BMC Genomics

RESEARCH

Open Access

Novel insights into the *Leishmania infantum* transcriptome diversity of protein-coding and non-coding sequences in both stages of parasite development using nanopore direct RNA sequencing

Jean-Guillaume Emond-Rheault[1,2†], Gabriel Reis Ferreira[1,2†], Camille Lavoie-Ouellet[1,2], Martin A. Smith[3,4,5,6] and Barbara Papadopoulou[1,2*]

Abstract

Background *Leishmania* relies on posttranscriptional control to regulate gene expression. Protein-coding genes are synthesised as polycistronic precursors that are processed into individual mRNAs by *trans*-splicing adding the spliced leader (SL) RNA to the 5'-end and 3' cleavage-polyadenylation. Here, we employ Nanopore direct RNA sequencing (DRS) combined with Illumina RNA-Seq to comprehensively interrogate the transcriptomes of *Leishmania infantum* developmental stages at single-molecule resolution.

Results Analysis of DRS full-length reads of poly(A)+-enriched RNA from *L. infantum* developmental stages enabled us to precisely determine the primary SL and poly(A) sites for 52% of the protein-coding transcripts and to accurately define their 5'- and 3'-end and the length of UTRs. In addition, our analysis confirmed the motifs '[C/A/T]A|G' being associated with 94.8% of the SL cleavage sites and better defined the genomic context for cleavage and polyadenylation. Overall, we observed more diversity for poly(A) than SL sites per transcript. The frequency of the primary SL and poly(A) sites was 64.2% and 24%, respectively, with most transcripts having additional poly(A) sites nearby. Alternative polyadenylation was detected in 11-13% of transcripts with ~20% of these having different primary poly(A) sites between promastigote and amastigote developmental stages. Furthermore, DRS uncovered multiple processing events occurring mostly within 3'UTRs, leading to the formation of long non-coding RNAs (lncRNAs). The *L. infantum* transcriptome expresses a rich repertoire of 1,825 lncRNAs, of which 98% were not previously annotated in *L. infantum* and only 21.5% were found in *L. major*. These lncRNAs exhibit generally distinct expression patterns from the 3'UTRs they derived and several are developmentally regulated, representing ~27%

†Jean-Guillaume Emond-Rheault and Gabriel Reis Ferreira contributed equally to this work.

*Correspondence:
Barbara Papadopoulou
barbara.papadopoulou@crchudequebec.ulaval.ca

Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

of the *L. infantum* stage-regulated transcriptome. Their expression was generally higher in amastigotes than in promastigotes, highlighting their importance in parasite intracellular development. Protein prediction tools combined to mass-spectrometry revealed that 7.6% of these lncRNAs have a limited protein-coding potential.

Conclusions This is the first comprehensive transcriptomic analysis of *L. infantum* developmental stages using single-molecule Nanopore DRS. Our findings advance knowledge on existing *Leishmania* expression datasets and provide new insights into the transcriptome complexity and dynamics of both protein-coding and non-coding sequences throughout the parasite development.

Keywords *Leishmania infantum*, Nanopore direct RNA sequencing, Illumina RNA-Seq, *trans*-splicing, Polyadenylation, Alternative RNA processing, Long non-coding RNAs, Stage-regulated gene expression

Background

Protozoan parasites from the *Leishmania* genus cause leishmaniasis, a complex vector-borne disease with an estimate of 700 000 to 1 million new cases annually encompassing a large spectrum of clinical manifestations depending on the species, ranging from self-healing cutaneous lesions to lethal visceral infections [1]. Visceral leishmaniasis (VL) caused by *L. infantum* and *L. donovani* species is the second most lethal tropical and subtropical disease [2]. *Leishmania* spp. are transmitted to the mammalian host as metacyclic promastigotes through the bite of infected sandflies and once inside the macrophage phagolysosome they differentiate into amastigote forms, which are responsible for establishing the disease.

Unlike most other eukaryotes, *Leishmania* and related trypanosomatids lack individual control of transcription initiation, and regulation of gene expression relies heavily on posttranscriptional regulation (PTR) (reviewed in [3, 4]). All mRNAs are synthesised as polycistronic precursors that are co-transcriptionally processed into individual mRNAs by *trans*-splicing and polyadenylation [5–11]. *Trans*-splicing catalyzed by the spliceosome cuts the polycistronic precursor and adds to the 5′-end of all mRNAs a 39-nt miniexon (spliced leader; SL) containing a highly modified 5′-cap structure [12–14]. Similarly to 3′ *cis* splice acceptor sites in other eukaryotes, *trans*-splicing sites in trypanosomatids consist of the consensus AG dinucleotide splice acceptor site preceded by a polypyrimidine tract of variable length [9, 15–19]. Unlike other eukaryotes (reviewed in [13]), no conserved motifs recognized for cleavage and polyadenylation at the 3′-end of mRNAs were found in *Leishmania* transcripts. Previous reports in *Leishmania* [10] and *Trypanosoma brucei* [9, 11, 20] have shown that *trans*-splicing and polyadenylation are temporally and mechanistically coupled, such that the SL acceptor site of the downstream gene determines the location of the polyadenylation site of the upstream gene.

Since the first sequenced genome of *L. major* in 2005 [8], the development of next generation sequencing (NGS) technologies led to a continuously growing number of *Leishmania* species that have their genome sequenced (https://tritrypdb.org/) with several full transcriptomes being already characterized [15, 21–25]. Following the first draft of the *L. infantum* (strain JPCM5) genome sequencing [26] and its improved genome assembly [27], genome-wide expression analysis [28] and full transcriptomes using NGS tools [22, 25] have been reported. Previous studies in *L. major* [15, 23] and *L. donovani* [24] have provided insights into the 5′- and 3′-untranslated region (3'UTR) boundaries of mRNAs and detected widespread alternative processing events.

Long non-coding RNAs (lncRNAs), a class of transcripts longer than 200 nt without protein-coding activity, have emerged in recent years as essential components of various biological processes across all domains of life regulating gene expression at the levels of transcription, pre-mRNA processing, mRNA stability and translation [29–31]. In addition, lncRNAs have been found to sequester miRNAs, control epigenetic processes, form biomolecular condensates, regulate organelles and encode micropeptides [32–36]. Similarly to mRNAs, lncRNAs usually harbor a 5' cap and are polyadenylated at their 3'-end [30, 31]. LncRNAs can be intergenic, antisense or intronic and can also include circular RNAs and *trans*-acting regulatory RNAs derived from 3'UTRs of mRNAs [29, 37, 38]. LncRNAs are often highly cell-specific [39, 40], they affect several cellular functions of great physiological relevance, and changes in their expression are inherent to numerous diseases, notably cancer [31, 41, 42]. Statistics from Human GENCODE release 47 report that the human genome contains 35,934 lncRNA genes and 191,106 loci transcripts, which is twice the amount of protein-coding genes, with other estimates exceeding 100,000 lncRNAs [43]. LncRNAs have also been described in the transcriptomes of *Leishmania* [21, 44, 45] and related *Trypanosoma* species [17, 46, 47]. However, despite some progress, lncRNAs are still not well characterized in *Leishmania*, and little is known about their molecular functions and regulatory roles throughout the parasite development.

In this study, we provide the first comprehensive transcriptomic analysis of *L. infantum* developmental stages at single-molecule resolution using Nanopore direct RNA sequencing (DRS). This analysis enabled us to accurately define the 5'- and 3'-ends of *L. infantum* transcripts, thereby determining the length of UTRs, and to gain further insights into the genomic context favoring pre-mRNA processing and the extent of alternative splicing and polyadenylation during the parasite development. Moreover, DRS analysis substantially advanced our knowledge of the non-coding transcriptome of *L. infantum* developmental stages. We showed that *L. infantum* expresses a rich repertoire of previously unreported lncRNAs that are processed mostly from 3'UTRs through *trans*-spicing and polyadenylation, and several are developmentally regulated or specific to the *L. infantum* transcriptome. Altogether, these findings provide important novel insights into the transcriptome complexity and expression dynamics of protein-coding and non-coding sequences in both stages of *L. infantum* development.

Methods

Leishmania culture

L. infantum (MHOM/MA/67/ITMAP-263) promastigotes were cultured at 25 °C in SDM-79 medium supplemented with 5 µg/ml hemin at pH 7.0 and 10% heat-inactivated Fetal Calf Serum (FCS) (Multicell Wisent Inc., Canada). Differentiation of *L. infantum* promastigotes into amastigote differentiating forms in culture (axenic amastigotes; AxeAMA) was induced by growing 1×10^6 log-phase promastigotes in MAA/20 medium [48] in 25 cm^2 flasks at 37 °C in the presence of 5% CO_2 for three sub-passages prior to RNA and protein extractions.

Oligo (dT) mRNA capture and library preparation

Leishmania infantum log-phase promastigotes (5×10^7 parasites/ml) and passage-3 axenic amastigotes (5×10^7 parasites/ml) were harvested in Trizol (Invitrogen), and total RNA was purified from two independent cultures using RNeasy® Plus Mini Kit (Qiagen, SAS, France). Oligo (dT) based enrichment (New England Biolabs) was used for mRNA purification. Approximately 6 µg of total RNA were used and processed according to the instruction manual. The RNA quality was assessed by Bioanalyzer chips before and after mRNA capture. The Oxford Nanopore Technologies (ONT) sequencing platform was used to generate long reads of the *L. infantum* promastigote and amastigote-like transcriptomes. Libraries were made using the SQK-RNA002 kit (ONT, Oxford, United Kingdom), as per the manufacturer's instructions. Initially, 60 ng of poly(A) + captured mRNA input was used and following an ATP boost, another 70 ng of purified poly(A) + RNA was added.

Nanopore direct RNA sequencing

Nanopore direct RNA sequencing (DRS) was performed on FLO-PRO002 (R9.4.1) flow cells on a high-output PromethION 24 sequencer (ONT, Oxford, United Kingdom). With long Nanopore reads, full-length transcripts can be sequenced end-to-end, facilitating their accurate, unambiguous analysis. Poly(A)+-enriched RNA from two independent *L. infantum* promastigote (P1 and P2) and axenic amastigote samples (A1 and A2) was sequenced with the SQK-RNA002 sample preparation protocol using manufacturer's recommendations. All four samples were processed and analyzed independently (from RNA extraction to poly(A) + mRNA enrichment, library preparation, sequencing, and analysis) to minimize batch-related effects and ensure the comparability of biological replicates. Raw FAST5 were basecalled using Guppy (v6.5.7) [49] (see Fig. 2). Following basecalling, only the reads in the "PASS" folder were used for subsequent analyses. The quality tag (*qc_tag*) "PASS" indicates that the Nanopore reads meet the Nanopolish criteria for containing a poly(A) tail [49]. 'PASS' reads with low q-score (< 7) and short in length (< 50 nt) were removed using NanoFilt (v2.8.0) [50]. Quality of the sequencing, as well as basecalled reads and alignment BAM files were evaluated by NanoStat (v1.5.0) [50]. The quality of the FASTQ basecalled and trimmed reads was assessed using fastQC (v0.12.0). Then, filtered reads were aligned to the *L. infantum* JPCM5 reference genome (v62) available at https://tritrypdb.org/tritrypdb/app/downloads using min imap2 (v2.24) with the parameters set as "-ax map-ont" [51]. Only the primary read alignments were retained for further analysis (see Table 1).

Analysis of poly(A) and full-length reads

The poly(A) length was evaluated using the poly(A) module of the Nanopolish software (v0.14.0) based on the ONT raw signals (FAST5) and the aligned reads (BAM) on the *L. infantum* JPCM5 genome (v62) (see Fig. 2) [52]. Next, ONT reads with the quality control tag (*qc_tag*) labeled as 'PASS' and a poly(A) length ≥ 20 A's, as determined by Nanopolish [49], were extracted from the aligned ONT reads BAM files by an in-house Python script utilizing samtools (v1.18) and the parameters view -b -N -F 2048 -q 1. Subsequently, the aligned poly(A) reads in BAM files were converted into FASTA files using samtools (v1.18). Reads were considered as 'Full-Length' when harbouring the SL sequence at the 5'-end and the poly(A) tail at the 3'-end. The 39-nt SL sequence (5'-AACTAACGCTATATAAGTATCAGTTTCTGTACTTTATTG-3') was assessed in aligned poly(A) reads in FASTA files using BLASTn (v2.14.1+) [53] with the following

Table 1 Oxford nanopore direct RNA sequencing reads of *L. infantum* promastigotes (P1, P2) and axenic amastigotes (A1, A2)

		P1	P2	A1	A2
Sequenced	# reads	6,924,673	6,304,455	6,791,066	5,251,284
	Total bases (Mb)	5,395.3	4,979.0	6,157.0	4,795.1
	Coverage (X) [a]	164	152	188	146
	Median read length (nt)	666	673	753	761
Basecalled & filtered	# reads	6,813,287	6,202,325	6,666,603	5,161,219
Aligned	# reads	6,604,760	6,024,113	6,402,659	4,981,118
	Primary alignment	6,602,769	6,022,563	6,400,153	4,979,826
	Secondary alignment	1,280,232	1,239,698	1,078,033	833,446
	Supplementary alignment	1,986	1,543	2,476	1,266
	Total bases (Mb)	4,969.1	4,586.2	5,667.8	4,429.8
	Coverage (X) [a]	152	140	173	135
	Median read length (nt)	684	689	781	785
	% reads aligned [b]	95,4	95,6	94,3	94,9

[a] # of total bases/*L. infantum* JPCM5 genome size (32,803,248 bp)

[b] % was calculated based on the number of sequenced reads

parameters: -task "blastn-short" -max_target_seqs "1000000" -outfmt "6 std qlen slen qcovs". The poly(A) reads were considered to harbor the SL when ≥ 15 nt long of the SL sequence was mapped in the first 40 nt of the reads with an identity of ≥ 90%. The aligned poly(A) reads containing the SL sequence were next retained by samtools (v1.18) (view -b -N) to generate the aligned read BAM files (Fig. 2, green panel). The BAM files (poly(A) and full-length) were converted into PAF files using the htsbox tool (vr345). The positions of the *L. infantum* JPCM5 protein-coding genes (n = 8,527) included in the GFF (v62) annotation file were used to determine whether the poly(A) or full-length reads overlapped completely or partially with one or more CDS. Poly(A) or full-length reads were classified as monocistronic when overlapping a single CDS and as polycistronic when they overlapped at least two CDSs. For full-length reads not aligning on a CDS, a second analysis was conducted using the localization of the 5'- and 3'-UTRs predicted previously by PRED-A-TERM (Fig. 2, green panel) [16]. Five potential alignments of full-length reads were assessed: i) within the 5'UTR; ii) within the 3'UTR; iii) partially overlapping the 5'UTR; iv) partially overlapping the 3'UTR, and v) located in the intergenic region (IR).

