communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-08331-1

Genetic diversity and comparative genomics across *Leishmania* (*Viannia*) species

Laura Natalia Gonzalez-Garcia[1], Maria Paula Rodriguez[1], Marcela Parra-Muñoz[2], Ana M. Clavijo[2], Laura Levy[1], Clemencia Ovalle-Bracho[3], Claudia Colorado[3], Carolina Camargo[3], Eyson Quiceno[4], Maria Juliana Moncada[4], Carlos Muskus[4], Daniel Alfonso Urrea[5], Felipe Baez-Aguirre[6], Silvia Restrepo[7,8], María Clara Echeverry[2] ✉ & Jorge Duitama[1] ✉

Leishmaniasis is an important public health problem worldwide, with a broad spectrum of clinical and epidemiological features partly associated with the diversity and complex life cycle of the *Leishmania* parasites. This study analyzes genomic data from 205 *Leishmania* (*Viannia*) samples, including 65 newly sequenced clinical isolates. It also provides chromosome-level genome assemblies for 10 isolates representing different species and populations. The observed distribution of *Leishmania* genomic diversity across the sampling locations suggests rapid adaptation to different ecosystems. The phylogenomic analysis provides new hypotheses challenging the current delimitation of species. Pangenomic analysis of high-quality assemblies shows consistent copy number variation between species for different gene families. Larger and more diverse amastin gene families were observed in the assembled genomes compared to previous reports based on the analysis of short-read data. This work provides genomic resources and helpful information regarding central problems in the biology of *Leishmania spp*, including species diversification, transmission dynamics, and the evolution of virulence mechanisms.

In Tropical and Subtropical regions, dipterous sandfly insects of the *Psychodidae* family can transmit protozoa parasites belonging to the *Leishmania* genus, producing a disease known as leishmaniasis. In humans, leishmaniasis has three major clinical manifestations: cutaneous (CL), mucocutaneous (MCL), and the widespread fatal visceral leishmaniasis (VL), which is the most devastating form of the disease. Approximately 700,000 to 1 million new leishmaniasis cases occur annually, and about 350 million people are at risk of infection, affecting populations from 92 countries[1]. The *Leishmania* genus includes approximately 20 species associated with human infections.

The fight against leishmaniasis presents challenges such as the absence of effective vaccines[2] and the toxicity and variable efficacy of the drugs used to treat it[3,4]. Additionally, the wide variation in clinical manifestations is attributed to the unpredictable host immune response and the genetic variability of the parasite[5,6]. All forms of the disease are present in the South American region, CL being the most prevalent and associated mainly with the *L.* (*V.*) subgenus. Intensification of the social internal conflict during the first decade of the present century led to the massive internal mobilization of civil and armed populations through sylvatic and rural areas in Colombia, increasing the CL incidence from 5000 cases per year in the nineties up to 20,000 new cases per year in 2005[7]. The CL cases in Colombia are produced mostly by parasites of the *Viannia* subgenus (hereafter referred to as *L.* (*V.*)), specifically from the species *L.* (*V.*) *panamensis, L.* (*V.*) *braziliensis*, and *L.* (*V.*) *guyanensis*. Less than 1% of the CL cases are associated with species of the subgenus *Leishmania* (hereafter *L.* (*L.*)), represented by the species *L.* (*L.*) *mexicana, L.* (*L.*) *amazonensis,* and *L.* (*L.*) *donovani*)[8–10].

Considering the importance of knowing the genetic variability of *Leishmania* parasites, in 1995, the World Health Organization (WHO)

[1]Systems and Computing Engineering Department, Universidad de los Andes, Bogotá, Colombia. [2]Departamento de Salud Pública, Facultad de Medicina, Universidad Nacional de Colombia – Sede Bogotá, Bogotá, Colombia. [3]Hospital Universitario Centro Dermatológico Federico Lleras Acosta E.S.E, Bogotá, Colombia. [4]Programa de Estudio y Control de Enfermedades Tropicales-PECET, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia. [5]Laboratorio de Investigaciones en Parasitología Tropical (LIPT), Universidad del Tolima, Ibagué, Colombia. [6]Applied Genomics Research Group, Vicerrectoría de Investigación y Creación, Universidad de los Andes, Bogotá, Colombia. [7]Universidad de los Andes, Bogotá, Colombia. [8]Boyce Thompson Institute, Ithaca, NY, USA. ✉e-mail: mcecheverryg@unal.edu.co; ja.duitama@uniandes.edu.co

decided to promote genome sequencing of medically important species[11]. The information at the genomic scale of the *L.* (*Leishmania*) subgenus has provided useful information to discuss *Leishmania* taxonomy and evolution and to study population dynamics[12]. In the case of the *L.* (*Viannia*) subgenus, it presents a population structure associated with geographic isolation[13,14], altitudinal segregation[15], and divergence between the transmission scenarios[16]. Likewise, genetic diversity varies among species[14,15,17], which is hypothesized to be related to host adaptation[14]. In addition, several cases of hybridization and recombination have been reported[15,18–20]. Few genomic studies included all *L.* (*Viannia*) group members[21]. Less than 50 Colombian isolates have been sequenced[14,22,23], none using long-read technologies. Furthermore, some of these studies found conflicting interspecific relationships compared to individual loci identification[9,24]. Specifically, controversy is related to whether some taxa are invalid species[25,26] and are subgroups of others, for example, (*L.* (*V.*) *peruviana* being a subspecies of *L.* (*V.*) *braziliensis*, as well as *L.* (*V.*) *panamensis*—and presumably *L.* (*V.*) *shawi*—as subspecies of *L.* (*V.*) *guyanensis*[24,27]. In addition, when multilocus assessments have resulted in concordance with individual loci[16,18], some species were underrepresented or not considered in the analyses. Therefore, in the absence of a genome-wide analysis encompassing most species of the group coming from diverse geographic origins, it is unclear whether some markers lack resolution at the species level[28,29] or if taxonomy needs revision.

This study presents a large-scale genomic assessment of the *L.* (*V.*) subgenus, providing novel insights into taxonomy, genome evolution, and the distribution of inter and intra-species genetic variability of the subgenus. We generated databases of over one million SNPs and hundreds of gene presence/absence variants—a rich resource to design population markers for epidemiological surveillance—including new tools for rapid species identification. We analyzed whole genome DNA sequencing data from 242 *Leishmania* samples, including 65 Colombian clinical isolates sequenced in this study. Furthermore, we used state-of-the-art long-read sequencing technologies to build and compare chromosome-scale de novo genome assemblies of ten isolates belonging to three species.

Results

Genomic diversity of *Leishmania* (*Viannia*) clinical isolates

Sixty-five *Leishmania* Colombian isolates were sequenced using Illumina whole-genome shotgun sequencing to assess some of the unexplored genetic diversity of *Leishmania* in Colombia. According to the Lab-based molecular characterization, the selected isolates included 34 *L.* (*V.*) *braziliensis* (1 from the Caribbean region, 9 from the Amazonian region, 9 from the Andean region, 15 from the Orinoquia), 12 *L.* (*V.*) *guyanensis* (3 from the Amazonian region, 3 from the Pacific region, 6 from the Andean region), 18 *L.* (*V.*) *panamensis* (2 from the Pacific region, 3 from the Caribbean region, 13 from the Andean region), and 1 *L.* (*Leishmania*) from the Andean region (See Supplementary Fig. S1 and Supplementary Data 1 for details). To compare the Colombian isolates sequenced in the present study with sequenced isolates from Latin America having publicly available data, reads sequenced from 139 isolates belonging to the *Leishmania* (*Viannia*) genera were also included in the analysis, as well as 38 *Leishmania* (*Leishmania*) isolates (See "Methods" and Supplementary Data 1 for details). Although the complete dataset included samples from different species, reads from each sample were mapped against the reference genome of the *L.* (*V.*) *braziliensis* strain M2904 to have a common reference for variant genotyping and comparison across the complete dataset. As expected, the percentage of mapped reads was larger than 60% for each *Leishmania* (*Viannia*) sample and lower than 6% for *L.* (*L.*) samples (Supplementary Fig. S2 and Supplementary Data 1). Variant discovery integrating all aligned reads resulted in 3,844,799 SNPs identified across the *Leishmania* species sampled in this study. A total of 419 aneuploidies were identified from read-depth data for samples with a good percentage of mapped reads and good read coverage, following the procedure suggested by Dumetz et al.[30] (Supplementary Data 1). All samples had predicted aneuploidies (mostly tetrasomies) on chromosome 31. The remaining 224 predicted events were mostly trisomies (197) scattered across chromosomes and samples.

After filtering genotype calls by minimum quality of 40, a total of 2,114,255 SNPs with Minimum Allele Frequency (MAF) > 0.01 were included in the global variation database (See statistics per isolate in the Supplementary Data 1). Approximately 48% of those SNPs were located in coding regions, and from these, 58% corresponded to synonymous variants, 41% to missense variants, and less than 0.2% affected start and stop codons. Because this database included species from the *Leishmania* and the *Viannia* subgenus, the genotyping missing data was close to 20% due to the low percentages of mapped reads obtained for *L.* (*L.*). Reads from *L.* (*L.*) samples that could be mapped are not distributed across the genome, but they are concentrated in conserved segments between species. After further filtering, keeping SNPs with at least 97.5% of genotyped individuals, removing missense and nonsense variants, and removing SNPs in chromosome 31, a maximum-likelihood (ML) phylogeny was reconstructed based on 50,543 SNPs (Supplementary Fig. S3). The separation between the *Leishmania* and the *Viannia* subgenus was highly supported and consistent with previous analyses[8,21]; within the *L.* (*L.*) subgenus, the separation among *L.* (*L.*) *amazonensis*, *L.* (*L.*) *mexicana, L.* (*L.*) *tropica*, *L.* (*L.*) *aethiopica*, *L.* (*L.*) *major*, *L.* (*L.*) *donovani*, and *L.* (*L.*) *infantum* was also supported. Within the *L.* (*V.*) subgenus, the *L.* (*V.*) *braziliensis*/*L.* (*V.*) *peruviana* complex and the *L.* (*V.*) *guyanensis*/*L.* (*V.*) *panamensis* complex formed two different clades with high support. However, the relationships within the two groups were difficult to visualize in this tree. From the newly sequenced data, seven samples presented inconsistencies between the clustering and the Lab-based species classification (Supplementary Data 1). The isolate LL0087, classified within the *Leishmania* subgenus without a species name, and the isolate LL0490, classified as *L.* (*V.*) *panamensis*, clustered within *L.* (*L.*) *amazonensis* (Supplementary Fig. S3). The samples W8252 and LL0725, classified as *L.* (*V.*) *braziliensis* and *L. guyanensis*/*panamensis*, respectively, turned out to have admixed DNA (details below). The isolates LL0732 and LL0775, classified as *L.* (*V.*) *guyanensis*, clustered within *L.* (*V.*) *panamensis*. The opposite situation was observed for the isolate W8104.

