Figure 12. PCAdmix local ancestry assignment to PAU, INP and HUP source populations of the 19 ADM isolates, 4 UNC isolates, 2 STC isolates and three randomly selected isolates from each source population. Ancestry was assigned in windows of 30 SNPs along each chromosome (here only chromosomes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



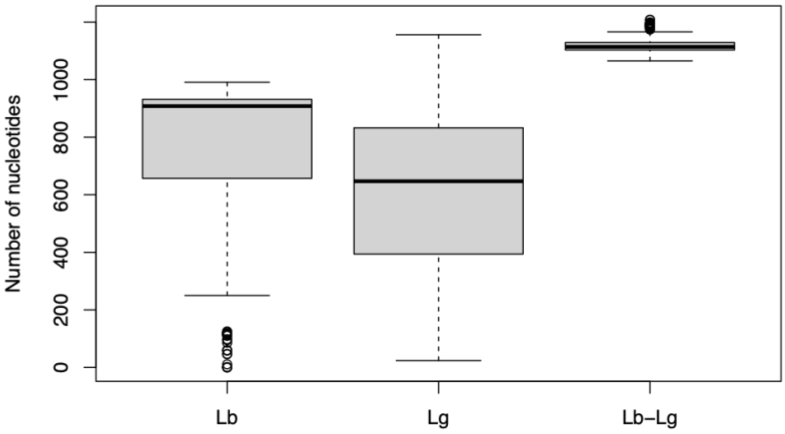
Supplementary Figure 13. Number of fixed SNP differences between *Lb* isolates of the same group or between *Lb* isolates of different groups.



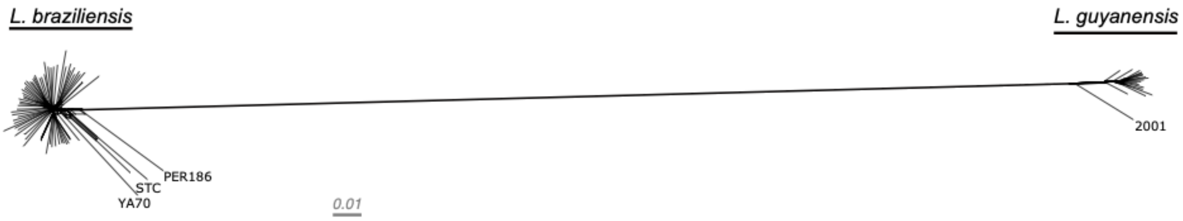
Supplementary Figure 14. Barplot shows the number of *Lb* isolates per sampling locality. Number on the right of each bar shows the number of *Lb* isolates that were positive for LRV1. Bars are coloured according to the Department. Inset reveals the number of LRV1 lineages versus the number of LRV1 isolates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



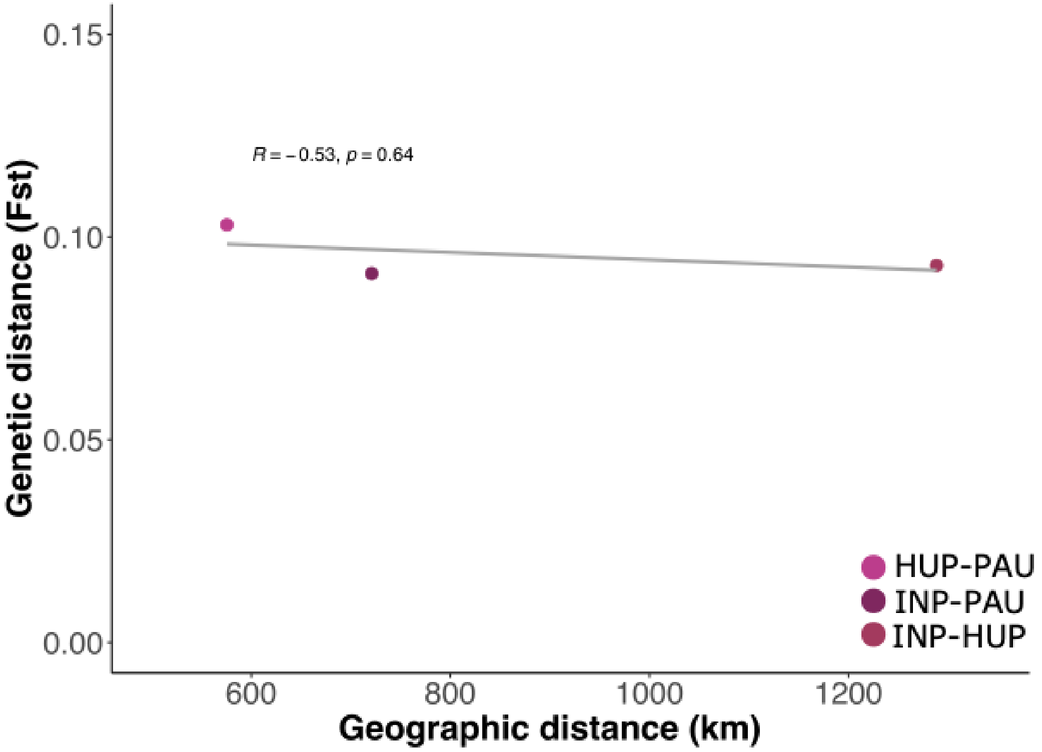
Supplementary Figure 15. Maximum likelihood tree based on partial LRV1 sequences (756bp) from *L. braziliensis*, *L. guyanensis* and *L. shawi* originating from Peru, Bolivia, Brazil, French Guiana and Suriname [17–20]. Colored lineages in *L. braziliensis* correspond to the LRV1 lineages described in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



