Nuclear and mitochondrial genome sequencing of North-African Leishmania infantum isolates from cured and relapsed visceral leishmaniasis patients reveals variations correlating with geography and phenotype

Giovanni Bussotti, Alia Benkahla, Fakhri Jeddi, Oussama Souiaï, Karim Aoun, Gerald Späth, Aïda Bouratbine

HAL Id: pasteur-03109058
https://pasteur.hal.science/pasteur-03109058

Submitted on 13 Jan 2021

Nuclear and mitochondrial genome sequencing of North–African *Leishmania infantum* isolates from cured and relapsed visceral leishmaniasis patients reveals variations correlating with geography and phenotype

Giovanni Bussotti[1,2], Alia Benkahla[3], Fakhri Jeddi[4,5], Oussama Souiaï[3], Karim Aoun[6], Gerald F. Späth[2] and Aïda Bouratbine[6,*]

Abstract

Although several studies have investigated genetic diversity of *Leishmania infantum* in North Africa, genome-wide analyses are lacking. Here, we conducted comparative analyses of nuclear and mitochondrial genomes of seven *L. infantum* isolates from Tunisia with the aim to gain insight into factors that drive genomic and phenotypic adaptation. Isolates were from cured (*n*=4) and recurrent (*n*=3) visceral leishmaniasis (VL) cases, originating from northern (*n*=2) and central (*n*=5) Tunisia, where respectively stable and emerging VL foci are observed. All isolates from relapsed patients were from Kairouan governorate (Centre); one showing resistance to the anti-leishmanial drug Meglumine antimoniate. Nuclear genome diversity of the isolates was analysed by comparison to the *L. infantum* JPCM5 reference genome. Kinetoplast maxi and minicircle sequences (1 and 59, respectively) were extracted from unmapped reads and identified by BLAST analysis against public data sets. The genome variation analysis grouped together isolates from the same geographical origins. Strains from the North were very different from the reference showing more than 34 587 specific single nucleotide variants, with one isolate representing a full genetic hybrid as judged by variant frequency. Composition of minicircle classes within isolates corroborated this geographical population structure. Read depth analysis revealed several significant gene copy number variations correlating with either geographical origin (amastin and Hsp33 genes) or relapse (CLN3 gene). However, no specific gene copy number variation was found in the drug-resistant isolate. In contrast, resistance was associated with a specific minicircle pattern suggesting *Leishmania* mitochondrial DNA as a potential novel source for biomarker discovery.

DATA SUMMARY

(1) The whole-genome-sequencing data sets were uploaded to the National Center for Biotechnology Information (NCBI) under the SRA accessions SRX7737288–SRX7737294, BioProject accession PRJNA607007 and BioSample accessions numbers SAMN14122793–SAMN14122799.

(2) Assembled maxi- and minicircle sequences are available at URL: http://tesla.pasteur.tn/Leishmania/

(3) Assembled maxi- and minicircle sequences are also available from the NCBI GenBank database (http://www.ncbi.nlm.nih.gov/genbank/sequenceids/) under accession numbers MT762281–MT762287 and MT598258–MT598591, respectively.

INTRODUCTION

Visceral leishmaniasis (VL) caused by *Leishmania infantum* is endemic in all countries of the Mediterranean basin [1]. Parasites are transmitted to humans by the bite of *Phlebotominae* sandflies, and canids serve as reservoir host [2, 3]. The disease is mainly diagnosed in young children and immunocompromised adults. It results from uncontrolled multiplication of the parasites in phagocytes of the reticuloendothelial system and is lethal if left untreated [4]. In Tunisia, North Africa VL remains primarily a paediatric disease that occurs in children less than 5 years of age. About 50 to 100 cases are reported each year mainly in the Northern and Central parts of the country where historical and emerging VL foci, respectively, are found [5–7]. Until 2018, VL treatment in Tunisia was based on pentavalent antimonials offered in the form of meglumine antimoniate (Glucantime) with no alternative drug options available. Using antimonials, the cure rate in immunocompetent children exceeds 95 % and cases of relapse are mainly re-treated with additional course(s) of the same drug [4]. Relapses are second symptomatic VL episodes within 6 to 12 months after initial apparent patient cure [8]. They can occur in immunocompetent patients harbouring Glucantime-sensitive strains in the case of sub-optimal antimonial treatment or high parasitic load [9, 10]. It is important to note that cases of relapse seem to be less responsive to additional treatment, which suggests that *Leishmania* strains may become resistant to antimonials [11]. This phenomenon, which has also been described in dogs infected with *L. infantum* [12] has been demonstrated by *in vitro* studies with *L. infantum* strains isolated from VL patients after several courses of treatment [9].

Molecular approaches are being used increasingly for epidemiological studies of VL. Several molecular markers have been developed to address key epidemiological and population genetic questions. Among them, multilocus microsatellite typing (MLMT) is able to discriminate below the zymodeme level and has been extensively used in population genetic studies [13]. MLMT revealed hierarchic population structure in *L. infantum* according to geography and the genetic diversity within this *Leishmania* species has been well documented in Northern and Southern Mediterranean countries as well as in America [14–17]. Although whole-genome-based genetic polymorphism studies gained importance in recent years [18–20], the use of whole-genome sequencing in investigating genetic polymorphism within *L. infantum* species are scarce [21] and genome-wide analyses of North African isolates are lacking. Furthermore, only few studies investigated the genetic markers associated with antimony resistance in *L. infantum* relapsing isolates [22]. In particular, there is no data on the genome structure of *Leishmania* parasites isolated from relapsed patients.

Impact Statement

Leishmania infantum is a protozoan parasite that causes life-threatening human and veterinary diseases in the Mediterranean basin and South America. A hallmark of this parasite species is its capacity to adapt to a variety of ecological environments leading to new emerging foci of visceral leishmaniasis and unpredictable phenotypic fluctuations inside its human host, notably after drug interventions. Our study generated insight into factors that may drive genomic and phenotypic adaptation of *L. infantum* through the comparative analysis of nuclear and mitochondrial genomes of Tunisian isolates collected in different geographical areas from cured and relapsed visceral leishmaniasis patients. It provides the first evidence that parasites have adapted through a combination of nuclear and kinetoplast genomic changes that could be used as epidemiological biomarkers. It also highlights the challenges in biomarker discovery using field isolates, with clinically relevant loci likely being masked by genetic heterogeneity and hybridization events as a result of local adaptation. On the other hand, neither gene nor chromosome copy number variations correlated with drug resistance. In contrast, resistance (*n*=1) was associated with a specific minicircle pattern suggesting *Leishmania* mitochondrial DNA as a potential novel source for biomarker discovery. Our study establishes and validates the computational framework for future environmental genomic analyses using large sample collections for the discovery of genetic loci associated with *Leishmania* geographic or phenotypic adaptation.