To investigate the species divergence within the *L.* (*Viannia*) subgenus, Maximum likelihood (ML) and Neighbor-joining (NJ) trees were constructed using a subset of the genotypic data, including exclusively *L.* (*Viannia*) isolates (Fig. 1, Supplementary Data 2, and Supplementary Figs. S4 and S5). Because this subset only included samples with good mapping percentages and genome coverage, over 13× more SNPs (673,944) could be used in this case after applying the same filters used for the complete dataset. The species *L.* (*V.*) *naiffi* and *L.* (*V.*) *lainsoni* were clearly separated in both trees. As previously observed, *L.* (*V.*) *panamensis* and *L.* (*V.*) *guyanensis* were separated from *L.* (*V.*) *braziliensis* and *L.* (*V.*) *peruviana*. Filtering of SNPs by allele frequency differences between these complexes revealed 108,131 SNPs that perfectly differentiated the two species complexes. Although these species complexes were recovered in two strongly supported groups, species within each group were not reciprocally monophyletic. Within the *L.* (*V.*) *panamensis*/*L.* (*V.*) *guyanensis* complex, the samples classified as *L.* (*V.*) *guyanensis* by Lab-based classical typing did not form a monophyletic group: nine samples were more closely related to the *L.* (*V.*) *shawi* sample, whereas the remaining four samples (UN0043, UN0049, UN0063, and W8134) formed a clade sister to most of *L.* (*V.*) *panamensis* isolates. Likewise, the reference *L.* (*V.*) *panamensis* isolate L13 did not cluster within the two major groups of this species but was more similar to the isolate UN0005, which was classified as *L.* (*V.*) *guyanensis* by classical markers. Regarding the *L.* (*V.*) *braziliensis*/*L.* (*V.*) *peruviana* complex, a group that included the *L.* (*V.*) *braziliensis* isolates from Brazil (except one) was sister to a clade with all other *L.* (*V.*) *braziliensis* isolates and *L.* (*V.*) *peruviana*. Thus, *L.* (*V.*) *braziliensis* was recovered as paraphyletic with *L.* (*V.*) *peruviana* nested within *L.* (*V.*) *braziliensis*.

To further analyze the structure within *L.* (*Viannia*) populations, the two main species complexes were analyzed separately, following population genomic approaches. Samples belonging to the two complexes were separately selected from the global genomic variation dataset, and

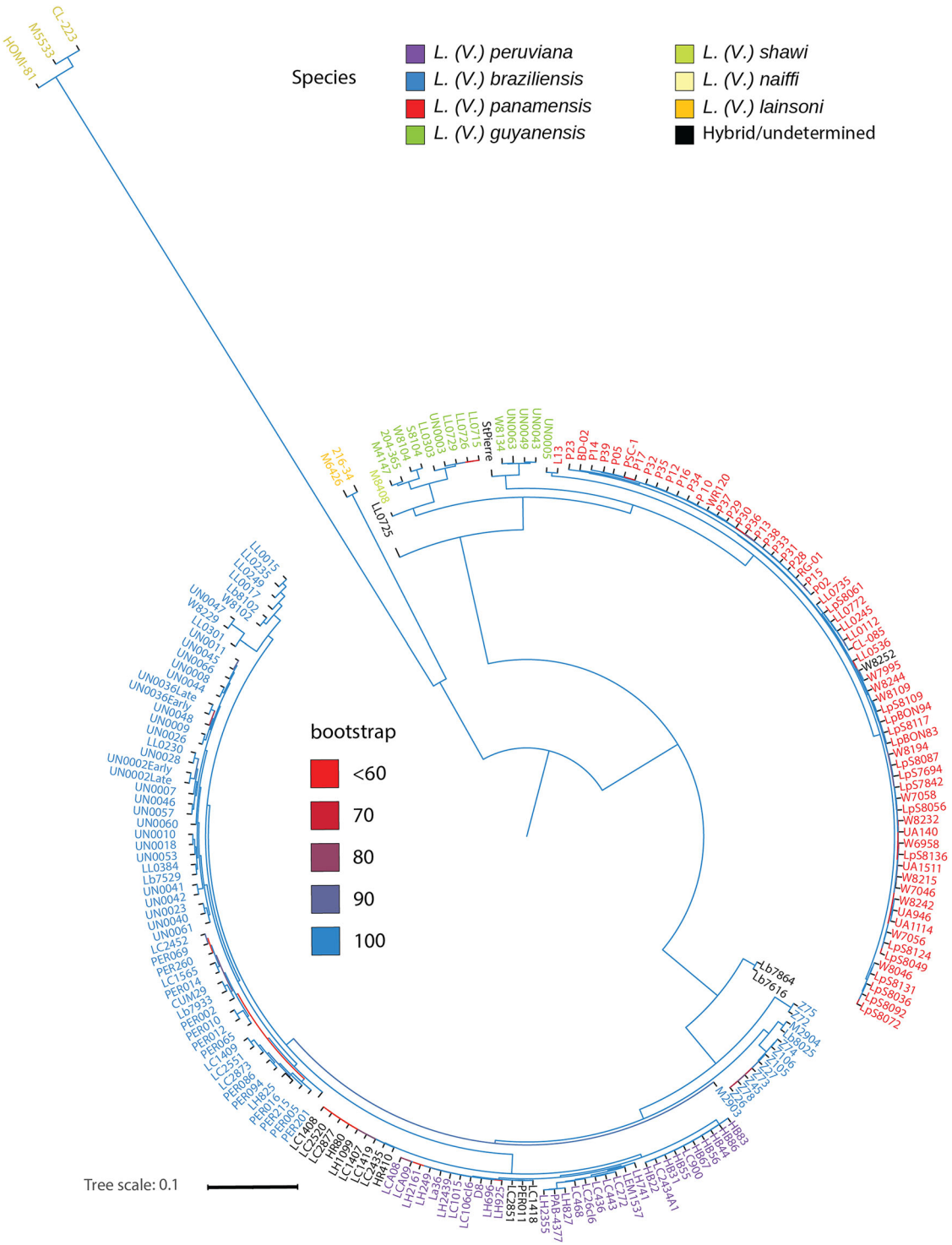

Fig. 1 | Maximum Likelihood phylogenetic reconstruction of *Leishmania* (*Viannia*) isolates. The tree was built from 673,944 SNPs segregating within *L.* (*Viannia*) after filtering to keep genotype calls with quality >40, biallelic sites, MAF > 0.01, and 200 genotyped individuals. Chromosome 31, missense, and nonsense variants were removed. The tree was rooted at the midpoint. Branch support was calculated with a 1000 ultrafast bootstrap on IQtree. Isolates are colored according to the species. A detailed visualization of branches for isolates within species can be seen in the visualization of the tree available at the Supplementary Fig. S4.

SNPs were further filtered within each group (See "Methods" for details). A model-based admixture analysis and a Neighbor Joining (NJ) tree were carried out for each group. Figure 2A, B show the sample clustering obtained for the *L.* (*V.*) *braziliensis/L.* (*V.*) *peruviana* complex (See Supplementary Data 1 and 3 for related data). Cross-validation (CV) analysis of the admixture analysis suggests K = 10 as the first local minimum of the CV error (Supplementary Fig. S6). The isolates of *L.* (*V.*) *peruviana*, the Brazilian isolates of *L.* (*V.*) *braziliensis* (population Lbra1), and the Peruvian isolates of *L.* (*V.*) *braziliensis* (population Lbra4) differentiate from K2, K3, and K4, respectively. Two populations within *L.* (*V.*) *peruviana* can be differentiated from K5 (Lper1 and Lper2), as previously reported[15]. Further subdivisions differentiate two populations,

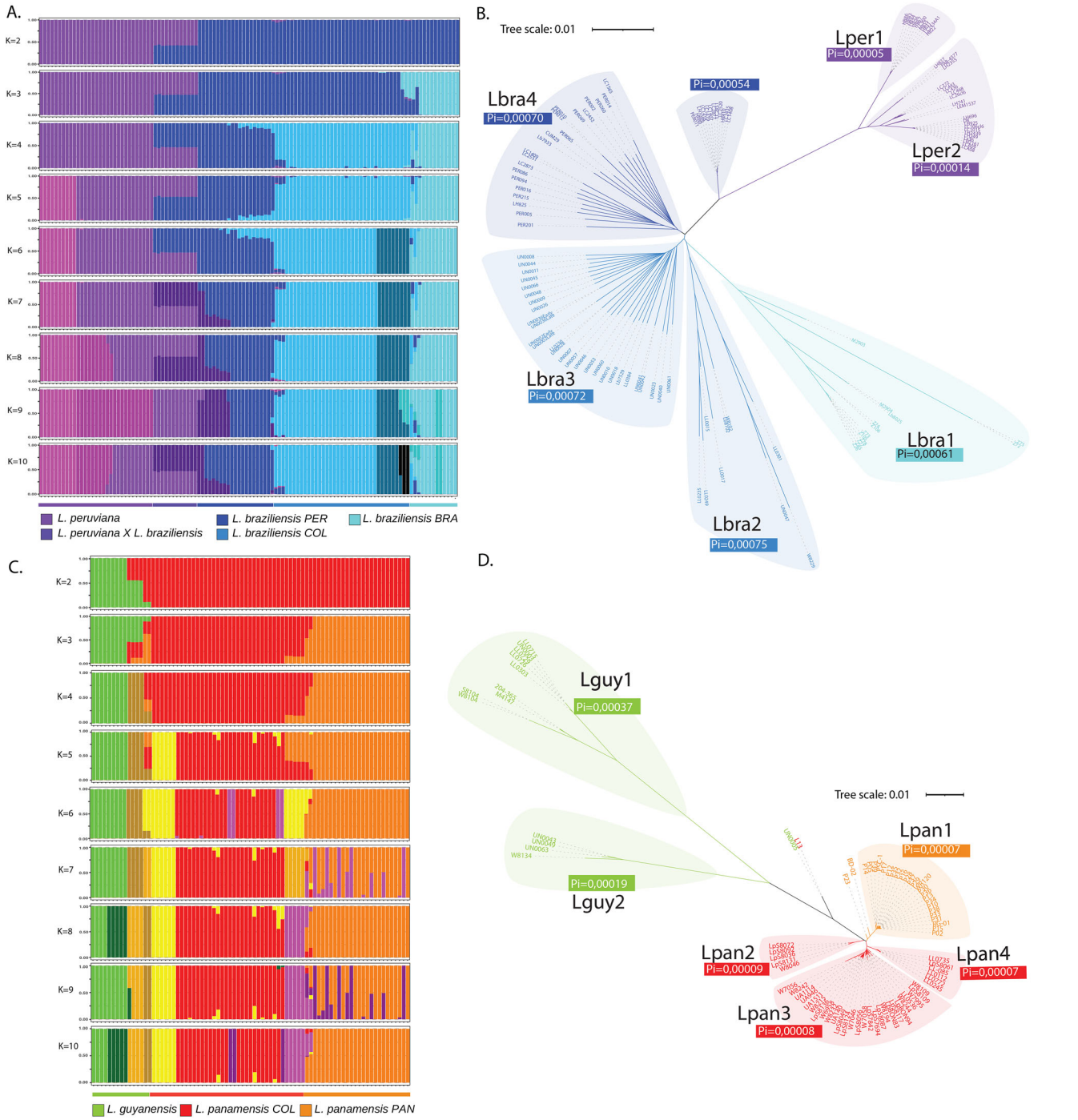

Fig. 2 | Genetic diversity of *Leishmania* (*Viannia*) species. A Admixture analysis for *L.* (*V.*) *braziliensis* and *L.* (*V.*) *peruviana*. **B** Distance tree based on the *L.* (*V.*) *braziliensis* and *L.* (*V.*) *peruviana* SNPs identified in comparison to the M2904 *L.* (*V.*) *braziliensis* strain. Labels show the Pi diversity statistic. **C** Admixture analysis for *L.* (*V.*) *guyanensis* and *L.* (*V.*) *panamensis*. **D** Distance tree based on the *L.* (*V.*) *guyanensis* and *L.* (*V.*) *panamensis* SNPs identified in comparison to the M2904 *L.* (*V.*) *braziliensis* strain. Labels show the Pi diversity statistic. Subpopulation names are assigned for the larger species: *L.* (*V.*) *braziliensis* is composed of Lbra1 (Brazilian samples), Lbra2 (Colombian samples from the Andean region), Lbra3 (Colombian samples from the Orinoco-Amazonian region), Lbra4 (Peruvian samples); *L.* (*V.*) *guyanensis* is composed of Lguy1 (Colombian samples from the Andean region (4 samples), and 4 samples from Venezuela, French Guyana and Brazil) and Lguy2 (Colombian samples from the Pacific region); *L.* (*V.*) *panamensis* is composed of Lpan1 (Panamanian samples), Lpan2 (Colombian samples from the Pacific-Andean region), Lpan3 (Colombian samples from the Andean-Caribbean region), Lpan4 (Colombian samples from the Andean-Amazonian region); *L.* (V.) *peruviana* is composed of Lper1 (Peruvian samples from Porculla region) and Lper2 (Peruvian samples from the Surco region).

including the Colombian isolates of *L.* (*V.*) *braziliensis* (Lbra2, Lbra3). This clustering is consistent with the geographical origin of the isolates. While Lbra2 includes isolates from the Andean region, Lbra3 includes isolates from the Orinoco-Amazon region (southeast of the country). Regarding hybrids, the *L.* (*V.*) *braziliensis/peruviana* hybrids are shown as such throughout the entire analysis and appear as a separate group in the NJ tree. The M2903 reference strain and the Colombian Lb8025 appeared as admixed isolates between Lbra1 and Lbra2.

