Article

Gene Annotation and Transcriptome Delineation on a *de novo* Genome Assembly for the Reference *Leishmania major* Friedlin Strain

Esther Camacho+, Sandra González-de la Fuente+, Jose C. Solana, Alberto Rastrojo, Fernando Carrasco-Ramiro, Jose M. Requena* and Begoña Aguado*

Centro de Biología Molecular "Severo Ochoa" (CBMSO, CSIC-UAM) Campus de Excelencia Internacional (CEI) UAM+CSIC, Universidad Autónoma de Madrid, 28049 Madrid, Spain

+ Equal contributors

* Correspondence: jmrequena@cbm.csic.es (J.M.R.), baguado@cbm.csic.es (B.A.)

Abstract: *Leishmania major* is the main causative agent of cutaneous leishmaniasis in humans. The Friedlin strain of this species (LmjF) was chosen when a multi-laboratory consortium undertook the objective of deciphering the first genome sequence for a parasite of the genus *Leishmania*. The objective was successfully attained in 2005, and this represented a milestone for *Leishmania* molecular biology studies around the world. Although the LmjF genome sequence was done following a shotgun strategy and using classical Sanger sequencing, the results were excellent and this genome assembly served as the reference for subsequent genome assemblies in other Leishmania species. Here, we present a new assembly for the genome of this strain (named LMJFC for clarity), generated by the combination of two high throughput sequencing platforms, Illumina short-read sequencing and PacBio Single Molecular Real-Time (SMRT) sequencing, which provides long-read sequences. Apart from resolving uncertain nucleotide positions, several genomic regions have been reorganized and a more precise composition of tandemly repeated gene loci was attained. Additionally, the genome annotation has been improved by adding 542 genes and more accurate coding-sequences defined for around two hundred genes, based on the transcriptome delimitation also carried out in this work. As a result, we are providing gene models (including untranslated regions and introns) for 11,238 genes. Genomic information ultimately determines the biology of every organism; therefore, our understanding of molecular mechanisms will depend on the availability of precise genome sequences and accurate gene annotations. In this regards, this work is providing an improved genome sequence and updated transcriptome annotations for the reference L. major Friedlin strain.

Keywords: genome; transcriptome; gene models; Leishmania; Illumina sequencing; PacBio sequencing; expression levels; untranslated regions (UTRs); SL-additions sites; polyadenylation sites

1. Introduction

Leishmaniasis is a group of neglected tropical diseases caused by parasitic protists of the genus *Leishmania*. This parasite has a digenetic life cycle, alternating between the alimentary tract of the sandfly vector, as an extracellular promastigote, and the phagolysosomal vacuole of macrophages, in which the parasite adopts the amastigote form. Transmissions to humans occur in nearly 100 countries, and around one million new cases of leishmaniasis are reported per year [1]. Unfortunately, there is no effective vaccine for prevention of human leishmaniasis [2], and the current treatments are based on chemotherapy, which relies on four drugs having problems of toxicity, cost, growing drug resistance and/or treatment failure [3].

Given the global relevance of leishmaniasis, in 1994, the WHO Leishmania Genome Initiative was launched, bringing together a large number of laboratories determined to get sequenced the full genome of a pathogenic *Leishmania* species [4]. *Leishmania major* was the selected one and the genome sequence was determined on a chromosome-by-chromosome basis. Firstly, a genome physical map was constructed from 9,216 genomic cosmids by DNA hybridizations using probes derived from the ends of contigs and chromosome specific probes [5]. Meanwhile, contigs were fragmented and sequenced by the classical Sanger's sequencing technique. After sequence assembling, the accuracy of sequence assemblies was assessed by comparison to optical maps for the 36 chromosomes of *L. major* genome [6]. Finally, in 2005, the complete genome sequence and gene annotations were reported [7]. This work represented a milestone that provided important insights about the gene content and genome architecture of this parasite and paved the way for genome-wide studies [8]. Soon after, the genome sequences for two other *Leishmania* species, *Leishmania infantum* and *Leishmania braziliensis*, were produced by whole-genome shotgun cloning and classical Sanger's sequencing [9], even though these assemblies did not achieve the completeness of that attained for the *L. major* genomic assembly.

The development of next generation sequencing (NGS) technologies has transformed the field of genomics, and genome sequencing became an affordable and indispensable technique for molecular biology studies. Hence, a continuously growing number of genomes are being sequenced and, particularly, within the genus *Leishmania*, most of the named species have their genome sequenced [10-18]. Remarkably, due to its high quality, the 2005-genome sequence of *L. major* Friedlin has remained as the reference genome, and has been used for template-guided assembly of those new sequenced genomes. However, despite its relevance, the *L. major* (Friedlin) genome assembly cannot be considered as a final product 'set on stone'. In fact, previous studies have documented some deficiencies in this assembly [19, 20]. Moreover, two features of the *Leishmania* genomes represent hurdles for a correct assembly based on sequencing short DNA fragments. On the one hand, the existence of a large number of repetitive DNA sequences, which are scattered along the different *Leishmania* chromosomes [21-23], is cause of conflict for assemblers. On the other hand, many loci in *Leishmania* genomes are comprised of multiple identical gene copies that are head-to-tail tandemly arranged [24]; in this case, assembly collapses lead to underestimation on the real number of gene copies. The third-generation/long-read sequencing methods have solved most of these issues, contributing to produce genome assemblies of unprecedented quality [25]. However, a drawback of these new sequencing methods is the high sequence error rate, around 15% [25]. Thus, a combination of high-accurate short-reads and less-accurate long-reads has allowed to produce new and improved genomes assemblies for several *Leishmania* species [26-31].

The generation of high quality genome assemblies is a basic step in the process of studying molecular mechanisms of gene expression, but additional information other than nucleotide sequences needs to be generated. There are dedicated bioinformatics tools, like Companion [32] that automatically and efficiently performs predictions of open reading frames (ORFs). However, protein-coding genes contain sequences other than ORFs, i.e. they also contain 5'- and 3'-untranslated regions (5'- and 3'-UTRs). Although some bioinformatics algorithms have been developed to delineate UTRs in *Leishmania* genes [33], the absence of conserved sequence motifs in the Leishmania gene boundaries precludes an accurate prediction of gene models, which only can be generated after obtaining the complete sequence of their transcripts. To date, experimental transcriptomes have been reported for *L. major* [34], *L. mexicana* [35] and *L. donovani* [30]. Genome wide gene expression studies require of precise gene models, being especially relevant in *Leishmania*, where a significant number of genes share identical ORFs but differ substantially in their UTRs [36-42].

In this work, the *L. major* (Friedlin) genome was re-sequenced using the Pacific Biosciences (PacBio) technology, which provides long reads reads able to span long repeats, and the Illumina technology to generate paired-end short-reads useful to join fragmented chromosomes, extend chromosomes ends and correct homopolymer indel errors. As a result, here is reported the complete and improved sequence of the 36 chromosomes comprising the *L. major* genome. Additionally, based on this improved genome, a re-annotation of the *L. major* transcriptome is provided. This is a valuable information aimed to guide future studies on gene expression in this parasite.

2. Materials and Methods

2.1. Leishmania parasites and DNA isolation

Promastigotes of *L. major* (Friedlin strain) were grown at 26 °C in M199 medium supplemented with foetal bovine serum (10%), HEPES (40 mM; pH 7.4), adenine (0.1 mM), hemin (10 µg/ml), biotin (1 µg/ml) biopterin (2 ng/ml), penicillin G (100 U/ml) and streptomycin sulphate (0.1 mg/mL). This strain was provided by Dr. Javier Moreno, Instituto de Salud Carlos III (Madrid, Spain), a WHO Collaborating Centre for Leishmaniasis.

DNA for Illumina sequencing was prepared from 2×10^8 promastigotes using the "High Pure PCR Template Preparation Kit" (Roche), following manufacturer's instructions. DNA for PacBio sequencing was prepared from a similar number of promastigotes but following a classical phenol extraction method [43].

2.2. Illumina sequencing

Library construction and paired-end sequencing were performed at the Centro Nacional de Análisis Genómico (CNAG-CRG, Spain; http://www.cnag.crg.eu/) using Illumina HiSeq 2000 technology. A total of 52,845,525 paired-end, 2×126 nucleotides (nt) sequence reads were generated. A median insert size of 305-bp was estimated. The reads were analysed using PrinseqQuality (http://prinseq.sourceforge.net/) and poor-quality reads (cut-off value, 20) were removed; additionally, only those reads having a length ≥ 60-nt were considered. Filtered reads were assembled using the CLC Genomics Workbench version 5.0 (CLC Bio).

2.3. PacBio sequencing and de novo assembly

The single-molecule real-time (SMRT) sequencing technology developed by PacBio [44] was used for generating long sequencing reads. A total of 285,082 pre-filtered reads were obtained on a PacBio RS II sequencing instrument. The Norwegian Sequencing Centre (www.sequencing.uio.no) provided the sequencing service.

A hierarchical genome-assembly process (HGAP) [45], using the HGAP3 (Pacific Biosciences, SMRT Analysis Software v2.3.0) and HGAP4 (Pacific Biosciences, SMRT Link 4.0.0) protocols, was carried out. Three strategies in the *de novo* genome assemblies were followed: i) and ii) HGAP3 and expected genome sizes of 34 and 35 Megabases (Mb), respectively, and iii) HGAP4 and an expected genome size of 35 Mb. Equivalent assemblies were obtained in all three strategies. PacBio contigs having low coverage (<40x) or short length (<15-Kb) were considered spurious and discarded.

2.4. Assembly refinements

Assembled contigs were compared by BLAST [46] against the reference *L. major* Friedlin genome sequence (Tritryp, v.46). Thirty-one of the PacBio contigs represented complete chromosomes. The other five chromosomes were assembled in two PacBio contigs each. Minimus2 software [47], which is based on NUCmer algorithm, was used to compute overlaps between contigs in order to join these contigs into a sole chromosome.

Additionally, Illumina contigs were aligned against these PacBio assembled chromosomes using LAST aligner (http://last.cbrc.jp/). These analyses served to extend the ends of chromosomes. For this purpose, three tools were used: MAFFT multiple-aligner [48], BLAST and SSPACE-standard [49]. Finally, taking into account the distance information of paired-reads, GapCloser (https://sourceforge.net/projects/soapdenovo2/files/GapCloser/) and Gapfiller [50] were used to determine the appropriate size and sequence of the chromosomal extensions.

On the final assembly, a further sequencing revision was done by using ARAMIS [51], a recent tool developed to correct sequences derived from PacBio genome assemblies. For this purpose, Illumina reads were used and an indel fraction of 0.8 was selected.

In order to check assembling defects in genomic regions with unexpected frameshifts, we used SAMTools [52] to extract those reads mapping into a defined region. Afterwards, these reads were re-assembled using the Canu assembler, a tool specifically designed for noisy single-molecule sequences [53].

2.5. Coverage and alignment

Coverage analyses on either the newly assembled chromosomes or the reference genome (LmjF) were performed using both Illumina and PacBio reads. Firstly, Illumina reads were aligned by Bowtie2 [54], and PacBio bax.h5 reads were aligned with BLASR [55]. Afterwards, coverage analysis was done from each alignment along the 36 chromosomes using the GenomeCoverageBed tool (http://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html). The graphical coverage plots files were generated using GNUPLOT (http://www.gnuplot.info/).

2.6. Somy analysis

Somy estimation was performed using the 2-loop method, as described elsewhere [56]. Somy graphs were generated from the median coverage values for each chromosome using the barplot function of the R package (https://cran.r-project.org).

2.7. Synteny Analysis

Synteny was evaluated via progressive algorithm MAUVE [57] and genoPlotR [58] using as reference the *L. major* Friedlin genome in which the seven new loci identified by Alonso et al [20] were included. Gepard tool (https://academic.oup.com/bioinformatics/article/23/8/1026/198110) was used to create graphical plots for visualization of changes in synteny.