Detection of *trans*-splicing and polyadenylation sites

ONT alignments from both biological replicates of *L. infantum* promastigotes and axenic amastigotes were pooled to assess the SL acceptor or polyadenylation sites. We used the read mapping coordinates on the reference genome *L. infantum* JPCM5 (v62) to annotate each full-length read based on the *L. infantum* JPCM5 GFF file (v62) (see Fig. 2). The start and end alignment coordinates on the JPCM5 reference genome (v62) were considered as SL and poly(A) sites. The SL acceptor sites were determined using the monocistronic aligned full-length reads for CDSs covered by ≥ 5 reads. The poly(A) sites were assessed by the 3'-end alignment position for a CDS completely or partially overlapping ≥ 5 poly(A) sequences using the same strategy applied for full-length reads by an in-house Python script (Fig. 2, blue panel). The most prevalent SL acceptor or poly(A) sites were termed as primary sites and sequences surrounding the primary sites were extracted using 'samtools faidx'. Specific motifs surrounding the cleavage sites were generated with the webserver WebLogo (https://weblogo.berkeley.edu/logo.cgi). The distance (nt) of the nearest SL ad dition site downstream with ≥ 5 full-length reads for each poly(A) site was determined by a Python script using the full-length reads start and end alignment positions. Only *trans*-splicing sites in a 1,500 nt window from the poly(A) site were considered. The same analysis was done for the closest poly(A) site upstream (≥ 5 full-length reads) of each SL site using also a 1,500 nt window. Alternative polyadenylation was defined when more ONT-DRS poly(A) reads were mapped upstream or downstream of the primary polyadenylation site at ± 75 nt. For genes covered by ≥ 5 reads and presenting stage-regulated alternative poly(A) or SL sites, the primary poly(A) or SL sites had to be conserved in both biological replicates and located at least 100 nt apart between promastigotes and axenic amastigotes.

Analysis of long non-coding RNAs

Full-length reads aligning to annotated 3'UTR regions but not to the CDS were assessed as ncRNAs. Briefly, each transcriptome was analyzed independently to identify primary SL and poly(A) sites within the 3'UTR with ≥ 5 full-length reads using an in-house Python script. The presence of the SL sequence was confirmed by Illumina RNA-Seq data generated from two *L. infantum* promastigote and axenic amastigote poly(A) + RNA-seq paired-end runs in triplicates (BioProject PRJNA1071080). Briefly, following the alignment process of the Illumina RNA-Seq to the *L. infantum* JPCM5 genome (v62) described in a section below, Illumina reads covering the SL addition site identified with ONT-DRS were extracted using samtools faidx (v1.18), converted into FASTA with seqtk seq (v1.2-r94), and SL sequence was searched using

BLASTn (v2.14.1+; option: -task 'blastn-short'). Illumina reads were considered to harbor the SL sequence when BLASTn alignment showed a perfect match of ≥8 nt in the first 30 nt of the miniexon. The conservation of the newly identified *L. infantum* ncRNAs in *L. major* was evaluated using the RNA-Seq data from SRA BioProject PRJEB27042 [54]. These RNA-Seq data were trimmed and aligned on the *L. infantum* JPCM5 reference genome (v62) to obtain the same SL position using BBMap default parameters (v.38.86). To be considered as the SL sequence, IGV should show a consecutive alignment of ≥7 nt starting from the 5'-end. Python library orffinder 1.8 was used to predict proteins encoded by lncRNAs (≥ 25 aa) (https://pypi.org/project/orffinder/). Predicted proteins encoded by lncRNAs were aligned using BLASTp against the UniProtKB (the reviewed database was downloaded at 10 h30 on 2024-04-19 from https://www.uniprot.org/help/downloads). Alignments with a bit score >50 were considered as significant [55]. The coding potential of each lncRNA was evaluated with the CPC web server [56]. The transcriptome of the *L. major* Friedlin strain 2021 [23] (release 68, available at https://tritrypdb.org/common/downloads/release-68/Lmajor/fasta) was used to investigate the conservation of *L. infantum* lncRNAs across other *Leishmania* species by comparing both transcriptomes using BLASTn (v2.14.1+). The criteria for considering lncRNAs homologous between *L. infantum* and *L. major* were ≥80% sequence identity within ≥75% coverage and ≤30% of difference in length using BLASTn. The number of lncRNAs contained in significant longer *L. major* transcripts was further determined by dropping the length difference threshold.

CRISPR-Cas9 based genome editing

L. infantum JPCM5 promastigotes expressing the vector pSP72αHYGαCas9 [57] were cultured in SDM-79 medium containing 10% heat-inactivated FBS, 5 µg/ml hemin and 20 µM biopterin. gRNAs targeting the end of selected putative protein-coding lncRNAs were designed using the Eukaryotic Pathogen CRISPR gRNA Design Tool [58] and verified with CRISPOR tool [59]. The primers used for genome editing are listed in Additional File 1: Table S11. The donor DNA introducing an HA-tag before the TGA stop codon was generated by PCR from *L. infantum* JPCM5 genomic DNA and 5 µg were co-transfected with 50 µM tracrRNA: gRNA (IDT) duplex in 5×10^7 cells using Amaxa-Nucleofector™ transfection kit (Lonza). The next day, cells were selected with hygromycin (80 µg/ml). Clones were next isolated from SDM-79 plates, resuspended in liquid SDM-79 and genomic DNA was extracted. Clones were PCR-screened for the presence of the HA-tag and positive PCRs were sequenced by Sanger. Next, proteins from 1×10^8 cells were extracted in Laemmli buffer and a Western blotting analysis was performed using an anti-mouse hemagglutinin (HA) tag monoclonal antibody (1:3000; Applied Biological Materials, Canada). The protein length was predicted with the Expasy Compute pI/Mw server [60].

Liquid chromatography-mass-spectrometry analysis

Protein digestion and LC-mass spectrometry analysis were performed by the Proteomics Platform of the CHU de Québec Research Center (Quebec, Qc, Canada) as described previously [61], with minor modifications. Three biological replicates of total protein lysates from *L. infantum* promastigotes were used for this analysis (experiments LP202, DDA010 and RD196-03; see Additional File 3). The LC-M/MS experiment RD196-03 of total protein lysates from *L. infantum* promastigotes was conducted on an Orbitrap Astral mass-spectrometer and resulted in the identification of ~91.9% (7,835/8,527) of the *L. infantum* proteome. Peptide and protein identifications were filtered at 1% False Discovery Rate (FDR) and only proteins identified with at least 2 peptides were considered for DE analysis. The z-score, the Limma *p*-value and q-value (Benjamini Hochberg adjusted *p*-value) were calculated. The following filters were applied to define differentially expressed proteins: q-value <0.05 and |z-score| > 1.96. The protein intensity ranking was calculated for each LC-MS/MS experiment by sorting protein intensities in decreasing order, with the most intense proteins ranked first.

Differential expression and statistical analyses

Illumina RNA-Seq paired-end runs in triplicates from *L. infantum* promastigote and axenic amastigote samples (BioProject PRJNA1071080 [62]) were pooled, trimmed using Trimmomatic (v0.39), and aligned to the *L. infantum* JPCM5 genome (v62) using the software HISAT2 (v2.2.1) [63]. On average, a total of 67.8 million RNA-Seq reads per sample were used to evaluate the transcriptome expression. Expression levels were quantified from the HISAT alignment with the software StringTie (v2.1.3) [64] based on the *L. infantum* JPCM5 (v64) transcriptome available from https://tritrypdb.org/tritrypdb/app/downloads to which the lncRNA coordinates were added. Differential gene expression analysis between *L. infantum* promastigotes and amastigotes was evaluated using the DESeq2 [65]. Protein-coding and non-coding transcripts were considered as DE when the log2 fold change (log2 FC) was ≥1 or ≤ −1 and the adjusted *p*-value ≤0.05. Gene expression levels were normalized using FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values calculated with the DESeq2 library. All statistical analyses were performed using RStudio (v2023.6.2.561; RStudio Team 2023).

Results

Oxford Nanopore direct RNA sequencing of *L. infantum* promastigote and amastigote developmental stages

In this study, we used Nanopore direct RNA sequencing (DRS) with the aim of better characterizing the boundaries of *Leishmania* transcripts and comprehensively investigating alternative pre-mRNA processing events in *L. infantum* promastigote and amastigote life stages. The Oxford Nanopore Technologies (ONT) long-read sequencing platform represents a third-generation sequencing technology that uses protein nanopores to detect changes in the current of RNA molecules as they pass through the nanopore [66]. Key advantages of ONT sequencing are its ability to observe full-length native RNA molecules and effectively capturing a range of transcriptomic properties including unbiased expression, poly(A) tail length, and RNA modifications. For this analysis, we used poly(A)+-enriched RNA from *L. infantum* promastigotes (PRO) and parasites subjected to amastigote differentiation in culture (AxeAMA). Two biological replicates of PRO (P1 and P2) and AxeAMA (A1 and A2) were subjected to DRS using the DRS kit (SQK-RNA002) on the ONT PromethION platform. A total of 6,924,673 and 6,304,455 reads was obtained for P1 and P2 samples and 6,791,066 and 5,251,284 reads for A1 and A2 samples, respectively (Fig. 1A; Table 1). After basecalling with Guppy and filtering out low-quality and short reads using NanoFilt, 94–96% of the sequenced reads successfully aligned to the reference genome of *L. infantum* JPCM5 v62 using minimap2, with a theoretical genome coverage ranging from 135X to 173X (Fig. 2; Table 1). Approximately 15.9–19.7% of the reads with primary alignment exhibited a secondary alignment, indicative of repetitive or homologous sequence content (Fig. 1B). The AxeAMA transcriptomes showed a lower frequency of secondary alignment (15.9%) compared to the PRO transcriptomes (18.5–19.7%) (Fig. 1B). A detailed summary of ONT sequencing metrics, read processing and read analysis results generated in this study are provided in Additional file 1: Table S1.

The longest ONT reads sequenced and successfully aligned to the reference genome for each transcriptome varied between 10,886 nt (P1) to 14,910 nt (P2). Median lengths for AxeAMA reads were slightly longer (781 nt and 785 nt) than for PRO reads (684 nt and 689 nt) (Table 1), with both biological replicates exhibiting the same trend (Table 1 and Additional file 2: Fig. S1A). The read's frequency across the four transcriptomes was calculated for each 50 nt interval. As depicted in Additional file 2: Fig. S1B, reads in AxeAMA transcriptomes were generally longer than 850 nt compared to PRO transcriptomes where reads were in average between 300 nt and 850 nt. This difference in reads length may be attributed to inherent biological variations between life stages, such as alternative RNA processing events.

Defining primary polyadenylation and spliced leader addition sites of *L. infantum* transcripts based on aligned poly(A) and full-length ONT-DRS reads

As the result of polycistronic transcription, synthesis of mature mRNAs in *Leishmania* requires two pre-mRNA processing reactions, namely *trans*-splicing adding the 39-nt spliced leader (SL) RNA sequence to the 5'-end of all known protein-coding RNAs [67, 68]

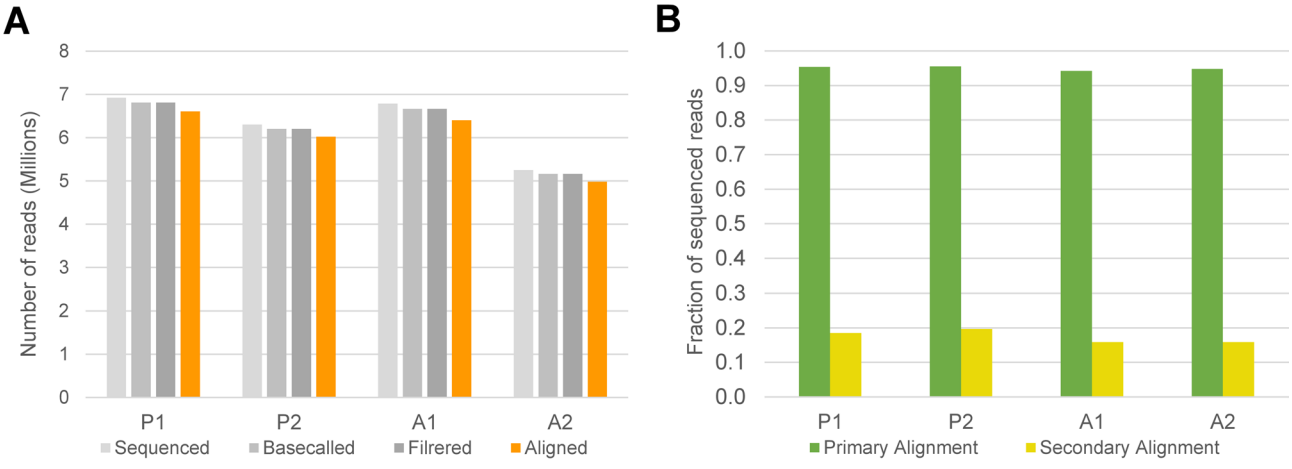

Fig. 1

Fig. 1 Oxford Nanopore direct RNA-sequencing runs of *L. infantum* promastigote and amastigote life stages. **A** Total number of reads analyzed for *L. infantum* promastigote (P1, P2) and axenic amastigote (A1, A2) transcriptomes. **B** Fraction of sequenced reads with primary or secondary alignment to the *L. infantum* JPCM5 reference genome. Reads with supplementary alignment account for 0.024–0.036%, rendering them invisible in the histogram

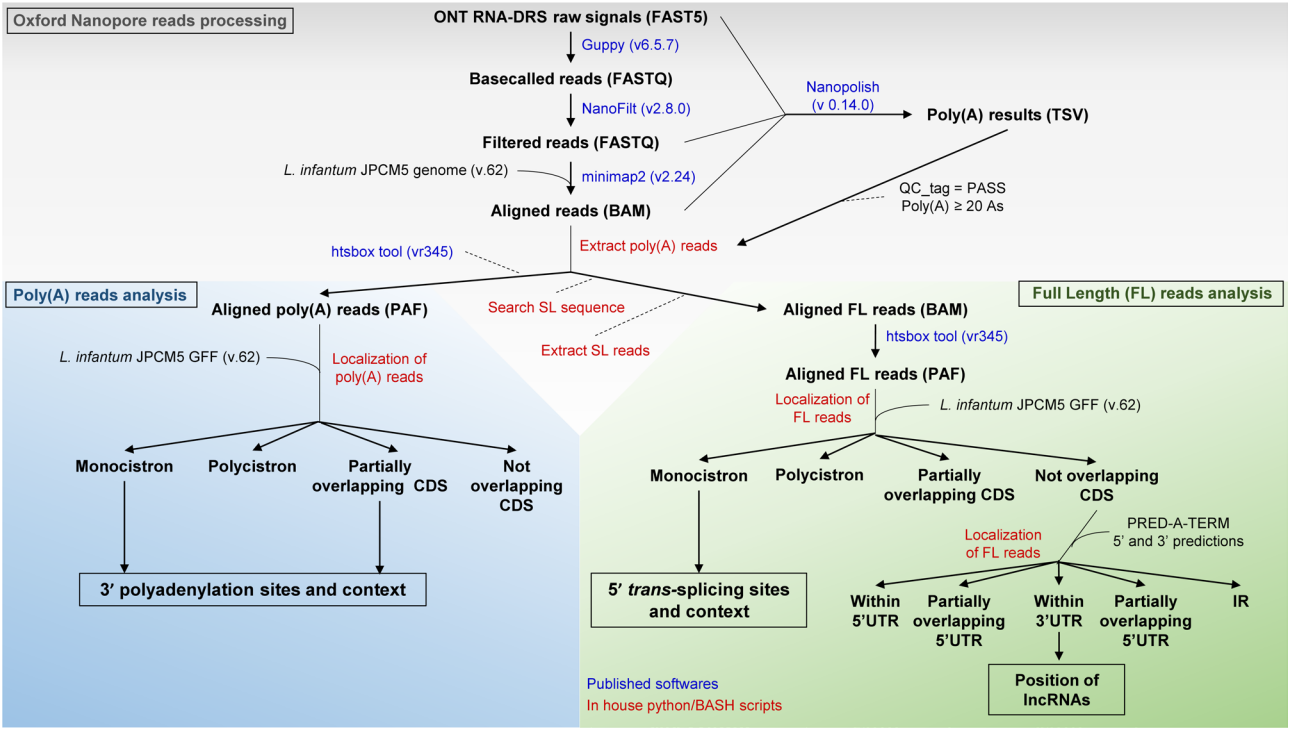

Fig. 2 Schematic representation of the workflow used for Nanopore direct RNA sequencing data analysis. Raw ONT-DRS reads were analyzed using published Nanopore softwares to generate an alignment on the *L. infantum* JPCM5 reference genome and calculate the poly(A) tail lengths (section in gray). Two paths were taken depending on the question being addressed. Poly(A) sites and the polyadenylation context (section in blue) were analyzed using the aligned poly(A) reads by an in-house Python script. To determine the *trans*-splicing sites and context (section in green), reads containing both poly(A) tails and the SL RNA sequence were analyzed (Full-length reads). Full-length reads were also used to assess the positions of lncRNAs in 3'UTRs predicted by PRED-A-TERM