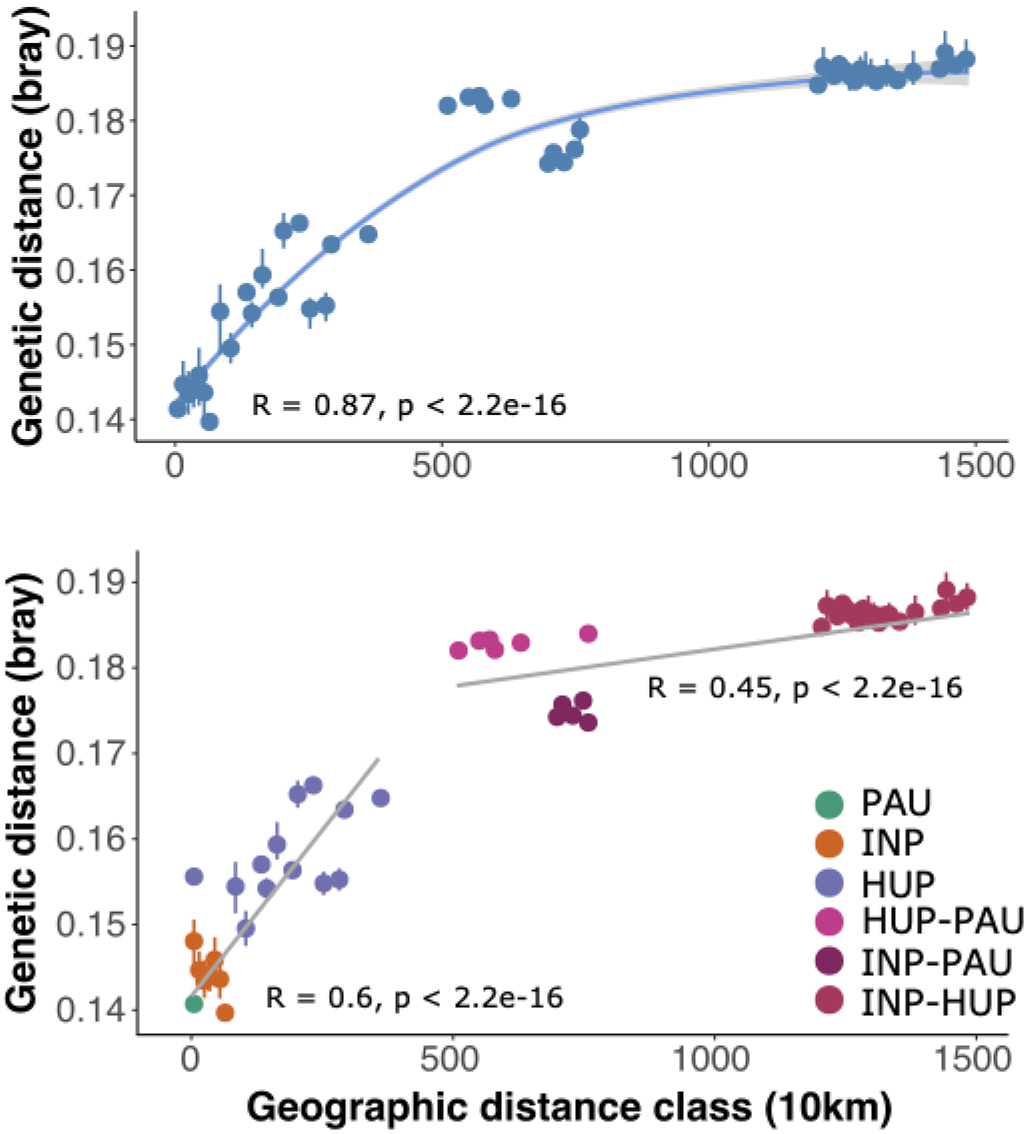
Supplementary Figure 16. Nucleotide differences - on average higher between *Lb* and *Lg* viral genomes then within *Lb* and *Lg*.