2.8. Haplotype detection

Pre-processing of Illumina alignment file was carried out with Picard tools (http://broadinstitute.github.io/picard/) to reduce bias introduced by PCR amplification. GATK HaplotypeCaller (version 4.1) [59] was chosen to detect variants. The resulting VCF file was used to reconstruct individual haplotypes across the whole *L. major* Friedlin genome assembly by HapCUT2 [60], a maximum-likelihood-based tool designed for assembling haplotypes from DNA sequence reads. IGV [61] and Jalview (https://www.jalview.org/) were used to visualize the variants and haplotype-blocks detected after the analyses.

2.9. Annotation of protein-coding sequences, known non-coding RNAs and structural RNAs

The *L. major* Friedlin genome, assembled in this work, was annotated using Companion web server (https://companion.sanger.ac.uk/) with default settings. The *L. major* Friedlin strain genome (LmjF) was used as a reference template. OrthoMCL [62] and BLAST software were used to further improve the gene annotations. Finally, the annotations were combined and used to create a GFF3 file using an in-house script in Python.

The automatic ID codes generated by Companion were maintained. The code structure was LMJFC_XXYYYYYYYY, in which the label LMJFC is common to all annotated elements, XX stands for the chromosome number and the set of Y corresponds to a serial number, starting from 5000 at the beginning of the chromosome and increasing by 100 units for the ID of the downstream-annotated element. For structural RNAs, the nomenclature for IDs was modified to indicate the RNA type (rRNA, tRNA and snoRNA), intercalated between the chromosome number and the serial number (LMJFC_XX.rRNA.YYYY).

2.10. Transcriptome definition and annotation

Poly-A+ RNA from *L. major* promastigotes was used for library construction and Illumina sequencing (HiSeq 2000 technology); details about RNA-seq data have been described previously [63]. A total of 88,315,069 (2 × 76-nt) stranded RNA-seq reads were used for transcriptome definition. Transcripts were generated from RNA-seq reads following the pipeline developed by Rastrojo and co-workers [34]. Briefly, quality-filtered RNA-seq reads were mapped to the *de novo* genome assembly generated in this work (LMJFC genome) using Bowtie2 aligner (parameters: : --np 0 --n-ceil L,0,0.02 --rdg 0,6 --rfg 0,6 --mp 6,2 y --score-min L,0,-0.24). Then, assembly of transcripts was performed using Cufflinks with default parameters [64]. Additionally, among those unaligned reads, a search for the presence of eight or more nucleotides identical to the 3′-end of the SL sequence (AACTAACGCT ATATAAGTAT CAGTTTCTGT ACTTTATTG) was performed. After trimming the SL-containing reads, the remaining sequences were mapped back to the LMJFC genome to define the SL-addition sites (SASs) and, therefore, the transcript start. Using a similar strategy, i.e. searching for unaligned reads containing a poly-A stretch (>5 nt in length) at their 3′-end, the poly-A addition sites (PASs) were defined. To increase the identification of PASs, we further used the huge amount of Illumina RNA-seq reads generated by Dillon and co-workers [65] from *L. major* (Friedlin) promastigote RNA samples. In this manner, most transcripts could be precisely delimited after mapping the SASs and PASs. Finally, annotated CDSs (see above) were associated with the corresponding transcripts, which were named according to the CDS code, but intercalating a 'T' between the chromosome number and the CDS serial number (LMJFC_XXTYYYYYYYY). For those transcripts lacking of an associated CDS, the intermediate serial number between the neighboring CDS-containing transcripts was used for naming them.

2.11. Generation and annotation of gene models

The CDS and transcripts coordinates were merged in order to create gene-models. For simplicity, genes maintained the corresponding transcript names, but excluding the 'T' from the transcript ID. A final manual revision of the annotations was carried by parallel IGV visualizations of CDSs, transcripts and RNA-seq reads distribution on the final LMJFC genome assembly.

2.12. Data availability

Genomic and transcriptomic raw reads have been deposited in the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/). Besides, the assembled genome and transcriptome sequences together with annotations files were uploaded under the Study accession number PRJEB25921. Additionally, the genome (fasta file), the annotations for the genome, transcriptome and gene models are downloadable at the Leish-ESP website (http://leish-esp.cbm.uam.es/).

3. Results and discussion

3.1. Re-sequencing and de novo assembly of the L. major (Friedlin strain) genome

As described above, the *L. major* genome was the first to be sequenced among species of the genus *Leishmania* [7]. This fact, together with the robustness of the assembly attained, justifies that this genome became a reference in the field of trypanosomatids. Nevertheless, after its publication, a few studies documented the existence of some inaccuracies in this reference genome assembly, mostly associated with the abundance of both tandemly reiterated genes and retroposon-derived repeated sequences in the *Leishmania* genomes [19-20]. Hence, given the relevance of this genome in the field, this work was aimed to *de novo* re-assembly of the genome for this strain, exploiting the advances in sequencing methodologies. Recently, we have succeeded in re-assembling the genomes for two other *Leishmania* reference species, *L. infantum* [29] and *L. braziliensis* [27], and the *de novo* assembly of the *L. donovani* (HU3 strain) genome [30] by using the PacBio single-molecule real-time (SMRT) sequencing technology [44].

A total of 285,082 high-quality reads with an average length of 16-kb were generated by PacBio sequencing, representing an estimated 140-fold coverage based on the 32.8-Mb size for the *L. major* genome [7]. As detailed in the Materials and Methods section, HGAP3 and HGAP4 assemblers were used to construct contigs from the PacBio reads. After filtering out contigs with low coverage and short length, a total of 41 contigs were further analysed: 31 of them represented complete chromosomes, whereas chromosomes 8, 19, 22, 27 and 35 appeared assembled in two contigs each. However, the pairs of contigs were easily joined by Minimus2 software [47]. Figure 1 shows the reads coverage along the five chromosomes that resulted from the joining of two contigs; the coverage of both PacBio and Illumina reads was continuous along the chromosomes, indicating that the assembly of these five chromosomes was correct. Illumina sequence reads (370-fold coverage) were also generated from the same *L. major* DNA sample and used for refinements of the final nucleotide sequence and to extend chromosomal ends. In particular, 27 out of the 36 chromosomes were extended using contigs assembled from the Illumina reads; as shown in figure 1, PacBio reads coverage drastically decreased at the chromosomal ends, suggesting structural constraints of the telomeres that affect PacBio library preparation. Additionally, the higher accuracy of Illumina reads served to correct 1,964 indel errors (1,894 insertions and 70 deletions) in the sequence assembled from PacBio reads by using ARAMIS [51].

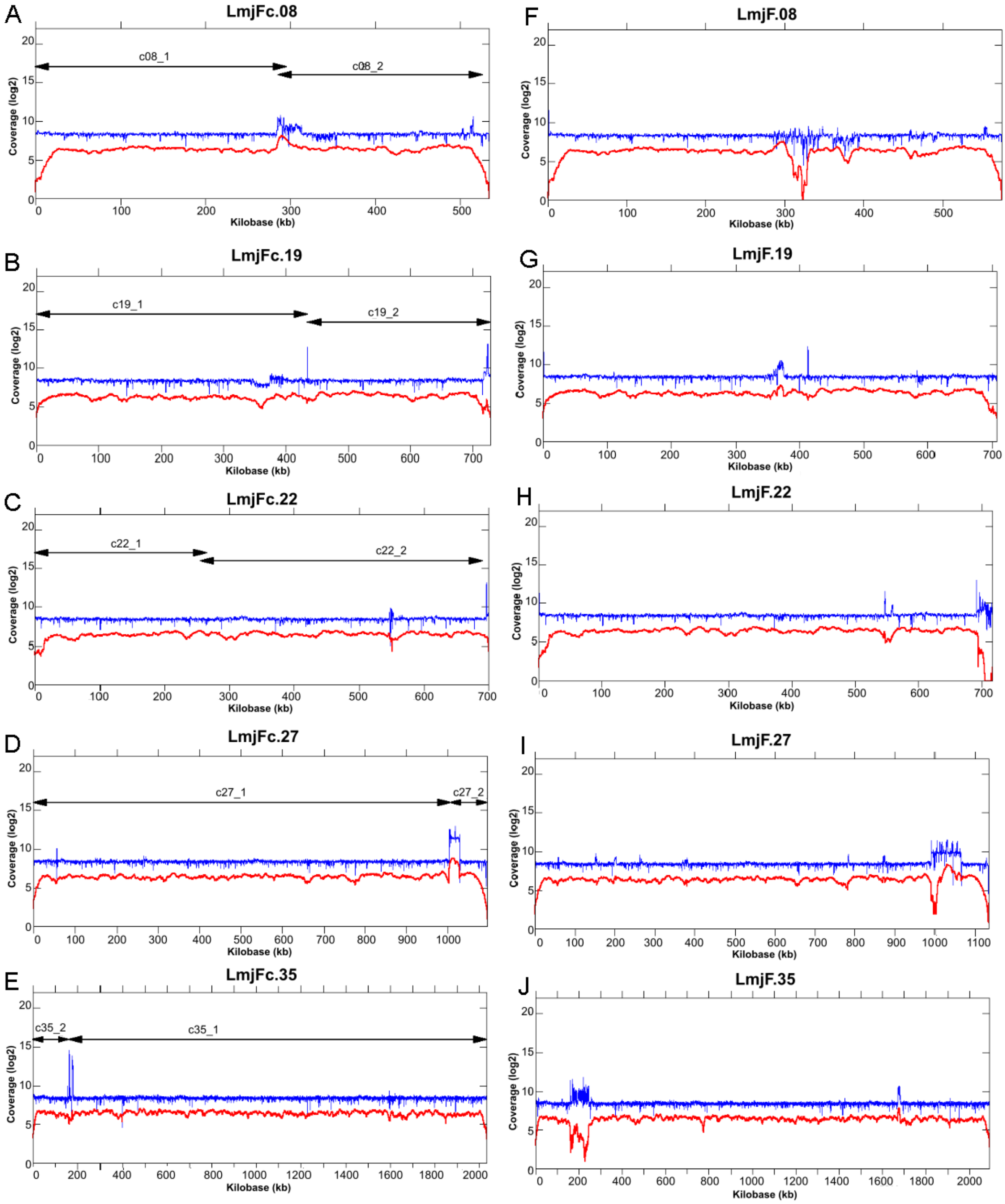

Figure 1. Read coverage along five chromosomes that resulted from joining two contigs. Illumina reads (in blue) and PacBio reads (in red) were mapped to the *de novo* assembled chromosomes (LmjFc) 8, 19, 22, 27 and 25, and to the corresponding chromosomes from the reference *L. major* genome (LmjF). The lines with arrow-heads denote the position of the two contigs joined to form the final chromosomes.

Apart from chromosome 22, in which no obvious reason for halting the assembly advance into a sole contig was found, the presence of a long repeated region would be the cause of blocking the assembly, preventing the other four chromosomes to be assembled into a single contig each. In particular, the two contigs forming chromosome 27 were stopped at the rDNA locus, composed of repetition units of about 20-kb [66]. Moreover, the sudden increase of reads coverage in the rDNA locus, observed after mapping of either PacBio or Illumina reads (Fig. 1D), would indicate that a sequence collapse remains yet in the final assembled chromosome 27. According to the read coverage on the assembled rDNA region regarding the median value along the entire chromosome, it was calculated that 15-16 rDNA units must exist in the locus. However, the assembled genome attained in this work (hereinafter named LMJFC) contains only two units, whereas 6 rDNA units are found in the current reference *L. major* genome assembly (LmjF). In a classical study on the *L. major* (Friedlin strain) rDNA locus, Martínez-Calvillo and co-workers estimated in 10-14 the number of rDNA units per chromosome [66]. The collapse assembly of the rDNA locus in the assembled LMJFC genome is expected, taking into account that the size of an rDNA unit is close to the mean size of the PacBio reads. Therefore, larger sequence reads would be needed to accurately determine the real number of rDNA units existing in the Leishmania genomes.