and 3' cleavage-polyadenylation. Here, we exploited the signal sequences for *trans*-splicing and polyadenylation to accurately map the 5'- and 3'-ends of *L. infantum* transcripts by comparing reads containing these signals to the reference genome sequence (see Fig. 2). Our analysis first focused on mapping the 3'-end of *Leishmania* transcripts. DRS of poly(A)+-enriched RNA enables improved poly(A) tail determination, as intact mRNA molecules are brought to the sequencing pore by the poly(A) tail and sequenced from the 3'- to the 5'-end. Poly(A)-tail lengths were calculated using Nanopolish's raw signal-based poly(A) length estimation [52]. Only reads with a poly(A) tail flagged as 'PASS' from Nanopolish software were retained for further analysis (see Methods). The mean and median poly(A) tail lengths were similar between the four transcriptomes, with mean values ranging from 77–78 adenines (A's) for promastigotes and 78–81 A's for amastigotes (Additional file 2: Fig. S2).

To assess the polyadenylation site and 3'UTR length of *L. infantum* transcripts, we analyzed 3'-end alignment positions on the reference genome using ONT-DRS reads with a poly(A) tail (≥ 20 A's) from combined biological replicates for each life stage (P1 + P2 and A1 + A2) that were completely or partially overlapping a CDS (see Fig. 2 and Methods). Based on these criteria, we identified 8,213 transcripts in PRO and 8,217 transcripts in AxeAMA fully or partially covered by ≥ 5 poly(A) reads out of 8,704 protein-coding transcripts annotated in the *L. infantum* JPCM5 genome (Table 2). Of these transcripts, 4,566 displayed the same primary poly(A) site across the four transcriptomes (Table 2 and Additional file 1: Table S2). The frequency of the primary poly(A) site was estimated at 24% (Additional file 1: Table S2), with most transcripts having additional poly(A) sites near the primary site. The average 3'UTR length estimated at 499 nt was calculated by measuring the distance from the CDS stop to the primary poly(A) site in all four transcriptomes (Table 2 and Additional file 1: Table S2). The smallest 3'UTRs ranged in length from 20 to 100 nt and the largest ones from 2,400 to 3,900 nt (Additional file 1: Table S2).

To determine the SL addition site for each *L. infantum* transcript, we searched for the miniexon sequence in the aligned poly(A) reads (≥ 20 A's) of combined PRO and AxeAMA biological replicates using BLASTn (see Fig. 2 and Methods). Reads harboring both the SL RNA and poly(A) sequences, referred as full-length reads, represented 11.2%–12.8% of all sequenced reads (Table 2 and Fig. 3A), reflecting the known difficulty

Table 2 Analysis of the ONT direct RNA aligned reads of *L. infantum* promastigotes (P1, P2) and amastigotes (A1, A2) to search for the spliced leader (SL) and poly(A) sequences

			P1	P2	A1	A2
Aligned	# reads		6,604,760	6,024,113	6,402,659	4,981,118
	% reads aligned [a]		95.4	95.6	94.3	94.9
Poly(A) tail (≥ 20 As)	# reads		4,530,839	4,244,590	4,407,812	3,403,132
	Total bases (Mb)		3,448.4	3,265.5	3,946.7	3,066.2
	Coverage (X) [b]		105	100	120	93
	Median read length (nt)		701	706	804	811
	% reads poly(A) [a]		65.4	67.3	64.9	64.8
	# transcripts ≥ 5 poly(A) reads		8,213 [c]		8,217 [c]	
	Total transcripts with the same primary poly(A) site among all four transcriptomes		4,566			
	Average 3'UTR length (nt) [d]		499			
Poly(A) tail and SL (Full-length)	# reads		773,496	761,908	868,119	660,056
	Total bases (Mb)		644.3	647.7	833.3	643.3
	Coverage (X)		19.6	19.7	25.4	19.6
	Median read length (nt)		780	790	879	897
	% Full-length reads [a]		11.2	12.1	12.8	12.6
	# transcripts ≥ 1 full-length reads		6,966		7,604	
	# transcripts ≥ 5 full-length reads		5,418		6,429	
	Total transcripts with the same primary SL among all four transcriptomes		4,487			
	Average 5'UTR length (nt) [e]		213			
	Monocistrons	#	587,685	578,312	588,564	453,993
		% [f]	76.0	75.9	67.8	68.8
	Polycistrons	#	1,365	1,431	3,865	2,906
		% [f]	0.176	0.188	0.445	0.440
	Partially overlapping CDS	#	26,944	26,298	29,809	22,986
		% [f]	3.48	3.45	3.43	3.48
	Within 3'UTR	#	80,483	78,104	137,247	99,385
		% [f]	10.4	10.3	15.8	15.1
	Partially overlapping 3'UTR	#	25,786	24,155	45,321	33,314
		% [f]	3.33	3.17	5.22	5.05

[a] % was calculated based on the number of sequenced reads

[b] # of total bases/*L. infantum* JPCM5 genome size (32,803,248 bp)

[c] 94% of the *L. infantum* transcriptome (8,748 genes)

[d] Average 3'UTR length calculated from Additional file 1: Table S2

[e] Average 5'UTR length calculated from Additional file 1: Table S4

[f] % was calculated based on the number of reads harboring both poly(A) and SL sequences (full-length reads)

of DRS to accurately capture 5' ends. Most of the full-length reads (68–76%) were monocistronic spanning a single coding sequence (Table 2 and Additional file 2: Fig. S3A). Only an average of 1,318 reads were polycistronic (predominantly bicistronic), representing 0.18–0.44% of the aligned reads (Table 2; Fig. 3B). The remaining full-length reads did not overlap or only partially a CDS and were mostly aligned within 3'UTRs (Table 2). We detected in total 7,719 transcripts with at least one full-length read in all transcriptomes combined (Additional file 1: Table S3), from which 5,418 and 6,429 with ≥ 5 full-length reads/transcript in PRO and AxeAMA transcriptomes, respectively (Table 2). Of these transcripts, 4,487 shared the same primary SL site (most prevalent) among promastigote and amastigote transcriptomes (Table 2 and Additional file 1: Table S4), accounting for 51.5% of all *L. infantum* protein-coding genes. The frequency of the primary SL site was estimated at 64.2%, however a secondary site was detected at 57 nt apart in 99.6% of the transcripts, but with a much lower frequency (Additional file 1: Table S4).

The average 5'UTR length was calculated by measuring the distance from the primary SL addition site shared by 4,487 transcripts to the CDS start (Additional file 1: Table S4). We detected a wide range of 5'UTR lengths varying from 1 to 2,000 nt with median and mean lengths of 162 nt and 213 nt, respectively

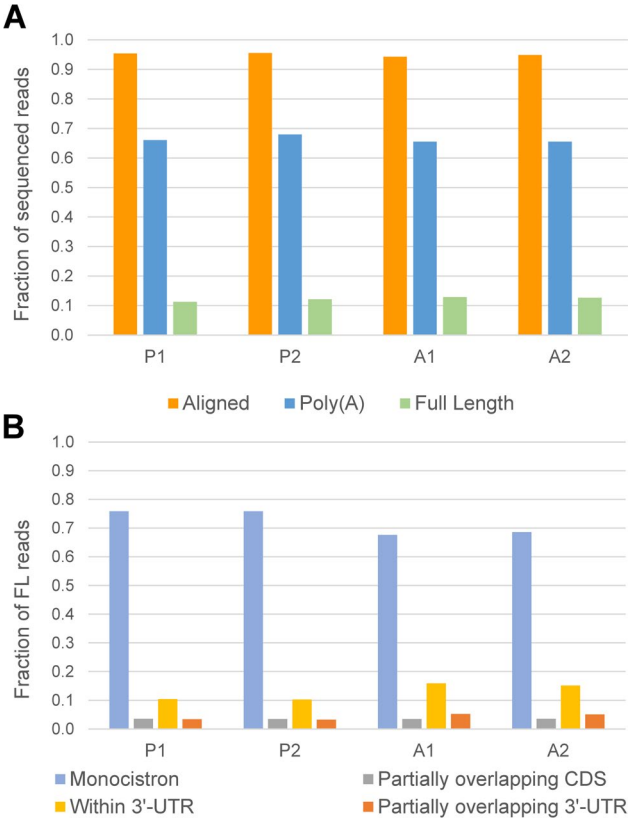

Fig. 3 Assessment of the spliced leader addition and polyadenylation sites by Nanopore direct RNA sequencing aligned reads for the *L. infantum* promastigote and axenic amastigote transcriptomes. **A** Fractions of sequenced reads containing either a poly(A) or both poly(A) and SL sequences (full-length reads). **B** Histogram depicting the four most represented groups of ONT-DRS full-length reads

(Table 2). These values closely match the 181–195 nt reported previously for 7,987 transcripts using the PRED-A-TERM software [16]. Indeed, comparative analysis of 5'UTR lengths estimated from ONT-DRS full-length reads vs. PRED-A-TERM predictions showed equivalent 5'UTR lengths in >60% of the transcripts (R^2 = 0.20) (Additional file 2: Fig. S4A). A visual inspection of native ONT-DRS read alignments for selected 5'UTRs of variable lengths using IGV confirmed these predictions (see Additional file 2: Fig. S4 B-D).

Alternative splicing and polyadenylation of *L. infantum* transcripts in promastigote and amastigote developmental stages

To assess the extent of alternative polyadenylation sites among *L. infantum* transcripts, we specifically looked for ONT-DRS reads (> 5 full-length reads/transcript) aligning to a CDS and ending at 75 nt upstream or downstream of the primary poly(A) site (see Methods). On average, 86.8% of all ONT-DRS poly(A) reads aligned into this 150 nt window. Alternative polyadenylation was detected in 11.1–13.3% of transcripts across all four transcriptomes (Additional file 1: Table S5 and Additional file 2: Fig. S5), thus leading to a higher 3'UTR length diversity.

Next, we compared the promastigote and amastigote transcriptomes using the Illumina RNA-Seq data of *L. infantum* PRO and AxeAMA from a previous study (see Methods) to specifically look at differences in the primary poly(A) or SL sites of transcripts between PRO and AxeAMA and whether these differences were associated with differential gene expression (DE). For the analysis of poly(A) sites, we considered transcripts with poly(A) sites conserved across ONT-DRS poly(A) reads in either promastigote or amastigote biological replicates and located at ≥ 100 nt from the primary poly(A) site. Based on these criteria, we identified 159 transcripts with evidence of alternative polyadenylation between promastigotes and amastigotes, of which 18 were upregulated and 26 downregulated in AxeAMA vs. PRO (Additional file 1: Table S6). This number is likely an underestimate due to the diversity of poly(A) sites for the same transcript, even between biological replicates. Examples of stage-alternatively polyadenylated transcripts are illustrated in Fig. 4A, B and Additional file 2: Fig. S6 A-D. Next, we searched for alternative splicing events between promastigote and amastigote life stages occurring within a region of ~100 nt away from the primary SL site. We identified 147 transcripts with different primary SL sites between PRO and AxeAMA (Additional file 1: Table S7), of which 19 were upregulated and 20 downregulated in AxeAMA vs. PRO (log2 fold change >0.58). Examples of stage-alternatively spliced transcripts are shown in Fig. 4C, D and Additional file 2: Fig. S6E-H.

Comparison with the whole *L. infantum* stage-regulated transcriptome (up- or down-regulated transcripts between promastigotes and amastigotes) indicated no enrichment of alternatively polyadenylated transcripts among the developmentally regulated transcripts (X^2 (1, N = 8,704) = 0.86, p = 0.35). Similarly, no enrichment of DE alternatively spliced transcripts was observed compared to the rest of the *L. infantum* stage-expressed transcriptome (X^2 (1, N = 8,704) = 0.32, p = 0.57).