Regarding diversity, *L.* (*V.*) *braziliensis* was the most diverse species, with a Pi higher than that of *L.* (*V.*) *panamensis* (Supplementary Table S1).

Four populations of *L.* (*V.*) *braziliensis* had similar genetic diversity (Pi ~0.0007), despite the lower number of isolates forming Lbra1 and Lbra2, compared to Lbra3 and Lbra4. The pairwise Fst statistic between the Colombian Lbra3 and the Peruvian Lbra4 clusters was lower (Fst 0.105) than the Fst between the Colombian Lbra2 and Lbra3 clusters (Fst 0.178, see Supplementary Table S2). Although the evolution of Leishmania diversity is not expected to follow the Wright-Fisher model, the distribution of minor allele frequency (MAF) follows a decreasing curve similar to the expected distribution of diversity under this model (Supplementary Fig. S7). Differences between expected (He) and observed (Ho) heterozygosity are centered at 0.1, suggesting a small reduction in heterozygosity produced by the population structure within *L.* (*V.*) *braziliensis*.

Figure 2C, D show the sample clustering obtained for the *L.* (*V.*) *panamensis/L.* (*V.*) *guyanensis* complex (See Supplementary Data 1 and 4 for related data). CV analysis of the admixture analysis identifies two local minimums of the CV error at K8 and K10 (Supplementary Fig. S6). The group of *L.* (*V.*) *guyanensis* isolates that appear more distant than the other samples in the NJ tree (Lguy1) can be observed as a separate group from K2. A second group of four *L.* (*V.*) *guyanensis* isolates forms a group from K4. Geographical segregation of the isolates was observed within *L.* (*V.*) *panamensis*. Isolates from Panama formed a cluster differentiated from the Colombian isolates from K3. This cluster was previously reported as Lpan1[13]. Furthermore, the Colombian isolates were separated into three clusters, two of them previously reported as Lpan2 and Lpan3[13], and one new cluster containing isolates mostly from the eastern Andean mountain chain (Lpan4). These clusters are evident from K5 and K6 in the admixture analysis. The cluster Lpan4 contained two new isolates classified as *L.* (*V.*) *guyanensis* and two classified as *L.* (*V.*) *panamensis* samples based on classical markers; it also includes one isolate typed as *L.* (*V.*) *guyanensis* and one typed as *L.* (*V.*) *panamensis* from public databases. The Colombian isolate UN0005, typed as *L.* (*V.*) *guyanensis*, and the Colombian L13 *L.* (*V.*) *panamensis* reference isolate appear as admixes until K6, but from K7, they form a separate cluster. Finally, Lguy1 splits into two subgroups at K8 and K10, one corresponding to Colombian isolates and the other corresponding to non-Colombian isolates, including the M4147 reference strain. Regarding admixes, the P23 and BD-02 isolates appear as admixed between Lpan1 and Lpan2. This is consistent with Llanes et al.[13]. Between three to five isolates within Lpan3 seem to have some level of admixture with Lpan4.

The global diversity within *L.* (*V.*) *panamensis* (0.000121) was much lower than the diversity within other species, including *L.* (*V.*) *guyanensis* (Supplementary Table S1). This can be evidenced in the MAF distribution, which shows that fewer SNPs are segregating at high frequencies compared to *L.* (*V.*) *braziliensis* (Supplementary Fig. S7). A small peak close to 0.5 can be observed, which can be explained by the differentiation between the population from Panama (Lpan1) and the Colombian populations. The distribution of differences between expected and observed heterozygosity is also centered at 0.1, but it shows two peaks at the extremes of the distribution. Around 3000 SNPs with differences close to 0.5 (Ho close to zero) correspond to SNPs differentiating the two major groups. Nearly 3,000 other SNPs with differences close to −0.5 correspond to SNPs with high observed heterozygosity (~1). Pairwise Fst values between populations of *L.* (*V.*) *panamensis* were higher than 0.5 except for the comparisons involving Lpan3. This can be attributed to the low number of isolates within Lpan2 and Lpan4. The Fst between Lpan1 and Lpan3 (0.27) was high in absolute numbers, but it was comparable to Fst values between populations of *L.* (*V.*) *braziliensis*.

A separate admixture analysis including all isolates was carried out to investigate possible admixed isolates between the two major species complexes (Supplementary Fig. S8). Five isolates were predicted as admixed between the two major groups (StPierre, LL0725, W8252, Lb7864, and Lb7616) in this analysis. The NJ tree (Supplementary Fig. S5) placed these isolates outside the two major groups, whereas the ML tree clustered StPierre within *L.* (*V.*) *guyanensis* and W8252 within *L.* (*V.*) *panamensis* (Fig. 1). The StPierre isolate was previously reported as a triploid hybrid between *L.* (*V.*) *braziliensis* and *L.* (*V.*) *guyanensis*[31]. Looking at heterozygosity, StPierre, W8252, and LL0725 had larger heterozygosity rates (>10%) in comparison to well-differentiated isolates. In contrast, Lb7616 and Lb7864 had heterozygosity rates lower than 1%. Heterozygosity percentages per sample increased to more than 40% for StPierre, W8252, and LL0725 when the analysis was restricted to species fixed SNPs (Supplementary Data 1). Metagenomic analysis of reads for the samples LL0725 and W8252 confirmed that no contamination of foreign DNA was causing this pattern. More than 99% of the reads corresponded to *L.* (*Viannia*) species (Supplementary Fig. S9). The admixture results suggest that W8252 is an admixed isolate between *L.* (*V.*) *panamensis* and *L.* (*V.*) *braziliensis*, and StPierre, LL0725, Lb7616, and Lb7864 are admixes between *L.* (*V.*) *guyanensis* and *L.* (*V.*) *braziliensis* (Supplementary Fig. S8). An introgression analysis looking for population assignment of local haplotypes confirmed that StPierre is an admixed isolate between *L.* (*V.*) *guyanensis* and *L.* (*V.*) *braziliensis*. The StPierre isolate contains introgressed regions summing up to 170 Kbp across the genome (Supplementary Table S3), and 88.44% of the fixed SNPs were heterozygous. A similar behavior was observed for the sample LL0725, for which 46.92% of the species fixed SNPs were heterozygous, and up to 10 Kbp introgressions were observed. In contrast, W8252 did not have introgression traces, as 98.7% of the fixed SNPs were heterozygotes, suggesting a coinfection scenario. Finally, Lb7616 and Lb7864 were assigned to *L.* (*V.*) *braziliensis* by half the regions. The placement of these samples in the phylogenetic trees (Fig. 1 and Supplementary Figs. S3 and S4) suggests that these could be isolates of a different population of *L.* (*V.*) *braziliensis* or even a close species.

Leishmania (*Viannia*) species genomes are highly conserved but are differentiated by multi-copy gene families

Given the observed diversity and differentiation of *Leishmania* (*Viannia*) populations in Colombia, ten Colombian isolates were sequenced using a long reads technology (see "Methods" for details). The sequenced samples correspond to five *L.* (*V.*) *braziliensis*, one *L.* (*V.*) *guyanensis*, two *L.* (*V.*) *panamensis*, and two isolates in between *L.* (*V.*) *guyanensis* and *L.* (*V.*) *panamensis*. Haploid genomes were assembled into 43 to 100 contigs per assembly, with an average N50 of 975 Kbp (Supplementary Data 5). The average genome size was 34 Mbp, which was more than 2 Mbp longer than the size of the current reference genome assemblies of *L.* (*V.*) *braziliensis* M2904 (32 Mbp) and *L.* (*V.*) *panamensis* PSC-1 (30.6 Mbp). The GC content ranged between 57.66% and 57.99%, close to the GC content of the references (~57.7%). The kinetoplast maxicircle was also identified and circularized using the 12s rRNA gene as the starting point of the molecule. On average, kinetoplasts were assembled into 27.5 Kbp molecules, as expected[32]. Consistent with previous reports[33], the alignment of the kinetoplasts showed conservation within the coding region and size variations in the divergent region (Supplementary Fig. S10).

Contigs were assigned to chromosomes according to the M2904 reference, followed by manual curation of the misassemblies between chromosomes. Each contig-level assembly was highly contiguous, having a range of 24 to 31 chromosomes reconstructed in one single contig. The number of gaps per assembly ranged between 3 and 13, according to the chromosome placement. For this reason, no scaffolding was attempted for these genome assemblies. The contigs assigned to chromosomes were consistently longer than the reference chromosomes (Supplementary Fig. S11), with large repetitions at the beginning and end of the contigs that could be related to a better resolution of the subtelomeric repetitive regions. However, due to ONT base calling errors[34], it was not possible to reconstruct and quantify repetitive sequences characteristic of the telomeres.