Another feature in figure 1 that caught our attention was the sudden decrease in the coverage of PacBio reads on the LmjF.08 chromosome assembly (panel F), suggesting the existence of a clear discrepancy in relation to the new assembly (LMJFC; panel A). In order to further analyze this finding, the genomic regions from both assemblies were analyzed at a per gene level (Fig. 2). When PacBio sequence reads were aligned against both assemblies, a lack of coverage was observed around coordinate 323-kb of chromosome LmjF.08, indicating that this region in the LmjF genome would be miss-assembled. In contrast, a continuous and smooth distribution of the PacBio reads was observed when they were mapped against the LMJFC assembly (Fig. 2, bottom panel). Two differences were found between both genome assemblies. Firstly, the LmjF assembly contains 11 copies of an amastin-like protein coding gene, whereas only 5 genes were assembled in the LMJFC genome. In fact, this region is composed by a repeated unit consisting in two alternating genes (amastin-like- and hypothetical protein-coding genes). The genes coding for this hypothetical protein, although existing in the LmjF.08 chromosome sequence, were not annotated previously. The other difference (minor) is that the tandem consisting of five genes coding for another amastin-like protein in the LmjF genome (IDs: LmjF.08.0810 to LmjF.08.0850) were reduced to four in the LMJFC assembly (Fig. 2).

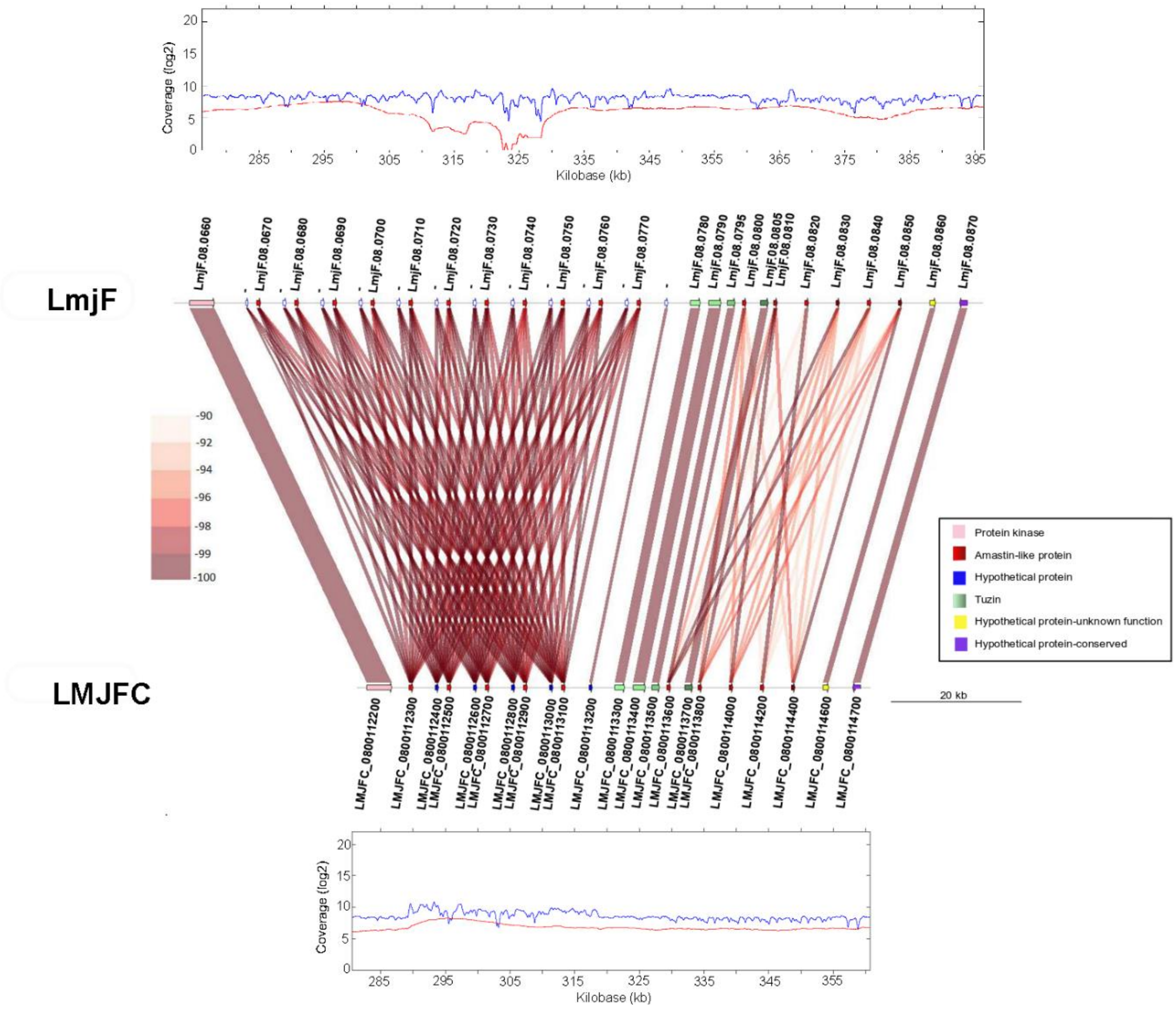

Figure 2. Analysis at the gene level of the differences existing between the LmjF and LMJFC assemblies in the middle of chromosome 8. LmjF corresponds to the gene organization existing in the current reference genome, whereas LMJFC corresponds to the equivalent region in the newly assembled genome. Gene sequence identity is shown according to a color-intensity scale (brow hue ranges from 90 to 100% of sequence identity). The upper and bottom graphs show the coverages of Illumina reads (in blue) and PacBio reads (in red) mapped to this chromosomal region in the LmjF and LMJFC assemblies, respectively.

Another chromosomal region in which the LmjF assembly contains more genes than those found in the LMJFC one is the HSP83/90 locus (Fig. 3). In the LmjF genome, there were annotated 17 HSP83/90 genes, but it is likely that the real number, according to the reads coverage, would be 12, as annotated in the LMJFC assembly.

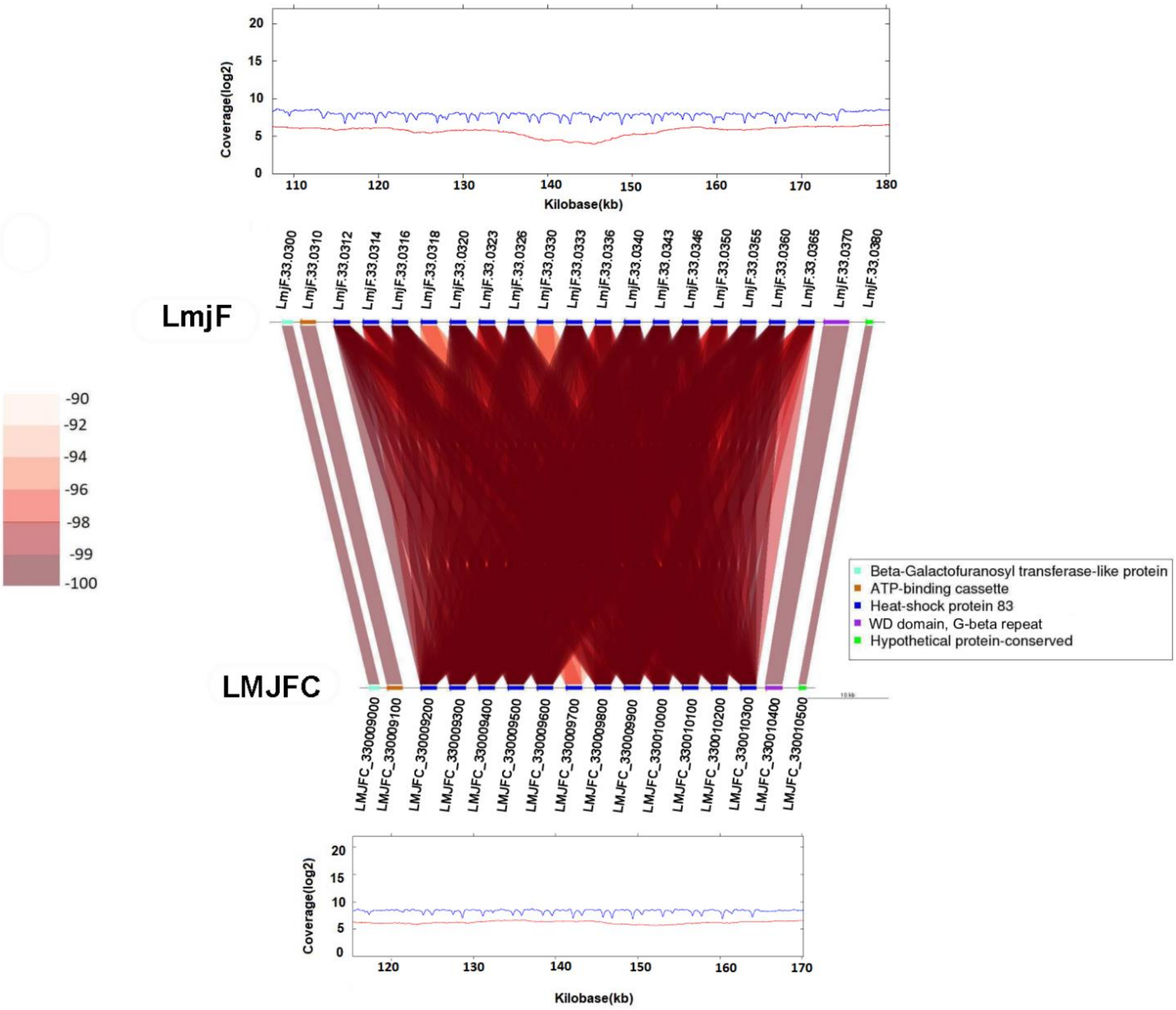

Figure 3. Different gene copy number in the HSP83/90 *locus* are assembled in either LmjF or LMJFC chromosome 33. Gene sequence identity is shown according to a color-intensity scale. The upper and bottom graphs show the coverages of Illumina reads (in blue) and PacBio reads (in red) mapped to this chromosomal region in the LmjF and LMJFC assemblies, respectively.

In addition to chromosomal *loci* in which the number of repeated genes appeared as overestimated in the LmjF assembly, in some other *loci* the situation was the converse. Figure 4 illustrates a region of chromosome 30 in which both assemblies are markedly different. On the one hand, in the LMJFC assembly, six genes coding for Ama1 protein were annotated, whereas only three are found in the LmjF assembly. On the other hand, in a region located downstream of the Ama1-protein locus, the number of genes coding for a family of class i-nuclease-like proteins was found to be larger in the LMJFC assembly (24 genes) than in the LmjF genome (4 genes). Moreover, in the LMJFC assembly, there were assembled four p1/s1 nuclease-encoding genes, whereas only two are present in the LmjF genome. The fair distribution of sequencing reads on the LMJFC assembly (Fig. 4, bottom), but uneven on the LmjF one (Fig. 4, upper) supports that the assembly attained for *L. major* (Friedlin) chromosome 30 in this work would be closer to the real one.