Revisiting the genomic context favoring *tran*-splicing and polyadenylation based on full-length ONT-DRS reads

The context for *trans*-splicing in kinetoplastids has been reported previously to require a stretch of polypyrimidines (CT tracts) preceding the AG splice acceptor site where cleavage occurs by the *trans*-splicing machinery to add the 39-nt miniexon sequence [12–14]. To determine the context for *trans*-splicing in *L. infantum* transcripts, we calculated the nucleotide prevalence of CT/GC motifs within 200 bases upstream and downstream of the primary SL addition site shared by 4,487 transcripts

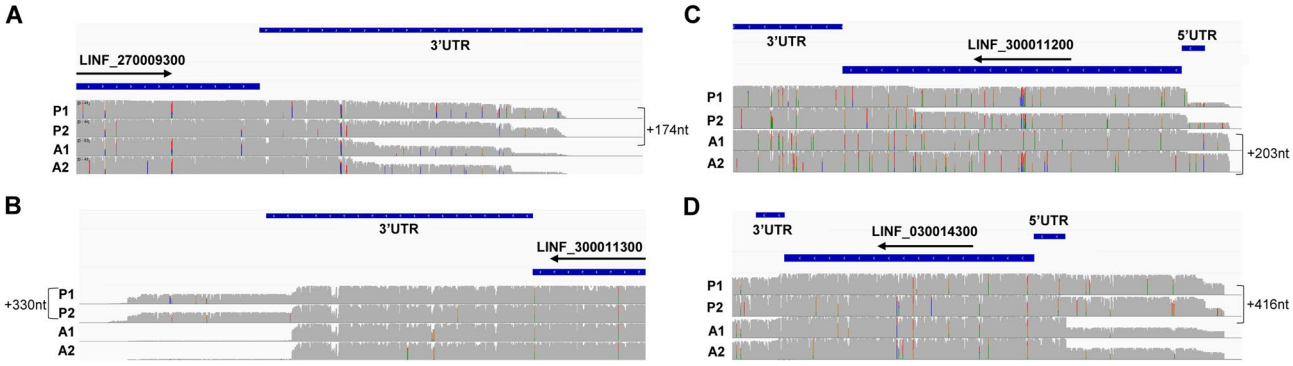

Fig. 4 Examples of stage-regulated alternatively polyadenylated or alternatively spliced *L. infantum* transcripts. The histogram coverages at the 3'-end of the ONT-DRS poly(A) reads from four transcriptomes (P1, P2, A1 and A2) show alternative polyadenylation of LINF_270009300 (**A**) and LINF_300011300 (**B**). Two examples of alternative *trans*-splicing are represented by histogram coverage at the 5'-end for each transcriptome using the full-length reads for LINF_300011200 (**C**) and LINF_030014300 (**D**). Black arrows indicate the direction of transcription

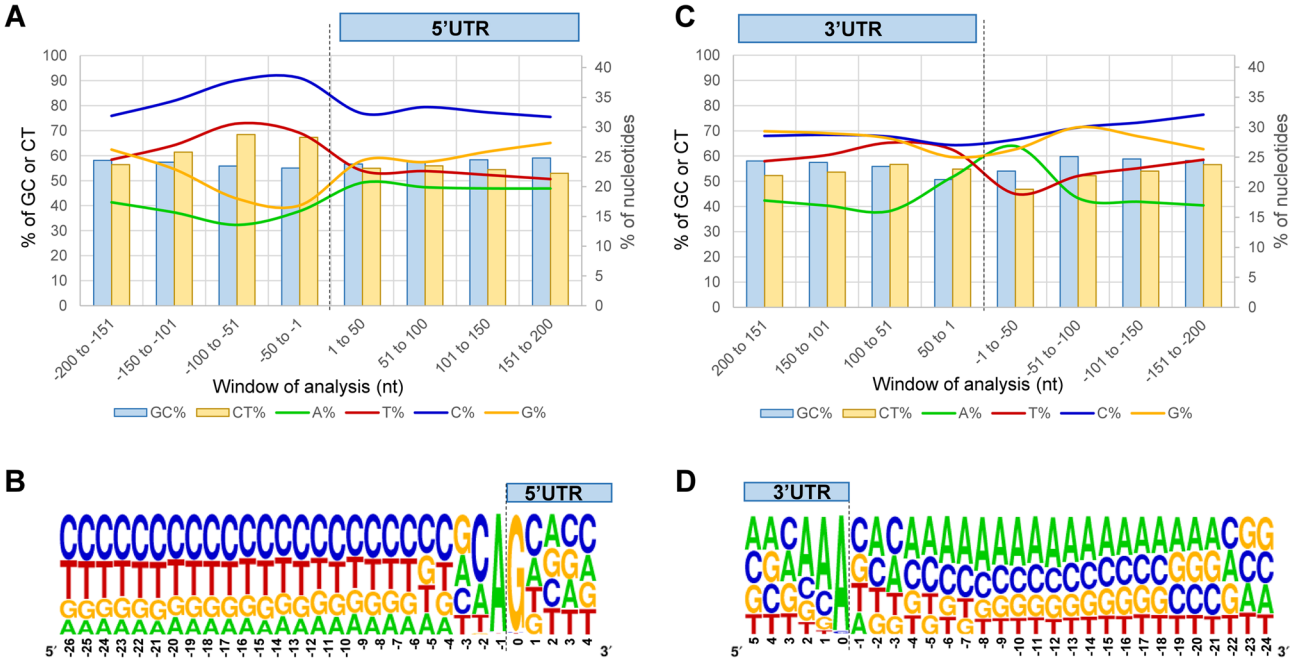

Fig. 5 Genomic context for the spliced leader (SL) and poly(A) addition sites. The percentage of each nucleotide, as well as the GC and CT contents were calculated for 200 bp upstream and downstream of the SL cleavage sites (**A**) and (**C**) poly(A) sites for 4,487 and 4,566 *L. infantum* transcripts, respectively. The genomic context associated with the cleavage sites for SL (**B**) and poly(A) addition (**D**) was generated using WebLogo (https://weblogo.berkeley.edu/logo.cgi). Cleavage occurs between the position 0 and – 1

in both life stages using a 50 nt window (Fig. 5A). Our analysis revealed an enrichment of pyrimidines (67.9%) and a decrease in the GC content (55.5%) within the first 100 nt upstream of the SL acceptor site (Additional file 1: Table S4), contrasting with averages of 50.3% and 59.8% across the JPCM5 genome, respectively. These observations are consistent with previous studies in *L. major* [10]. The 'CA|G' motif (where '|' indicates the cleavage site) accounted for 55.6% of the transcripts, followed by 'AA|G' (25.0%) and 'TA|G' (14.2%) (Fig. 5B and Additional file 1: Table S4). Only 62 'GA|G' motifs were annotated, suggesting that the presence of a guanine upstream of the AG splice acceptor site drastically reduces the probability of cleavage by the *trans*-splicing machinery. Overall, the motif '[C/A/T] A|G' was associated with 94.8% of the SL cleavage sites (Additional file 1: Table S4). While there may be other less frequent motifs for SL addition, it is most likely that the absence of the '[C/A/T] A|G' motif in the 5.2% remaining CDSs is due to full-length read misalignment attributed to genomic complexity.

In *Leishmania* like in other kinetoplastidae, the context for poly(A) addition is yet poorly defined. Here, we examined the primary poly(A) positions from 3'UTRs

delimited by ONT-DRS poly(A) reads in 4,566 transcripts to better define the context for polyadenylation (Fig. 2 and Additional file 1: Table S2). Our analysis revealed a decrease in the GC content 50 nt upstream of the primary poly(A) site, followed by an enrichment of adenines and a thymine decrease within the 50 nt downstream (Fig. 5C). Although no consensus motif was found surrounding the poly(A) site, polyadenylation often occurred after a stretch of adenines (> 3 A's) (Fig. 5D).

To evaluate the average distance between the poly(A) site of an upstream transcript and the SL addition site of a transcript downstream, we looked for the poly(A) position relative to the nearest SL site (≥ 5 full-length reads) in transcripts sharing the same primary poly(A) site in all transcriptomes examined (see Methods). On average, we found that this distance varied between transcriptomes from 379 nt to 387 nt (P1: 387 nt, P2: 383, A1: 379 and A2: 382) (Additional file 1: Table S2). A reverse analysis looking for the distance between the SL site of a downstream transcript and the nearest primary poly(A) site upstream similarly estimated it at ~360 nt. Previous studies in *L. major* [10] and *Trypanosoma* species [67] established this distance at ~400–500 nt.

The *L. infantum* transcriptome expresses a rich repertoire of long non-coding RNAs derived mostly from 3'UTRs

While exploring the ONT DRS data, it became apparent that many processed reads in the *L. infantum* transcriptome did not overlap CDS regions, suggesting the presence of unannotated non-coding RNAs (ncRNAs). Therefore, we searched for ncRNAs by identifying full-length ONT-DRS reads that do not overlap a CDS, taking advantage of the 5'UTR and 3'UTR positions predicted by PRED-A-TERM [16]. Our data indicated that a significant proportion of these reads were located within predicted 3'UTRs (Table 2). While 3'UTRs predicted by PRED-A-TERM were initially used for this analysis, 3'UTR annotations were robustly supported by ONT-DRS data that allowed us to accurately map the poly(A) site of ~8,200 out of the 8,704 *L. infantum* transcripts and the 3'UTRs for 4,566 transcripts (~ 52% of the *L. infantum* genome) (see Table 2 and Additional file 1: Tables S2 and S5). To determine the boundaries of the identified ncRNAs, we processed the four transcriptomes independently to retain loci with a minimum of 5 full-length reads per 3'UTR and the established primary SL and poly(A) sites by ONT-DRS (Additional file 1: Tables S2 and S4). Using this approach, we identified 1,892 ncRNAs processed from 3'UTRs, including 67 short ncRNAs (sncRNAs; <200 nt) and 1,825 lncRNAs (> 200 nt) with mean and median lengths of 843 nt and 655 nt, respectively (Additional file 1: Table S8). This study focuses on lncRNAs. Among the 1,825 identified lncRNAs, only 39 (2.1%) were already annotated in the *L. infantum* JPCM5 (V62) genome: 20 as pseudogenes and 19 as ncRNAs (Table 3 and Additional file 1: Table S8). Figure 6A, B illustrates the aligned ONT reads for two selected lncRNAs located in 3'UTR.

To validate whether the identified lncRNAs all harbor the SL sequence at their 5'-end, we took advantage of the Illumina RNA-Seq data of three biological replicates for *L. infantum* PRO and AxeAMA. We extracted Illumina RNA-Seq reads overlapping the lncRNA SL addition site using samtools, converted them to FASTA with seqtk and then searched for the SL sequence (see Methods). The presence of the SL sequence was confirmed for 82% of the lncRNAs (Additional file 1: Table S8). Figure 6C, D illustrates the conservation of SL sequence between the ONT and Illumina reads for two selected lncRNAs. The difficulty of aligning Illumina short reads containing the SL sequence to the reference genome may be the reason why we were unable to confirm the SL site in the remaining lncRNAs. The context for SL addition in lncRNAs was predominantly CA|G with a CT percentage increasing up to 68.3% within a region of 100 nt upstream of the SL site (Additional file 2: Fig. S7 A-B), similarly to that of protein-coding genes (Fig. 5A, B). LncRNAs displayed sequence composition surrounding poly(A) sites like mRNAs (Fig. 5C, D and Additional file 2: Fig. S7 C-D). Overall, our analysis indicated that lncRNAs are mostly derived from 3'UTRs harboring a favorable context for *trans*-splicing.

Table 3 Genome-wide expression and properties of *L. infantum* long non-coding RNAs

L. infantum lncRNAs (> 200 nt)								
#	**Mean length (nt)**	**# Annotated**		**# Protein-coding [a]**	**# DE log2 FC ≥ +/−1**			
		L. infantum	*L. major*		**AxeAMA vs. PRO**		**MφAMA vs. PRO**	
1825	843	39 (2.1%)	392 (21.5%)	138 Predicted: 67 LC-MS:107	287 123 Up 164 Down		105 47 Up 58 Down	
					lncRNAs vs. upstream CDS			
					Similar expression	Different expression	Similar expression	Different expression
					115	158	24	79

[a] lncRNAs predicted as protein-coding by UniProtKB, CPC or *L. major* annotation (67/138) and/or detected by LC-MS/MS analysis (107/138)

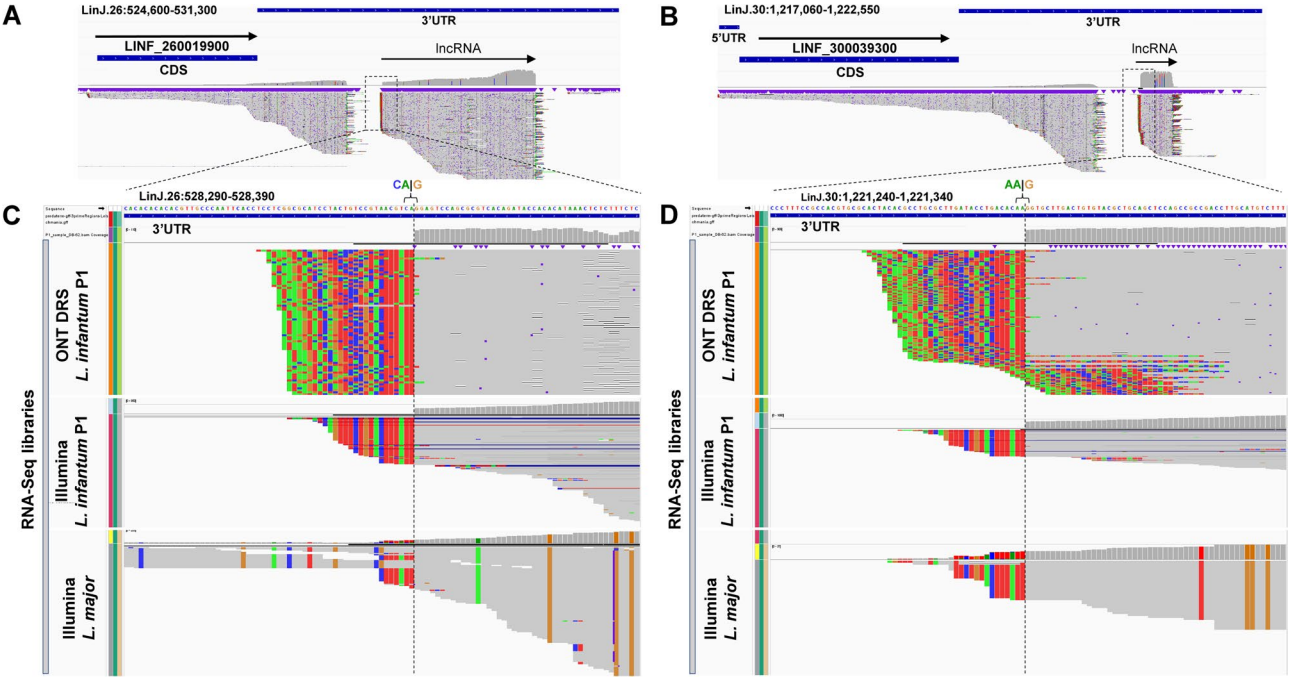

Fig. 6 Examples of long non-coding RNAs sharing conserved SL addition sites between *L. infantum* and *L. major,* as determined by combined Nanopore direct RNA sequencing and Illumina RNA-Seq data. IGV representation of three different RNA-Seq libraries aligned against the *L. infantum* JPCM5 reference genome. Native ONT-DRS reads aligning on the CDS in UTR regions of (**A**) LINF_260019900 and (**B**) LINF_300039300. The SL addition sites of two lncRNAs derived from the 3'UTR of LINF_260019900 (**C**) and LINF_300039300 (**D**) are conserved between *L. infantum* RNA-Seq libraries (ONT-DRS and Illumina), as well as *L. major* Illumina RNA-Seq. Colored bases (green, red, blue or orange) in the IGV image represent mismatched bases corresponding to the SL sequence upstream of the black dashed line

Most of the newly identified lncRNAs are specific to the *L. infantum* transcriptome

To determine whether the newly identified *L. infantum* lncRNAs were also found in other *Leishmania* species, we initially investigated *L. major* using the *L. major* Friedlin 2021 transcriptome available from https://tritrypdb.org/tritrypdb/app/downloads [23]. Out of the 1,825 *L. infantum* lncRNAs, only 21.5% (392/1,825) were annotated in *L. major* as non-coding transcripts of equivalent length present on the same chromosome, of which 331 as ncRNAs, 12 as pseudogenes and 49 as protein-coding genes (Table 3 and Additional file 1: Table S8). To assess whether the SL sequence in lncRNAs was added at the same position between *L. major* and *L. infantum*, we manually examined 70 randomly selected lncRNAs identified with ONT-DRS (see Methods). Interestingly, 32.9% (23/70) of these lncRNAs shared the SL sequence at the same position between *L. infantum* and *L. major* (Fig. 6C, D and Additional file 1: Table S8), suggesting similar processing events in both species. Considering that the *L. major* Friedlin 2021 transcriptome was sequenced using the Illumina short read technology [23], some of the complex *trans*-splicing events likely to occur in *L. major* 3'UTRs might not be detected. Long-read sequencing could address this question and enable a better understanding of the *L. major* lncRNA landscape.