The use of next-generation-sequencing (NGS) technology transformed *Leishmania* genome studies. It allowed the description of genetic diversity at karyotype, gene copy number and single nucleotide levels within *Leishmania* strains and provided valuable insights into the complexity of the genome and gene regulation [20, 23–27]. NGS was also applied successfully in drug-resistance studies in *Leishmania* identifying changes in chromosome copy number, alterations in gene dosage and single nucleotide polymorphisms that correlated with resistance in *Leishmania* strains derived from the laboratory and from the field [18, 28–35]. However, despite increasing information on nuclear genome sequences, the mitochondrial genome sequence of these parasites is often ignored in NGS analyses. To our knowledge, only one report exploited the raw data generated by NGS to study *Leishmania* kinetoplast maxicircle and minicircle DNA sequences [36]. Here, we conducted comparative analyses of the nuclear and mitochondrial genomes of *L. infantum* isolates from both cured and relapsed immunocompetent VL patients originating from Tunisia. Our aims were to gain the first insights into the genetic diversity of *L. infantum* in Tunisia across various geographic areas, into genomic differences associated with parasite persistence under drug pressure and

Table 1. *L. infantum* isolates from VL cases

Isolate	Geographical origin (governorate)	Circumstances of sampling	Previous Glucantime treatment	Clinical outcome after Glucantime treatment	Zymodem*	EC50 (µg ml⁻¹)†
ZK28	North (Zaghouan)	First VL episode	No	Definitive cure‡	MON-1	30
LIPA83	North (Nabeul)	First VL episode	No	Definitive cure	MON-80	50
ZK5	Centre (Kairouan)	First VL episode	No	Definitive cure	MON-1	50
ZK25	Centre (Kairouan)	First VL episode	No	Definitive cure	MON-1	30
LIPA60	Centre (Kairouan)	Relapse§	Yes	Improvement	MON-1	45
ZK43	Centre (Kairouan)	Relapse	Yes	Improvement	MON-1	50
ZK47	Centre (Kairouan)	Relapse	Yes	Non response to treatment	MON-1	120

*The zymodem corresponds to the iso-enzyme typing code (Rioux *et al.* 1990) [37].

†The 50% effective concentration (EC50) corresponds to the concentration of Glucantime that reduced the intracellular survival of *Leishmania* parasites by half and was calculated after a semi-logarithmic transformation to obtain a linear dose-response relationship.

‡Definitive cure: no recurrence of VL symptoms and absence of relapse by the 12-month follow-up visit.

§Relapse: recurrence of clinical symptoms and visualization of parasites in bone marrow aspirate any time between end of treatment up to the 12-month follow-up visit.

potentialmitochondrial DNA modifications associated with treatment failure. Our analysis proposes a new computational framework for future environmental genomics analyses using large sample collections for the discovery of genetic loci associated with geographic or phenotypic adaptation.

METHODS

Leishmania isolates, species identification and sensitivity to antimonials

We included in this study seven *L. infantum* isolates collected from confirmed VL cases and available in the LR 16-IPT-06 *Leishmania* collection at Institut Pasteur of Tunis. The clinical strains were isolated on NNN medium from peripheral blood of immunocompetent children in the setting of VL diagnosis; four before treatment from patients with definitive cure and three during relapse from patients who had completed the full-course of Glucantime. Two isolates were from the historical VL focus of the North-East (Zaghouan and Nabeul governorates, respectively) and five from the emerging VL focus of central Tunisia (Kairouan governorate) (Table 1). All *Leishmania* isolates were typed using multilocus enzyme electrophoresis as described by Rioux *et al.* [37]. *In vitro* sensitivity to antimonials were assessed as described previously [22] by exposing infected THP1 cells (ATCC TIB-202) to various concentrations of Glucantime (0, 7.5, 15, 30, 45 and 60 µg ml⁻¹). Only one isolate was highly resistant under *in vitro* conditions. This latter was isolated from a patient who failed to respond to a second Glucantime treatment (Table 1).

DNA extraction and NGS

Cryopreserved isolates were stabilized and cultured at 23 °C in RPMI 1640 medium supplemented with 2 mM L-Glutamine, 15 % foetal bovine serum and antibiotics (500 IU ml⁻¹ penicillin, 500 µg ml⁻¹ streptomycin, 500 µg ml⁻¹ kanamycin). DNA extraction was performed using DNeasy Blood and Tissue kit (Qiagen, Germany) according to manufacturer instructions. Between 2 to 5 µg of DNA were used for sequencing. Whole genome, short-insert, paired-end libraries were prepared for each sample (Table S1, available in the online version of this article). Samples ZK28, LIPA83, ZK5, ZK25, LIPA60 and ZK43 were sequenced by the Biomics sequencing platform (https://research.pasteur.fr/en/team/biomics/) with HiSeq 2500 rapid runs, resulting in 2×108 bp reads using the NEXTflex PCR-Free kit. Sample ZK47 was sequenced with the KAPA Hyper Prep Kit (Kapa Biosystems) at Centro Nacional de Análisis Genómico (CNAG, http://www.cnag.crg.eu/) using the TruSeq SBS Kit v3-HS (Illumina). Multiplex sequencing was performed according to standard Illumina procedures, using a HiSeq 2000 flowcell v3 generating 2×101 bp paired-end reads.

Read alignment to *L. infantum* reference genome

All reads were aligned to the *L. infantum* JPCM5 ASM287v2 reference genome (GenBank assembly accession: GCA_000002875.2) [38, 39]. Alignments were carried out using BWA mem (version 0.7.12) [40] with the flag -M to mark shorter split hits as secondary. Samtools fixmate, sort and index (version 1.3) [41] were used to process the
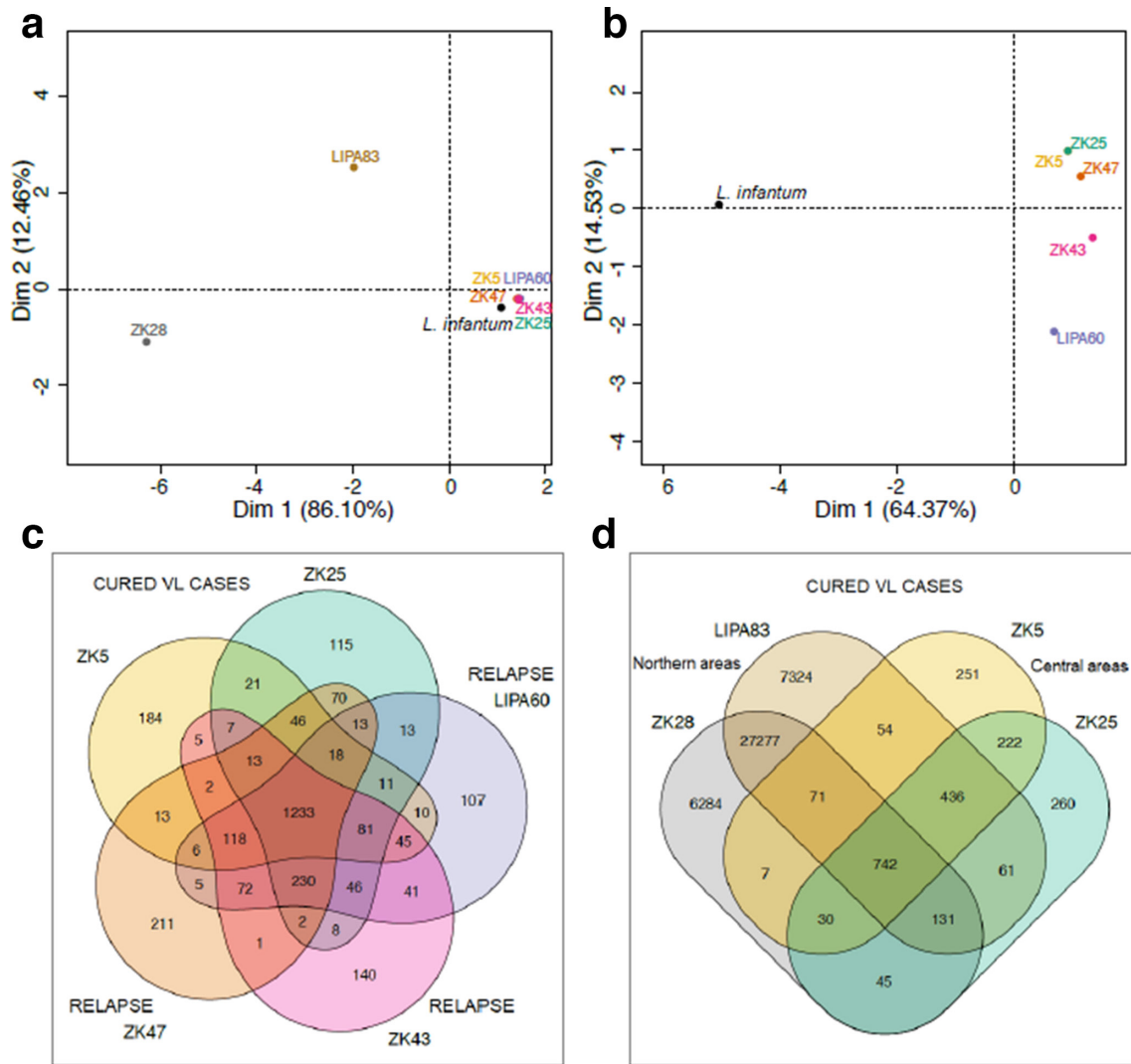
Fig. 1. Genetic similarities between Tunisian *L. infantum* isolates from VL cases. Principal-component analyses (PCAs) of average nucleotide identity to *L. infantum* reference genome included in (a) isolates from Northern (ZK28, LIPA83) and Central (ZK5, ZK25, LIPA60, ZK43, ZK47) Tunisia and in (b) isolates from cured (ZK5 and ZK25) and relapsed (LIPA60, ZK43, ZK47) VL cases. Venn diagrams show the number of unique and shared SNVs with respect to the reference genomes among cured and relapsed VL isolates from Central Tunisia in (c) and cured VL isolates from Central (ZK5, ZK25) and Northern Tunisia (ZK28, LIPA83) in (d).