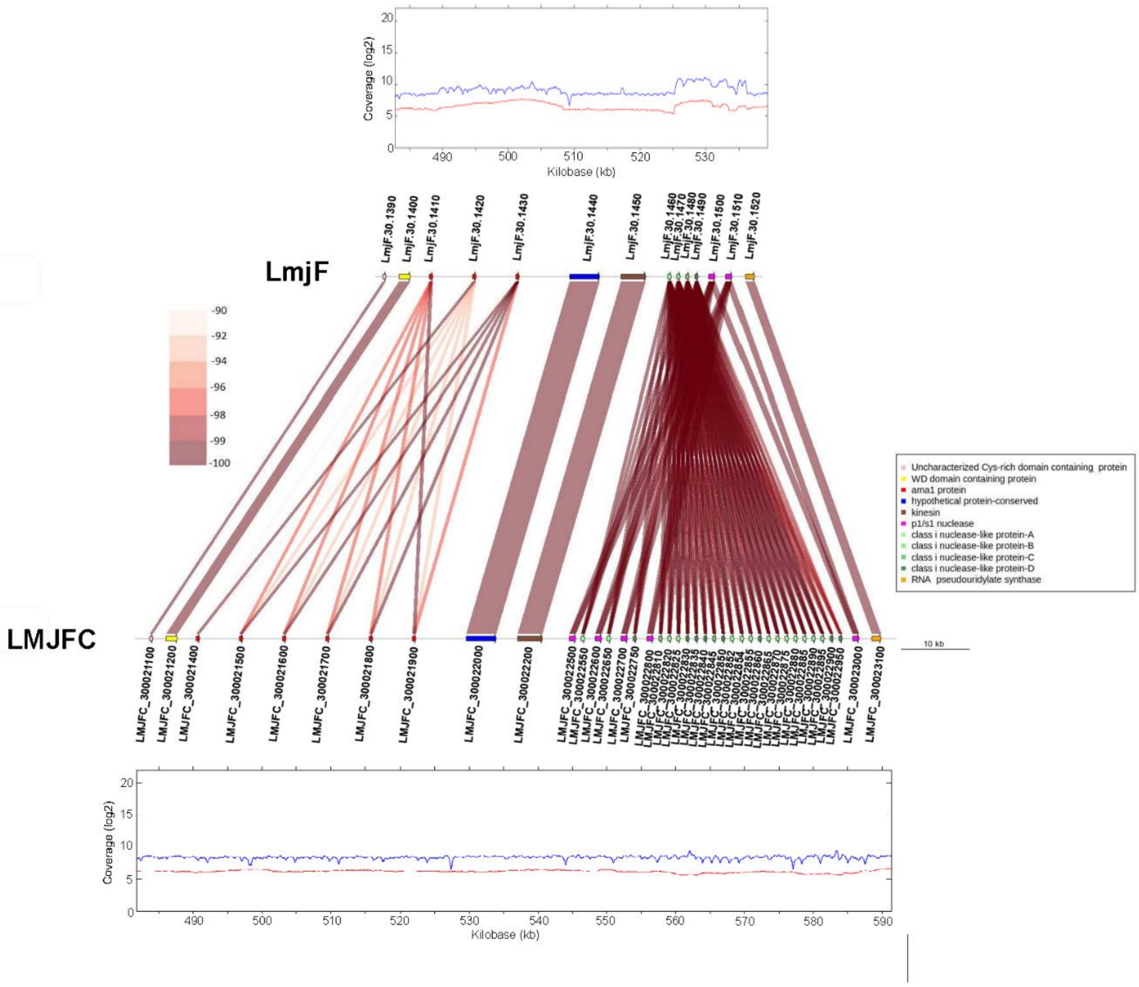

Figure 4. Different gene copy numbers exist in a central region of chromosome 30 in the LmjF and LMJFC assemblies. Gene sequence identity is shown according to a color-intensity scale. The upper and bottom graphs show the coverages of Illumina reads (in blue) and PacBio reads (in red) mapped to this chromosomal region in the LmjF and LMJFC assemblies, respectively.

The MAUVE tool [57] was used to visualize changes in synteny between the two *L. major* genome assemblies. Interestingly, few alterations were found, and the most remarkable one is that illustrated in figure 5. In this region of chromosome 29, two inverted segments were found. According to the LMJFC assembly, the gene LmjF.29.1420 would be inverted and miss-located in current LmjF genome, and its real position would be adjacent to another identical gene copy (LmjF.29.1520). The IDs for these genes in the LMJFC genome are LMJFC_290023500 and LMJFC_290023600 (Fig. 5, panel C). Also, genes LmjF.29.1430 and LmjF.29.1440 were found to be inverted in the LMJFC genome (IDs LMJFC_290022500 and LMJFC_290023400, respectively). The correctness of the LMJFC assembly in this region is supported by the smooth distribution of both PacBio and Illumina sequence reads, a fact contrasting with the abrupt decrease in coverage when these sequencing reads were aligned to the LmjFc genome (Fig. 5, panel B).

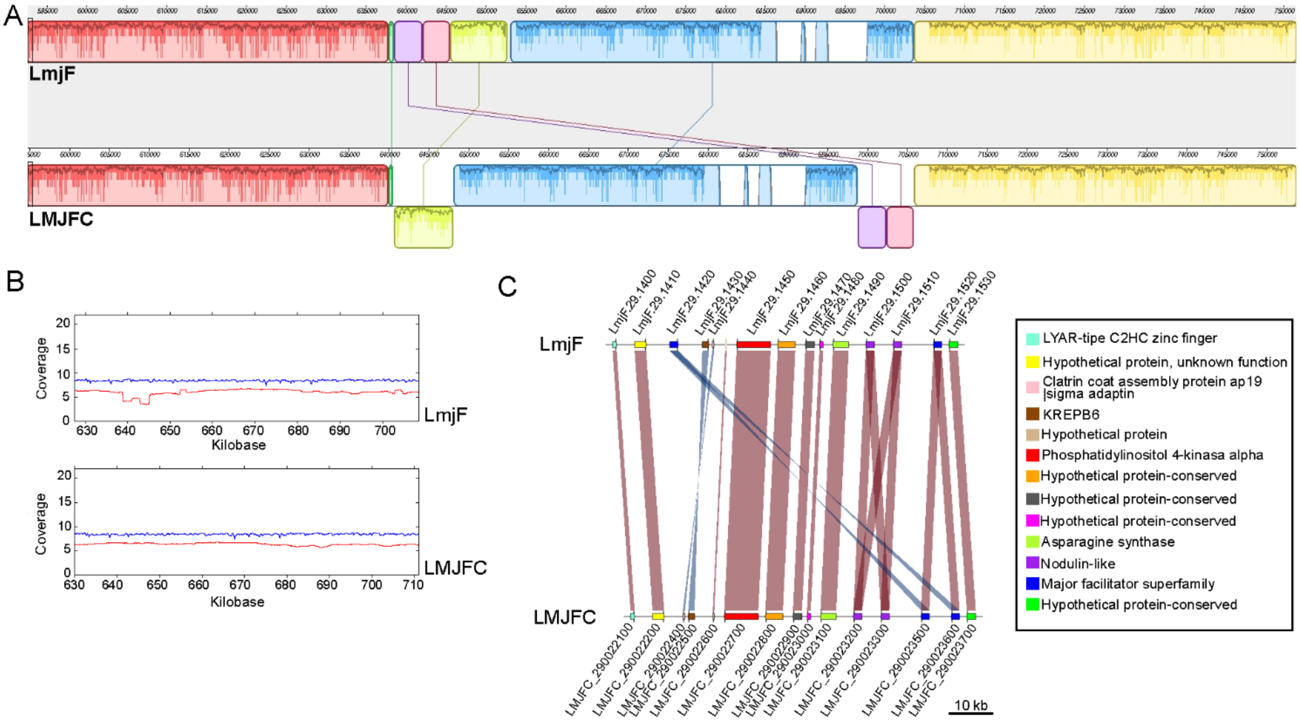

Figure 5. Reorganization of a region in chromosome 29 according to the LMJFC assembly regarding the reference LmjF genome. (A) Synteny blocks (represented by different colors that identified the conserved genomic regions) determined by pairwise comparisons between the LmjF genome (upper scheme) and the newly assembled LMJFC genome (bottom scheme), using the MAUVE tool. Blocks located underneath the X-asis denote inversion events. (B) The upper and bottom graphs show the coverages (log2) of Illumina reads (in blue) and PacBio reads (in red) mapped to this chromosomal region in the LmjF and LMJFC assemblies, respectively. (C) Schemes show the reorganization at per gene level. Genes with sequence identity and identical orientation are coloured in brown, whereas blue hues were used to denote an inverted orientation between the LmjF and LMJFC assemblies.

In addition, the seven genomic regions documented as absent from the reference *L. major* genome (LmjF) by Alonso et al [20] were verified in the new assembly (LMJFC); these findings reinforce the improving of the assembly attained in this work regarding the reference genome currently available.

Table 1 summaries some features (metrics) of the *L. major* (Friedlin) genome assembled in this study (LMJFC), and how they have varied regarding the current genome available at TrytripDB (LmjF). To note, the LMJFC assembly does not contain any sequence gap and nucleotide uncertainties, which, even though in low numbers, remained in the LmjF genome. A remarkable difference was found in the number of annotated genes, in the LMJFC genome 9,847 genes were annotated (excluding pseudogenes, listed in Supplementary file 1-Table S1), whereas in the LmjF genome (version 44) their number was 9,293. However, most of the differences are due to the use of different annotation procedures, as many of the newly annotated genes in the LMJFC genome could be also annotated in the LmjF genome sequence. Thus, for annotations on the LMJFC genome, apart from the automatic annotation generated by Companion, a manual curation was carried out in order to incorporate genes annotated in the genomes of other *Leishmania* species and related trypanosomatids. In fact, in a strict sense, only 183 genes (mainly protein-coding genes) may be categorized as new genes, as they exist only in the LMJFC assembly (see Supplementary file 1, Table S2). Another source contributing to increase the total number of annotated genes in the LMJFC genome is the existence of two or more copies for 74 genes that were single-copy genes in the LmjF assembly (see Supplementary file 1, Table S3). On the contrary, a hundred of protein-coding genes annotated on the LmjF assembly were not maintained in the new assembly (LMJFC) due to either an excessive copy number or lack of perfect matching with sequences in the LMJFC genome (these genes are listed in Supplementary file 1, Table S4).

Table 1. Parameters in the new (LMJFC) and previous (LmjF) genome assemblies

Parameters/genome	LMJFC [this work]	LmjF [v44-TritrypDB]
Number of chromosomes	36	36
Protein-coding genes	8,596	8,400
Pseudogenes	88	88
rRNAs	30	63
tRNAs	93	83
snoRNA+snRNA+slRNA	1,128	747
Number of gaps	0	9
Number of Ns	0	13
Genome size (bp)	32,792,963	32,855,082

Although the *L. major* genome sequence has been substantially improved after the re-assembling carried out in this work, we realized that a few chromosomal regions might not be assembled in a definitive manner. Apart from the rDNA locus (discussed above, and figure 1D), it was apparent that the 3’-end sequence of the chromosome 8 should be extended to accommodate the excess of Illumina reads mapping in this region (Fig. 6, panels A-C). This region contains a block of four genes that are tandemly repeated: genes LMJFC_080019000 to LMJFC_080019300 have high sequence identity with genes LMJFC_080019800 to LMJFC_080020100 (Fig. 6, panel B). According to the Illumina reads coverage, this region was found to be more accurately assembled in current LmjF genome, in which the repeated block consist of six genes (the additional genes are LmjF.08.1270 and LmjF.08.1280; fig. 6B). Even more, the Illumina reads coverage on the LmjF and LMJFC assemblies would be indicating that the beta-tubulin gene forms also part of the repeated block. Remarkably, this region in the *L. donovani* (HU3 strain) chromosome 8 [30] contains two times the block with these seven genes (from the gene coding for the Zn-finger-containing protein to the gene coding for the beta-tubulin).