A significant portion of *L. infantum* lncRNAs are developmentally regulated

Considering that the majority of lncRNAs were predicted to derive from 3'UTRs, we first compared their expression to that of the upstream coding sequence (CDS). To perform this analysis, Illumina RNA-Seq data from three biological replicates of *L. infantum* promastigote (PRO), axenic amastigote (AxeAMA) and macrophage-derived amastigote (MΦAMA) transcriptomes (see Methods) to which we added the lncRNAs identified in this study were aligned to the *L. infantum* JPCM5 transcriptome (v.64; 8,748 transcripts). We detected >10 reads per RNA for 99.5% (8,704/8,748) of the *L. infantum* coding sequences and 99.9% for lncRNAs. As illustrated by the volcano plot in Fig. 7A, lncRNAs showed in general distinct expression patterns from their upstream CDS (Fisher's exact test; p-value <0.0001), suggesting a different mode of regulation from the 3'UTRs they derived. Moreover, a linear regression analysis comparing FPKM (fragments per kilobases of transcript per million mapped reads) mean values of lncRNAs and their upstream CDS showed a low correlation for either PRO (R^2 = 0.22) (Fig. 7B) or AxeAMA (R^2 = 0.21) (Additional file 2: Fig. S8), indicating that expression levels of most lncRNAs were distinct from those of their upstream CDS.

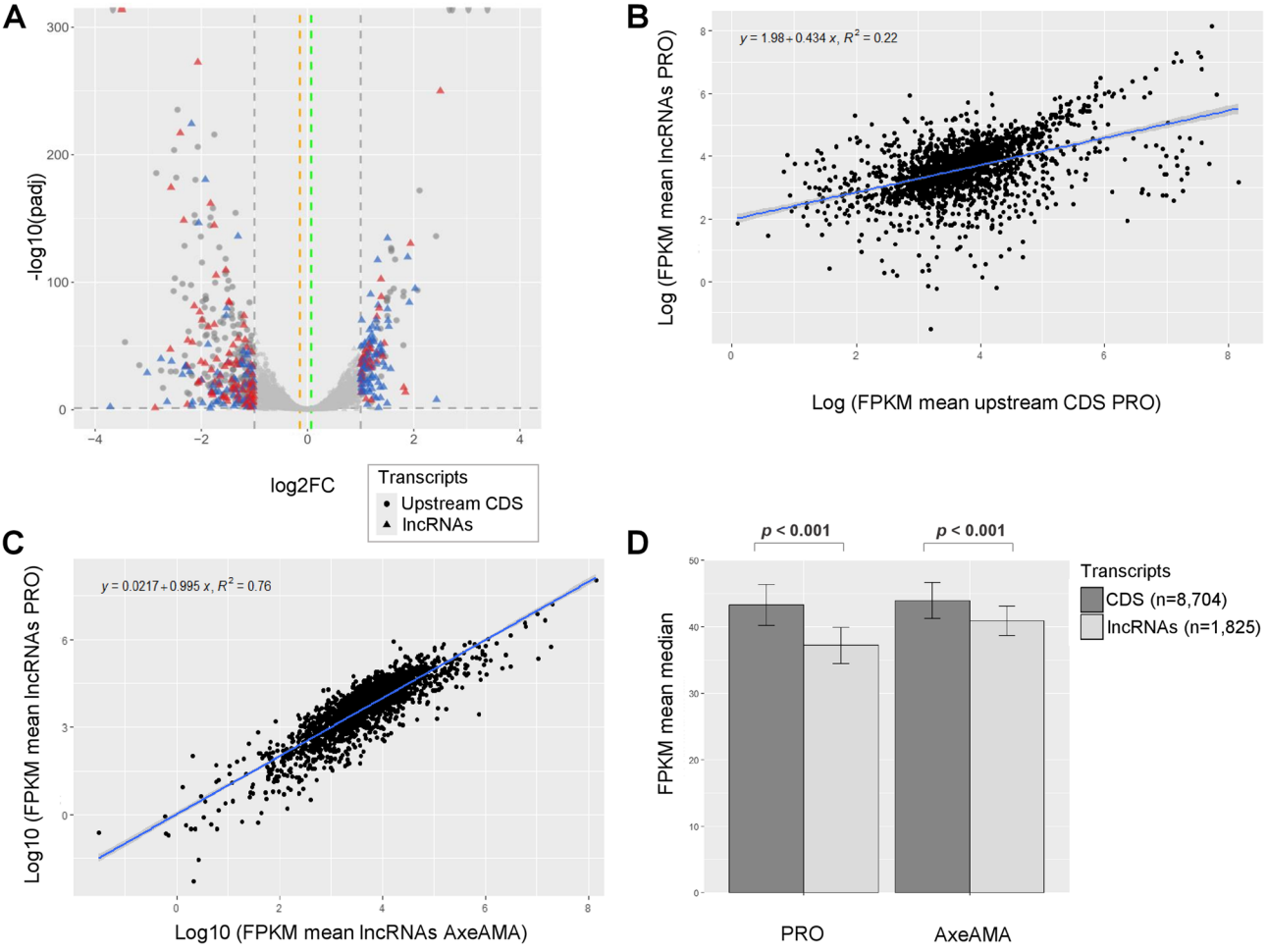

Fig. 7 Differential expression analysis of lncRNAs between *L. infantum* developmental stages. **A** Volcano plot of lncRNAs ($n = 1,825$) and their upstream CDS. Each small circle (CDS) or triangle (lncRNA) represents one transcript. The log2 FC indicates the change in differential expression (DE) of transcripts between *L. infantum* promastigotes (PRO) and axenic amastigotes (AxeAMA), while the −log10 (*p* adjust) shows the statistical significance of the difference in lncRNA expression. CDSs are represented by grey circles. LncRNAs displaying different expression than their upstream CDS are represented by blue triangles ($n = 158$), whereas lncRNAs of similar expression with their upstream CDS are represented by red triangles ($n = 115$) (see Table 3). Orange and green lines show the log2 FC median of the CDS and 3'UTR-derived lncRNA, respectively, while grey lines represent thresholds for considering transcripts as DE (log2 FC ≥ 1 or ≤ −1 and *p* adjust. ≤ 0.05). **B** Expression patterns between lncRNAs and their upstream CDS. Linear regression analysis of the FPKM mean for CDSs and the downstream lncRNAs in *L. infantum* PRO. **C** Linear regression of the log10 FPKM mean for lncRNAs between PRO and AxeAMA. **D** Comparative expression analysis between CDSs ($n = 8,704$) and lncRNAs ($n = 1,825$) in PRO and AxeAMA. Significance levels were assessed using the Wilcoxon rank-sum test

Differences in lncRNAs expression between *L. infantum* promastigote and amastigote developmental stages were assessed by comparing Illumina RNA-Seq data from both life stages using DESeq2 [65]. Most lncRNAs were similarly expressed between promastigotes and parasites grown under conditions triggering amastigote differentiation (AxeAMA) ($R^2 = 0.76$) (Fig. 7C). However, our analysis revealed 287 lncRNAs that were differentially expressed, with 123 being upregulated (log2 FC ≥ 1) and 164 downregulated (log2 FC ≤ −1) in AxeAMA vs. PRO (Table 3 and Additional file 1: Table S8-1). Comparative analysis of lncRNA expression indicated a higher expression in AxeAMA than in PRO (Fig. 7C). Similarly to most lncRNAs (Fig. 7A, B and Additional file 2: Fig. S8), the majority of developmentally regulated lncRNAs exhibited expression levels distinct from their upstream CDS (Table 3 and Additional File 1: Table S8-1). We next looked for lncRNAs differentially expressed in intracellular amastigotes isolated from in vitro macrophage infections (MϕAMA). Hundred and five (105) lncRNAs were differentially expressed, with 47 being upregulated and 58 downregulated in MϕAMA vs. PRO (Table 3 and Additional file 1: Table S8-1). A linear regression analysis of lncRNA expression between PRO and MϕAMA showed a stronger correlation ($R^2 = 0.86$) (Additional file 1: Table S8-1) than between PRO and AxeAMA. The higher correlation between PRO and MϕAMA was also reflected by fewer DE lncRNAs with a log2 FC ≥ 1 or ≤

−1. Interestingly, ~63% of DE lncRNAs between PRO and MϕAMA were also differentially expressed between PRO and AxeAMA (Additional file 1: Table S8-1).

Comparison between the *L. infantum* protein-coding transcriptome (8,704 mRNAs) and lncRNA transcriptome indicated that lncRNAs were on average less expressed than mRNAs (Fig. 7D). Furthermore, our analysis revealed that the percentage of DE lncRNAs between PRO and AxeAMA (15.7%; 287/1,825) (Additional file 1: Table S9) was significantly higher to that of developmentally regulated mRNAs (8.0%; 696/8,704) analyzed under the same experimental conditions. Similarly, the proportion of developmentally regulated lncRNAs between PRO and MϕAMA (5.75%; 105/1,825) (Table 3 and Additional file 1: Table S9) was significantly higher from that of DE mRNAs (3.7%; 325/8,704) (Fisher's exact test; p-value < 0.01). Altogether, these data highlight the importance of lncRNAs in the parasite development.

L. infantum 3'UTR processed transcripts with limited protein coding potential

To explore the protein coding potential, if any, of the identified lncRNAs in *L. infantum*, we first used a Python library ORFFinder (v1.8). Our analysis annotated 16,450 small proteins (mean = 66 aa) encoded by the 1,825 lncRNAs (Additional file 2: Fig. S9 A). Even when analyzing the longest predicted protein per lncRNA, the average length was 122 amino acids, significantly shorter than the average of 617 amino acids observed in canonical *L. infantum* proteins (Welch's t-test, $p <$ 0.0001) (Additional file 2: Fig. S9B-C). Subsequently, to investigate the potential functionality of these predicted proteins, we first aligned them against the UniProtKB database using BLASTp. Significant matches (≥ 50 bit scores) were obtained only for 19 lncRNAs encoding putative proteins of 108 to 734 aa (Additional file 1: Table S8). In addition, the online Coding Potential Calculator (CPC) [56] (CPC score ≥ 1) predicted 38 lncRNAs to code for a protein, of which 28 were annotated as protein-coding genes in the *L. major* Friedlin 2021 transcriptome or identified by UniProtKB as lncRNA-encoded proteins (Additional file 1: Table S8 and Additional File 2: Fig. S10 A). Overall, using three different tools (UniProtKB, CPC or *L. major* annotation), we predicted 67 lncRNAs with protein-coding potential, mostly encoding hypothetical proteins in *L. major* (Fig. 8A and Additional file 1: Table S8). Thirty-nine (39) were also annotated in the *L. infantum* JPCM5 genome: 20 as pseudogenes and 19 as ncRNAs potentially involved in rRNA processing, modification and biogenesis. Moreover, 4 lncRNAs were predicted by UniProtKB to encode novel proteins, not previously annotated in *Leishmania*. These include the

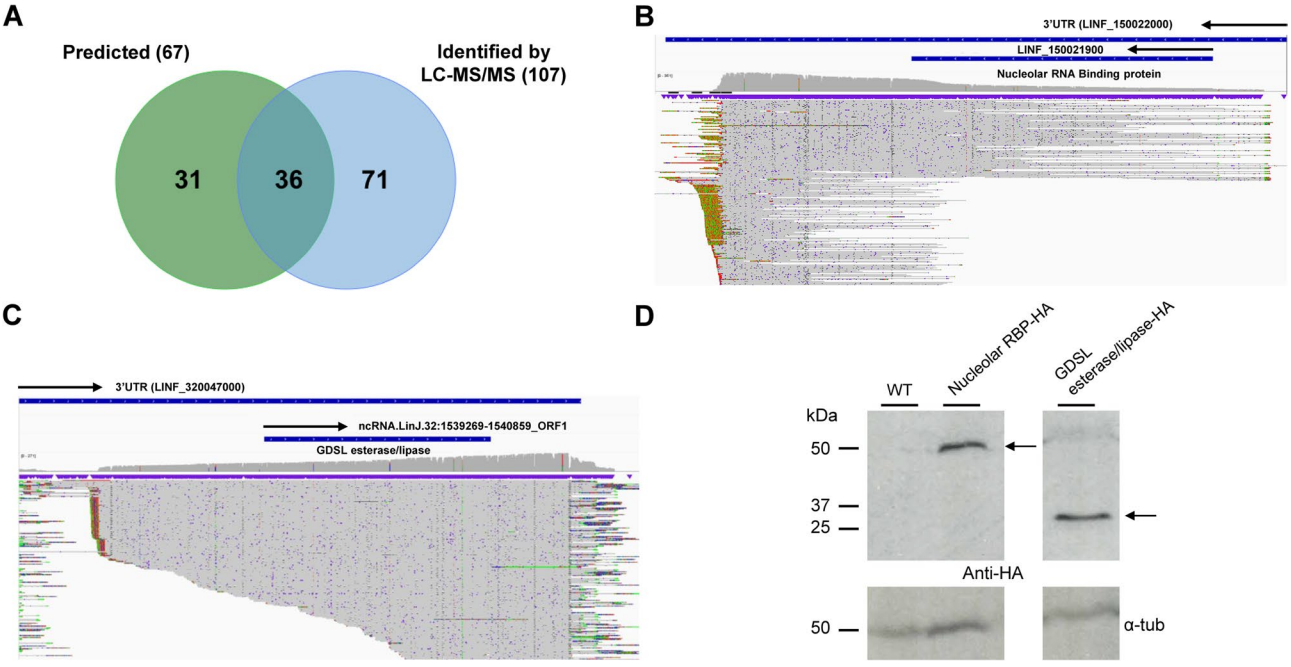

Fig. 8 *L. infantum* lncRNAs with protein-coding potential and validation studies using CRISPR-Cas9 genome editing. **A** Venn diagram showing the number of lncRNAs potentially encoding proteins, as determined by different bioinformatic tools (UniProtKB, CPC, *L. major* annotation) and mass-spectrometry analysis (MS). IGV visualization of native ONT-DRS reads for lncRNAs encoding a nucleolar RNA-binding protein (ncRNA.LinJ.15:590018 – 587628; LINF_150021900) (**B**) and a GDSL esterase/lipase (ncRNA.LinJ.32:1539269–1540859) (**C**) for which peptides were identified by LC-MS/MS. Black arrows indicate the direction of transcription. **D** Western blot analysis of protein extracts (1×10^8 cells) from *L. infantum* recombinant parasites generated by CRISPR-Cas9 genome editing to introduce an HA epitope tag at the C-terminus of the protein-coding lncRNAs shown in **B** and **C** using an anti-HA antibody. The alpha-tubulin antibody was used as protein loading control

1-phosphatidylinosotol 3-phosphate 5-kinase, diaminopimelate epimerase, GDSL esterase/lipase CPRD49 and malate dehydrogenase (Table 4 and Additional file 1: Table S8).

To expand our analysis and to validate the predicted lncRNA protein-coding potential, we screened three independent LC-MS/MS experiments of *L. infantum* promastigotes (see Methods and Additional file 3) to search for the presence of corresponding peptides. We identified peptides for 107 lncRNAs, of which 36 were predicted as protein-coding by UniProtKB or CPC or were annotated as protein-coding in *L. major* (Fig. 8A, Additional file 1: Table S10 and Additional file 2: Fig. S10B). The remaining 71 lncRNAs for which prediction tools failed to identify a protein but mass-spectrometry revealed the presence of peptides (1 to 8) encoded generally small proteins (25–296 aa) (Additional file 1: Table S10). Remarkably, 17 of the novel 3'UTR processed transcripts with protein-coding potential belong to the 50% most prevalent proteins identified by mass-spectrometry (Table 4), suggesting they may have biological roles.