alignment files and turn them into bam format. RealignerTargetCreator and IndelRealigner from the GATK suite [42] were run to homogenize indels. Eventually, PCR and optical duplicates were labelled with Picard MarkDuplicates [version 1.94 (1484)] (https://broadinstitute.github.io/picard/) using the option 'VALIDATION_STRINGENCY_LENIENT'. Picard CollectAlignmentSummaryMetrics was used to estimate sequencing and mapping statistics (Table S1).

Comparative genome analysis

In a first lenient filtering step, all mapped reads were processed with Trimmomatic (version 0.35) [43] to remove low-quality bases (options LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15) and adapter contaminations (ILLUMINACLIP option, with values 2:30:12:1:true). Reads

that were shorter than 36 bases after filtering were discarded (option MINLEN:36). The trimmed reads were assembled with SPAdes (version 3.7.0) [44] with option 'careful'. The resulting contigs were used to estimate the average nucleotide identity (ANI) with the dnadiff part of the MUMme*r* system (version 3.23) [45]. The analysis included the reference genomes of *L. donovani* LdBPKv2 [26], *L. infantum* (see above) and *L. major* strain Friedlin (GenBank assembly accession: GCA_000002725.2 ASM272v2) together with the reference genomes of *L. braziliensis, L. mexicana* and *L. panamensis* available from ENSEMBL Protists release 29. The ANI values were converted to a matrix of distances, which in turn were used for principal component analysis (PCA) and hierarchical clustering (R hclust function, https://www.R-project.org/).
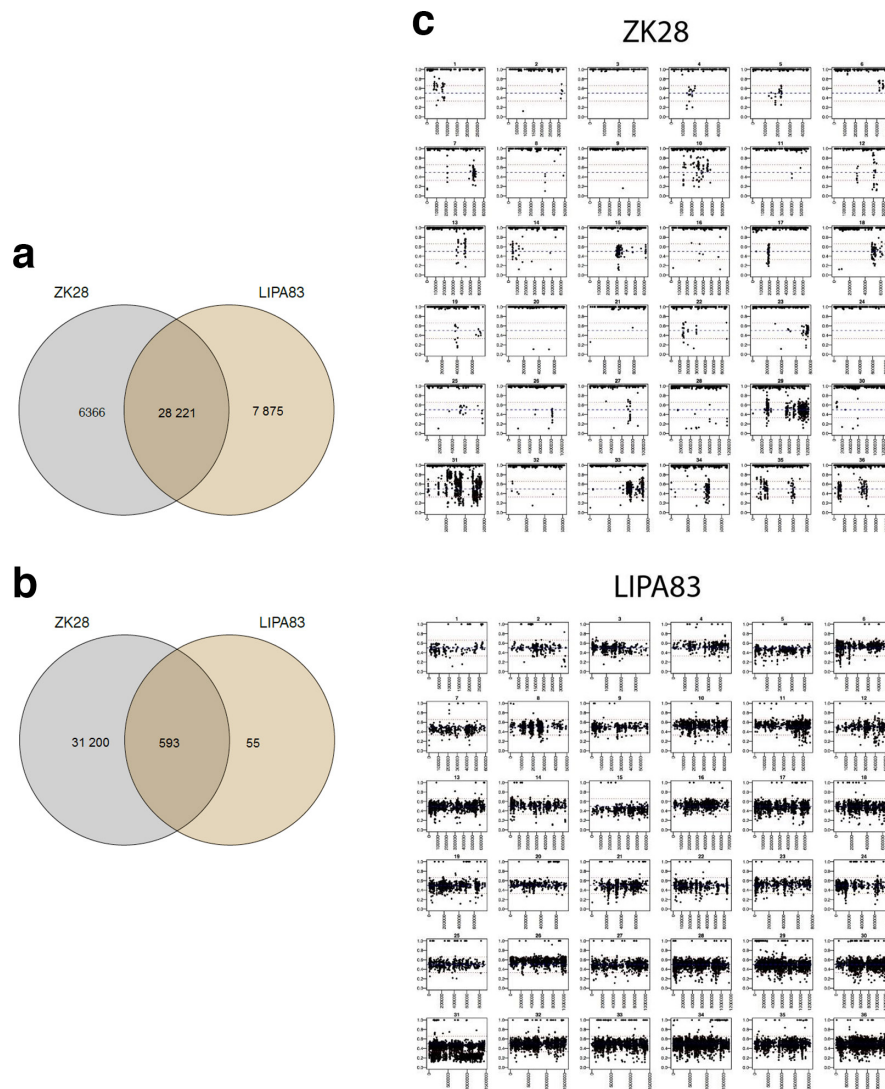
Fig. 2. SNVs in *L. infantum* isolates from Northern Tunisia. The two Venn diagrams on the left show the number of unique and shared SNVs among the two Northern isolates ZK28 and LIPA83. All SNVs are reported in (a) and only high frequency SNVs with VRF≥0.9 are reported in (b). (c) The position (x-axis) and the VRF (y-axis) of the predicted SNVs in the two isolates. Individual panels represent different chromosomes.