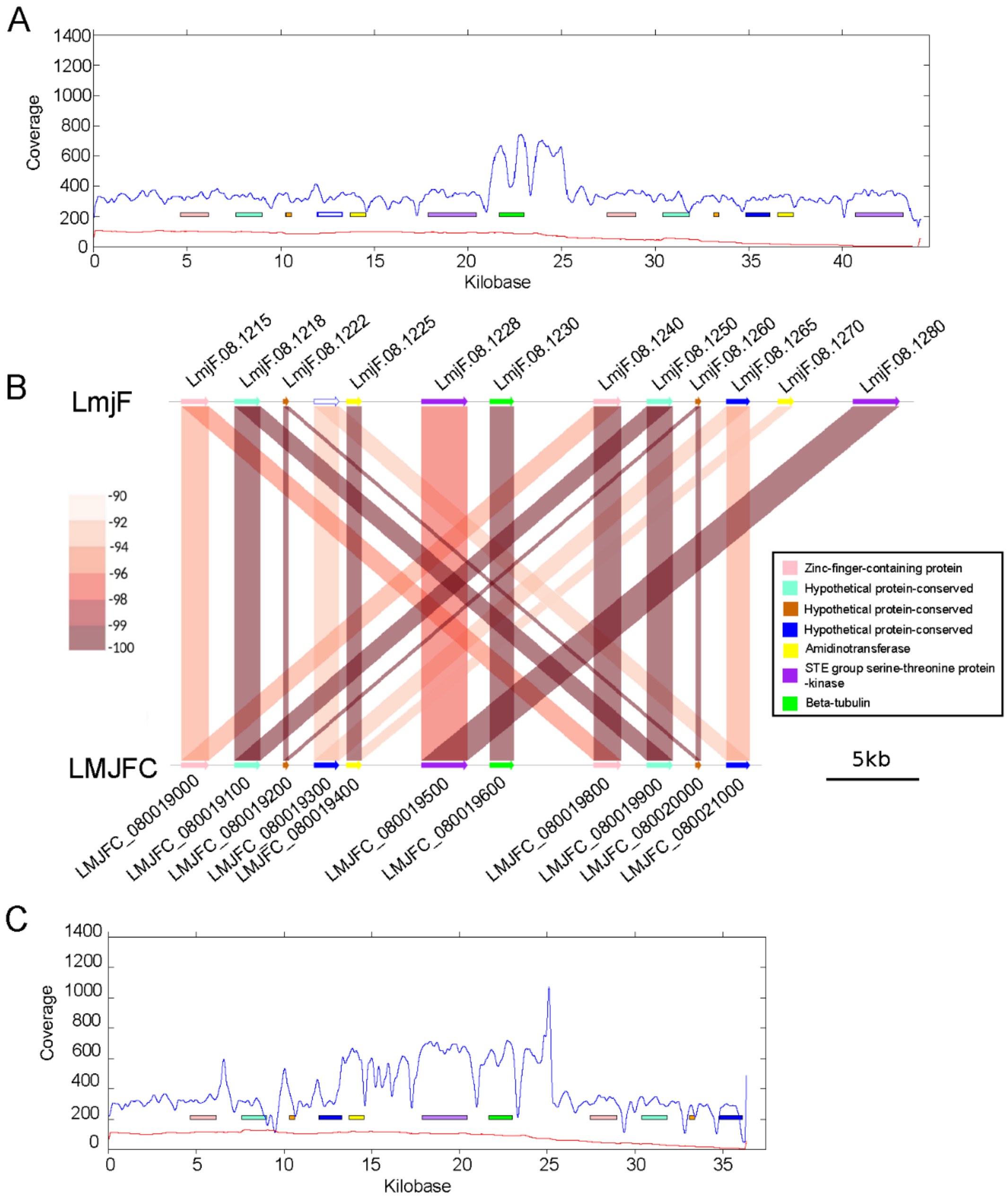

Figure 6. The right end of chromosome 8 has not been assembled in a definitive manner. (A) Linear coverages of Illumina reads (in blue) and PacBio reads (in red) mapped to this chromosomal region in the LmjF assembly. (B) Schemes show the gene-oganization and sequence identity of the genes annotated at the right end of chromosome 8 in the LmjF and LMJFC assemblies. Gene sequence identity is shown according to a color-intensity scale (brow hue ranges from 90 to 100% of sequence identity). (C) Linear coverages of Illumina reads (in blue) and PacBio reads (in red) mapped to this chromosomal region in the LMJFC assembly.

The somy of the chromosomes was calculated based on the Illumina reads coverage using the 2-loop method [56]. Most of the chromosomes in this strain were found to be diploid (Fig. 7) with the exception of chromosome 23 and 31 that appeared as trisomic and tetrasomic, respectively. This karyotype is very similar to that reported for this strain by Rogers et al [18], the sole difference was that chromosome 23 was reported as diploid and, according to our calculations, the somy of this chromosome would be triploid.

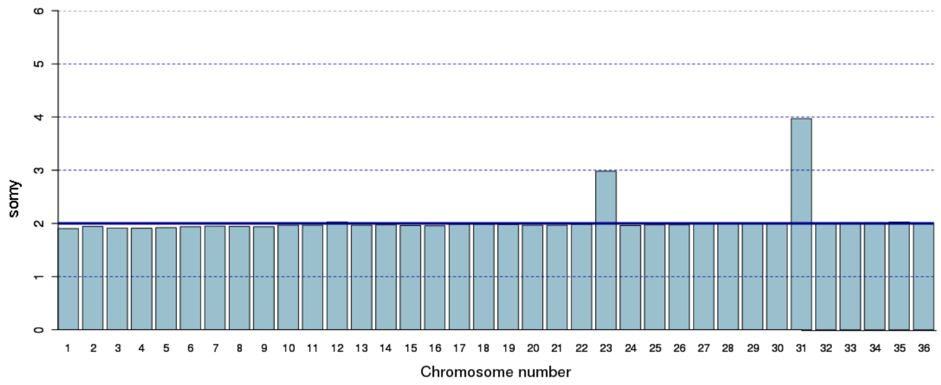

Figure 7. Somy analysis of the 36 chromosomes composing the *L. major* (Friedlin strain) genome. The somy values were calculated from the coverage of Illumina reads mapped to the LMJFC assembly, as detailed in Methods section. Graphs were generated with R, using the barplot function (https://cran.r-project.org).

Additionally, we searched for allelic polymorphisms in the assembled LMJFC genome by using HapCUT2 software (see Materials and Methods for further details). A total of 2,904 positions were found to be polymorphic; listed in Supplementary file 1, Table S5. Most were found to be nucleotide variations (Single nucleotide polymorphisms, SNPs), but insertions and deletions (InDels) were also frequent. Additionally, as the HapCUT software allows reconstructing individual haplotypes in diploid genomes [60], we looked for possible haplotype blocks, and 138 were indeed identified. The term haplotype block refers to a combination of consecutive variant sites (SNPs and/or small InDels) that are linked in a single chromosome. The delimitation of these haplotype blocks adds a valuable information regarding the gene structure, as this allows deducing the precise sequence of the two allelic genes co-existing in the genome. Figure 8 illustrates a haplotype block mapping on gene LMJFC_070017800. This block is constituted by three nucleotide transitions and one InDel of three nucleotides. As expected for a disomic chromosome, around 50% of the Illumina DNA-seq reads mapping on this region correspond to each allele (Fig. 8, panel A). Although two of the SNPs represent silent changes, an SNP and the InDel would be indicating the co-existence of two proteins differing in two amino acids (Fig. 8, panel B). If this change in sequence has a functional role merits to be analysed, in view of a recent article in which a single amino-acid change in the L. donovani RagC protein was found to dramatically affect the virulence of this parasite [67]. According to the Companion annotation, gene LMJFC_070017800 codes for a putative nucleolar RNA-binding protein, but no additional studies on this protein have been reported to date.

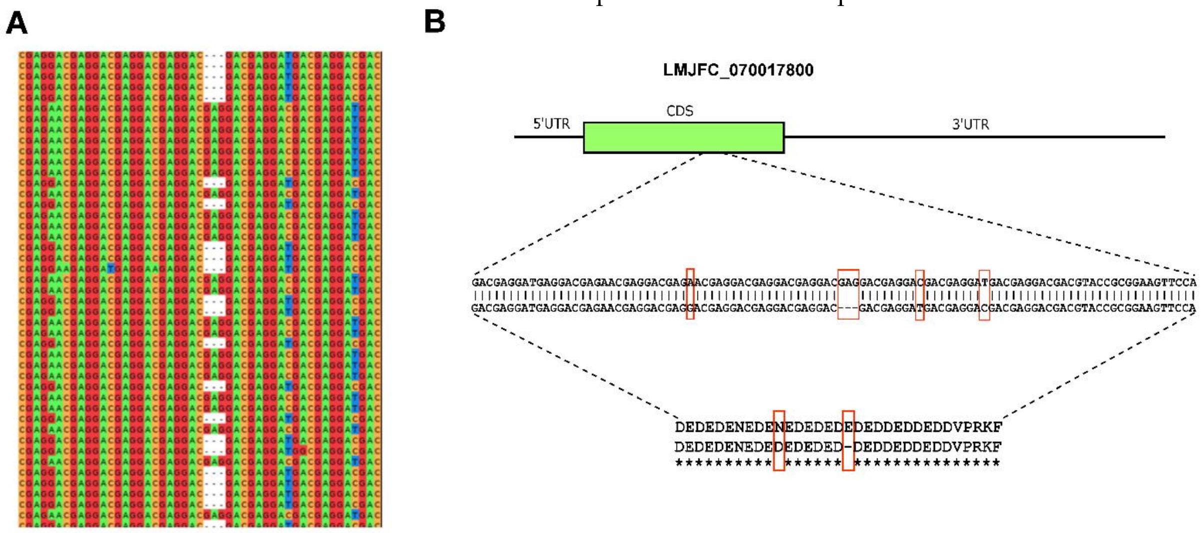
A
B
LMJFC_070017800
CDS
5'UTR
3'UTR
GACGAGGATGAGGACGAGAACGAGGACGAGAACGAGGACGAGGACGAGGACGAGGACGAGGACGACGAGGATGACGAGGACGACGTACCGCGGAAGTTCCA
GACGAGGATGAGGACGAGAACGAGGACGAGGACGAGGACGAGGACGAGGAC---GACGAGGATGACGAGGACGACGAGGACGACGTACCGCGGAAGTTCCA
DEDEDENEDENEDEDEDEDEDDEDDEDDVPRKF
DEDEDENEDEDEDEDED-DEDDEDDEDDVPRKF
********** ****** ***************

Figure 8. Identification of two distint alleles for LMJFC_070017800 gene. (A) Two haplotype blocks were identified by the HapCUT2 tool and confirmed when Illumina reads mapping to this gene were visualized. (B) Scheme of gene structure, location of allelic polymorphisms and differences in the amino acid sequences between both alleles.

3.2. Transcriptome of L. major Friedlin strain based on the new assembly (LMJFC)

The first annotated transcriptome for a species of the genus *Leishmania* was generated by our group, and that was the transcriptome for the *L. major* Friedlin strain based on the reference LmjF genome [34]. Here, we have refined the transcriptome of this strain, using the new genome assembly (LMJFC) and RNA-seq data derived from both previous [34] and more recent [63] studies. Table 2 summarizes the main features of the *L. major* transcriptome. A total of 9,828 transcripts were annotated in the LMJFC genome, and the complete list is provided in the Supplementary file (Table S6). The vast majority of the annotated coding sequences (CDS) are associated with defined transcripts, but transcripts could not be delimited for the following annotated CDS: LMJFC_020007050, LMJFC_020009550, LMJFC_070019450, LMJFC_090013750, LMJFC_100007750, LMJFC_170022050, LMJFC_270007950, LMJFC_270026300, LMJFC_280034650, LMJFC_290011550, LMJFC_310033150, LMJFC_350010450 and LMJFC_350034300. In these cases, Cufflinks software failed in generating transcripts due to the very low number of RNA-seq reads mapping to these genes, suggesting that they are not expressed in the *L. major* promastigote form. On the other hand, 47 transcripts were categorized as polycistronic (43 of them bicistronic), as they contained two or more annotated CDS. Further experimental approaches will be required to determine whether these CDS also exist as individual transcripts.

Table 2. The poly-A+ transcriptome of *L. major* Friedlin strain.

Annotated transcripts	9,828
Protein-coding transcripts	8,517
Transcripts with mapped SL addition (SLA) site	9,745 (99.1 %)
Transcripts with alternative SLA sites	9,341 (95 %)
Transcripts with mapped poly-A addition site (PAS)	8,677/9,336[1]
Transcripts with alternative PASs	6,668
Annotated CDS lacking a defined transcript	10
Transcripts with two or more CDS	47 (43 are bicistronic)

[1]PAS mapped by the analysis of RNA-seq data from Dillon et al [65].