Next, we experimentally validated the protein coding potential of two randomly selected lncRNAs encoding the nucleolar RNA-binding protein (ncRNA.LinJ.15:590018 – 587628; LINF_150021900) and the previously unannotated GDSL esterase/lipase (ncRNA.LinJ.32:1539269–1540859; LINF_320047000), respectively (Fig. 8B, C). Using CRISPR-Cas9-mediated genome editing successfully adapted to *L. infantum* [57], we introduced an HA-tag at the C-terminus of one of the genomic alleles for the nucleolar RBP and the GDSL esterase/lipase. Following transfection, clones were screened by PCR to detect the presence of the HA-tag and confirmed by sequencing. Western blot analysis using an anti-HA antibody visualized a protein of ~50 kDa for the nucleolar RBP (446 aa) and of ~29 kDa (255 aa) for the GDSL esterase/lipase (Fig. 8D), confirming the protein-coding capacity of these new 3'UTR-processed transcripts.

Taken together, our analysis demonstrated that the vast majority of the newly identified lncRNAs in *L. infantum* were indeed non-coding, and only 138 out of 1,825 (7.6%) (Table 3) were predicted or confirmed by LC-MS/MS to potentially code for a protein. Moreover, our analysis identified new proteins of unknown function encoded by 3'UTR processed transcripts with a significant intensity in the *L. infantum* proteome (Table 4 and Additional file 1: Table S10).

Discussion

Here, we provide the first comprehensive study combining Nanopore direct RNA-sequencing (DRS) and Illumina RNA-Seq technologies to gain novel insights into the transcriptome diversity of *L. infantum* developmental stages. Contrary to previous studies that rely solely on short-read RNA sequencing data [15, 21, 54], ONT-DRS enables the sequencing of full-length native RNA molecules and enhances the accuracy of defining the 5'- and 3'-end boundaries of *Leishmania* transcripts and the UTR lengths. Our analysis provides further insights into the genomic context favoring pre-mRNA processing and the extent of stage-regulated alternative splicing and polyadenylation, hence emphasizing the role that alternative RNA processing could play in gene expression dynamics throughout the parasite development. This study also advances knowledge on the non-coding transcriptome of *L. infantum* developmental stages, providing novel insights into the rich repertoire of previously unreported 3'UTR-derived lncRNAs, their stage-regulated expression and parasite species-specificity.

One of the goals in this study was to better delineate the 5′UTR and 3′UTR boundaries of *L. infantum* protein-coding transcripts. In *Leishmania*, several studies have evidenced the relevance of 3′UTRs in post-transcriptional control, regulating mRNA stability, turnover, translation, transport, and subcellular localization [69–77]. Thus, delimiting 3′UTRs would be pivotal for understanding post-transcriptional gene regulation in *Leishmania.* Mature mRNAs in *Leishmania* and related trypanosomatids are generated by coordinated *trans*-splicing and polyadenylation cleavage reactions which add the SL (39-nt miniexon) sequence at the 5'-end and the poly(A) tail at the 3'-end of all transcripts [9, 10, 68]. Here, we accurately mapped 3′ and 5′ ends providing the poly(A) and SL genomic contexts for more than 52% of the *L. infantum* protein-coding transcripts in both stages of parasite development. Delineating the boundaries of each transcript and looking for differences in RNA processing between promastigotes and amastigotes turned out to be challenging due largely to the high number of alternative processing events, most notably, alternative poly(A) sites. Among the transcripts analyzed, the frequency of the primary SL site was 64.2%, which was expected given the high conservation of the genomic context for *trans*-splicing, but a secondary SL site with a much lower frequency was detected at ~57 bp apart in 99.6% of the transcripts. The context for *trans*-splicing in kinetoplastids was previously reported to require a stretch of polypyrimidines preceding the AG splice acceptor site where cleavage occurs. Consistent with previous data in *L. major* [15, 21], our analysis revealed an enrichment of pyrimidines at ~100 nucleotides upstream of the AG splice acceptor site with a slight preference for cytosine over thymine residues in 55.6% of the transcripts. Interestingly, a higher diversity for the poly(A) site per transcript was depicted by our analysis. The frequency for the primary poly(A) site was ~24% and explained by the lack of a well-defined motif for cleavage and polyadenylation in *Leishmania*. Indeed, there

Table 4 *L. infantum* 3'UTR processed transcripts (lncRNAs) with limited protein-coding potential, as determined by three independent liquid chromatography-mass-spectrometry (LC-MS/MS) experiments

lncRNA ID	Protein length (aa)	Putative function	# peptides MS/MS (3 exp.)	PRO Intensity rank [a]
ncRNA.LinJ.01:249215–250,109	61	-	3	4946
ncRNA.LinJ.04:211941 – 211,522	35	-	2	7513
ncRNA.LinJ.05:422834 – 421,082	509	hypothetical protein – conserved [b]	30	1606
ncRNA.LinJ.06:175809–177,662	546	SpoU rRNA Methylase family - putative	12	6276
ncRNA.LinJ.06:97407 – 96,175	269	-	6	7104
ncRNA.LinJ.09:7001–7392	62	-	2	6743
ncRNA.LinJ.10:377541 – 376,934	63	-	3	3896
ncRNA.LinJ.10:550354 – 549,741	135	-[b]	7	2937
ncRNA.LinJ.15:425133 – 423,348	515	hypothetical protein - conserved	20	5708
ncRNA.LinJ.15:590018 – 587,628	**446**	**nucleolar RNA binding protein – putative**[b]	**26**	**536**
ncRNA.LinJ.15:633077 – 631,775	353	presenilin-like aspartic peptidase clan AD family A22 A putative	2	7030
ncRNA.LinJ.16:660443–661,654	240	Putative snoRNA binding domain containing protein - putative	10	5635
ncRNA.LinJ.18:178433 – 175,461	370	PAP2 superfamily - putative	5	4623
ncRNA.LinJ.19:288698–290,243	362	U3 snoRNA-associated protein UTP11 - putative	15	4573
ncRNA.LinJ.19:354991–355,730	122	ATG8/AUT7/APG8/PAZ2 - putative	5	2664
ncRNA.LinJ.20:239215–241,929	734	rRNA biogenesis protein-like protein [b]	41	2218
ncRNA.LinJ.21:616314 – 614,926	284	-	6	1174
ncRNA.LinJ.22:417098–417,629	63	-	2	4893
ncRNA.LinJ.22:459913–460,701	67	-[b]	3	898
ncRNA.LinJ.23:109094–109,454	75	-	3	4042
ncRNA.LinJ.23:127767–129,970	258	hypothetical protein	9	6262
ncRNA.LinJ.23:374769 – 373,909	125	-	3	7308
ncRNA.LinJ.24:794802 – 793,754	184	-	2	7595
ncRNA.LinJ.25:110081 – 108,690	180	Fcf2 pre-rRNA processing – putative [b]	8	643
ncRNA.LinJ.26:1011110–1,011,813	155	hypothetical protein	6	7204
ncRNA.LinJ.26:15710 – 14,915	160	1-phosphatidylinositol 3-phosphate 5-kinase	4	5986
ncRNA.LinJ.26:161954 – 159,535	672	spliced leader RNA PSE-promoter transcription factor –putative [b]	33	3261
ncRNA.LinJ.26:528346–530,231	156	-[b]	4	1941
ncRNA.LinJ.27:729538–730,207	63	-[b]	4	238
ncRNA.LinJ.28:197115 – 192,988	75	-	2	5245
ncRNA.LinJ.28:359912–362,530	149	hypothetical protein	2	5649
ncRNA.LinJ.28:859336–862,018	334	Galactose oxidase - central domain containing protein - putative	3	6331
ncRNA.LinJ.29:1181708–1,182,314	105	-	2	7591
ncRNA.LinJ.30:1243771–1,244,573	168	-	8	5611
ncRNA.LinJ.30:1344003–1,344,495	82	hypothetical protein-conserved [b]	2	287
ncRNA.LinJ.31:1210696 – 1,205,180	241	-	6	6521
ncRNA.LinJ.32:1174989–1,175,638	186	-	7	5533
ncRNA.LinJ.32:1385314–1,386,378	90	-[b]	3	932
ncRNA.LinJ.32:1539269–1,540,859	**255**	**GDSL esterase/lipase CPRD49**[b]	**12**	**3389**
ncRNA.LinJ.32:189645–190,544	78	-	2	7541
ncRNA.LinJ.32:248736–249,993	117	U6 snRNA-associated Sm-like protein LSm4p	6	5452
ncRNA.LinJ.32:500800–501,368	95	hypothetical protein [b]	6	3317
ncRNA.LinJ.33:351963–353,649	435	hypothetical protein - conserved	20	5036
ncRNA.LinJ.34:17355–18,519	237	-	6	7377
ncRNA.LinJ.34:310524 – 310,085	93	-	5	4678
ncRNA.LinJ.35:1108480 – 1,106,329	608	SpoU rRNA Methylase family - putative	27	5584
ncRNA.LinJ.35:1360086 – 1,358,934	222	Pre-rRNA-processing protein PNO1 – putative [b]	13	2796
ncRNA.LinJ.35:1777657–1,779,295	468	small nuclear RNA gene activation protein (SNAP) 50- putative [b]	20	3613
ncRNA.LinJ.35:1965872–1,966,370	83	-	3	7261
ncRNA.LinJ.35:673269–675,508	275	Eukaryotic rRNA processing protein EBP2, putative [b]	13	2074

Table 4 (continued)

lncRNA ID	Protein length (aa)	Putative function	# peptides MS/MS (3 exp.)	PRO Intensity rank [a]
ncRNA.LinJ.35:788824–792,825	81	-	3	7322
ncRNA.LinJ.36:1043969–1,045,173	351	small nuclear RNA-activating protein	10	6306
ncRNA.LinJ.36:1274347 – 1,273,229	197	-	6	6749
ncRNA.LinJ.36:1432207 – 1,430,531	475	hypothetical protein	18	5284
ncRNA.LinJ.36:2251943 – 2,251,121	48	-	2	3866
ncRNA.LinJ.36:2523994 – 2,523,256	195	hypothetical protein	5	5718
ncRNA.LinJ.36:2707985 – 2,706,764	115	hypothetical protein	2	6369

[a] *L. infantum* promastigote (PRO) intensity rank from three independent LC-MS/MS experiments (see Methods and Supplementary file 3). Only lncRNAs potentially encoding a protein for which more that 2 peptides were identified were included. The protein intensity ranking was calculated for each LC-MS/MS experiment by sorting protein intensities with the most abundant proteins having the lower intensity rank values

[b] Protein-coding lncRNAs belonging to the 50% most abundant *L. infantum* proteins detected by LC-MS/MS

In bold are two protein-coding lncRNAs validated by Western blot (see Fig. 8D)

is no consensus motif surrounding the primary poly(A) site in *L. infantum* transcripts, such as the core AAUAAA cytosolic signal in higher eukaryotes that together with the upstream UGUA and downstream GU-rich sequences defines the strength of a given poly(A) site in a combinatorial fashion [78]. However, our data indicated that poly(A) sites were not randomly selected but that cleavage and polyadenylation often occurred after a stretch of adenines (> 3) preceded by a lower GC content and followed by an adenine-rich sequence. This context is different from the (A/G)(A/G) motif preceded by 1–2 thymines abutting the poly(A) addition site reported in *L. major* [15], with our consensus being closer to that described by Rastrojo et al. [21]. Moreover, our data demonstrated that the distance between the poly(A) site of an upstream transcript and the SL addition site of a transcript downstream was on average around 370 nt, which is close to the 400–500 nt previously reported for *L. major* mRNAs [10] and in agreement with previous studies in *Leishmani*a and *T. brucei* that *trans*-splicing is mechanistically coupled to polyadenylation [9–11, 20]. When considering only the primary polyadenylation and *trans*-splicing sites for each *L. infantum* mRNA, we estimated the average length of 5'UTRs and 3'UTRs at 213 nt and 499 nt, respectively, akin to what has previously been reported in *L. major* (233 and 517 nt) [15]. In most cases, 5′UTR and 3′UTR lengths were similar between promastigote and amastigote developmental stages, which suggests that differences in UTR lengths do not contribute significantly to developmental gene regulation.

Alternative cleavage and polyadenylation are co-transcriptional molecular processes to generate mRNA isoforms with different 3'UTR lengths and sequence content that greatly affect gene expression through mRNA stability, translation efficacy, subcellular localization, or mRNA seeded protein-protein interactions [79, 80]. Our DRS analysis revealed alternative polyadenylation in 11.1–13.3% of *L. infantum* transcripts. A heterogeneity in the poly(A) addition sites was also commonly observed for most transcripts in *L. donovani* [24], *L. major* [15, 21] and *L. mexicana* [81], suggesting that alternative polyadenylation could represent an additional level of regulation for controlling gene expression in *Leishmania*. Accordingly, we observed that about a quarter of the alternatively polyadenylated or *trans*-spliced transcripts were differentially expressed between promastigote and amastigote developmental stages, although differences in expression were relatively low. There may be some relationship between alternative processing and stage-specific regulation, however further analysis showed no specific enrichment of alternatively spliced or polyadenylated transcripts among the developmentally regulated transcripts in *L. infantum*. Moreover, instances of regulation through alternative RNA processing between *Leishmania* developmental stages have not been systematically identified, nevertheless this does not rule out the possibility that alternative processing could be regulatory.

An important advancement to our understanding of the *L. infantum* transcriptome complexity is our finding that it is enriched with many unannotated 3'UTR processed RNA transcripts. These lncRNAs display different chromosomal distributions and locations and a wide range of lengths, with a mean of 843 nt. Among the 1,825 lncRNAs identified in the transcriptomes of *L. infantum* developmental stages, only 2.1% were already annotated in the *L. infantum* genome and ~21.5% in *L. major*. Thus, the majority of these newly identified lncRNAs seem to be preferentially expressed in *L. infantum*, the species responsible for visceral leishmaniasis, and less or not expressed in *L. major*, the species causing cutaneous leishmaniasis. However, further studies would be necessary to assess whether some of these lncRNAs could play a role in parasite tropism. To date, even though a few studies have reported a large number

of ncRNAs in several *Leishmania* species, little is known about their function and expression during the parasite development, with most data arising from Illumina RNA-Seq studies and a fewer from PacBio Iso-Seq long read sequencing of the promastigote stage. A recent study in *L. infantum* combining Illumina RNA-Seq with PacBio identified 1,175 ncRNAs that were not fully characterized [22]. Independent studies annotated 1,128 [21, 54] to 1,884 non-coding transcripts [21] in *L. major*, 2,410 transcripts not containing annotated ORFs in *L. donovani* [24] and 2,334 ncRNAs in *L. braziliensis* [45] Here, using the Oxford Nanopore long-read DRS technology, we have generated important new knowledge on the repertoire of *L. infantum* lncRNAs, their chromosomal distribution and location, their 5'- and 3'-end boundaries and expression profiles during the main developmental stages of the parasite. Our analysis revealed that lncRNAs are derived from ~22% of *L. infantum* 3'UTRs through controlled processing events and are all *trans*-spliced and polyadenylated, similarly to mRNAs. Thus, our results suggest that 3'UTR sequences can function not only in *cis* to regulate protein expression, but also intrinsically and independently in *trans* likely as ncRNAs fulfilling biological roles different from their normally associated mRNA. Large numbers of 3'UTRs in human, mouse and fly are also expressed separately from the associated protein-coding sequences to which they are linked, likely by post-transcriptional cleavage [29, 38, 82–84]. Similarly, we showed that most 3'UTR-derived *L. infantum* lncRNAs exhibit distinct expression patterns from their upstream protein-coding sequence. This indicates that they are subjected to independent regulatory processes, possibly through interactions with other proteins or RNAs, as lncRNAs often contain embedded sequence motifs that can recruit certain factors [29]. Examples of ncRNAs transcribed from the untranslated regions of coding genes have also been described in the transcriptomes of *Trypanosoma brucei* [17, 47], *L. major* [44], *L. donovani* [44] and *L. braziliensis* [45].