Figure 3 shows the major genomic features of representative genomes of *L.* (*V.*) *braziliensis* (LL0249), *L.* (*V.*) *guyanensis* (UN0003), and *L.* (*V.*) *panamensis* (LL0536), respectively (See the related data in the Supplementary Data 5). Gaps were observed, especially in chromosomes 2, 20, and 34, and they can be related to an increased GC content (Fig. 3A–C, tracks i). The gap at chromosome 2 was consistently present across all assemblies, and it was caused by a repetitive region of 40–80 Kbp in size, which could not be resolved with the nanopore reads (Supplementary Fig. S12). Interestingly,

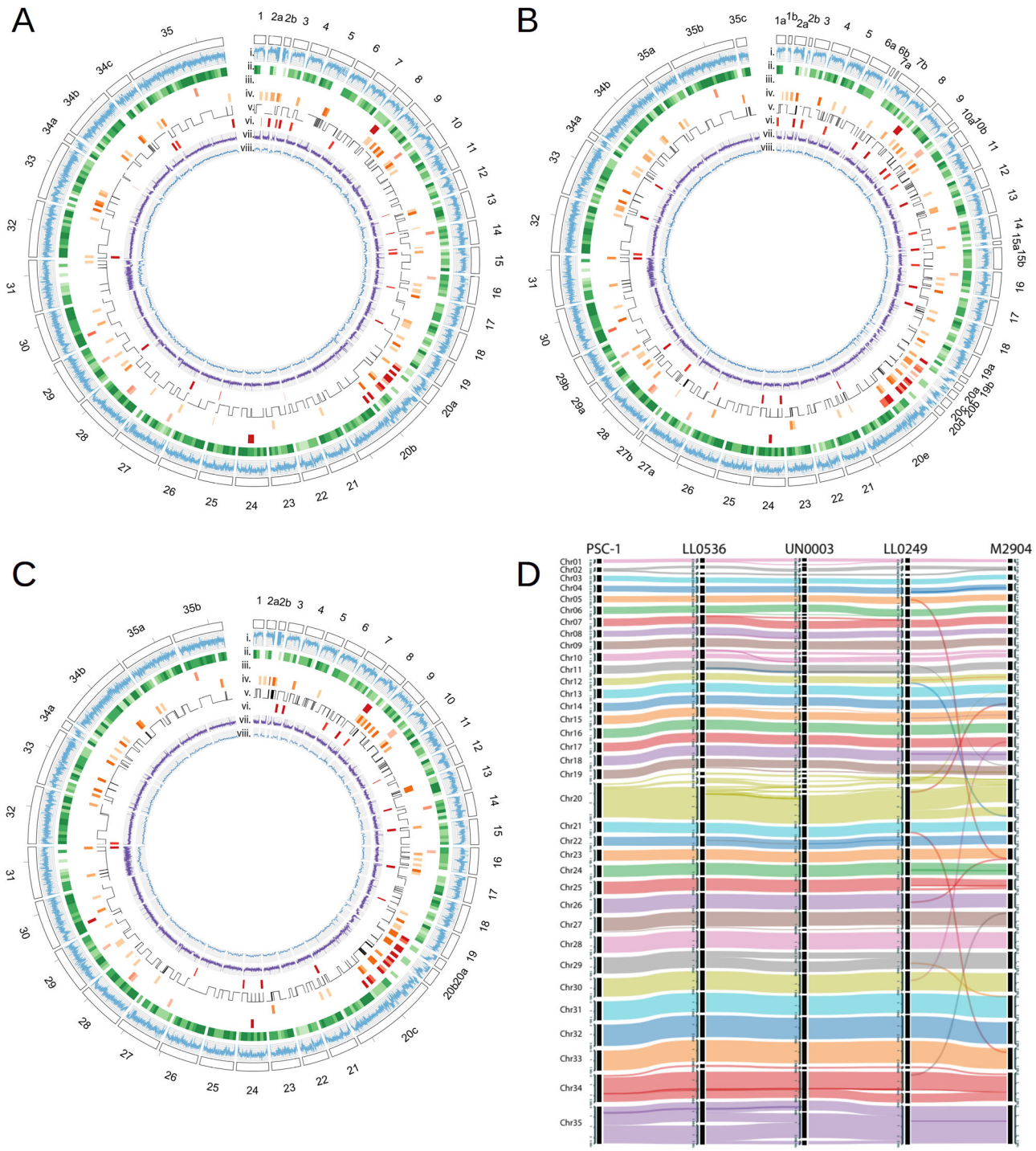

Fig. 3 | Genomic structure and main characteristics of *Leishmania* (*Viannia*) genomes. A–C Circos plots with main genomic features for **A** the LL0249 *L.* (*V.*) *braziliensis* isolate, **B** the UN0003 *L.* (*V.*) *guyanensis* isolate, and **C** the LL0536 *L.* (*V.*) *panamensis* isolate. The tracks show (i) GC content, (ii) density of core-genes, (iii) the density of genes coding for amastin surface glycoproteins, (iv) distribution of multi-copy gene families (at least 5 copies), (v) strand switches, (vi) distribution of TATE retroposons, (vii) Illumina sequencing depth (0–300×), (viii) Nanopore sequencing depth (0–100×). **D** Genome alignment of the M2904 and PSC-1 reference genomes and the assembled genomes of the isolates LL0249, UN0003, and LL0536.

this region is not observed in the reference genomes. The same behavior was observed in chromosome 34, where the assemblies were broken into two to three contigs, and two large repetitive regions were observed (Supplementary Fig. S13). The remaining chromosomes were assembled into 1 to 3 contigs, except for chromosome 20, which was broken into 5 contigs in some cases. Genomes were annotated considering evidence from the *L.* (*V.*) *braziliensis* M2904 and the *L.* (*V.*) *panamensis* L13 genomes (See methods for details). The total number of genes ranged between 9172 and 10,242, representing an increase of around 20% over the number of genes annotated in the current reference genomes (close to 8500 genes). Compared to the references, the increase in the gene number can be attributed to a better resolution of paralogous genes, achieved through the assembly of long reads.

Synteny conservation among isolates and species was assessed by aligning all genomes using the synteny relationships inferred from gene ortholog relationships. Figure 3D shows the alignment among the three representative genomes, compared to the *L.* (*V.*) *braziliensis* M2904 and *L.*

(*V.*) *panamensis* PSC-1 references (see all assemblies in Supplementary Fig. S14 and the blocks in the Supplementary Data 5). Each contig could be assigned to one chromosome in the reference M2904; however, two misassemblies in the reference were identified (Supplementary Fig. S15); all the contigs assigned to chromosome 11 consistently contained a region previously reported as part of chromosome 19, and a similar misassembly was observed between chromosome 12 and the end of chromosome 20. These misassemblies were also resolved in the non-referenced M2904 assembly reported in 2019[22], available at TriTrypDB[35]. Chromosome 20 of *L.* (*Viannia*) species has been reported as a fusion between chromosomes 20 and 34, compared to old-world *L.* (*Leishmania*) species[36].

Although the chromosome number and general genomic structure are conserved within the species, a gene-based pangenome analysis within *Leishmania* (*Viannia*) revealed important presence-absence variation affecting gene content. This pangenome, comprising 8749 orthogroups, also allowed us to identify the core-genome, as well as multicopy genes and species-specific genes. The exact core-genome comprised 6635 orthogroups containing genes present in all assemblies, including the references. Among these orthogroups, 4934 were single-copy genes, 35 were duplicated, and the remaining had three or more copies per genome. The size of this core-genome corresponds to nearly 60% of the genome, supporting the conservation of the gene content across the species in terms of functionality. These core-genes are distributed across the entire genome (Fig. 3A–C, tracks ii). Also, 25 additional orthogroups were identified as single-copy genes in our isolates but were absent in the references.

On the other hand, 122 multicopy orthogroups were identified. Each orthogroup could be mapped to a gene family (determined by functional domains); however, some highly divergent gene families were divided into several orthogroups. Only one gene family had, on average, more than 100 genes per genome, whereas the majority had, on average, between 5 and 37 genes per genome. A larger number of copies were annotated in the new assemblies for all gene families, compared to the references: the amastin surface glycoproteins gene family (190 vs 51) mainly located in chromosomes 8 and 20, the GP63 leishmanolysin gene family (37 vs 18) located in chromosomes 10 and 31, the alpha and beta tubulin families (41 vs 8) located in chromosomes 8, 13 and 33, the Autophagy ATG8 gene family (30 vs 4) located in chromosomes 9 and 19, the Tuzin coding genes (85 vs 13) located in chromosome 20 and associated with amastins, and the EF-1alpha coding genes (16 vs 1) located in chromosome 17. Also, more TATE DNA retroposons were found in our assemblies compared to the reference genomes (74 vs 15). These repetitive elements were mainly found in subtelomeric regions (Fig. 3A–C tracks vi).

A principal component analysis (PCA) of the number of copies in multicopy orthogroups differentiated species within the *Leishmania* (*Viannia*) subgenus, except for the previously assembled reference genomes (Supplementary Fig. S16). The first principal component separated the current reference genomes (PSC-1, L13, and M2904) from the isolates sequenced in this study. The second principal component separated the new assemblies of *L.* (*V.*) *panamensis*/*L.* (*V.*) *guyanensis* isolates from the new assemblies of *L.* (*V.*) *braziliensis*. As mentioned above, this result is a consequence of the lower number of copies that could be recovered for multicopy orthogroups in the previous assemblies. Excluding the reference genomes, the counts of 35 orthogroups were statistically different between *L.* (*V.*) *braziliensis* and *L.* (*V.*) *panamensis/guyanensis* complex, based on a non-paired Wilcoxon test (Supplementary Data 5). Besides amastin coding genes (described below), gene families with functional annotations and different copies between the *L.* (*V.*) *panamensis/guyanensis* complex and the *L.* (*V.*) *braziliensis*, were those encoding an ATP-dependent DEAD-box helicase (7 vs 3), a ubiquitin conjugation enzyme (7 vs 3), a kinesin (6 vs 3), an elongation factor (5 vs 1) and a sugar transporter (9 vs 4). On the other hand, genes encoding a peptidase (29 vs 51), the leishmaniolysin (18 vs 49), an epidermal growth factor (10 vs 17), a phosphatidic acid phosphatase (5 vs 11), and a glycerol uptake protein (3 vs 7) presented more copies within the *L.* (*V.*) *braziliensis* genomes.

Investigation of gene nucleotide and amino acid evolution through the rate of synonymous mutations (Ks) and non-synonymous to synonymous mutations (ka/ks) on core genes revealed patterns consistent with the phylogenetic relationships among species (Supplementary Fig. S17). Ks values below 0.05 were observed in all pairwise comparisons between *L.* (*Viannia*) genomes, suggesting recent diversification. In contrast, values above 0.5 were observed when comparing *L.* (*Viannia*) against *L.* (*Leishmania*) species. Regarding protein evolution, the Ka/Ks ratio is lower than 0.5 in all comparisons, suggesting that purifying selection acts on core genes across the species. Outliers are differentiated from the distribution and have Ka/Ks values close to or above 1. Although, theoretically, genes with Ka/Ks values above 1 should be investigated as genes under diversifying selection, we observed that in most of these cases, the overall number of mutations was low, which produced an artificial inflation of ratios. Only three genes had $Ks > 0.05$ and $Ka/Ks > 1$ within the *L.* (*Viannia*) comparisons, two annotated as hypothetical proteins and one annotated as a zinc finger.

Genomic organization and evolution of amastin genes in *Leishmania* (*Viannia*)

Amastin surface glycoproteins are reported as the largest gene-family in *Leishmania* species[37] and are mostly expressed during the parasite infective amastigote stage[38]. Genes annotated as amastins or amastin-like proteins were found in 31 different orthogroups. These orthogroups are consistent with amastin subfamilies and with the genomic location of each copy. Five of the 31 orthogroups discriminated between the two major species complexes, reflecting the divergence in sequence between species. The total number of genes annotated as amastins varied from 152 to 291 in our assemblies, consistently higher than previous reports and expected counts in previous studies[37,39,40]. This difference was produced by a large expansion of amastin paralogs located in chromosome 20 (Supplementary Table S4 and Supplementary Fig. S18).

A phylogeny across the gene family, based on amino acid sequences, shows that amastin proteins are grouped by sub-families (Fig. 4A and Supplementary Data 6). Close paralogs also seem to be clustered by genomic location. Copies belonging to the same species complex tend to group but are widely distributed along the phylogeny. Amastins belonging to α, β, and γ sub-families are closely related and differentiated from the genes belonging to the δ sub-family. A tandem pair of structurally distinct α amastins was found in chromosome 28 in all isolates. These amastins were previously described, and within them, a tandem pair in chromosome 14 clustered together. Although these isoforms in chromosome 14 were previously described as δ amastin, the inclusion of copies in ten genomes allowed us to improve their classification as α amastins. Chromosome 24 included a large array of between three and fifteen copies of γ amastins per genome assembly. Also, a tandem gene array comprising multiple copies of two β isoforms of the amastin was identified in chromosome 30. Finally, the δ amastin subfamily was distributed across the genome. The largest clusters comprised gene arrays in chromosomes 8 and 20 (over 100 copies per genome). These arrays varied in the copy number and the size of amastin genes, and some of them in chromosome 20 were intercalated with Tuzin genes. A single copy in chromosome 29 was identified and classified also as δ amastin. Amastin proteins vary in size and structure, even between tandem paralogs; however, most of them contain four predicted transmembrane regions (TM) and the amastin 11-aa signature[40] (Fig. 4B, C). Four single-copy genes in chromosomes 8, 16, 27, and 35 were annotated by Companion as amastins. Although these sequences contain similar transmembranal domains and match the PFAM motif PF07344, the amastin signature was not identified in the sequences (Supplementary Fig. S19).