A common feature to most of the *Leishmania* transcripts is the presence of the 39-nucleotides SL sequence at their 5′-end. In our study, this sequence was found in the vast majority of the transcripts (Table 2). Moreover, for 9,341 out of the 9,828 annotated transcripts, a multiplicity of SL addition sites was evidenced. This finding means that around 95% of the genes are transcribed into two or more RNA species differing in the length of their 5′-UTRs. In the Supplementary file (Table S6) a complete list of main and alternative SL addition sites (SASs) for every gene is provided. At the 3′-end, transcripts were delimited based on the presence of a non-encoded poly-A tail. Thus, the poly-A addition site (PAS) was identified in 8,677 of the annotated transcripts. Again, a heterogeneity in the PAS usage was evidenced, as 6,668 of those transcripts were found to be polyadenylated using two or more alternative PAS.

Additionally, the analysis of SAS allowed correction of miss-annotated CDS. As mentioned above, CDS annotation was done automatically on the LMJFC genome sequence by the Companion tool [32]. This software is designed to annotate a new genome based on a reference genome (in our case, we selected the LmjF genome). When the SASs detected in this study were positioned relative to the automatically annotated CDSs, we found that main SASs were located, in some cases, within the predicted CDS. Therefore, CDS annotation of those genes had to be shortened in order to establish the first in-phase

doi:10.20944/preprints202107.0562.v1

ATG, downstream from the SAS, as the initiation codon. This modification was introduced for 247 genes. An example, based on gene LMJFC_360016000, is shown in figure 9. This figure also illustrates the process followed in the definition of the gene models that consisted of four steps: i) the transcripts are created from the distribution of RNA-seq reads; ii) transcripts are trimmed at their ends by mapping SAS and PAS; iii) Companion-annotated CDS are placed on the transcript; iv) a gene model including 5′- and 3′-UTR is created. All the gene models generated in this study are listed in supplementary file (Table S7).

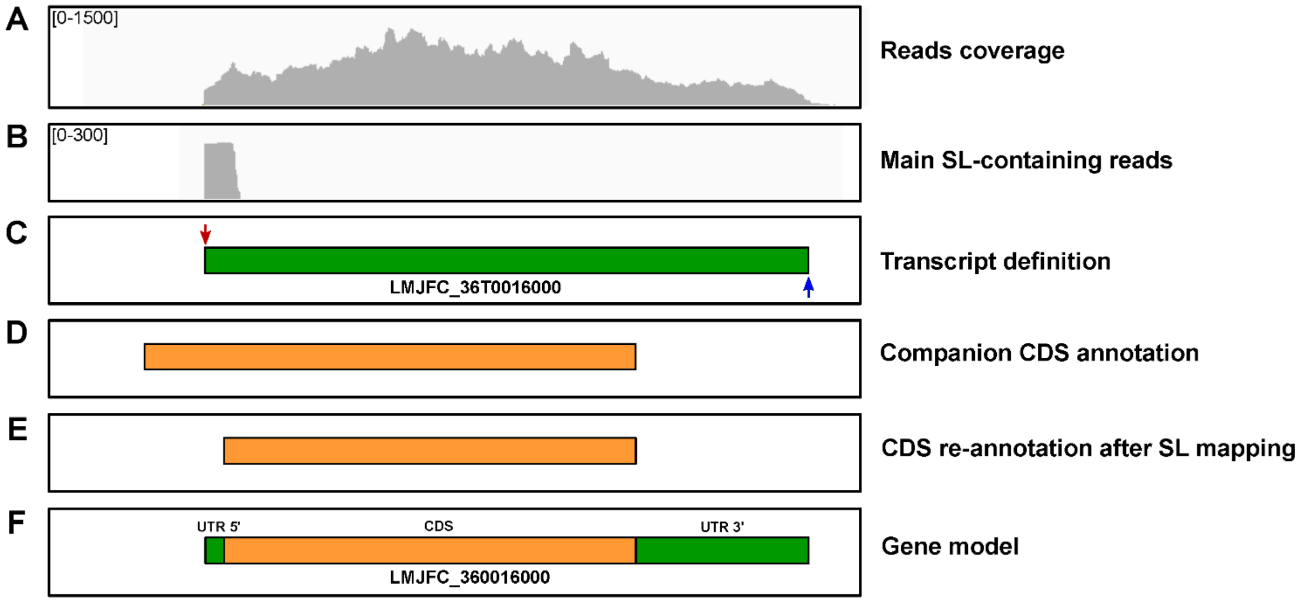

Figure 9. Process followed in the generation of the gene models. (A) RNA-seq reads distribution. (B) Mapping of RNA-seq containing SL-sequences. (C) Definition of the transcript boundaries based on the SAS (supported by 213 reads) and PAS (4 reads) positions. (D) CDS as annotated by Companion. (E) CDS re-annotated according to SAS position. (F) Final gene model for LMJFC_360016000 gene.

The availability of gene models results crucial for studying gene expression, either for individual genes or by whole-genome approaches. Thus, the ectopic expression of a given gene may vary according to the regulatory sequences surrounding the coding sequence and this may explain that phenotypic defects in deletion mutant could be not restored by add-back plasmids containing the coding regions without their regulatory sequences [42]. On the other hand, *Leishmania* genomes contain many repeated genes having identical CDS but different UTRs. In this situation, when expression studies are conducted at the genomic-scale based solely in CDS coordinates, it is not possible distinguish whether differential expression exists among the repeated genes [63].

Upon establishment of gene models, we used RNA-seq data from a recent study [63] to quantify relative transcript levels in *L. major* promastigotes. Measuring of the relative RNA abundance was performed by the TPM (transcripts per million) method [68]. A wide range of expression levels was observed, from 3,660 TPM for transcript LMJFC_27T0019600 to near zero (transcript LMJFC_36T0057500). In Table S8 (Supplementary file), the relative levels for every one of the transcripts delineated in this study are listed. Table 3 shows the 40 transcripts with the highest expression levels. Two transcripts coding for histone H1 were found to be the most expressed, and 11 additional transcripts coding for nucleosomal histones (H2A, H2B, H3 and H4) were ranked among the top-40 expressed genes. This is an expected finding given the abundance of these proteins in the cell. In addition, half of the more expressed genes code for ribosomal proteins, again abundant cellular constituents. Therefore, it is worthy to discuss about the presence of the other transcripts listed among the most abundant poly-A+ RNA molecules in the *L. major* promastigotes. The sixth most abundant transcript (LMJFC_36T0028500) codes for an inosine-guanosine transporter named NT2. Purine transporters are essential in *Leishmania* and other trypanosomatids, since they are incapable of purine biosynthesis and have to acquire purines from the host milieu [69]. The ninth transcript in the list is LMJFC_13T0009700, which codes for ALBA protein 1. Alba proteins are RNA-binding proteins that participate in mechanisms controlling developmentally regulated gene expression and have been reported to regulate translational efficiency and turnover rate of particular transcripts in *Leishmania* [70, 71]. A transcript coding for a nucleoside diphosphate kinase (LMJFC_32T0040100) occupied the twenty-first position, which also agrees with the relevant role played by these enzymes, i.e. they catalyze the transfer of the γ-phosphate moiety from a nucleoside triphosphate (NTP) donor to an NDP acceptor in order to maintain appropriate cellular levels of NTPs [72]. The 25th transcript (LMJFC_25T0016400) encodes for a cyclophilin having peptidyl-prolyl cis/trans isomerase activity [73]; this isomerase activity is essential for protein folding after translation. Another abundant transcript, LMJFC_32T0014200 (position 29th; Table 3) encodes for a protein-binding protein; the homologous protein in the *Leishmania*-related trypanosomatid *Trypanosoma cruzi* has been described as an abundant protein associated with polysomes [74]. Finally, position 39th is occupied by transcript LMJFC_35T0029900, which codes for the kinetoplastid membrane protein 11 (KMP11); accordingly, several millions of KMP11 molecules have been estimated to exist per promastigote cell [75]. In summary, according to the functional relevance and literature data, all the transcripts listed in Table 3 would be coding for abundant *Leishmania* proteins. In this regards, in the list, there is a transcript (LMJFC_31T0016500, ranked 35th) that encodes for a short protein (79 amino acids in length) of unknown function. This protein is well-conserved among different trypanosomatids, and according to the expression levels of its transcript, might be also a highly abundant molecule in *L. major* promastigotes.

Table 3. The 40 most abundant transcripts in L. major (Friedlin strain) promastigotes.

Transcript ID	TPM ± SD[1]	Name of the encoded protein
LMJFC_27T0019600	3660.01 ± 344.91	histone H1
LMJFC_27T0019100	3629.05 ± 386.87	histone H1
LMJFC_35T0007800	3477.13 ± 722.67	ribosomal protein L30
LMJFC_06T0005100	3034.31 ± 646.98	histone H4
LMJFC_29T0026200	3027.80 ± 656.33	histone H2A
LMJFC_36T0028500	2876.01 ± 261.79	inosine-guanosine transporter (NT2)
LMJFC_19T0005400	2862.36 ± 308.82	histone H2B
LMJFC_19T0005500	2684.51 ± 221.06	histone H2B
LMJFC_13T0009700	2578.58 ± 230.03	ALBA-domain protein 1 (ALBA1)
LMJFC_19T0005600	2555.44 ± 211.16	histone H2B
LMJFC_30T0045400	2430.21 ± 261.54	ribosomal protein L9
LMJFC_15T0005100	2294.60 ± 265.01	histone H4
LMJFC_16T0012400	2285.04 ± 250.45	histone H3
LMJFC_35T0048600	2246.30 ± 380.22	ribosomal protein L23
LMJFC_31T0048800	2078.23 ± 130.10	histone H4
LMJFC_29T0026000	2068.24 ± 167.92	histone H2A
LMJFC_24T0032000	2060.45 ± 60.25	ribosomal protein L12
LMJFC_19T0005700	2057.42 ± 51.88	ribosomal protein S2
LMJFC_13T0011100	2052.88 ± 103.60	ribosomal protein S12
LMJFC_09T0020600	1994.99 ± 193.32	histone H2B

LMJFC_32T0040100	1964.77 ± 53.15	nucleoside diphosphate kinase b
LMJFC_35T0011400	1939.67 ± 357.53	ribosomal protein L18a
LMJFC_28T0032200	1935.03 ± 70.94	ribosomal protein S29
LMJFC_35T0026300	1899.65 ± 350.17	ribosomal protein L15
LMJFC_25T0016400	1898.52 ± 326.43	cyclophilin A | CyP1
LMJFC_29T0039100	1898.13 ± 80.33	ribosomal protein S19-like protein
LMJFC_14T0020500	1887.76 ± 63.85	ubiquitin/ribosomal protein S27a
LMJFC_32T0010100	1857.84 ± 87.17	ribosomal protein L17
LMJFC_32T0014200	1836.37 ± 60.13	RNA binding protein
LMJFC_32T0010300	1833.43 ± 53.95	ribosomal protein S2
LMJFC_35T0027400	1814.80 ± 300.02	ribosomal protein S6
LMJFC_36T0051800	1798.18 ± 96.35	ribosomal protein L34
LMJFC_35T0042900	1794.95 ± 276.06	ribosomal subunit protein L31
LMJFC_32T0016000	1782.48 ± 98.87	ribosomal protein L18a
LMJFC_31T0016500	1780.58 ± 476.99	hypothetical protein-conserved
LMJFC_35T0048200	1779.35 ± 288.58	ribosomal protein L27A/L29
LMJFC_29T0034600	1779.35 ± 55.64	ribosomal protein L13
LMJFC_35T0048400	1770.18 ± 217.83	ribosomal protein L27A/L29
LMJFC_35T0029900	1754.88 ± 188.03	kinetoplastid membrane protein 11 (KMP11)
LMJFC_25T0035900	1745.73 ± 104.64	histone H4

[1] Standard deviation (SD).

4. Conclusions

Combination of second (Illumina) and third (PacBio) NGS tools has proved to be a powerful strategy for attaining complete assemblies of genomes. In this work, based on the use of both technologies, a *de novo* assembly of the *L. major* reference strain (Friedlin) genome is reported. Although the previous reference genome for this strain (LmjF; [7]) may be categorized as an outstanding assembly and has been widely used for a long time, the assembly attained in this work represents an improved version that should replace the LmjF genome. It should be bore in mind that genomic-whole studies, either transcriptomics or proteomics, depend on the accuracy of the sequence and annotations of the reference genome.