LncRNAs constitute a large portion of the transcriptome in many eukaryotic organisms playing crucial roles in controlling the expression of neighbouring protein-coding genes or genes away from their loci at the epigenetic, transcriptional and posttranscriptional levels [31–36]. Our comparative transcriptomic analyses revealed that *L. infantum* lncRNAs were on average less expressed than mRNAs and that their expression was generally higher in amastigotes than in promastigotes. The biological relevance of these lncRNAs is further emphasized by their differential expression between parasite life stages. While only 3.7–8.0% of mRNAs (depending on the stage) were developmentally regulated, this proportion increased to 15.7% for lncRNAs, suggesting their potential involvement in stage-specific biological processes. LncRNAs account for 21% of the whole *L. infantum* transcriptome (protein-coding and non-coding combined), but interestingly they represent more than 27% of all differentially expressed transcripts. Together, these data highlight the importance that lncRNAs could play in the parasite development. Additional studies would be required, however, to determine the mechanisms regulating developmental expression of lncRNAs and their role in the parasite differentiation and pathogenesis. A recent study in *L. braziliensis* reported the role of a short ncRNA preferentially expressed in amastigotes in modulating gene expression and contributing to stress response and parasite differentiation [85].

Recent advances in high-throughput techniques, including ribosome profiling and mass-spectrometry have demonstrated that many lncRNAs retain small open reading frames and can potentially encode small proteins or micropeptides. Many of these lncRNA-encoded micropeptides have been shown to perform fundamental biological functions, including cell division, transcription regulation, and cell signaling within organisms ranging from bacteria to flies to humans [36, 86–88]. Here, by combining *L. infantum* transcriptomic datasets to protein prediction tools and proteomic data, we found that 138 of the 1,825 3'UTR processed transcripts encode putative proteins of mostly unknown functions, and a few are novel proteins not previously annotated in *Leishmania* spp. Approximately 40% of these novel proteins are smaller than 100 aa and a few belong to the 50% most prevalent proteins identified by mass-spectrometry, implying that they may have important biological functions. Integrating transcriptomic and proteomic datasets was shown previously in humans to help in identifying novel proteins [89]. In *L. major*, the use of mass-spectrometry data led to the identification of eight novel proteins encoded by long intergenic non-coding RNAs [90].

In summary, this comprehensive transcriptomic analysis of *L. infantum* developmental stages at single-molecule resolution enabled us not only to advance knowledge on existing *Leishmania* expression datasets, but also to accurately define the boundaries of *Leishmania* transcripts, thereby determining the length of UTRs, and to gain further insights into the genomic context favoring pre-mRNA processing and alternative splicing and polyadenylation events throughout the parasite development. Moreover, this study has considerably advanced our understanding of the *L. infantum* non-coding transcriptome by providing important novel insights into previously unannotated 3'UTR-derived lncRNAs and their expression during the parasite development or between *Leishmania* species.

Conclusions

This study provides the first comprehensive transcriptomic analysis of *Leishmania infantum* developmental stages at single-molecule resolution using Nanopore direct RNA sequencing (DRS). Sequencing full-length native RNA molecules enabled us to accurately define the 5'- and 3'-end boundaries of *Leishmania* transcripts and the length of UTRs. Considering, the central role 3′UTRs play in regulating gene expression in *Leishmania*, delimiting 3′UTRs would be pivotal for understanding post-transcriptional control in these parasites. Moreover, DRS analysis provided further insights into the genomic context favoring pre-mRNA processing and the extent of alternative splicing and polyadenylation depending on the stage of development, thus emphasizing the dynamics of gene regulation across both parasite life stages. In addition, this study substantially advanced our knowledge of the non-coding transcriptome of *L. infantum* developmental stages by providing novel insights into the rich repertoire of previously unannotated lncRNAs, their level of expression during development and species-specificity. Of importance, we showed that lncRNAs are processed mostly from 3'UTRs through *trans*-splicing and polyadenylation, like mRNAs, and exhibit generally distinct expression profiles from their normally associated mRNA, suggesting a different mode of regulation from the 3'UTRs they derived. Several of these lncRNAs are developmentally regulated, representing more than a quarter of the *L. infantum* stage-regulated transcriptome, hence highlighting their importance in parasite development. Overall, these findings not only improve our knowledge of existing *Leishmania* expression datasets, but also provide new important insights into the transcriptome complexity and dynamics of protein-coding and non-coding sequences in both stages of parasite development and pave the way for future investigations of new regulatory pathways.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11767-8.

Additional file 1: Tables S1-S11. Table S1. Detailed summary of Oxford Nanopore direct RNA-sequencing reads of ***L. infantum*** promastigotes (P1, P2) and axenic amastigotes (A1, A2). Table S2. The poly(A) context and position for ***L. infantum*** transcripts with the same primary poly(A) site in the four transcriptomes (P1, P2, A1, A2). Table S3. This table includes the number of full-length reads for each ***L. infantum*** transcript and their average length in each life stage of the parasite: P1 + P2 for promastigotes and A1 + A2 for axenic amastigotes. Table S4. Genomic context and position for ***trans***-splicing of ***L. infantum*** transcripts having the same primary spliced leader (SL) site in all four transcriptomes (P1, P2, A1, A2). Table S5. This table provides data on alternative poly(A) sites per transcript within a 150 nt window surrounding the primary poly(A) site (~75 nt upstream and/or ~75 nt downstream of the primary poly(A) site) for ***L. infantum*** promastigote and amastigote transcriptomes. Transcripts for which two primary poly(A) sites were detected (i.e., two equally prevalent sites with the same number of poly(A) reads) were not included. Highlighted in grey are alternatively polyadenylated transcripts shown in Fig. 4 A-B and Additional file 2: Fig. S6 A-D. Table S6. This table lists transcripts for which alternative poly(A) sites were identified between promastigotes and axenic amastigotes. A poly(A) site was considered as alternative when the primary poly(A) site was conserved between both biological replicates (P1 and P2 for promastigotes; A1 and A2 for axenic amastigotes) and located at > 100 nt apart. Highlighted in blue are alternatively polyadenylated transcripts downregulated in AxeAMA vs. PRO (log2 FC >-0.58) and highlighted in grey those that are upregulated in AxeAMA vs. PRO (log2 FC >0.58). In bold are alternatively polyadenylated transcripts between PRO and AxeAMA shown in Fig. 4 A-B or Additional file 2: Fig. S6 A-D. Table S7. This table lists the alternatively spliced ***L. infantum*** transcripts between the two life stages of the parasite. Highlighted in blue are alternatively spliced transcripts downregulated in AxeAMA vs. PRO (log2 FC >-0.58) and in grey those that are upregulated in AxeAMA vs. PRO (log2 FC >0.58). In bold black characters are differentially spliced transcripts between PRO and AxeAMA shown in Fig. 4 C-D. In bold blue characters are alternatively spliced transcripts shown in Additional file 2: Fig. S6 E-H. Table S8. This table includes the list of 1892 non-coding RNAs (ncRNAs) identified in this study from which 1825 were long non-coding RNAs (lncRNAs (>200 nt in length). The data were derived from the full length (FL) ONT_DRS reads that aligned within the 3'UTR of protein coding genes. Table S8-1. List of differentially expressed lncRNAs between ***L. infantum*** promastigotes (PRO), axenic amastigotes (AxeAMA) and macrophage-derived amastigotes (MφAMA). In bold are differentially expressed lncRNAs with log2 FC <1 (Down AMA vs. PRO; blue) or >1 (Up AMA vs. PRO; black) between promastigotes and amastigotes. Not indicated in bold are down- or up-regulated lncRNAs between promastigotes and amastigotes with log2 FC from (+ or -) 0.54 to 1. Table S9. This table summarizes the differential expression data from Illumina RNA-Seq analysis of ***L. infantum*** promastigotes and axenic amastigotes (BioProject PRJNA1071080).The analysis conducted using DESeq2 and includes 8,704 ***L. infantum*** JPCM5 genes (v.64) and the 1,825 lncRNAs. Highlighted in grey are lncRNAs with their upstream CDS. Table S10. This table lists***L. infantum*** lncRNAs for which peptides were identified based on three independent LC-MS/MS experiments (see Methods and Additional file 3). Highlighted in grey are lncRNAs predicted to encode a protein in the ***L. major*** genome or by the UniProtKB tool. Table S11. List of primers used in this study.

Additional file 2: Figs. S1-S10. Fig. S1. Oxford Nanopore direct RNA sequencing runs of ***L. infantum*** promastigotes and amastigotes.A Graphical representation of the read length distribution of aligned reads. B Length frequency of aligned reads across samples. To improve graphical resolution in panels A and B, only reads comprised between 50 and 2,500 nt were shown. Fig. S2. Poly(A) tail lengths of the aligned direct RNA-Seq reads. Box plot representing the poly(A) tail lengths (number of A residues) evaluated using Nanopolish PolyA for the four ONT RNA-Seq libraries (P1, P2, A1 and A2). The mean and median poly(A) tail lengths were 77.2 and 72.8 nt (P1), 77.7 and 73.4 nt (P2), 81.1 and 74.4 nt (A1), and 78.0 and 70.7 nt (A2), respectively. Fig. S3. Mapping of the full-length reads against the ***L. infantum*** JPCM5 reference genome.A IGV representation of full-length reads mapped to the JPCM5 reference genome (v62) using minimap2 based of their classification when regarding gene annotation. Soft-clipped bases were shown in IGV. Full-length reads were obtained from the A1 transcriptome. A subset of 160 on the 498 monocistronic reads (80 per transcript) was used to generate the BAM alignment. B Pairwise alignment of the 39-nt spliced leader (SL) sequence against 45 full-length reads visualized with Jalview (15 monocistrons, 15 polycistrons and 15 not or partially overlapping a CDS). Fig. S4. Comparison of 5'UTR lengths estimated from ONT-DRS reads or predicted using the PRED-A-TERM software.A Linear regression of the 5'UTR lengths predicted by PRED-A-TERM and evaluated from the ONT Nanopore full-length reads. IGV visualization of three 5'UTRs of variable size (4 nt, 213 nt and 1,840 nt) revealed by the native ONT-DRS reads for LINF_280014200 (B), LINF_140019900 (C) and LINF_290007400 (D). Fig. S5. Diversity of polyadenylation sites within the promastigote (P1) transcriptome.IGV representation of the P1 ONT-DRS poly(A) reads aligning on the ***L. infantum*** JPCM5 reference genome for four genes exhibiting alternative polyadenylation: LINF_360081400 (A) and LINF_090019600) (B) with a primary poly(A) site accounting for most of the reads, LINF_180008500 (C) and LINF_280030900 (D). Fig. S6. Examples of alternatively polyadenylated or ***trans***-spliced transcripts between ***L. infantum*** promastigotes and axenic amastigotes. Coverage plots for

each transcriptome (P1, P2, A1 and A2) of the ONT-DRS poly(A) reads at the 3'-end show alternative polyadenylation for LINF_320038800 (A), LINF_280022100 (B), LINF_310012800 (C) and LINF_150021600 (D). Four examples of alternative spliced transcripts are represented by histogram coverage at the 5'-end using the full-length reads for LINF_260006000 (E), LINF_250011200 (F), LINF_110018900 (G) and LINF_050010300 (H). Fig. S7. Genomic context for the SL and poly(A) addition sites of lncRNAs. The percentage of each nucleotide, as well as the GC and CT content were calculated for 200 nt upstream and downstream of the SL cleavage site (A) and poly(A) site (n=1,825) (C). Nucleotide patterns upstream or downstream were generated using WebLogo (https://weblogo.berkeley.edu/logo.cgi) for the SL addition site (B) and poly(A) sites (D). The cleavage occurs between the position 0 and -1. Fig. S8. Expression patterns between lncRNAs and their upstream CDS. Linear regression analysis of the FPKM mean for CDSs and the downstream lncRNAs for axenic amastigotes (AxeAMA). Fig. S9. Protein size histograms of ***L. infantum*** CDSs and proteins encoded by lncRNAs. Distribution of protein lengths for 16,450 predicted lncRNA-proteins (A), 8,527 JPCM5 ***L. infantum***-encoded proteins (B) and the longest protein annotated per lncRNA (C). The red vertical lines indicate the mean length in log.Fig. S10. Number of lncRNAs potentially encoding a protein as predicted by different tools or identified by LC-MS/MS analysis. Venn diagrams showing the number of lncRNAs predicted to potentially encode a protein as determined by three bioinformatics tools (CPC, ***L. major*** annotation, or UniProtKB) (A), and or identified in three independent LC-MS/MS experiments (LF202, DDA010 and RD196-03) (B).

Additional file 3. LC-MS/MS analysis of three independent experiments (LF202, DDA010 and RD196-03) using total protein extracts from***L. infantum*** promastigotes (PRO) (see Methods).

Additional file 4.

Acknowledgements
We would like to thank Carole Dumas for helping us with the parasite culture and Dr. Philippe Leprohon for initial assistance with the analysis of Illumina RNA-sequencing data. We also acknowledge TriTrypDB for providing indispensable resources and tools.

Authors' contributions
JGER designed and conducted computational experiments, ran most of the analysis and wrote the initial manuscript draft. GRF performed RNA sampling and sequencing, analyzed most of the data and revised the manuscript. MAS contributed to the study design, data interpretation and revised the manuscript. CLO optimized and generated the CRISPR-Cas9 clones. BP conceived and supervised the study, wrote the manuscript with inputs from all the other co-authors and raised funding. All authors have read and approved the final version of the manuscript.

Funding
This work was supported by the Canadian Institutes of Health Research (CIHR) operating grant [MOP-12182] awarded to BP. GRF was the recipient of the Pierre J Durant doctoral fellowship from the faculty of medicine/Université Laval and the CHU de Québec Foundation and Desjardins doctoral award.