Similar to the analysis of core genes, we calculated nucleotide and protein evolution statistics for amastin paralogs within species (Fig. 4D, E and Supplementary Data 7) and orthologs between species (Fig. 4F, G and Supplementary Data 7). The Ks values for paralogs that could be aligned were, on average, below 0.06 for close paralogs. In contrast, the average Ks for pairwise comparisons of paralogs between chromosomes

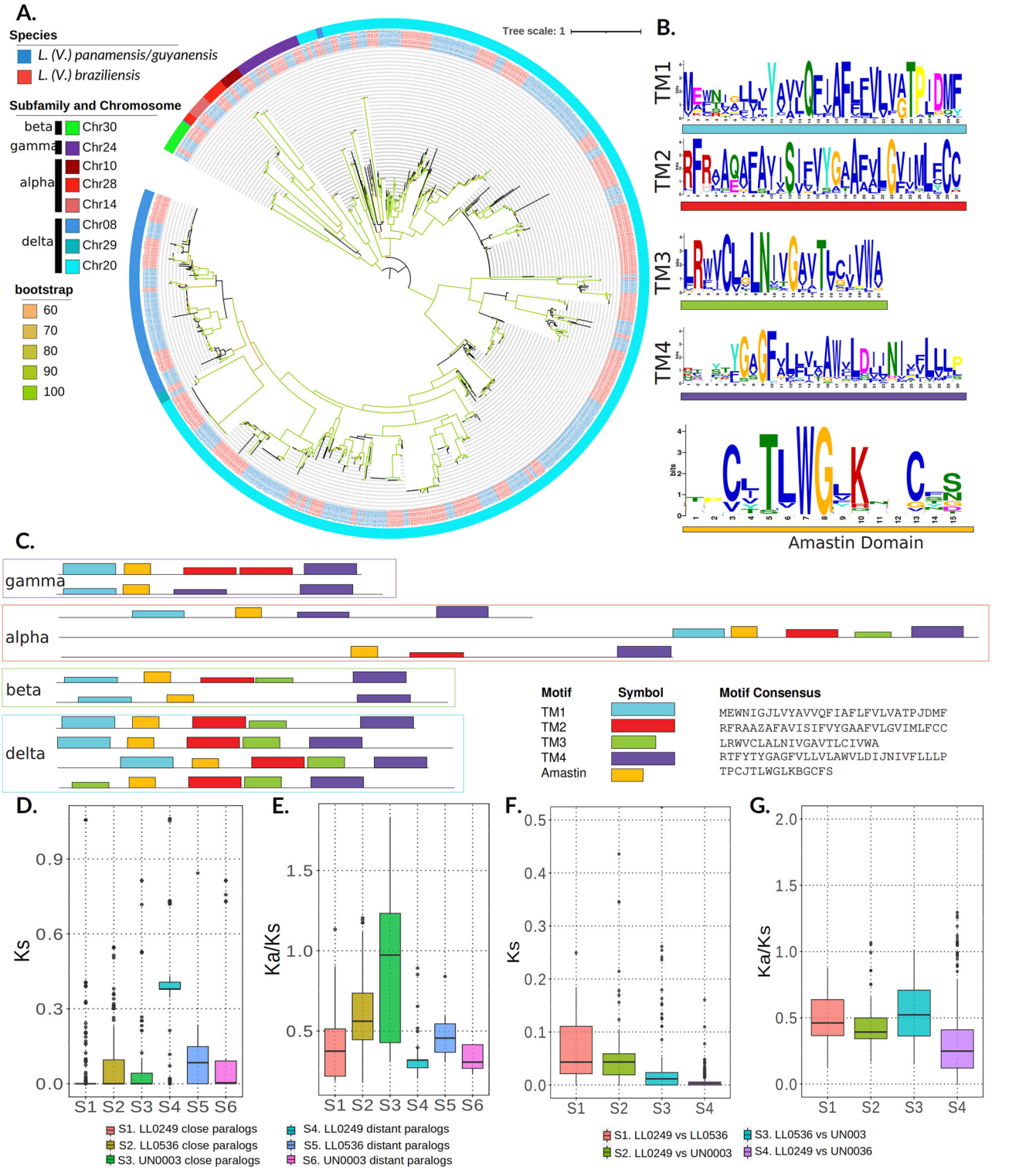

Fig. 4 | Amastin surface glycoproteins evolution within *Leishmania* (*Viannia*) isolates. A Amastins maximum-likelihood phylogeny using the entire repertoire of amastins in 10 *Leishmania* (Viannia) genomes. Branches colors show the bootstrap values. The color of the nodes corresponds to the species complex, and the color of the outer strip corresponds to the chromosome where each sequence was found. Chr30 corresponds to β-amastins, Chr28 and 14 correspond to α-amastins, Chr24 corresponds to γ-amastins. The remaining chromosomes correspond to δ-amastins. **B** Conserved domains between amastin proteins represented as the MEME logo for each domain. **C** Pattern of conserved domains in each amastin subfamily. **D**, **E** Ks and Ka/Ks respectively among close and distant amastin paralogs of *Leishmania* (Viannia) isolates: LL0249 and UN0036 (*L.* (*V.*) *braziliensis*), UN0003 (*L.* (*V.*) *guyanensis*), LL0536 (*L.* (*V.*) *panamensis*). See sample sizes and *p* values in the Supplementary Table S5. **F**, **G** Ks and Ka/Ks respectively among amastin orthologs of *Leishmania* (*Viannia*) isolates: LL0249 (*L.* (*V.*) *braziliensis*), UN0003 (*L.* (*V.*) *guyanensis*), LL0536 (*L.* (*V.*) *panamensis*). See sample sizes and *p* values in the Supplementary Table S6. Middle lines are medians and box limits represent first (Q1) and third (Q3) quartiles. Lines are drawn from Q1 minus 1.5 of the interquartile range (IQR) to Q3 + 1.5*IQR.

was 0.44 for *L.* (*V.*) *braziliensis*, 0.09 for *L.* (*V.*) *panamensis*, and 0.17 for *L.* (*V.*) *guyanensis*. All pairwise differences between groups, including *L.* (*V.*) *braziliensis* genes, were significant (p value < 0.001 for a Wilcoxon rank test). Also, the Ks values for close paralogs of both *L.* (V.) *panamensis* and *L.* (V.) *guyanensis* were significantly lower than the Ks values for distant paralogs of *L.* (V.) *panamensis*. The distribution of Ka/Ks ratios suggests some level of purifying selection, except for the group of close paralogs within *L.* (V.) *guyanensis*. Comparisons of amastin orthologs between species (Fig. 4F, G) resemble those obtained from core orthologs (Supplementary Fig. S17).

Discussion

This manuscript presents the main results of our work toward a comprehensive characterization of the genomic variability of *Leishmania* parasites circulating in Colombia. Building on previous research, we sequenced uncharacterized populations of *L.* (*V.*) *braziliensis* circulating across eastern and central Colombia and *L.* (*V.*) *panamensis* isolates from northwestern Colombia. In both cases, the genetic diversity observed within Colombia was larger than expected and comparable to that observed in the *L.* (*V.*) *panamensis* samples collected in Panama and the *L.* (*V.*) *braziliensis* samples collected in Peru and Brazil. The relatively low genetic distance among *Leishmania* (*Viannia*) species allowed us to build a common database of genomic variants for the complete species group. This database enabled us to compare genetic diversity between and within species using the same background of genetic markers. Comparing genetic diversity between species, *L.* (*V.*) *braziliensis* populations seem to have a larger genetic variation compared to *L.* (*V.*) *panamensis* and *L.* (*V.*) *guyanensis*. The location of the participant centers influenced the inclusion of the clinical isolates analyzed in the present study. However, the species-level geographic distribution of the sequenced parasites agrees with previous reports[41]. Interestingly, the two clusters observed in the Colombian *L.* (*V.*) *braziliensis* coincide with the geographical separation between the parasites circulating in the Andean region vs parasites from the same species circulating in the Amazon and Orinoco regions. This last group is genetically closer to the *L.* (*V.*) *braziliensis* population circulating in Peru, which suggests that the vectors and hosts supporting transmission cycles could differ between the divergent ecosystems present in South America[16]. That idea agrees with the differential biogeographic distribution of the sublineages in *L.* (*V.*) *peruviana*[15], based on a phylogenomic analysis that was corroborated in this work. We acknowledge that the current sampling only includes clinical isolates. Larger sampling, including different reservoirs and vectors, is needed to explore this idea further. Our sampling was also consistent with a previous clustering and group delimitation within the *L.* (*V.*) *panamensis* species[13]; however, a new group containing isolates from the Colombian eastern Andean Mountain chain was identified for this species.

Phylogenomic analysis of over 500 thousand SNPs segregating within *L.* (*Viannia*) revealed results consistent with previous studies that separate *L.* (*V.*) *braziliensis* from *L.* (*V.*) *panamensis* and *L.* (*V.*) *guyanensis*[18,29]. However, it did not show strong support for the separation of *L.* (*V.*) *panamensis* and *L.* (*V.*) *guyanensis*. Assuming that species classification by classical markers was correct, *L.* (*V.*) *guyanensis* did not form a monophyletic group. Likewise, the isolate L13, commonly used as a reference genome for *L.* (*V.*) *panamensis*, did not cluster within a *L. panamensis* population, but it clustered with the isolate UN0005, classified as *L.* (*V.*) *guyanensis*. Reciprocal monophyly has not been highly supported in previous single-locus and multi-loci studies[18,24,27,29]. The only *L.* (*V.*) *shawi* isolate in our study was closer to some *L.* (*V.*) *guyanensis* samples, a pattern reported by previous works[18,21,27]. Different hypotheses can be drawn from this phylogeny. First, as already suggested by the aforementioned studies, isolates that are currently classified as *L.* (*V.*) *panamensis* and *L.* (*V.*) *guyanensis* could better be represented as a single species complex. However, this single species must include the isolate currently classified as *L.* (*V.*) *shawi*. Future studies should include more *L. shawi* samples to elucidate the relationship between *L.* (*V.*) *shawi, L.* (*V.*) *guyanensis*, and *L.* (*V.*) *panamensis*. Second, assuming a two-species scenario, but given the relatively scarce genomic information on *L.* (*V.*) *guyanensis*, it is not surprising that current classification methods cannot be accurate to separate *L.* (*V.*) *guyanensis* from *L.* (*V.*) *panamensis*. To recognize these species based on reciprocal monophyly, the isolates UN0005, UN0043, UN0049, UN0063, and W8131 could be reclassified as *L.* (*V.*) *panamensis*. Finally, given the branch topology and the observed distances, further experiments could be conducted to evaluate if the clade including UN0043, UN0049, UN0063, and W8131 could be proposed as a new species. We acknowledge that the sampling covered in this study misses an important amount of species within the *Viannia* subgenus and that the number of isolates of *L.* (*V.*) *guyanensis* is small. Further sampling, sequencing, genome assemblies, and phenotypic screening of isolates within the different clades are needed to understand the phylogenetic history and to reconstruct the taxonomy within the *Viannia* subgenus fully. If the differentiation between these species has clinical or epidemiological relevance, new lab protocols and markers should be developed to redefine the separation between species in this scenario. In clinical practice, quick and accurate identification of parasite species constitutes an essential factor for the treatment response in cases of American cutaneous leishmaniasis[42–44]. Therefore, using a cost-effective and rapid technology, such as qPCR, would allow species discrimination from a clinical sample to be used in making clinical decisions. The genomic information generated in this study can be used as a source of information for the development of such protocols.