Another contribution of this work is the generation of the poly-A+ transcriptome for the *L. major* promastigote stage. Essentially, all the protein-coding genes are represented in the transcriptome. Moreover, remarkable heterogeneities in the SL and polyadenylation sites were observed for around 95% of the transcripts. The combination of gene annotations and transcript delimitation have allowed the generation of gene models for the entire *L. major* genome. A precise determination of the 5'- and 3'-UTRs is mandatory for studies dealing with gene expression, mainly in organisms like *Leishmania*, in which gene expression regulation occurs almost exclusively at the post-transcriptional level.

To our knowledge, this is the first report in which the existence of haplotype blocks has been analysed in the *L. major* genome. The coexistence of different alleles for a given gene adds another layer of complexity that might have phenotypic implications in this parasite. Moreover, this finding may explain that *Leishmania* is mostly diploid when sexual reproduction is a rare event in this group of protists.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Supplementary file containing: Table S1.- List of putative pseudogenes found in the LMJFC genome; Table

S2.- New genes annotated in the LMJFC assembly that could not be found in the LmjF genome sequence; Table S3.- Single-copy genes in the LmjF assembly that were found to be repeated in the LMJFC assembly; Table S4.- List of genes in the LmjF assembly that are not present in the LMJFC assembly; Table S5.- Polymorphic positions found in the LMJFC genome; Table S6.- Transcriptome: names, coordinates, SL addition sites (SAS, both main and alternatives), poly-A addition site (PAS, both main and alternatives) and associated function of the encoded protein; Table S7.- Gene models defined in the LMJFC genome; Table S8.- Relative expression levels (TPM) of the transcripts annotated in the LMJFC genome.

Author Contributions: Conceptualization, E.C., S.G.-F, J.C.S., A.R., F.C.-R., J.M.R., and B.A.; methodology, E.C., S.G.-F, J.C.S., and A.R.; formal analysis, E.C., S.G.-F, J.C.S., and A.R.; data curation, E.C., S.G.-F, and J.M.R.; writing—original draft preparation, E.C., S.G.-F, and J.M.R.; writing—review and editing, F.C.-R., J.M.R., and B.A.; funding acquisition, B.A., and J.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants (to B.A. and J.M.R.) from Proyecto del Ministerio de Economía, Industria y Competitividad SAF2017-86965-R, and by the Network of Tropical Diseases Research RICET (RD16/0027/0008); both grants are co-funded with FEDER funds. J.C.S. was funded by a postdoctoral contract based on project UAM/133 from Universidad Autonoma de Madrid (granted to J.M.R.). The CBMSO receives institutional grants from the Fundación Ramón Areces and from the Fundación Banco de Santander.

Data Availability Statement: Genomic and transcriptomic raw reads have been deposited in the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/). Besides, the assembled genome and transcriptome sequences together with annotations files were uploaded under the Study accession number PRJEB25921. Additionally, the genome (fasta file), the annotations for the genome, transcriptome and gene models are downloadable at the Leish-ESP website (http://leish-esp.cbm.uam.es/).

Acknowledgments: We thank the Genomics and NGS Core Facility at the Centro de Biología Molecular Severo Ochoa (CBMSO, CSIC-UAM) for helping with the bioinformatics analysis. Also, wa acknowledge to the sequencing service provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the South-Eastern Regional Health Authorities.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Burza, S.; Croft, S.L.; Boelaert, M. Leishmaniasis. Lancet 2018, 392, 951–970, doi:10.1016/S0140-6736(18)31204-2.
2. Iborra, S.; Solana, J.C.; Requena, J.M.; Soto, M. Vaccine candidates against leishmania under current research. Expert Rev. Vaccines 2018, 17, 323–334, doi:10.1080/14760584.2018.1459191.
3. Hefnawy, A.; Berg, M.; Dujardin, J.C.; De Muylder, G. Exploiting Knowledge on Leishmania Drug Resistance to Support the Quest for New Drugs. Trends Parasitol 2017, 33, 162–174, doi:10.1016/j.pt.2016.11.003.
4. Ivens, A.C.; Blackwell, J.M. Unravelling the Leishmania genome. Curr Opin Genet Dev 1996, 6, 704–10.
5. Ivens, A.C.; Lewis, S.M.; Bagherzadeh, A.; Zhang, L.; Chan, H.M.; Smith, D.F. A physical map of the Leishmania major Friedlin genome. Genome Res 1998, 8, 135–45.
6. Zhou, S.; Kile, A.; Kvikstad, E.; Bechner, M.; Severin, J.; Forrest, D.; Runnheim, R.; Churas, C.; Anantharaman, T.S.; Myler, P.; et al. Shotgun optical mapping of the entire Leishmania major Friedlin genome. Mol Biochem Parasitol 2004, 138, 97–106.
7. Ivens, A.C.; Peacock, C.S.; Worthey, E.A.; Murphy, L.; Aggarwal, G.; Berriman, M.; Sisk, E.; Rajandream, M.A.; Adlem, E.; Aert, R.; et al. The Genome of the Kinetoplastid Parasite, Leishmania major. Science (80-.). 2005, 309, 436–442.
8. Cohen-Freue, G.; Holzer, T.R.; Forney, J.D.; McMaster, W.R. Global gene expression in Leishmania. Int J Parasitol 2007, 37, 1077–1086.
9. Peacock, C.S.; Seeger, K.; Harris, D.; Murphy, L.; Ruiz, J.C.; Quail, M.A.; Peters, N.; Adlem, E.; Tivey, A.; Aslett, M.; et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat Genet 2007, 39, 839–847.
10. Butenko, A.; Kostygov, A.Y.; Sadlova, J.; Kleschenko, Y.; Becvar, T.; Podesvova, L.; Macedo, D.H.; Zihala, D.; Lukes, J.; Bates, P.A.; et al. Comparative genomics of Leishmania (Mundinia). BMC Genomics 2019, 20, 726, doi:10.1186/s12864-019-6126-y 10.1186/s12864-019-6126-y [pii].

11. Real, F.; Vidal, R.O.; Carazzolle, M.F.; Mondego, J.M.; Costa, G.G.; Herai, R.H.; Wurtele, M.; de Carvalho, L.M.; Carmona e Ferreira, R.; Mortara, R.A.; et al. The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and extended analysis of gene models. DNA Res 2013, 20, 567–581, doi:10.1093/dnares/dst031 dst031 [pii].
12. Llanes, A.; Restrepo, C.M.; Del Vecchio, G.; Anguizola, F.J.; Lleonart, R. The genome of Leishmania panamensis: insights into genomics of the L. (Viannia) subgenus. Sci Rep 2015, 5, 8550, doi:10.1038/srep08550.
13. Gupta, A.K.; Srivastava, S.; Singh, A.; Singh, S. De Novo Whole-Genome Sequence and Annotation of a Leishmania Strain Isolated from a Case of Post-Kala-Azar Dermal Leishmaniasis. Genome Announc 2015, 3, doi:10.1128/genomeA.00809-15.
14. Downing, T.; Imamura, H.; Decuypere, S.; Clark, T.G.; Coombs, G.H.; Cotton, J.A.; Hilley, J.D.; de Doncker, S.; Maes, I.; Mottram, J.C.; 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, doi:gr.123430.111 [pii]10.1101/gr.123430.111.
15. Raymond, F.; Boisvert, S.; Roy, G.; Ritt, J.F.; Legare, D.; Isnard, A.; Stanke, M.; Olivier, M.; Tremblay, M.J.; Papadopoulou, B.; et al. Genome sequencing of the lizard parasite Leishmania tarentolae reveals loss of genes associated to the intracellular stage of human pathogenic species. Nucleic Acids Res 2012, 40, 1131–1147, doi:gkr834 [pii]10.1093/nar/gkr834.
16. Valdivia, H.O.; Almeida, L. V; Roatt, B.M.; Reis-Cunha, J.L.; Pereira, A.A.S.; Gontijo, C.; Fujiwara, R.T.; Reis, A.B.; Sanders, M.J.; Cotton, J.A.; et al. Comparative genomics of canine-isolated Leishmania (Leishmania) amazonensis from an endemic focus of visceral leishmaniasis in Governador Valadares, southeastern Brazil. Sci Rep 2017, 7, 40804, doi:10.1038/srep40804.
17. Urrea, D.A.; Duitama, J.; Imamura, H.; Alzate, J.F.; Gil, J.; Munoz, N.; Villa, J.A.; Dujardin, J.C.; Ramirez-Pineda, J.R.; Triana-Chavez, O. Genomic Analysis of Colombian Leishmania panamensis strains with different level of virulence. Sci Rep 2018, 8, 17336, doi:10.1038/s41598-018-35778-6.
18. Rogers, M.B.; Hilley, J.D.; Dickens, N.J.; Wilkes, J.; Bates, P.A.; Depledge, D.P.; Harris, D.; Her, Y.; Herzyk, P.; Imamura, H.; et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. Genome Res 2011, 21, 2129–2142, doi:gr.122945.111 [pii]10.1101/gr.122945.111.
19. Johner, A.; Kunz, S.; Linder, M.; Shakur, Y.; Seebeck, T. Cyclic nucleotide specific phosphodiesterases of Leishmania major. BMC Microbiol 2006, 6, 25.
20. Alonso, G.; Rastrojo, A.; López-Pérez, S.; Requena, J.M.; Aguado, B. Resequencing and assembly of seven complex loci to improve the Leishmania major (Friedlin strain) reference genome. Parasites and Vectors 2016, 9, 74, doi:10.1186/s13071-016-1329-4.
21. Pita, S.; Diaz-Viraque, F.; Iraola, G.; Robello, C. The Tritryps Comparative Repeatome: Insights on Repetitive Element Evolution in Trypanosomatid Pathogens. Genome Biol Evol 2019, 11, 546–551, doi:10.1093/gbe/evz017.
22. Ubeda, J.M.; Raymond, F.; Mukherjee, A.; Plourde, M.; Gingras, H.; Roy, G.; Lapointe, A.; Leprohon, P.; Papadopoulou, B.; Corbeil, J.; et al. Genome-wide stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite Leishmania. PLoS Biol 2014, 12, e1001868, doi:10.1371/journal.pbio.1001868PBIOLOGY-D-13-04965 [pii].
23. Requena, J.M.; Rastrojo, A.; Garde, E.; López, M.C.; Thomas, M.C.; Aguado, B.; Lopez, M.C.; Thomas, M.C.; Aguado, B. Genomic cartography and proposal of nomenclature for the repeated, interspersed elements of the Leishmania major SIDER2 family and identification of SIDER2-containing transcripts. Mol Biochem Parasitol 2017, 212, 9–15, doi:10.1016/j.molbiopara.2016.12.009.
24. Requena, J.M. Lights and shadows on gene organization and regulation of gene expression in Leishmania. Front. Biosci. 2011, 16, 2069–85, doi:10.2741/3840.
25. van Dijk, E.L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. Trends Genet 2018, 34, 666–681, doi:10.1016/j.tig.2018.05.008.
26. Lypaczewski, P.; Hoshizaki, J.; Zhang, W.W.; McCall, L.I.; Torcivia-Rodriguez, J.; Simonyan, V.; Kaur, A.; Dewar, K.; Matlashewski, G. A complete Leishmania donovani reference genome identifies novel genetic variations associated with virulence. Sci Rep 2018, 8, 16549, doi:10.1038/s41598-018-34812-x.
27. Gonzalez-de la Fuente, S.; Camacho, E.; Peiro-Pastor, R.; Rastrojo, A.; Carrasco-Ramiro, F.; Aguado, B.; Requena, J.M. Complete and de novo assembly of the Leishmania braziliensis (M2904) genome. Mem Inst Oswaldo Cruz 2018, 114, e180438, doi:10.1590/0074-02760180438.
28. Lin, W.; Batra, D.; Narayanan, V.; Rowe, L.A.; Sheth, M.; Zheng, Y.; Juieng, P.; Loparev, V.; de Almeida, M. First Draft Genome Sequence of Leishmania (Viannia) lainsoni Strain 216-34, Isolated from a Peruvian Clinical Case. Microbiol Resour Announc 2019, 8, e01524, doi:10.1128/MRA.01524-18.
29. Gonzalez-de la Fuente, S.; Peiro-Pastor, R.; Rastrojo, A.; Moreno, J.; Carrasco-Ramiro, F.; Requena, J.M.; Aguado, B. Resequencing of the Leishmania infantum (strain JPCM5) genome and de novo assembly into 36 contigs. Sci Rep 2017, 7, 18050, doi:10.1038/s41598-017-18374-y.
30. Camacho, E.; Gonzalez-de la Fuente, S.; Rastrojo, A.; Peiro-Pastor, R.; Solana, J.C.; Tabera, L.; Gamarro, F.; Carrasco-Ramiro, F.; Requena, J.M.; Aguado, B. Complete assembly of the Leishmania donovani (HU3 strain) genome and transcriptome annotation. Sci Rep 2019, 9, 6127, doi:10.1038/s41598-019-42511-4.
31. Batra, D.; Lin, W.; Narayanan, V.; Rowe, L.A.; Sheth, M.; Zheng, Y.; Loparev, V.; de Almeida, M. Draft Genome Sequences of Leishmania (Leishmania) amazonensis, Leishmania (Leishmania) mexicana, and Leishmania (Leishmania) aethiopica, Potential Etiological Agents of Diffuse Cutaneous Leishmaniasis. Microbiol Resour Announc 2019, 8, e00269, doi:10.1128/MRA.00269-19.
32. Steinbiss, S.; Silva-Franco, F.; Brunk, B.; Foth, B.; Hertz-Fowler, C.; Berriman, M.; Otto, T.D. Companion: a web server for annotation and analysis of parasite genomes. Nucleic Acids Res 2016, 44, W29-34, doi:10.1093/nar/gkw292.
33. Smith, M.; Blanchette, M.; Papadopoulou, B. Improving the prediction of mRNA extremities in the parasitic protozoan Leishmania. BMC Bioinformatics 2008, 9, 158.