Single nucleotide variant (SNV) analysis

To call SNVs we used Freebayes (version v1.0.1–2-g0cb2697) with options '--no-indels --read-indel-limit 0 --no-mnps --no-complex --read-mismatch-limit 3 --read-snp-limit 3 --hwe-priors-off --binomial-obs-priors-off --allele-balance-priors-off --min-alternate-fraction 0.05 --min-base-quality 5 --min-mapping-quality 50 --min-alternate-count 2 --pooled-continuous'. The output was filtered to retain the positions with just one alternative allele with a minimum frequency of 0.1, and a minimum mean mapping quality of 20 for the reads supporting the reference or the alternative allele. SNVs mapping inside homopolymers (i.e. simple repeats of the same nucleotide) were filtered using a more stringent parameter, requiring at least 20 reads supporting the variant. The homopolymers were defined as the DNA region spanning +/-5 bases from the SNV, with over 40% of identical nucleotides. We discarded SNVs with sequencing coverage above or below four median absolute deviations (MADs) from the median chromosome coverage. For each isolate, SNVs and their corresponding read frequency (VRF) are reported in Table S2. To predict the effect of SNVs on genes we used SnpEff [46]. We detected a total of 43705 SNVs, with 26417 (60%) located in intergenic regions. The remaining SNVs (17288, or 40%) were distributed on 4571 genes, with a mean number of SNVs per gene of 3.8. Genic SNVs accounted overall for 7812 synonymous substitutions and 9476 non-synonymous substitutions. SNVs predicted by snpEff to have any of the following effects were counted as non-synonimous substitutions: 'NON_SYNONYMOUS_CODING', 'NON_SYNONYMOUS_START', 'START_LOST',
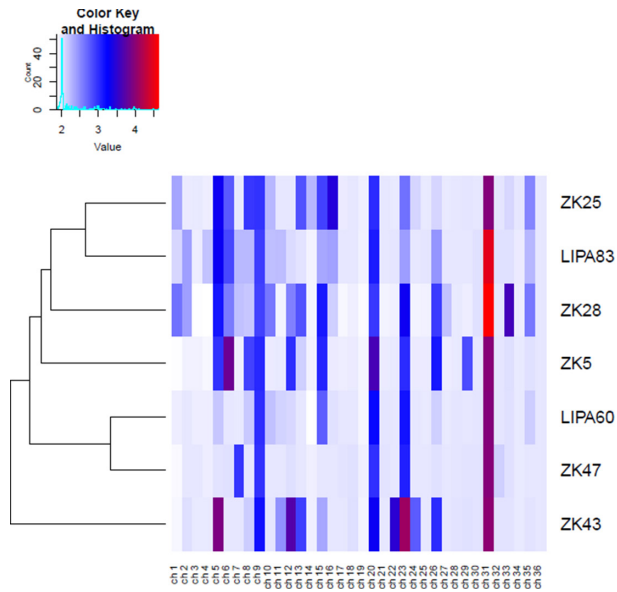
Fig. 3. Aneuploidy among isolates from VL cases. The heat map shows the estimated chromosome somy for each chromosome and each isolate with cluster analysis. Chromosome numbers and isolates are listed across the bottom and on the right side of the heat map, respectively. The dendrogram on the left side shows clusters of the isolates based on similarity of aneuploidy profiles and was generated according to Euclidean distances by R statistics (R packages; stats, gplots). Colour key indicates ploidy of chromosomes and the distribution frequency.

'STOP_GAINED', 'STOP_LOST'. Vice versa, SNVs with 'SYNONYMOUS_CODING' or 'SYNONYMOUS_STOP' snpEff predicted effects were counted as synonymous substitutions. The d*N*/d*S* ratio scores for individual isolates were computed and reported in Table S2.

Chromosome copy number variation

For each read alignment file, Samtools view (version 1.3) [41] and BEDTools genomecov (version 2.25.0) [47] were used to measure the sequencing depth of each nucleotide. Samtools was run with options '-q 50 F 1028' to discard reads with low map quality score or potential duplicates, while BEDTools genomecov was run with options '-d -split'. In order to compare chromosome copy numbers, we first normalized nucleotide coverage by the median genomic coverage. This normalization was performed to avoid bias related to library size and make sample sequencing coverage comparable. Following this normalization and under the assumption that most of the genome is disomic, normalized values close to one likely reflect disomic chromosomes. To reduce noise and complexity, we moved from single nucleotide to genomic windows and estimated the median coverage of each chromosome by computing the median sequencing coverage for contiguous windows of 2500 bases. The window median coverage was further normalized by the median coverage of chromosome 36 (showing a stable disomic level across samples) and multiplied by 2. This measure was evaluated to

estimate the chromosome copy number, under the assumption that the most frequent somy level is disomy

Gene copy number variation

The gene function was extracted from TriTrypDB (https://tritrypdb.org/tritrypdb/) [48] on 25 March 2019. Genes missing functional definitions were annotated using the GAF file available from the Wellcome Sanger institute FTP server and eventually the description available from NCBI Gene (https://www.ncbi.nlm.nih.gov/gene) [49]. Samtools view (version 1.3) and BEDTools coverage (version 2.25.0) [47] were used to measure the mean sequencing depth of every annotated gene. Possible intragenic gap regions were excluded from the calculation of the mean. To account for GC content sequencing bias, the coverage values were corrected using a LOESS regression with a fivefold cross validation to optimize the model 'span' parameter. Genes supported by reads with a mean mapping quality (MAPQ) score <50 were filtered. To estimate gene copy number, mean coverage of each gene was normalized by the median coverage of its chromosome (Table S3). For each isolate and each chromosome every outlier gene copy number value larger than the upper quartile +3 inter quartile ranges (IQRs) was considered as significantly increased. Outlier detection was performed with MedCalc Statistical software (version 13.0.6).

Processing of non-aligned reads

The unaligned reads were extracted from the BAM files using Samtools view with option '-f 4'. All unaligned reads were then processed with Trimmomatic (version 0.36) [43] with parameterization (options 'LEADING:3 TRAILING:3 SLID-INGWINDOW:4:20') to remove low-quality bases. Reads that were unpaired or shorter than 78 bases after filtering were discarded (option: MINLEN:78). The remaining reads were processed with SPAdes (version 3.13.0) [44] for *de novo* assembly. Scaffolds larger than 500 nt were used for maxicircle and minicircle BLAST analysis.