Regarding the *L.* (*V.*) *braziliensis*/*L.* (*V.*) *peruviana* complex, the SNP-based phylogenetic tree also did not support reciprocal monophyletic clades for these species. These findings are in agreement with a previous phylogeny obtained using the gene *HSP70*[24], in which *L.* (*V.*) *peruviana* is described as a subspecies of *L.* (*V.*) *braziliensis*. It also agrees with the topology of the neighbor net built from sequences of the genes *G6PD*, *6PGD*, *MPI*, and *ICD*[18]. Although the tree shown in the original publication describing the diversity of *L.* (*V.*) *braziliensis*/*L.* (*V.*) *peruviana* in Peru looks like a reciprocal monophyly[15], their analysis did not include *L.* (*V.*) *braziliensis* isolates from Brazil. During the review of this paper, a new manuscript was published describing population genomics of *Leishmania braziliensis* from Brazil[45]. This manuscript also describes *L.* (*V.*) *peruviana* as a distant population of *L.* (*V.*) *braziliensis*.

The sequencing effort presented in this work included samples that could be considered interspecies hybrids. The clustering analysis performed in this study agrees with the preceding report, which located the previously sequenced StPierre isolate as a hybrid between the two major groups[31]. In the set of clinical isolates sequenced in this study, we found an isolate that proved challenging for determining the parasite species by classical Lab-based methods (LL0725). The admixture analysis and the NJ tree showed a similar hybridization pattern, compared to the StPierre strain. The percentage of heterozygous sites in both cases is higher than that of a typical isolate; still, at the same time, it is much lower than the expected heterozygosity that could be produced by a co-infection event or by contamination. Although sexual reproduction is considered to be facultative in *Leishmania*[15,46–48], the more frequently reported occurrence of natural hybrids and the distribution of allele frequencies and heterozygosity suggest that sexual reproduction plays a role in the distribution of genetic variability within the species[49].

The initial analysis of the sequenced samples revealed that two isolates had low mapping rates to the *L.* (*V.*) *braziliensis* reference and formed a separate cluster. A phylogenomic analysis including variants obtained from aligned reads taken from publicly available *L.* (*L.*) isolates allowed us to classify the two samples as *L. (L.) amazonensis*. As expected, the mapping percentages for reads from *L.* (*L.*) isolates were small because only reads sequenced from regions conserved between species could be mapped. As a consequence, only 50,543 SNPs could be genotyped across the samples. However, the phylogenomic analysis of these SNPs clearly separated *L.* (*L.*) from *L.* (*V.*) samples. Although mapping reads between species can be considered counterintuitive, the variants that can be obtained in this analysis are similar to those obtained by assembly and alignment of conserved genes.

However, these variants should not be further filtered to perform population genomics within species. Future work in population genomics of *L.* (*L.*) should include more sampling in relevant locations and a population genomic analysis of variants obtained from reads aligned to a *L.* (*L.*) reference.

In this study, we also present the results of de novo assembly, annotation, and analysis of representatives of the major groups observed in the diversity analysis for which we had access to genetic material. The analysis confirmed the synteny conservation among species. The most important structural event was a previously reported chromosome fusion between chromosomes 20 and 34 of *L.* (*L.*) *major*[36]. The technology used for sequencing generated better-quality genomes than the reference genomes widely used in the last years. In particular, the new assemblies include nearly complete repertories of multicopy gene families. The largest family is the amastins, for which more than 150 copies per genome were reconstructed in our assemblies. This number of copies is between 2 and 4 times larger than that identified in current reference genomes. A large array of tandem paralogs in chromosome 20 explains this difference, which could be produced by a combination of biological and technical reasons. First, it is known that *Leishmania* isolates can gain and lose gene copies as a fitness gain for in vitro adaptation[50]. Second, the long-read sequencing technology allowed us to resolve genomic regions with large, nearly identical tandem copies. Paralogs of this family evolved through localized tandem duplications, so most copies are clustered at a few genomic loci. The analysis of nucleotide divergence showed that tandem copies are relatively younger than paralog pairs between chromosomes. This divergence pattern is consistent with amastin subfamilies reported across the Trypanosomatidae family[51].

Phylogenetic analysis revealed at least two new groups of amastins, compared to the previous studies[37], mostly based on the already sequenced genomes. A proper annotation and characterization of multicopy gene clusters through the use of long sequencing reading technologies have already been achieved in *L.* (*L.*) *donovani*, producing amastin genes copy number increased in the no reference genome, which also has a better annotation of A2 cluster, one important visceralization factor of Leishmania. Re-assembly of the A2 region indicated that evolution between cutaneous and visceral pathologies is associated with SNPs, pseudogenes, and copy number variation and not from chromosome rearrangements or large INDEL regions. Therefore, complete and reliable genomic information on amastins could be of great significance, given that they are preferentially expressed in amastigotes and involved in the interaction of the intracellular parasite and host cell membranes[37]. Furthermore, differences in amastin gene content have been associated with higher virulence in *L.* (*V.*) *braziliensis* vs *L. (V.) peruviana*[17].

Pangenomic analysis of de novo genome reconstruction leads to a better characterization of gene families than bioinformatic approaches guided by the alignment of short reads. Most previous studies in *Leishmania* calculated the copy number variation of genes based on the depth of reads aligned against a reference genome (usually the M2904 assembly). However, the reference misassemblies and the contraction of gene copies lead to errors and the overestimation of differences[52]. Our results indicate that the number of copies predicted by the depth of mapped short reads was inferior in most cases compared to the estimate obtained from genome assemblies. Thus, the analysis of de novo genome assemblies removed the bias produced by using a reference genome, especially when some gene families are absent in the reference. This fact was evidenced by a clear separation of samples according to the species of origin using only the estimated number of copies of the gene families identified by the pangenomic analysis. In contrast, we found that even in species with relatively simple genomes, such as *Leishmania*, automated gene annotation pipelines still can miss some genes, leading to overestimating gene presence/absence events. Given the current explosion of genome assemblies generated by the availability of long read sequencing technologies, the accuracy of annotation pipelines must be improved to minimize manual curation of gene annotations.

In summary, the work described in this manuscript represents an important step forward in the availability of genomic resources to understand the population dynamics of *Leishmania* species. Focusing on *Leishmania* species circulating in Colombia, our work complements previous efforts focused on *L. panamensis* circulating in Panama and *L. braziliensis* circulating in Peru, achieving a wide view of the distribution of variability through the northwest of South America. The results open interesting hypotheses on evolution and population genomics, motivating further sampling and sequencing on a wider range of geographical regions, ecosystems, and host species. The resources available with this study will also provide a rich source of information to develop new strategies for diagnosis, genomic surveillance, and treatment of the different forms of Leishmaniasis prevalent in tropical regions of the world.

Methods

Parasite culture, DNA preparation, and sequencing

65 *Leishmania* clinical isolates belonging to the biobanks of the Parasitology Laboratories from Facultad de Medicina- Universidad Nacional de Colombia, Hospital Universitario Centro Dermatológico Federico Lleras Acosta, and the PECET group at Universidad de Antioquia, derived from cutaneous lesions of Colombian patients diagnosed with CL or MCL from 2000 to 2021, were defrosted and grown for this study. All parasite isolates were derived from clinical samples collected from patients as part of routine diagnosis and treatment. All samples included written informed consent signed by the patient at the time of sampling. This consent authorizes the use of the stored sample. The ethical board at Universidad de los Andes granted ethical approval for the study. All ethical regulations relevant to human research participants were followed.

The isolates included 1 *L.* (*Leishmania*), 35 *L.* (*Viannia*) *braziliensis*, 12 *L.* (V.) *guyanensis*, and 17 *L.* (*V.*) *panamensis* according to molecular characterization (See Supplementary Data 1 for details). The parasites were cultured in Schneider's Insect Medium (Sigma-Aldrich) supplemented with 10% fetal bovine serum (FBS, Gibco) and 1% Penicillin-Streptomycin (p/s, Lonza, cat 17-602) at 27 °C. Parasite growth was evaluated by daily cell counting in the Neubauer chamber. Genomic DNA was prepared from 1 × 10e8 promastigotes using a QIAmp DNA mini kit (Qiagen, Cat #51306). DNA concentration was assessed by Qubit™ dsDNA HS and BR Assay Kits (Invitrogen, Cat # Q32851), and DNA quality was assessed by NanoDrop (Thermo Scientific™ NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer, Cat # ND-ONE-W) and Ethidium Bromide-stained agarose gel.

Short-read whole-genome shotgun sequencing was performed by Macrogen generating paired-end Illumina Hiseq reads, with a fragment length of around 350 bp and a read length of 100 bp. Raw reads were cleaned using Trimmomatic[53] v0.38, removing Illumina sequencing adapters and setting the parameters LEADING:20, SLIDINGWINDOW:5:20, and MINLEN:50. Oxford Nanopore (ONT) long-read sequencing was performed at Universidad de los Andes following the native barcoding protocol for R9 flow cells and a GridION device. ONT raw data basecalling was carried out using the Guppy v6 software for GPU using the Super Accurate model. We obtained about 2 Gbp (66×) for each sample in reads with a median (N50) length of around 12 Kbp. The quality distribution for these reads is shown in the Supplementary Fig. S20. After base calling using the super accurate model, we performed an error correction step with NECAT[54] v0.0.1 using default parameters.

Population analysis

To conduct population genetics analyses of the *Leishmania* (*Viannia*) group, we performed an integrated analysis of the sequenced samples with publicly available *Leishmania* Illumina datasets of isolates of *L.* (*V.*) *braziliensis* ($n = 38$), *L.* (*V.*) *peruviana* ($n = 31$), *L.* (*V.*) *braziliensis* × *L.* (*V.*) *peruviana* ($n = 12$), *L.* (*V.*) *guyanensis* ($n = 4$), *L.* (*V.*) *guyanensis* × *L.* (*V.*) *braziliensis* ($n = 1$), *L.* (*V.*) *panamensis* ($n = 47$), *L.* (*V.*) *shawi* ($n = 1$), *L.* (*V.*) *lainsoni* ($n = 2$), *L.* (*V.*) *naiffi* ($n = 3$), *L.* (*L.*) *mexicana* ($n = 3$), *L.* (*L.*) *amazonensis* ($n = 3$), *L.* (*L.*) *donovani* ($n = 9$), *L.* (*L.*) *infantum* ($n = 9$), *L.* (*L.*) *major* ($n = 6$), *L.* (*L.*) *tropica* ($n = 3$), *and L.* (*L.*) *aethiopica* ($n = 4$) (See Supplementary Data 1 for details). Reads obtained from all samples were

mapped to the *L.* (*V.*) *braziliensis* MHOM/BR/M2904 reference genome (GCF_000002845.2 NCBI accession number) using the ReadsAligner command of NGSEP[55] v5.0.0 with default parameters. Samples with a low percentage of mapping reads were also mapped to the *L.* (*L.*) *mexicana* MHOM/GT/2001/U1103 genome (GCF_000234665.1 NCBI accession number) to validate if the samples corresponded to *L.* (*Leishmania*) isolates. No further analysis was performed on reads aligned to the *L.* (*L.*) *mexicana* genome. Aligned reads were sorted and indexed using Picard[56] and were used as input to identify and to genotype variants using the MultiSampleVariantsDetector command of NGSEP[57]. Default parameters were used for this command, taking into account that these parameters are tuned for the analysis of Illumina WGS reads[57]. The population VCF file was filtered to keep only biallelic SNPs, variants with a minor allele frequency (MAF) < 0.01, and excluding genotype calls with low-quality scores (<40).