34. Rastrojo, A.; Carrasco-Ramiro, F.; Martín, D.; Crespillo, A.; Reguera, R.M.; Aguado, B.; Requena, J.M. The transcriptome of Leishmania major in the axenic promastigote stage: Transcript annotation and relative expression levels by RNA-seq. BMC Genomics 2013, 14, 223, doi:10.1186/1471-2164-14-223.
35. Fiebig, M.; Kelly, S.; Gluenz, E. Comparative Life Cycle Transcriptomics Revises Leishmania mexicana Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. PLoS Pathog 2015, 11, e1005186, doi:10.1371/journal.ppat.1005186.
36. Soto, M.; Requena, J.M.; Alonso, C. Isolation, characterization and analysis of the expression of the Leishmania ribosomal PO protein genes. Mol. Biochem. Parasitol. 1993, 61, 265–74, doi:10.1016/0166-6851(93)90072-6.
37. Soto, M.; Requena, J.M.; Garcia, M.; Gómez, L.C.; Navarrete, I.; Alonso, C. Genomic organization and expression of two independent gene arrays coding for two antigenic acidic ribosomal proteins of Leishmania. J. Biol. Chem. 1993, 268, 21835–43.
38. Requena, J.M.; Soto, M.; Quijada, L.; Alonso, C. Genes and Chromosomes of Leishmania infantum. Mem. Inst. Oswaldo Cruz 1997, 92, 853–858, doi:10.1590/s0074-02761997000600022.
39. Zilka, A.; Garlapati, S.; Dahan, E.; Yaolsky, V.; Shapira, M. Developmental Regulation of Heat Shock Protein 83 in Leishmania. 3′ Processing and mRNA stability control transcript abundance, and translation is directed by a determinant in the 3′-untranslated region. J Biol Chem 2001, 276, 47922–9.
40. Soto, M.; Quijada, L.; Larreta, R.; Iborra, S.; Alonso, C.; Requena, J.M. Leishmania infantum possesses a complex family of histone H2A genes: Structural characterization and analysis of expression. Parasitology 2003, 127, 95–105, doi:10.1017/S0031182003003445.
41. Larreta, R.; Soto, M.; Quijada, L.; Folgueira, C.; Abanades, D.R.; Alonso, C.; Requena, J.M. The expression of HSP83 genes in Leishmania infantum is affected by temperature and by stage-differentiation and is regulated at the levels of mRNA stability and translation. BMC Mol. Biol. 2004, 5, 3, doi:10.1186/1471-2199-5-3.
42. Folgueira, C.; Quijada, L.; Soto, M.; Abanades, D.R.; Alonso, C.; Requena, J.M. The translational efficiencies of the two Leishmania infantum HSP70 mRNAs, differing in their 3′-untranslated regions, are affected by shifts in the temperature of growth through different mechanisms. J. Biol. Chem. 2005, 280, 35172–35183, doi:10.1074/jbc.M505559200.
43. Requena, J.M.; López, M.C.; Jimnez-Ruiz, A.; de la Torre, J.C.; Alonso, C. A head-to-tail tandem organization of hsp70 genes in Trypanosoma cruzi. Nucleic Acids Res. 1988, 16, 1393–406, doi:10.1093/nar/16.4.1393.
44. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. Real-time DNA sequencing from single polymerase molecules. Science (80-.). 2009, 323, 133–138, doi:10.1126/science.1162986.
45. Chin, C.S.; Alexander, D.H.; Marks, P.; Klammer, A.A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E.E.; et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 2013, 10, 563–569, doi:10.1038/nmeth.2474.
46. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: architecture and applications. BMC Bioinformatics 2009, 10, 421, doi:1471-2105-10-421 [pii] 10.1186/1471-2105-10-421.
47. Sommer, D.D.; Delcher, A.L.; Salzberg, S.L.; Pop, M. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 2007, 8, 64, doi:10.1186/1471-2105-8-64.
48. Katoh, K.; Kuma, K.; Toh, H.; Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 2005, 33, 511–518, doi:10.1093/nar/gki198.
49. Boetzer, M.; Henkel, C. V; Jansen, H.J.; Butler, D.; Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2011, 27, 578–579, doi:10.1093/bioinformatics/btq683.
50. Nadalin, F.; Vezzi, F.; Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics 2012, 13 Suppl 1, S8, doi:10.1186/1471-2105-13-S14-S8.
51. Sacristán-Horcajada, E.; González-de la Fuente, S.; Peiró-Pastor, R.; Carrasco-Ramiro, F.; Amils, R.; Requena, J.; Berenguer, J.; Aguado, B. ARAMIS: From systematic errors of NGS long reads to accurate assemblies. Brief. Bioinform. 2021, doi:10.1093/bib/bbab170.
52. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25, 2078–2079, doi:btp352 [pii]10.1093/bioinformatics/btp352.
53. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017, 27, 722–736, doi:10.1101/gr.215087.116.
54. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012, 9, 357–359, doi:10.1038/nmeth.1923nmeth.1923 [pii].
55. Chaisson, M.J.; Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 2012, 13, 238, doi:10.1186/1471-2105-13-238.
56. Mondelaers, A.; Sanchez-Cañete, M.P.; Hendrickx, S.; Eberhardt, E.; Garcia-Hernandez, R.; Lachaud, L.; Cotton, J.; Sanders, M.; Cuypers, B.; Imamura, H.; et al. Genomic and Molecular Characterization of Miltefosine Resistance in Leishmania infantum Strains with Either Natural or Acquired Resistance through Experimental Selection of Intracellular Amastigotes. PLoS One 2016, 11, e0154101, doi:10.1371/journal.pone.0154101.
57. Darling, A.E.; Mau, B.; Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 2010, 5, e11147, doi:10.1371/journal.pone.0011147.
58. Guy, L.; Kultima, J.R.; Andersson, S.G. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 2010, 26, 2334–2335, doi:10.1093/bioinformatics/btq413.
59. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013, 43, 11 10 1-33, doi:10.1002/0471250953.bi1110s43.

60. Edge, P.; Bafna, V.; Bansal, V. HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017, 27, 801–812, doi:10.1101/gr.213462.116.
61. Thorvaldsdottir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Br. Bioinform 2013, 14, 178–192, doi:10.1093/bib/bbs017.
62. Li, L.; Stoeckert Jr., C.J.; Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003, 13, 2178–2189, doi:10.1101/gr.1224503.
63. Rastrojo, A.; Corvo, L.; Lombraña, R.; Solana, J.C.; Aguado, B.; Requena, J.M. Analysis by RNA-seq of transcriptomic changes elicited by heat shock in Leishmania major. Sci Rep 2019, 9, 6919, doi:10.1038/s41598-019-43354-9.
64. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010, 28, 511–515, doi:nbt.1621 [pii]10.1038/nbt.1621.
65. Dillon, L.A.L.; Okrah, K.; Hughitt, V.K.; Suresh, R.; Li, Y.; Fernandes, M.C.; Belew, A.T.; Corrada Bravo, H.; Mosser, D.M.; El-Sayed, N.M. Transcriptomic profiling of gene expression and RNA processing during Leishmania major differentiation. Nucleic Acids Res 2015, 43, 6799–6813, doi:10.1093/nar/gkv656.
66. Martinez-Calvillo, S.; Sunkin, S.M.; Yan, S.; Fox, M.; Stuart, K.; Myler, P.J. Genomic organization and functional characterization of the Leishmania major Friedlin ribosomal RNA gene locus. Mol Biochem Parasitol 2001, 116, 147–57.
67. Lypaczewski, P.; Zhang, W.W.; Matlashewski, G. Evidence that a naturally occurring single nucleotide polymorphism in the RagC gene of Leishmania donovani contributes to reduced virulence. PLoS Negl. Trop. Dis. 2021, 15, e0009079, doi:10.1371/journal.pntd.0009079.
68. Wagner, G.P.; Kin, K.; Lynch, V.J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012, 131, 281–285, doi:10.1007/s12064-012-0162-3.
69. Carter, N.S.; Drew, M.E.; Sanchez, M.; Vasudevan, G.; Landfear, S.M.; Ullman, B. Cloning of a novel inosine-guanosine transporter gene from Leishmania donovani by functional rescue of a transport-deficient mutant. J Biol Chem 2000, 275, 20935–41.
70. da Costa, K.S.; Galucio, J.M.P.; Leonardo, E.S.; Cardoso, G.; Leal, E.; Conde, G.; Lameira, J. Structural and evolutionary analysis of Leishmania Alba proteins. Mol Biochem Parasitol 2017, 217, 23–31, doi:10.1016/j.molbiopara.2017.08.006.
71. Dupe, A.; Dumas, C.; Papadopoulou, B. An Alba-domain protein contributes to the stage-regulated stability of amastin transcripts in Leishmania. Mol Microbiol 2014, 91, 548–561, doi:10.1111/mmi.12478.
72. Mishra, A.K.; Singh, N.; Agnihotri, P.; Mishra, S.; Singh, S.P.; Kolli, B.K.; Chang, K.P.; Sahasrabuddhe, A.A.; Siddiqi, M.I.; Pratap, J. V Discovery of novel inhibitors for Leishmania nucleoside diphosphatase kinase (NDK) based on its structural and functional characterization. J Comput Aided Mol Des 2017, 31, 547–562, doi:10.1007/s10822-017-0022-9.
73. Rascher, C.; Pahl, A.; Pecht, A.; Brune, K.; Solbach, W.; Bang, H. Leishmania major parasites express cyclophilin isoforms with an unusual interaction with calcineurin. Biochem J 1998, 334, 659–67.
74. Oliveira, C.; Carvalho, P.C.; Alves, L.R.; Goldenberg, S. The role of the trypanosoma cruzi TcNRBD1 protein in translation. PLoS One 2016, 11, e0164650, doi:10.1371/journal.pone.0164650.
75. Bates, P.A. The lipophosphoglycan-associated molecules of Leishmania. Parasitol. Today 1995, 11, 317–318.