Maxicircle and minicircle identification

All scaffolds larger than 500 nt were searched with NCBI-BLASTn (using the standalone tool BLASTall 2.2.26) on 7 July 2019 against (i) the kinoplastid maxicircle sequences available in the Nucleotide database at GenBank (https://www.ncbi.nlm.nih.gov/genbank/) [50]: KJ803830.1 – *Trypanosoma rangeli*; DQ343646.1 – *Trypanosoma cruzi* strain Esmeraldo; DQ343645.1 – *Trypanosoma cruzi* strain CL Brener; M10126.1 – *Leishmania tarentolae*; KM386508.1 – *Trypanosoma vivax* strain MT1; MG948557.1 – *Trypanosoma copemani* isolate G1,and (ii) the assembled maxicircle sequences of *L. infantum*, *Leishmania major* and *Leishmania braziliensis* deposited by Camacho *et al.* [36] in http://leish-esp.cbm.uam.es/l_infantum_downloads.html. Scaffolds not matching any maxicircle were scanned with NCBI-BLASTn (BLASTall 2.2.26) against (i) kinetoplastid minicircle sequences downloaded from the Nucleotide database at GenBank (the list is provided in Supplementary Material) and (ii) assembled minicircle sequences of *L. infantum*, *L. major*
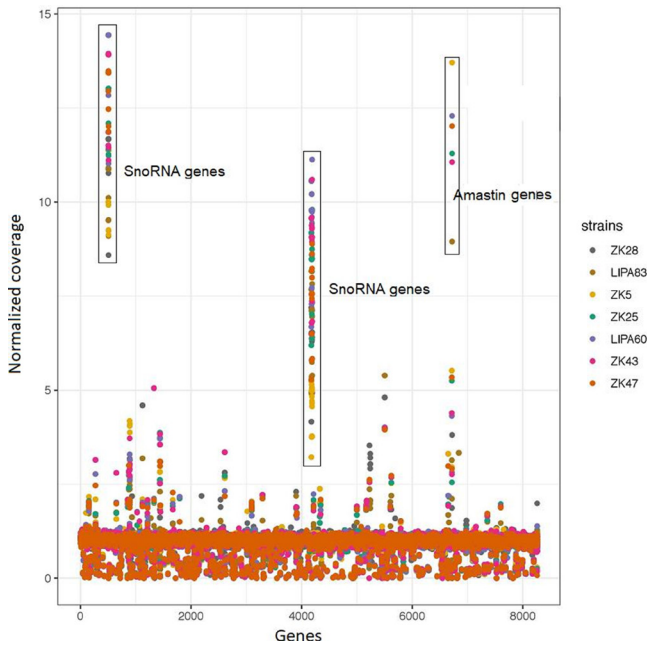
Fig. 4. Gene-amplification profiles among isolates from VL cases. The figure shows gene copy number variation (see methods) for the seven isolates. Genes of the seven isolates are plotted on the same figure, each isolate being identified by a specific colour. SnoRNA genes on chromosome 5 and 26 and amastin genes on chromosome 34 show the highest gene copy numbers. This increase concerns almost all isolates as shown by the different dot colours.

and *L. braziliensis* deposited by Camacho *et al.* [36] in http://leish-esp.cbm.uam.es/l_infantum_downloads.html.

Minicircle class copy number variation

All unmapped paired reads were aligned to SPAdes assembled maxi/minicircle scaffolds following the same protocol as the one used to map reads to the nuclear genome. Mean nucleotide coverages were extracted per maxi/minicircle using Samtools and BEDtools (with options '-q 50 F 1028' and '-d -split'). For each isolate each maxicircle/minicircle coverage was normalized by the median genomic coverage. To have an estimate copy number of each minicircle class within isolates, the mean coverage of each minicircle was further normalized by the mean coverage of the isolate's maxicircle.

RESULTS AND DISCUSSION
Comparative genome and SNV analyses

All isolates belonged to the *L. infantum* reference cluster (Fig. S1), which confirmed prior species identification by isoenzyme typing. Based on PCA analysis, ZK5, ZK25, LIPA60, ZK43 and ZK47 from the same emerging VL focus of Central Tunisia (Kairouan governorate) grouped together with the *L. infantum* reference strain, whereas ZK28 and LIPA83 from, respectively, the Zaghouan and Nabeul governorates (historical VL focus of the North-East) were distant from the former group and from each other (Fig. 1a). Of the

seven, the most distant isolate was ZK28 (Fig. 1a). To assess genetic similarity between the five isolates from the same geographical location but presenting differences according to Glucantime exposition and drug sensitivity, we excluded ZK28 and LIPA83 and reanalysed the data by PCA. Isolates from Kairouan governorate clustered closer to each other than to the reference strain. However, there was no specific clustering in relation with relapse or Glucantime resistance (Fig. 1b).

SNV analysis corroborated the PCA results. SNVs calling in Kairouan isolates (ZK5, ZK25, LIPA60, ZK43 and ZK47) recorded only 2887 SNVs in comparison to the reference genome, with less than 2053 SNVs observed for a given isolate (Fig. 1d). Furthermore, 1233 SNVs (42.7%) were shared between relapse and non-relapse isolates and only 72 (2.5%) were shared between relapse strains (Fig. 1c). On the other hand, a much greater number of SNVs was identified in LIPA83 and ZK28 in comparison to the reference genome (36096 and 34587 specific SNVs, respectively). Moreover, the two isolates shared 28221 SNVs (79.8%) suggesting their strong genetic relationship (Fig. 1d).

These NGS results support previous epidemiological information obtained by MLMT in Tunisia [16]. In fact, the very diverse isolate ZK28 likely belongs to the North African MON-1 population reported in southern and eastern Mediterranean regions, which was not fully sequenced until now [13, 16]. This population is highly endemic in Northern Tunisia and is genetically different from the European MON-1 population found in South-West Europe and South America, which include the strain JPCM5 used as a reference in this study [13]. In contrast, as reported with the South American *L. infantum* isolates [21], the five isolates from Kairouan Governorate closely matched the reference and likely belong to the European *L. infantum* MON1 population that has already been isolated in Northern Tunisia, especially from dogs [16]. These latter probably introduced this population to the Kairouan area, an arid zone bordering the Northern VL foci and where intensive agriculture projects have taken place since the 1980s, leading to the establishment of a stable *L. infantum* transmission cycle [7]. It is important to note in this context that the European *L. infantum* MON1 population has been introduced in the Americas in the paths of human migration from endemic areas of Spain and Portugal, which highlights its high capability for adaptation to new environments [51, 52].

Analysis of variant read frequency (VRF) of the predicted SNVs showed that most SNVs in ZK28 (31793 out of 34587, 91.9%) were homozygous variants with high read frequency (VRF≥0.9), whereas in LIPA83 most SNVs (35448 out of 36096, 98%) were largely heterozygous variants with VRF<0.9 (Fig. 2a, b). Moreover, most VRF of predicted SNVs recorded in LIPA83 plotted in the 0.25–0.75 range and were observed on entire chromosomes and distributed genome-wide, a pattern that suggests outcrossing by full-genome hybridization (Fig. 2c). Thus, the LIPA83 isolate typed as MON80 was a full genetic hybrid as judged by

Table 2. Genes differently amplified according to isolates

Gene	Annotation	Gene copy no.						
		ZK28	LIPA83	ZK5	ZK25	LIPA60	ZK43	ZK47
LinJ.06.1360	CLN3 protein - putative	1,7	1,8	1,6	1,7	2,0	2,8*	2,0
LinJ.09.0360	hypothetical protein	2,2*	1,6*	1,0	1,0	1,2	1,0	1,1
LinJ.10.0521	hypothetical protein	4,6*	3,2*	2,1	1,9	2,0	2,0	1,9
LinJ.11.1120	hypothetical protein	2,0	1,4	2,0	2,0	2,1	5,1*	2,1
LinJ.30.1520	p1/s1 nuclease	3,5*	2,4	2,6	2,6	2,6	2,5	2,6
LinJ.30.1630	ferric reductase	3,0*	2,0	1,0	1,0	1,0	1,0	1,0
LinJ.30.1640	hypothetical protein	2,9*	2,0	1,0	1,0	0,9	1,0	1,0
LinJ.30.1650	hypothetical protein	3,3*	2,1	1,2	1,0	1,1	1,1	1,0
LinJ.30.1660	Protein (DUF1077)	3,2*	1,9	0,9	1,0	0,9	0,8	1,0
LinJ.30.1670	Hsp33 protein putative	3,0*	2,1	1,0	1,1	0,9	1,0	1,1
LinJ.34.1020	Amastin	0,0**	1,7	3,3	2,0	2,2	1,9	3,0
LinJ.34.1150	Amastin	2,2*	1,5	1,0	1,1	1,1	1,0	1,1
LinJ.34.1680	Amastin	0,0†	3,1	5,5	5,3	4,3	4,4	5,3
LinJ.34.1730	Amastin	3,8†	9,0	13,7	11,3	12,3	11,1	12,0

*Superior to (average +2 sd).
†Inferior to (average −2 sd).

the SNV frequency. This finding corroborates previous MLMT results that showed evidence for the existence of hybrids and gene flow between genetically different *L. infantum* populations in Tunisia and identified *L. infantum* MON80 as a hybrid [16].