The genomic variation database was further filtered to select SNPs for the construction of phylogenetic trees and for population genomic analysis within the *L.* (*Viannia*) subgenus. SNPs for phylogenetic trees were selected, removing SNPs genotyped in less than 98% of the samples, removing nonsense and missense variants, and removing SNPs in the aneuploid chromosome 31. Maximum-likelihood phylogenies were constructed using IQTREE[58] v2.1.4, with 1000 bootstrap replicates and the GTR + F LG model, which was the best substitution model based on the lowest Bayesian Information Criterion (BIC) according to the results of ModelFinder. Given the low mapping rate of *L.* (*L.*) samples to the *L.* (*V.*) *braziliensis* genome and the consequent reduction of SNPs genotyped across the population, the analysis of the complete dataset was limited to the construction of the phylogenetic tree (Supplementary fig. 3). A distance matrix and a neighbor-joining (NJ) tree were also computed for *Viannia* isolates using the commands DistanceMatrixCalculator and Neighbor-Joining of NGSEP with default parameters. The Maximum-likelihood and Neighbor-Joining trees were visualized in iTOL[59] v5.

To examine the population structure within species groups, two VCF files were constructed, selecting the samples belonging to each group. The MAF filter was applied again to each group separately to remove variants that were monomorphic within each group. To reduce possible biases related to linkage disequilibrium, the VCF files were further filtered, keeping SNPs separated by at least 250 bp using VCFtools[60] v0.1.16, as previously suggested[12]. ADMIXTURE[61] v1.3 was run with a K ranging from 2 to 20 and specifying a 20-fold cross-validation procedure to select the most supported K. Genetic differentiation between populations was estimated with Fst statistics, and the diversity within populations was assessed with Pi statistics. Both statistics were calculated using VCFtools[60].

The SNPs perfectly differentiating species or subpopulations were calculated using the VCFIntrogressionAnalysis functionality of NGSEP, setting 0.99 as the minimum difference between reference allele frequencies of at least two populations to consider a variant discriminative. Potential hybrid isolates were assigned to one population afterward, and the introgression was rerun using default parameters to identify introgressed regions.

Genome assembly and annotation

Each genome was assembled using corrected ONT reads with the Assembler command of NGSEP[55], with a window size of 20, and considering only reads with lengths larger than 5000 bp. After that, each genome was polished with the Illumina reads using NGSEP. Briefly, short reads were mapped to the assembled genome (ReadsAligner command), and variants were called using the SingleSampleVariantsDetector command. Homozygous alternative variants between the mapped reads and the assemblies were assumed to be errors and were corrected using the IndividualGenomeBuilder command of NGSEP. All steps were run with default parameters. A first genome annotation was carried out in Companion[62] using *L.* (V.) *braziliensis* M2904 as the reference genome and avoiding contiguation of the contigs.

To assign contigs to chromosomes, each genome was aligned against the *L.* (V.) *braziliensis* M2904 according to gene synteny using the GenomesAligner command of NGSEP[63]. Each contig was assigned to a chromosome or as an unplaced contig after manual curation of identified misassemblies. Small redundant contigs were identified with minimap2[64] v2.24 and removed from the genome. Then, a BLASTn[65] v.2.15.0 search was performed to identify the Kinetoplast of each genome using the *L.* (V.) *braziliensis* M2904 maxicircle as reference (OY748430.1). The matching contig was circularized using the 12 s sequence as the starting gene of the entire sequence. BUSCO[66] v5.6.1 and QUAST[67] v5.2.0 were used to assess the base pair quality and contiguity of the assembled genomes. Merqury[68] v1.3 was used to estimate the QV score and error rate of each assembly based on Illumina reads.

Gene annotation for each curated genome was performed by combining the annotations of two independent runs of Companion[62]: one setting the reference genome as *L.* (V.) *braziliensis* M2904 and another using the *L.* (V.) *panamensis* L13 genome as reference. The annotations were combined using a custom script included in the NGSEP distribution (class ngsep.transcriptome.io.GFF3CombineAnnotations).

Comparative genomics and pangenome reconstruction

The GenomesAligner command of NGSEP[63] was run over the ten assembled genomes, the M2904 *L.* (V.) *braziliensis*, and the L13 and PSC-1 *L.* (V.) *panamensis* reference genomes to generate homolog gene clusters (orthogroups) and pairwise synteny blocks. The following parameters were modified from the defaults after running experiments comparing the clusters with OrthoMCL groups available in the TriTrypDB database (https://tritrypdb.org): k-mer length (-k) was set to 5, weighted percentage of shared k-mers (-p) was set to 20, and the Markov clustering step was not executed (-s). The genomes aligner generates a presence/absence matrix with the counts of genes belonging to orthogroups within the genomes. These counts were used to identify the exact-core genome as single-copy genes present in all the genomes, also including genes with two or three copies in all the genomes. Clusters were matched to OrthoMCL orthogroups available at the TriTrypDB database (https://tritrypdb.org), using the reference genes as anchors. Expression of genes in clusters not including reference genes was validated for the three representative genomes of *L.* (*V.*) *braziliensis*, *L.* (*V.*) *panamensis*, and *L.* (*V.*) *guyanensis*, mapping publicly available RNA-seq reads for samples of the three species (Supplementary Data 5).

The remaining clusters were filtered to identify multicopy gene-families as the families where at least one genome has more than two copies. An independent two-tailed Wilcoxon-test was applied to identify families differentiating *L.* (V.) *braziliensis* from *L.* (V.) *panamensis/guyanensis*. Because the references contained fewer genes than our assemblies due to technical issues, they were discarded from this analysis. Functional annotations of genes belonging to each cluster provided by Companion were used to assign functionality to gene clusters.

Amastins analysis

To clarify the evolutionary history of the amastin family, a maximum likelihood phylogenetic tree was inferred by using IQTREE[58] v2.1.4 with the LG (Le Gascuel) model, which was the best substitution model based on the lowest Bayesian Information Criterion (BIC) revealed by ModelFinder. Branch supports were assessed by bootstrapping with 1,000 replicates. The multiple sequence alignment used for this analysis was generated using Clustal Omega[69]. We used 1795 protein sequences of all gene-clusters annotated as amastins for the alignment. In addition, all sequences were scanned to identify common motifs with MEME[70] v5.5.

Statistics and reproducibility

Statistical analyses were conducted to determine significance of pairwise differences between multicopy orthogroups and between nucleotide and protein evolution statistics among groups. Non-parametric unpaired Wilcoxon rank tests were performed for each comparison. Exact *p* values for these comparisons are provided in the Supplementary Data 5 and Supplementary Tables S5 and S6.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this study is available at the NCBI sequence read archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra) with bioproject accession number PRJNA1095027. The genome assemblies are available at the Assembly database of NCBI (https://www.ncbi.nlm.nih.gov/assembly/) under the bioproject accession number PRJNA1095027. The reference genomes generated by previous studies and used in this work are available at TriTrypDB (https://tritrypdb.org/tritrypdb/app) genomes repository under the accession numbers GCA_000444285.2 (*L. aethiopica* L147), GCA_000438535.1 (*L. amazonensis* M2269), GCA_025688915.1 (*L. amazonesis* PH8), GCA_000410695.2 (*L. arabica* LEM1108), GCA_000340355.2 (*L. braziliensis* M2903), GCA_000002845.2 (*L. braziliensis* M2904), GCA_000227135.2 (*L. donovani* BPK282A1), GCA_003719575.1 (*L. donovani* CL-SL), GCA_900635355.2 (*L. donovani* HU3), GCA_017916305.1 (*L. enrietti* CUR178), GCA_000410755.2 (*L. enrietti* LEM3045), GCA_000443025.1 (*L. gerbilli* LEM452), GCA_900500625.2 (*L. infantum* JPCM5), GCA_916722125.1 (*L. major* Friedlin 2021), GCA_000002725.2 (*L. major* Friendlin), GCA_000331345.1 (*L. major* LV39c5), GCA_000250755.2 (*L. major* SD75.1), GCA_000409445.2 (*L. martiniquensis* LEM2494), GCA_017916325.1 (*L. martiniquensis* LSCM1), GCA_000234665.4 (*L. mexicana* U1103), GCA_017916335.1 (*L. orientalis* LSCM4), GCA_000340495.1 (*L. panamensis* L13), GCA_000755165.1 (*L. panamensis* PSC-1), GCA_009731335.1 (*L. tarentolae* Parrot 2019), GCA_000410715.1 (*L. tropica* L590), GCA_000441995.1 (*L. turanica* LEM423).

Code availability

The data analysis of short and long DNA sequencing reads was performed running open source software tools as detailed in the "Methods" section and in the reporting summary. In particular, reference-based analysis of short reads and genome assemblies based on long reads were performed running different functionalities of NGSEP v5.0. Public releases of NGSEP are available at sourceforge (http://ngsep.sf.net) and live development is available at github (https://github.com/NGSEP). Custom scripts for specific data management tasks are also available with the distribution of NGSEP.