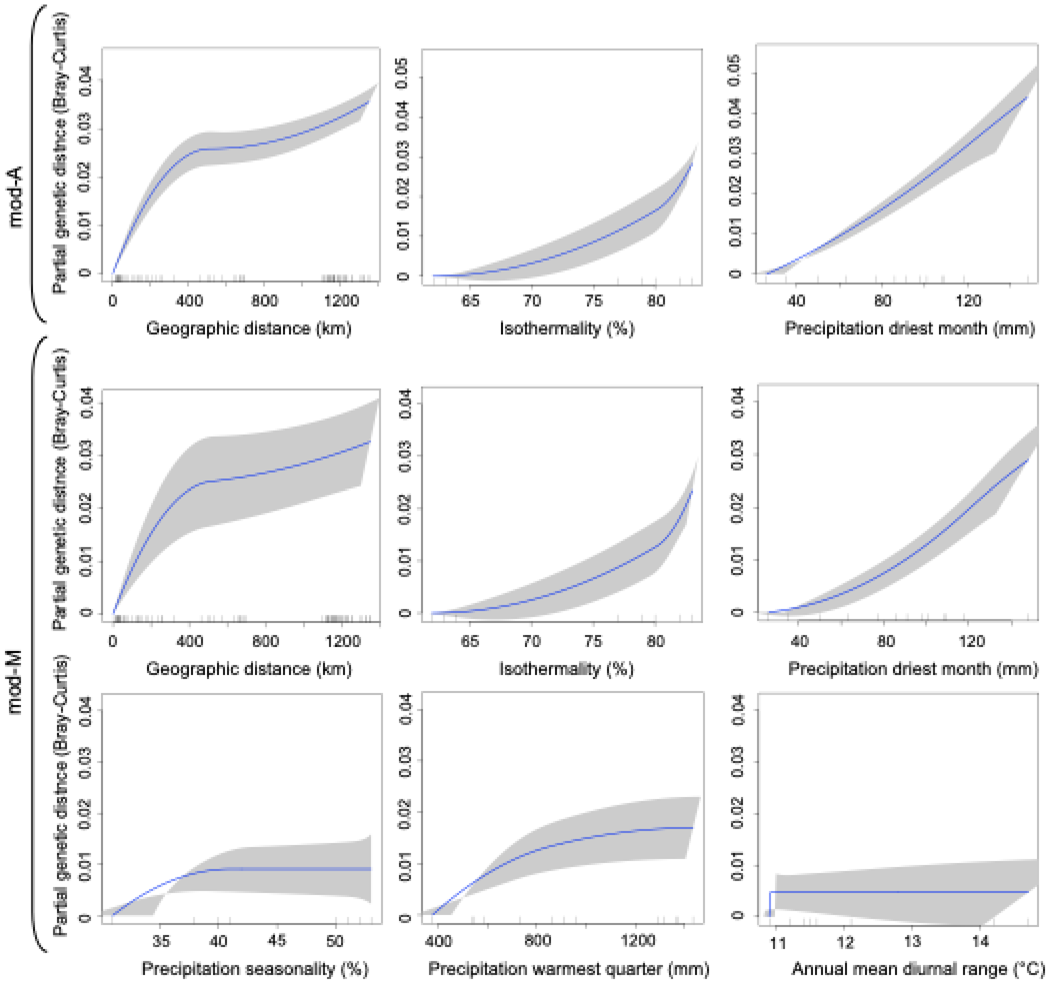
Supplementary Figure 17. A phylogenetic network, inferred with SPLITSTREE, based on uncorrected p-distances between 77 *Lb* and 19 *Lg* isolates typed at 7,571 bi-allelic SNPs.



Supplementary Figure 18. Linear regression of the interdeme great-circle distance (km) and the Weir & Cockerham's Fst 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.