Overall, when considering each individual sample, we identified 4.141 genes showing traces of selection pressure with a d*N*/d*S* ratio score >1 or <1 (Table S2). More than 95 % of these genes were identified within ZK28 and LIPA83 isolates which exhibited fairly similar gene d*N* and d*S* substitutions, with more than 80 % of shared genes presenting identical d*N*/d*S* ratio scores (Table S2). These results corroborate previous ones and showed that these two isolates from the same geographical area (Northern Tunisia) probably evolved under a similar environmental selection pressure. On the other hand, among the 738 genes measuring a d*N*/d*S* ratio score >1 or <1 and recorded within at least one Kairouan isolate (ZK5, ZK25, LIPA60, ZK43 and ZK47), 101 (66 with d*N*/d*S* >1 and 35 with d*N*/d*S* <1) were specifically recorded in isolates from relapse with 26 (25.7%) presenting the same pattern of d*N*/d*S* variability in more than one isolate (Table S2). Studying these encoding genes and related proteins may help understanding biological adaptations of *L. infantum* in relation to drug pressure.

Chromosome copy number variation

Comparison of chromosome somies between isolates is shown in Fig. 3. As reported by others, chromosome 31 displayed at least tetrasomy in all isolates [18]. Chromosomes 5, 23, 9 and 20 were more than disomic in almost all strains whereas chromosomes 3, 4, 14, 17, 18, 19, 21, 25, 27, 28, 30, 32, 34 and 36 were uniformly disomic. Somies of the other chromosomes were diverse and showed a strain-specific distribution. Hierarchical clustering of isolates according to their chromosome somy patterns grouped together the two relapse isolates ZK47 and LIPA60. They showed the lowest degree of mosaic aneuploidy among strains. However, despite the low aneuploidy, ZK47 displayed a strain-specific polysomy for chromosome 7. The third relapse isolate ZK43 was apart. It showed a unique pattern of somy variations with specific polysomy for several chromosomes. The lower tendency of relapsed strains to show karyotypic variation in culture could indicate reduced karyotypic diversity in these isolates as they went through the bottle neck of treatment or *in vivo* selection [53]. However, the presence of specific chromosome aneuploidy highlights the high genome plasticity and adaptability of the *Leishmania* parasite to *in vitro* culture [26, 27].

Gene copy number variation

Overall, we identified 97 genes (including 15 gene clusters) that showed copy number increase with respect to the reference genome in at least one of the isolates (Table S3). Among these, two snoRNAs gene clusters on chromosomes 5 and 26 and an amastin gene on chromosome 34 showed the highest gene copy numbers (Fig. 4). These fluctuations observed for genes implicated in RNA modification (snoRNAs) and
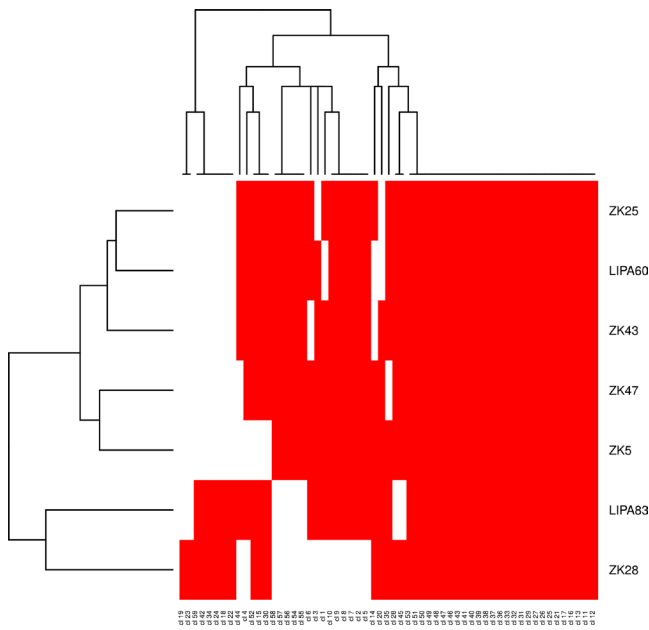
Fig. 5. Minicircle class networks of isolates from VL cases. The binary heat map shows the minicircle classes present in each isolate with cluster analysis. Minicircle class numbers and isolates are listed at the bottom and on the right side of the heat map, respectively. The presence of a given minicircle class is indicated in red. The dendrogram on the left side shows clusters of the isolates based on similarity of minicircle class composition and was generated according to Euclidean distances by R statistics (R packages; stats, gplots).

Fig. 6. Coverage of minicircle classes in isolates from VL cases. The graph generated by MedCalc statistical sotware (version 13.0.6) shows for each isolate the normalized coverage of all its minicircle classes. Normalization was done according to the mean coverage of the maxicircle in the same isolate. Means are indicted by the horizontal lines and error bars (2 SD) are shown.

virulence may indicate their possible role in driving fitness gain in the field. However, these high gene copy numbers concerned almost all isolates suggesting that either all isolates amplified this region with respect to the reference genome or that the latter one is mis-annotated and underestimates the gene copy number [54]. Alternatively, the reference strain may have been selected for lower gene copies during long-term culture.

Several genes showed significant differences in copy number between isolates Table 2(Table 2). Most of them namely (i) a gene array on chromosome 30 coding for the stress-activated chaperone Hsp33 protein (among others), (ii) the secretory nuclease (p1/s1) gene LinJ.30.1520 and 4 amastin genes on chromosome 34 showed significant differences in read depth according to geography, suggesting that these genes/gene arrays may be hot spots of environment–genotype interactions that drive geographical adaptation in a strain-specific manner [55]. Other significant variations between isolates concerned ZK43. Two genes showed significant copy number increase in this relapse isolate suggesting strain-specific adaptation. Among these, the CLN3 gene was incriminated in infectivity of *Leishmania* parasites. Its orthologue in *L. mexicana* (LmxM.22.0010) encodes a homologue of BTN1, a protein involved in Batten disease in humans, also known as CLN3 in other organisms. Although the function of this protein remains elusive, it is apparent that it may have a
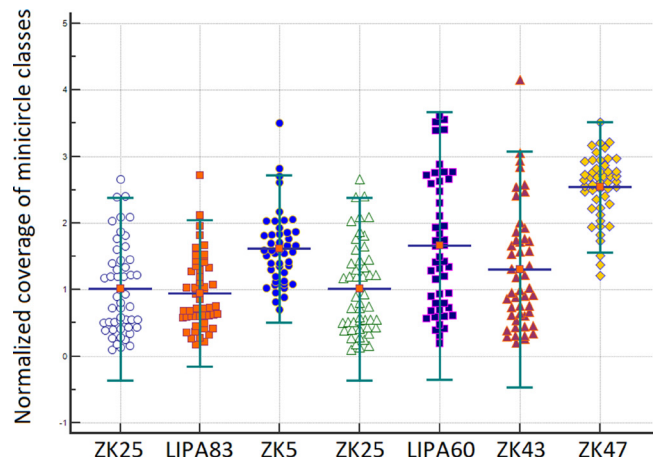
direct effect on lysosomal function [56]. Significantly, we did not identify any copy number variation in genes previously linked to drug resistance, suggesting that treatment failure may be governed by more complex genetic interactions or post-transcriptional adaptation.