Received: 16 April 2024; Accepted: 3 June 2025;
Published online: 14 June 2025

References

1. World Health Organization. *Neglected Tropical Diseases 2022* (World Health Organization, 2022).
2. Dumonteil, E., Herrera, C. & Buekens, P. A therapeutic preconceptional vaccine against Chagas disease: a novel indication that could reduce congenital transmission and accelerate vaccine development. *PLoS Negl. Trop. Dis.* **13**, e0006985 (2019).
3. Ponte-Sucre, A. et al. Drug resistance and treatment failure in leishmaniasis: a 21st century challenge. *PLoS Negl. Trop. Dis.* **11**, e0006052 (2017).
4. Rassi, A. & Marin, J. Chronic Chagas cardiomyopathy: a review of the main pathogenic mechanisms and the efficacy of aetiological treatment following the BENznidazole Evaluation for Interrupting Trypanosomiasis (BENEFIT) trial. *Mem. Inst. Oswaldo Cruz* **112**, 224–235 (2017).
5. Cunha, J. et al. Characterization of the biology and infectivity of *Leishmania infantum* viscerotropic and dermotropic strains isolated from HIV+ and HIV- patients in the murine model of visceral leishmaniasis. *Parasite Vectors* **6**, 122 (2013).
6. Jimenez, P., Jaimes, J., Poveda, C. & Ramirez, J. D. A systematic review of the *Trypanosoma cruzi* genetic heterogeneity, host immune response and genetic factors as plausible drivers of chronic chagasic cardiomyopathy. *Parasitology* **146**, 269–283 (2019).
7. Rodríguez, J. A. I., Rodríguez, S. N. I. & Olivera, M. J. Leishmaniasis in the Colombian post-conflict era: a descriptive study from 2004 to 2019. *Rev. Soc. Bras. Med. Trop.* **54**, e06122020 (2021).
8. Espinosa, O. A., Serrano, M. G., Camargo, E. P., Teixeira, M. M. G. & Shaw, J. J. An appraisal of the taxonomy and nomenclature of trypanosomatids presently classified as Leishmania and Endotrypanum. *Parasitology* **145**, 430–442 (2018).
9. Ramirez, J. D. et al. Taxonomy, diversity, temporal and geographical distribution of cutaneous leishmaniasis in Colombia: a retrospective study. *Sci. Rep.* **6**, 28266 (2016).
10. Salgado-Almario, J., Hernández, C. A. & Ovalle-Bracho, C. Geographical distribution of Leishmania species in Colombia, 1985-2017. *Biomédica* **39**, 278–290 (2019).
11. El-Sayed, N. M. et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309**, 409–415 (2005).
12. Franssen, S. U. et al. Global genome diversity of the *Leishmania donovani* complex. *eLife* **9**, e51243 (2020).
13. Llanes, A. et al. Genomic diversity and genetic variation of *Leishmania panamensis* within its endemic range. *Infect. Genet. Evol.***103**, 105342 (2022).
14. Patiño, L. H., Muñoz, M., Cruz-Saavedra, L., Muskus, C. & Ramírez, J. D. Genomic diversification, structural plasticity, and hybridization in *Leishmania (Viannia) braziliensis*. *Front. Cell. Infect. Microbiol.* **10**, 582192 (2020).
15. Van den Broeck, F. et al. Ecological divergence and hybridization of neotropical Leishmania parasites. *Proc. Natl. Acad. Sci. USA***117**, 25159–25168 (2020).
16. Figueiredo de Sá, B., Rezende, A. M., Melo Neto, O. P. D., Brito, M. & Brandao Filho, S. P. Identification of divergent *Leishmania (Viannia) braziliensis* ecotypes derived from a geographically restricted area through whole genome analysis. *PLoS Negl. Trop. Dis.* **13**, e0007382 (2019).
17. Valdivia, H. O. et al. Comparative genomic analysis of *Leishmania (Viannia) peruviana* and *Leishmania (Viannia) braziliensis*. *BMC Genomics* **16**, 715 (2015).
18. Boité, M. C., Mauricio, I. L., Miles, M. A. & Cupolillo, E. New insights on taxonomy, phylogeny and population genetics of *Leishmania (Viannia)* parasites based on multilocus sequence analysis. *PLoS Negl. Trop. Dis.* **6**, e1888 (2012).
19. Delgado, O. et al. Cutaneous leishmaniasis in Venezuela caused by infection with a new hybrid between *Leishmania (Viannia) braziliensis* and *L.(V.) guyanensis*. *Mem. Inst. Oswaldo Cruz* **92**, 581–582 (1997).
20. Jennings, Y. L. et al. Phenotypic characterization of *Leishmania* spp. causing cutaneous leishmaniasis in the lower Amazon region, western Pará state, Brazil, reveals a putative hybrid, *Leishmania (Viannia) guyanensis* × *Leishmania (Viannia) shawi shawi*. *Parasite* **21**, 39 (2014).
21. Harkins, K. M., Schwartz, R. S., Cartwright, R. A. & Stone, A. C. Phylogenomic reconstruction supports supercontinent origins for Leishmania. *Infect. Genet. Evol.***38**, 101–109 (2016).
22. Patiño, L. H. et al. Major changes in chromosomal somy, gene expression and gene dosage driven by SbIII in *Leishmania braziliensis* and *Leishmania panamensis*. *Sci. Rep.* **9**, 9485 (2019).
23. Urrea, D. A. et al. Genomic analysis of Colombian *Leishmania panamensis* strains with different level of virulence. *Sci. Rep.* **8**, 17336 (2018).
24. Fraga, J., Montalvo, A. M., De Doncker, S., Dujardin, J. C. & Van der Auwera, G. Phylogeny of Leishmania species based on the heat-shock protein 70 gene. *Infect. Genet. Evol.* **10**, 238–245 (2010).
25. Bañuls, A. L. et al. Genetic analysis of Leishmania parasites in Ecuador: are *Leishmania (Viannia) panamensis* and *Leishmania (V.) guyanensis* distinct taxa? *Am. J. Trop. Med. Hyg.* **61**, 838–845 (1999).
26. Bañuls, A. L. et al. Is *Leishmania (Viannia) peruviana* a distinct species? A MLEE/RAPD evolutionary genetics answer. *J. Eukaryot. Microbiol.* **47**, 197–207 (2000).
27. Coughlan, S. et al. *Leishmania naiffi* and *Leishmania guyanensis* reference genomes highlight genome structure and gene evolution in the Viannia subgenus. *R. Soc. Open Sci.* **5**, 172212 (2018).
28. Schönian, G., Mauricio, I. & Cupolillo, E. Is it time to revise the nomenclature of Leishmania? *Trends Parasitol.* **26**, 466–469 (2010).

29. Hoyos, J., Rosales-Chilama, M., León, C., González, C. & Gómez, M. A. Sequencing of hsp70 for discernment of species from the *Leishmania (Viannia) guyanensis* complex from endemic areas in Colombia. *Parasites Vectors* **15**, 1–11 (2022).
30. Dumetz, F. et al. Modulation of aneuploidy in *Leishmania donovani* during adaptation to different in vitro and in vivo environments and its impact on gene expression. *mBio* **8**, e00599-17 (2017).
31. Van den Broeck, F. et al. Genome analysis of triploid hybrid Leishmania parasite from the Neotropics. *Emerg. Infect. Dis.* **29**, 1076–1078 (2023).
32. Jensen, R. E. & Englund, P. T. Network news: the replication of kinetoplast DNA. *Annu. Rev. Microbiol.* **66**, 473–491 (2012).
33. Akhoundi, M. et al. Leishmania infections: molecular targets and diagnosis. *Mol. Asp. Med.* **57**, 1–29 (2017).
34. Tan, K. T. et al. Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. *Genome Biol.* **23**, 180 (2022).
35. Aslett, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic acids Res.* **38**, D457–D462 (2010).
36. Britto, C. et al. Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World Leishmania genomes. *Gene* **222**, 107–117 (1998).
37. Cardoso de Paiva, R. M. et al. Amastin knockdown in *Leishmania braziliensis* affects parasite-macrophage interaction and results in impaired viability of intracellular amastigotes. *PLoS Pathog.* **11**, e1005296 (2015).
38. Kaushal, R. S. et al. Leishmania species: a narrative review on surface proteins with structural aspects involved in host–pathogen interaction. *Chem. Biol. Drug Des.* **102**, 332–356 (2023).
39. Albanaz, A. T. S. et al. Genome analysis of *Endotrypanum* and *Porcisia* spp., closest phylogenetic relatives of Leishmania, highlights the role of amastins in shaping pathogenicity. *Genes* **12**, 444 (2021).
40. Rochette, A. et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in *Leishmania* spp. *Mol. Biochem. Parasitol.* **140**, 205–220 (2005).
41. Ovalle-Bracho, C., Londoño-Barbosa, D., Salgado-Almario, J. & González, C. Evaluating the spatial distribution of Leishmania parasites in Colombia from clinical samples and human isolates (1999 to 2016. *PLoS ONE***14**, e0214124 (2019).
42. Perez-Franco, J. E. et al. Clinical and parasitological features of patients with American cutaneous leishmaniasis that did not respond to treatment with meglumine antimoniate. *PLoS Negl. Trop. Dis.* **10**, e0004739 (2016).
43. Romero, G. A., Guerra, M. V., Paes, M. G. & Macêdo, V. O. Comparison of cutaneous leishmaniasis due to *Leishmania (Viannia) braziliensis* and *L. (V.) guyanensis* in Brazil: therapeutic response to meglumine antimoniate. *Am. J. Trop. Med. Hyg.* **65**, 456–465 (2001).
44. Rugani, J. N., Quaresma, P. F., Gontijo, C. F., Soares, R. P. & Monte-Neto, R. L. Intraspecies susceptibility of *Leishmania (Viannia) braziliensis* to antileishmanial drugs: antimony resistance in human isolates from atypical lesions. *Biomed. Pharmacother.* **108**, 1170–1180 (2018).
45. Heeren, S. et al. Evolutionary genomics of *Leishmania braziliensis* across the neotropical realm. *Commun. Biol.* **7**, 1587 (2024).
46. Cotton, J. A. et al. Genomic analysis of natural intra-specific hybrids among Ethiopian isolates of *Leishmania donovani*. *PLoS Negl. Trop. Dis.* **14**, e0007143 (2020).
47. Glans, H. et al. High genome plasticity and frequent genetic exchange in Leishmania tropica isolates from Afghanistan, Iran and Syria. *PLoS Negl. Trop. Dis.* **15**, e0010110 (2021).
48. Shaik, J. S., Dobson, D. E., Sacks, D. L. & Beverley, S. M. Leishmania sexual reproductive strategies as resolved through computational methods designed for aneuploid genomes. *Genes* **12**, 167 (2021).
49. Sádlová, J. et al. Comparative genomics of *Leishmania donovani* progeny from genetic crosses in two sand fly species and impact on the diversity of diagnostic and vaccine candidates. *PLoS Negl. Trop. Dis.* **18**, e0011920 (2024).
50. Bussotti, G. et al. Leishmania genome dynamics during environmental adaptation reveal strain-specific differences in gene copy number variation, karyotype instability, and telomeric amplification. *mBio* **9**, e01399-18 (2018).
51. Jackson, A. P. The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol. Biol. Evol.***27**, 33–45 (2010).
52. Valiente-Mullor, C. et al. One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput. Biol.* **17**, e1008678 (2021).
53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
54. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **12**, 60 (2021).
55. González-Garcia, L. et al. New algorithms for accurate and efficient de novo genome assembly from long DNA sequencing reads. *Life Sci. Alliance* **6**, e202201719 (2023).
56. Broad Institute. Picard Toolkit. GitHub Repository. https://broadinstitute.github.io/picard/ (2019).
57. Tello, D. et al. NGSEP3: accurate variant calling across species and sequencing protocols. *Bioinformatics* **35**, 4716–4723 (2019).
58. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.***37**, 1530–1534 (2020).
59. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
60. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
61. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
62. Steinbiss, S. et al. Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* **44**, W29–W34 (2016).
63. Tello, D. et al. NGSEP 4: efficient and accurate identification of orthogroups and whole-genome alignment. *Mol. Ecol. Resour.* **23**, 712–724 (2023).
64. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
65. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics***10**, 421 (2009).
66. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.***38**, 4647–4654 (2021).
67. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
68. Rhie, A. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
69. Madeira, F. et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
70. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).

Acknowledgements

The work presented in this manuscript was supported by MinCiencias [Patrimonio autónomo del Fondo Nacional de Financiamiento para la ciencia, la tecnología y la innovación Francisco José de Caldas. Contract 80740-441-2020], awarded to J.D. We acknowledge the High-Performance Computing Service at Universidad de los Andes, for providing HPC resources that have contributed to the research results reported in this paper. We also acknowledge the financial support provided by the Vice

Presidency of Research & Creation publication fund at Universidad de los Andes. Finally, we acknowledge Marcela Guevara, Mariana Restrepo, Fidias Gonzalez, and the personnel at the GenCore sequencing facility at Uniandes for their technical support to carry on the experiments.

Author contributions

J.D., D.A.U., and M.C.E. conceived the study and coordinated the project. C.O.-B., C. Colorado, C. Camargo, E.Q., M.J.M., and C.M. collected the samples. L.N.G.-G., M.P.-M., A.M.C., C.O.-B., C. Colorado, C. Camargo, E.Q., M.J.M., and F.B.-A. performed lab work for DNA sequencing. L.N.G.-G., M.P.R., L.L., and J.D. performed bioinformatic analysis. S.R., M.C.E., C.M., and D.A.U. provided scientific guidance and interpretation of the results. L.N.G.-G., M.P.R., L.L., M.C.E., and J.D. drafted the manuscript. All authors reviewed and approved the latest version of the manuscript.

Competing interests

The authors declare that they have no competing interests regarding the results presented in this manuscript. J.D. is an Editorial Board Member for *Communications Biology*, but was not involved in the editorial review of, nor the decision to publish this article.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08331-1.

Correspondence and requests for materials should be addressed to María Clara Echeverry or Jorge Duitama.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary handling editors: Johannes Stortz and Mengtan Xing.