Mitochondrial genome diversity

Seven large scaffolds (one per isolate, average size 17653±125 nucleotides and median coverage 46.9) fully matched the complete coding region of the maxicircle of *L. infantum*. Multiple sequence alignment showed remarkable similarity between the seven isolates and highlighted the high level of intra-specific conservation of the kinetoplast maxicircle (Fig. S2). In contrast, NCBI-BLASTn analysis identified a highly divergent set of 59 distinct minicircle classes (Table S4) with each isolate harbouring a specific minicircle network of 43 to 50 different classes. Isolates from the same geographical origin showed the highest number of similar minicircle classes as shown by cluster analysis that clearly distinguished ZK28 and LIPA83 from the five Kairouan isolates (Fig. 5). These results corroborate recent high-throughput sequencing of single minicircle PCR products that revealed high minicircle class diversity in *Leishmania* strains and the presence of patterns that fit with the current hypothesis of the phylogenetic relationship between *Leishmania* species [57].

Minicircle class copy number variation

Within each isolate minicircle network, some classes exhibited evident copy number variation (Table S5, Fig. 6) and might be preferentially amplified for enhanced expression of required guide RNA (gRNA) involved in RNA editing of maxicircle-encoded transcripts [58]. The resistant strain ZK47 however, exhibited a particular minicircle pattern, with more minicircle classes affected by copy number variation as assessed by a

significant increase of average minicircle classes' coverage (Table S5, Fig. 6). Increased copy number of certain minor classes within the minicircle network was reported previously during stepwise drug-resistance selection *in vitro* [59]. It may be explained by selection of particular minicircles required for RNA editing of maxicircle-encoded transcripts that are involved in mitochondrial proteomic changes observed in antimony-resistant strains [60].

CONCLUSION

This study reveals the effects of geography on the *Leishmania* nuclear and mitochondrial genomes and highlights the challenges in biomarker discovery using field isolates, with clinically relevant loci likely being masked by genetic heterogeneity and hybridization events as a result of geographic adaptation. Our results further highlight limits and contributions to use *Leishmania* promastigotes from *in vitro* culture in *Leishmania* studies given the changes in genomic features linked to chromosomal and gene copy number observed during *in vitro* culture adaptation [23, 25, 26]. However, the level of genome instability observed during culture adaptation may inform on the genetic heterogeneity of the isolates. Indeed, the relapse strains showed lower karyotypic diversity, which may be explained by drug and immune selection that may have led to loss of heterozygosity. Finally, our results indicate that several gene/gene array copy number variations may be either under macro- or microenvironment-specific selection and that drug resistance in the field is probably multifactorial and needs to be explored at the DNA level investigating both the nuclear and mitochondrial genomes.

References

1. Desjeux P. Leishmaniasis: current situation and new perspectives. *Comp Immunol Microbiol Infect Dis* 2004;27:305–318.

2. Gramiccia M, Gradoni L. The current status of zoonotic leishmaniases and approaches to disease control. *Int J Parasitol* 2005;35:1169–1180.

3. Killick-Kendrick R. Phlebotomine vectors of the leishmaniases: a review. *Med Vet Entomol* 1990;4:1–24.

4. Gradoni L, Soteriadou K, Louzir H, Dakkak A, Toz SO *et al*. Drug regimens for visceral leishmaniasis in Mediterranean countries. *Trop Med Int Heal* 2008;13:1272–1276.

5. Aoun K, Jeddi F, Amri F, Ghrab J, Bouratbine A. Current epidemiological data on visceral leishmaniasis in Tunisia. *Médecine Mal Infect* 2009;39:775–779.

6. Ben-Ahmed K, Aoun K, Jeddi F, Ghrab J, El-Aroui MA *et al*. Visceral leishmaniasis in Tunisia: spatial distribution and association with climatic factors. *J Trop Med Hyg* 2009;81:40–45.

7. Ben Salah AB, Ben Ismail R, Amri F, Chlif S, Ben Rzig F *et al*. Investigation of the spread of human visceral leishmaniasis in central Tunisia. *Trans R Soc Trop Med Hyg* 2000;94:382–386.

8. Ponte-Sucre A, Gamarro F, Dujardin JC, Barrett MP, López-Vélez R *et al*. Drug resistance and treatment failure in leishmaniasis: a 21st century challenge. *PLoS Negl Trop Dis* 2017;11:e0006052–24.

9. Faraut-Gambarelli F, Piarroux R, Deniau M, Giusiano B, Marty P *et al*. *In vitro* and *in vivo* resistance of *Leishmania infantum* to meglumine antimoniate: a study of 37 strains collected from patients with visceral leishmaniasis. *Antimicrob Agents Chemother* 1997;41:827–830.

10. Aoun K, Alaya NB, Mary C, Amri F, Raies A *et al*. Contribution of quantitative real-time polymerase chain reaction to follow-up of visceral leishmaniasis patients treated with meglumine antimoniate. *Am J Trop Med Hyg* 2009;81:1004–1006.

11. Bryceson ADM, Chulay JD, Ho M, Mugambii M, Were JB *et al*. Visceral leishmaniasis unresponsive to antimonial drugs I. Clinical and immunological studies. *Trans R Soc Trop Med Hyg* 1985;79:700–704.

12. Gramiccia M, Gradoni L, Orsini S. Decreased sensitivity to meglumine antimoniate (Glucantime) of *Leishmania infantum* isolated from dogs after several courses of drug treatment. *Ann Trop Med Parasitol* 1992;86:613–620.

13. Schönian G, Kuhls K, Mauricio IL. Molecular approaches for a better understanding of the epidemiology and population genetics of *Leishmania*. *Parasitology* 2011;138:405–425.

14. Pomares C, Marty P, Bañuls AL, Lemichez E, Pratlong F *et al*. Genetic diversity and population structure of *Leishmania infantum* from Southeastern France: evaluation using multi-locus microsatellite typing. *PLoS Negl Trop Dis* 2016;10:e0004303–0004328.

15. Seridi N, Amro A, Kuhls K, Belkaid M, Zidane C *et al*. Genetic polymorphism of Algerian *Leishmania infantum* strains revealed by multilocus microsatellite analysis. *Microbes Infect* 2008;10:1309–1315.

16. Chargui N, Amro A, Haouas N, Schönian G, Babba H *et al*. Population structure of Tunisian *Leishmania infantum* and evidence for the existence of hybrids and gene flow between genetically different populations. *Int J Parasitol* 2009;39:801–811.

17. Kuhls K, Alam MZ, Cupolillo E, Ferreira GEM, Mauricio IL *et al*. Comparative microsatellite typing of new world *Leishmania infantum* reveals low heterogeneity among populations and its recent old world origin. *PLoS Negl Trop Dis* 201;5:1–16.

18. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH *et al*. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res* 2011;21:2143–2156.

19. Imamura H, Downing T, Van den Broeck F, Sanders MJ, Rijal S *et al*. Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *eLife* 2016;5:1–39.

20. Iantorno SA, Durrant C, Khan A, Sanders MJ, Beverley SM *et al*. Gene Expression in *Leishmania* Is Regulated Predominantly by Gene Dosage. *MBio* 2017;8:1–20.

21. Teixeira DG, Monteiro GRG, Martins DRA, Fernandes MZ, Macedo-Silva V *et al*. Comparative analyses of whole genome sequences of *Leishmania infantum* isolates from humans and dogs in north-eastern Brazil. *Int J Parasitol* 2017;47:655–665.

22. Jeddi F, Mary C, Aoun K, Harrat Z, Bouratbine A *et al*. Heterogeneity of molecular resistance patterns in antimony-resistant field isolates of *Leishmania* species from the western mediterranean area. *Antimicrob Agents Chemother* 2014;58:4866–4874.

23. Prieto Barja P, Pescher P, Bussotti G, Dumetz F, Imamura H *et al*. Haplotype selection as an adaptive mechanism in the protozoan pathogen *Leishmania donovani*. *Nat Ecol Evol* 2017;1:1961–1969.

24. Zackay A, Cotton JA, Sanders M, Hailu A, Nasereddin A *et al*. Genome wide comparison of Ethiopian *Leishmania donovani* strains reveals differences potentially related to parasite survival. *PLoS Genet* 2018;14:e1007133.

25. Bussotti G, Gouzelou E, Côrtes Boité M, Kherachi I, Harrat Z *et al*. *Leishmania* genome dynamics during environmental adaptation reveal strain-specific differences in gene copy number variation, karyotype instability, and telomeric amplification. *mBio* 2018;9.

26. Dumetz F, Imamura H, Sanders M, Seblova V, Myskova J *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* 2017;8:1–14.

27. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA *et al*. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. *Genome Res* 2011;21:2129–2142.

28. Leprohon P, Fernandez-Prada C, Gazanion Élodie, Monte-Neto R, Ouellette M. Drug resistance analysis by next generation sequencing in *Leishmania*. *Int J Parasitol* 2015;5:26–35.

29. Leprohon P, Légaré D, Raymond F, Madore E, Hardiman G *et al*. Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. *Nucleic Acids Res* 2009;37:1387–1399.

30. Rastrojo A, García-Hernández R, Vargas P, Camacho E, Corvo L *et al*. Genomic and transcriptomic alterations in *Leishmania donovani* lines experimentally resistant to antileishmanial drugs. *Int J Parasitol* 2018;8:246–264.

31. Carnielli JBT, Crouch K, Forrester S, Silva VC, Carvalho SFG *et al*. A *Leishmania infantum* genetic marker associated with miltefosine treatment failure for visceral leishmaniasis. *EBioMedicine* 2018;36:83–91.

32. Laffitte M-CN, Leprohon P, Papadopoulou B, Ouellette M. Plasticity of the *Leishmania* genome leading to gene copy number variations and drug resistance. *F1000Research* 2016;5:2350.

33. Yasur-Landau D, Jaffe CL, David L, Doron-Faigenboim A, Baneth G. Resistance of *Leishmania infantum* to allopurinol is associated with chromosome and gene copy number variations including decrease in the S-adenosylmethionine synthetase (METK) gene copy number. *Int J Parasitol* 2018;8:403–410.

34. Monte-Neto R, Laffitte MCN, Leprohon P, Reis P, Frézard F *et al*. Intrachromosomal amplification, locus deletion and point mutation in the aquaglyceroporin AQP1 gene in antimony resistant *Leishmania (Viannia) guyanensis*. *PLoS Negl Trop Dis* 2015;9:e0003476.

35. Dumetz F, Cuypers B, Imamura H, Zander D, D'Haenens E *et al*. Molecular Preadaptation to Antimony Resistance in *Leishmania donovani* on the Indian Subcontinent. *mSphere* 2018;3:e00548-17–17.

36. Camacho E, Rastrojo A, Sanchiz África, González-de la Fuente S, Aguado B *et al*. *Leishmania* mitochondrial genomes: Maxicircle structure and heterogeneity of minicircles. *Genes* 2019;10:758.

37. Rioux JA, Lanotte G, Serres E, Pratlong F, Bastien P *et al*. Taxonomy of *Leishmania*. Use of isoenzymes. Suggestions for a new classification. *Annales de Parasitologie Humaine et Comparée* 1990;65:111–125.

38. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC *et al*. Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nat Genet* 2007;39:839–886.

39. Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde G *et al*. GeneDB-an annotation database for pathogens. *Nucleic Acids Res* 2012;40:D98–D108.

40. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.

42. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K *et al*. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.

43. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.

44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.

45. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004;5:12.

46. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T *et al*. A program for annotating and predicting the effects of single nucleotide-polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin)2012;6:80–92.

47. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–842.

48. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP *et al*. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* 2010;38:D457–D462.

49. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res* 2004;33:D54–D58.

50. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. Genbank. *Nucleic Acids Res* 2016;44:D67–D72.

51. Ferreira GEM, dos Santos BN, Dorval MEC, Ramos TPB, Porrozzi R *et al*. The genetic structure of *Leishmania infantum* populations in Brazil and its possible association with the transmission cycle of visceral leishmaniasis. *PLoS One* 2012;7:e36242.

52. Boité MC, Späth GF, Bussotti G, Porrozzi R, Morgado FN *et al*. Trans-Atlantic spillover: deconstructing the ecological adaptation of *Leishmania infantum* in the Americas 2019;11.

53. Sterkers Y, Lachaud L, Crobu L, Bastien P, Pagès M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. *Cell Microbiol* 2011;13:274–283.

54. Lypaczewski P, Hoshizaki J, Zhang WW, McCall LI, Torcivia-Rodriguez J *et al*. A complete *Leishmania donovani* reference genome identifies novel genetic variations associated with virulence. *Sci Rep* 2018;8:1–14.

55. Drini S, Criscuolo A, Lechat P, Imamura H, Skalický T *et al*. Species- and strain-specific adaptation of the HSP70 super family in pathogenic trypanosomatids. *Genome Biol Evol* 2016;8:1980–1995.

56. Ishemgulova A, Hlaváčová J, Majerová K, Butenko A, Lukeš J *et al*. CRISPR/Cas9 in *Leishmania mexicana*: A case study of LmxBTN1. *PLoS One* 2018;13:e0192723–17.

57. Kocher A, Valière S, Bañuls AL, Murienne J. High-throughput sequencing of kDNA amplicons for the analysis of *Leishmania* minicircles and identification of Neotropical species. *Parasitology* 2018;145:585–594.

58. Yurchenko VYY, Merzlyak EMM, Kolesnikov AAA, Martinkina LPP, Vengerov YYY *et al*. Structure of *Leishmania* minicircle kinetoplast DNA classes. *Journal of Clinical Microbiology* 1999;37:1656–1657.

59. Lee ST, Liu HY, Lee SP, Tarn C. Selection for arsenite resistance causes reversible changes in minicircle composition and kinetoplast organization in *Leishmania mexicana. Mol Cell Biol* 1994;14:587–596.

60. Vincent IM, Racine G, Légaré D, Ouellette M. Mitochondrial proteomics of antimony and miltefosine resistant *Leishmania infantum. Proteomes* 2015;3:328–346.