Data availability
The raw ONT-DRS RNA-Seq sequencing data for the two biological replicates of *L. infantum* promastigote and amastigote forms have been deposited to the NCBI Bioproject ID PRJNA1209106 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1209106) and are retrievable under the accession numbers SRX27352282, SRX27352283, SRX27352284 and SRX27352285.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Research Center in Infectious Diseases and Axis of Infectious and Immune Diseases, Research Center of the Centre Hospitalier Universitaire de Québec-Université Laval, 2705 Laurier Blvd, Quebec, QC G1V 4G2, Canada
[2]Department of Microbiology, Infectious Disease and Immunology, Faculty of Medicine, University Laval, Quebec, QC G1V 0A6, Canada
[3]School of Biotechnology and Molecular Bioscience, Faculty of Science, UNSW Sydney, Sydney, NSW 2052, Australia
[4]RNA Institute, UNSW Sydney, Sydney, NSW 2052, Australia
[5]CHU Sainte-Justine Research Centre, Montreal, QC H3T 1C5, Canada
[6]Department of Biochemistry and Molecular Medicine, Faculty of Medicine, University of Montreal, Montreal, QC H3T 1J4, Canada

Received: 12 March 2025 / Accepted: 30 May 2025
Published online: 01 July 2025

References

1. Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide and global estimates of its incidence. PLoS ONE. 2012;7:e35671.
2. Costa CHN, Chang KP, Costa DL, Cunha FVM. From infection to death: an overview of the pathogenesis of visceral leishmaniasis. Pathogens. 2023;12:969.
3. Haile S, Papadopoulou B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. Curr Opin Microbiol. 2007;10:569–77.
4. Clayton C. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. Open Biol. 2019;9:190072.
5. Myler PJ, Audleman L, Devos T, Hixson G, Kiser P, Lemley C, et al. Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc Natl Acad Sci U S A. 1999;96:2902–06.
6. Worthey EA, Martinez-Calvillo S, Schnaufer A, Aggarwal G, Cawthra J, Fazelinia G, et al. Leishmania major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene. Nucleic Acids Res. 2003;31:4201–10.
7. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ. Transcription of leishmania major Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell. 2003;11:1291–99.
8. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, leishmania major. Science. 2005;309:436–42.
9. Matthews KR, Tschudi C, Ullu E. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. Genes Dev. 1994;8:491–501.
10. LeBowitz JH, Smith HQ, Rusche L, Beverley SM. Coupling of poly(A) site selection and trans-splicing in leishmania. Genes Dev. 1993;7:996–1007.
11. Ullu E, Matthews KR, Tschudi C. Temporal order of RNA-Processing reactions in trypanosomes: rapid *trans* splicing precedes polyadenylation of newly synthesized tubulin transcripts. Mol Cell Biol. 1993;13:720–5.
12. Sutton RE, Boothroyd JC. Evidence for trans splicing in trypanosomes. Cell. 1986;47:527–35.
13. Perry KL, Watkins KP, Agabian N. Trypanosome mRNAs have unusual cap 4 structures acquired by addition of a spliced leader. Proc Natl Acad Sci U S A. 1987;84:8190–94.
14. Michaeli S. *Trans*-Splicing in trypanosomes: machinery and its impact on the parasite transcriptome. Future Microbiol. 2011;6:459–74.
15. Dillon LAL, Okrah K, Hughitt VK, Suresh R, Li Y, Fernandes MC, et al. Transcriptomic profiling of gene expression and RNA processing during leishmania major differentiation. Nucleic Acids Res. 2015;43:6799–813.
16. Smith M, Blanchette M, Papadopoulou B. Improving the prediction of mRNA extremities in the parasitic protozoan leishmania. BMC Bioinformatics. 2008;9:158.
17. Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen trypanosoma brucei at Single-Nucleotide resolution. PLoS Pathog. 2010;6:e1001090.
18. Huang J, Van Der Ploeg LH. Requirement of a polypyrimidine tract for trans-splicing in trypanosomes: discriminating the PARP promoter from the immediately adjacent 3' splice acceptor site. EMBO J. 1991;10:3877–85.
19. Sun Y, Hamilton K, Tong L. Recent molecular insights into canonical pre-mRNA 3'-end processing. Transcription. 2020;11:83–96.

20. Vassella E, Braun R, Roditi I. Control of polyadenylation and alternative splicing of transcripts from adjacent genes in a Procyclin expression site: a dual role for polypyrimidine tracts in trypanosomes? Nucleic Acids Res. 1994;22:1359–64.
21. Rastrojo A, Carrasco-Ramiro F, Martín D, Crespillo A, Reguera RM, Aguado B, et al. The transcriptome of leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. BMC Genomics. 2013;14:223.
22. Camacho E, González-de la Fuente S, Solana JC, Tabera L, Carrasco-Ramiro F, Aguado B, et al. Leishmania infantum (JPCM5) transcriptome, gene models and resources for an active curation of gene annotations. Genes (Basel). 2023;14:866.
23. Camacho E, Fuente SG, de la, Solana JC, Rastrojo A, Carrasco-Ramiro F, Requena JM, et al. Gene annotation and transcriptome delineation on a de Novo genome assembly for the reference leishmania major Friedlin strain. Genes (Basel). 2021;12:1359.
24. Camacho E, González-de la Fuente S, Rastrojo A, Peiró-Pastor R, Solana JC, Tabera L, et al. Complete assembly of the leishmania donovani (HU3 strain) genome and transcriptome annotation. Sci Rep. 2019;9:6127.
25. González-de la Fuente S, Requena JM, Aguado B. Datasets of Iso-Seq transcripts for decoding transcriptome complexity in four leishmania species. Data Brief. 2024;52:109838.
26. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative genomic analysis of three leishmania species that cause diverse human disease. Nat Genet. 2007;39:839–47.
27. González-de la Fuente S, Peiró-Pastor R, Rastrojo A, Moreno J, Carrasco-Ramiro F, Requena JM, et al. Resequencing of the leishmania infantum (strain JPCM5) genome and de Novo assembly into 36 contigs. Sci Rep. 2017;7:18050.
28. Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of leishmania infantum. Mol Biochem Parasitol. 2009;165:32–47.
29. Mercer TR, Dinger ME, Mattick JS. Long non-coding rnas: insights into functions. Nat Rev Genet. 2009;10:155–9.
30. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016;17:47–62.
31. Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021;22:96–118.
32. Mattick JS, Gagen MJ. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol. 2001;18:1611–30.
33. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, et al. Long non-coding rnas: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023;24:430–47.
34. Kobayashi A, Takeiwa T, Ikeda K, Inoue S. Roles of noncoding RNAs in regulation of mitochondrial Electron transport chain and oxidative phosphorylation. Int J Mol Sci. 2023;24:9414.
35. Hartford CCR, Lal A. When long noncoding becomes protein coding. Mol Cell Biol. 2020;40:e00528–19.
36. Xiao Y, Ren Y, Hu W, Paliouras AR, Zhang W, Zhong L, et al. Long non-coding RNA-encoded micropeptides: functions, mechanisms and implications. Cell Death Discov. 2024;10:450.
37. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol. 2012;30:99–104.
38. Patop IL, Wüst S, Kadener S. Past, present, and future of circrnas. EMBO J. 2019;38:e100836.
39. Clark MB, Mattick JS. Long noncoding RNAs in cell biology. Semin Cell Dev Biol. 2011;22:366–76.
40. Flynn RA, Chang HY. Long noncoding RNAs in Cell-Fate programming and reprogramming. Cell Stem Cell. 2014;14:752–61.
41. Deveson IW, Hardwick SA, Mercer TR, Mattick JS. The dimensions, dynamics, and relevance of the mammalian noncoding transcriptome. Trends Genet. 2017;33:464–78.
42. Rafiee A, Riazi-Rad F, Havaskary M, Nuri F. Long noncoding rnas: regulation, function and cancer. Biotechnol Genet Eng Rev. 2018;34:153–80.
43. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. Nat Rev Genet. 2018;19:535–48.
44. Freitas Castro F, Ruy PC, Nogueira Zeviani K, Freitas Santos R, Simões Toledo J, Kaysel Cruz A. Evidence of putative non-coding RNAs from leishmania untranslated regions. Mol Biochem Parasitol. 2017;214:69–74.
45. Ruy PDC, Monteiro-Teles NM, Miserani Magalhães RD, Freitas-Castro F, Dias L, Aquino Defina TP, et al. Comparative transcriptomics in leishmania braziliensis: disclosing differential gene expression of coding and putative noncoding RNAs across developmental stages. RNA Biol. 2019;16:639–60.
46. Guegan F, Shanmugha Rajan K, Bento F, Pinto-Neves D, Sequeira M, Gumińska N, et al. A long noncoding RNA promotes parasite differentiation in African trypanosomes. Sci Adv. 2022;8:eabn2706.
47. Kim HC, Jolly ER. LncRNAs are differentially expressed between wildtype and cell line strains of African trypanosomes. Noncoding RNA. 2022;8:7.
48. Sereno D, Lemesre JL. Vitro life cycle of Pentamidine-Resistant amastigotes: stability of the chemoresistant phenotypes is dependent on the level of resistance induced. Antimicrob Agents Chemother. 1997;4:1898–903.
49. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford nanopore sequencing. Genome Biol. 2019;20:129.
50. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–9.
51. Li H. Minimap2: pairwise alignment for nucleotide sequences. https://doi.org/10.1093/bioinformatics/bty191
52. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de Novo using only nanopore sequencing data. Nat Methods. 2015;12:733–5.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
54. Rastrojo A, Corvo L, Lombraña R, Solana JC, Aguado B, Requena JM. Analysis by RNA-seq of transcriptomic changes elicited by heat shock in leishmania major. Sci Rep. 2019;9:6919.
55. Pearson WR. An introduction to sequence similarity (homology) searching. Curr Protoc Bioinf. 2013;Chapter 3:3.1.1-3.1.8.
56. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007;35:W345–9.
57. Bhattacharya A, Leprohon P, Bigot S, Padmanabhan PK, Mukherjee A, Roy G, et al. Coupling chemical mutagenesis to next generation sequencing for the identification of drug resistance mutations in leishmania. Nat Commun. 2019;10:5627.
58. Peng D, Tarleton R, Eupagdt. A web tool tailored to design crispr guide Rnas for eukaryotic pathogens. Microb Genom. 2015;1:e000033.
59. Concordet JP, Haeussler M, CRISPOR. Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res. 2018;46:W242–5.
60. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the expasy server. The proteomics protocols handbook. Totowa, NJ: Humana; 2005. pp. 571–607.
61. Padmanabhan PK, Ferreira GR, Zghidi-Abouzid O, Oliveira C, Dumas C, Colaç O, Mariz F, et al. Genetic depletion of the RNA helicase DDX3 leads to impaired elongation of translating ribosomes triggering co-translational quality control of newly synthesized polypeptides. Nucleic Acids Res. 2021;49:9459–78.
62. Ferreira GR, Emond-Rheault J-G, Alves L, Leprohon P, Smith MA, Papadopoulou B. Evolutionary divergent clusters of transcribed extinct truncated retroposons drive low mRNA expression and developmental regulation in the protozoan leishmania. BMC Biol. 2024;22:249.
63. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, stringtie and ballgown. Nat Protoc. 2016;11:1650–67.
64. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads HHS public access author manuscript. Nat Biotechnol. 2015;33:290–5.
65. Love MI, Huber W, Anders S. Moderated Estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
66. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15:201–6.
67. Liang XH, Haritan A, Uliel S, Michaeli S. Trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. Eukaryot Cell. 2003;2:830–40.
68. Ullu E, Tschudi C, Gunzl A. Trans-splicing in trypanosomatid protozoa. In: Smith D, Parsons M, editors. Molecular Biology of Parasitic Protozoa. Oxford: Oxford University Press; 1996. p. 115–3.

69. Boucher N, Wu Y, Dumas C, Dube M, Sereno D, Breton M, et al. A common mechanism of stage-regulated gene expression in leishmania mediated by a conserved 3′-untranslated region element. J Biol Chem. 2002;277:19511–20.
70. McNicoll F, Müller M, Cloutier S, Boilard N, Rochette A, Dubé M, et al. Distinct 3′-untranslated region elements regulate stage-specific mRNA accumulation and translation in leishmania. J Biol Chem. 2005;280:35238–46.
71. Müller M, Padmanabhan PK, Rochette A, Mukherjee D, Smith M, Dumas C, et al. Rapid decay of unstable leishmania mRNAs bearing a conserved retroposon signature 3′UTR motif is initiated by a site-specific endonucleolytic cleavage without prior deadenylation. Nucleic Acids Res. 2010;38:5867–83.
72. Azizi H, Romão TP, Charret KS, Padmanabhan PK, De Melo Neto OP, Müller-McNicoll M, et al. RNA secondary structure and nucleotide composition of the conserved hallmark sequence of leishmania SIDER2 retroposons are essential for endonucleolytic cleavage and mRNA degradation. PLoS ONE. 2017;12:e0180678.
73. Bringaud F, Müller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NMA, et al. Members of a large retroposon family are determinants of Post-Transcriptional gene expression in leishmania. PLoS Pathog. 2007;3:e136.
74. Folgueira C, Quijada L, Soto M, Abanades DR, Alonso C, Requena JM. 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;280:35172–83.
75. Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M. Developmental regulation of heat shock protein 83 inleishmania. J Biol Chem. 2001;276:47922–9.
76. David M, Gabdank I, Ben-David M, Zilka A, Orr I, Barash D, et al. Preferential translation of Hsp83 in *Leishmania* requires a thermosensitive polypyrimidine-rich element in the 3′ UTR and involves scanning of the 5′ UTR. RNA. 2010;16:364–74.
77. Quijada L, Soto M, Alonso C, Requena JM. Identification of a putative regulatory element in the 3′-untranslated region that controls expression of HSP70 in leishmania infantum. Mol Biochem Parasitol. 2000;110:79–91.
78. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol. 2017;18:18–30.
79. Gallicchio L, Olivares GH, Berry CW, Fuller MT. Regulation and function of alternative polyadenylation in development and differentiation. RNA Biol. 2023;20:908–25.
80. Hao Y, Cai T, Liu C, Zhang X, Fu X-D. Sequential polyadenylation to enable alternative mRNA 3′ end formation. Mol Cells. 2023;46:57–64.
81. 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.
82. Mayr C. What are 3′ UTRs doing?? Cold Spring Harb Perspect Biol. 2019;11:a034728.
83. Mercer TR, Wilhelm D, Dinger ME, Soldà G, Korbie DJ, Glazov EA, et al. Expression of distinct RNAs from 3′ untranslated regions. Nucleic Acids Res. 2011;39:2393–403.
84. Ni W-J, Xie F, Leng X-M. Terminus-Associated Non-coding rnas: trash or treasure?? Front Genet. 2020;11:552444.
85. Quilles JC, Espada CR, Orsine LA, Defina TA, Almeida L, Holetz F, et al. A short NcRNA modulates gene expression and affects stress response and parasite differentiation in leishmania Braziliensis. Front Cell Infect Microbiol. 2025;15:1513908.
86. Wu P, Mo Y, Peng M, Tang T, Zhong Y, Deng X, et al. Emerging role of tumor-related functional peptides encoded by LncRNA and circrna. Mol Cancer. 2020;19:22.
87. Huang J-Z, Chen M, Chen D, Gao X-C, Zhu S, Huang H, et al. A peptide encoded by a putative LncRNA HOXB-AS3 suppresses Colon cancer growth. Mol Cell. 2017;68:171–e1846.
88. Ji Z, Song R, Regev A, Struhl K. Many lncrnas, 5′utrs, and pseudogenes are translated and some are likely to express functional proteins. Elife. 2015;4:e08890.
89. Gascoigne DK, Cheetham SW, Cattenoz PB, Clark MB, Amaral PP, Taft RJ, et al. Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. Bioinformatics. 2012;28:3042–50.
90. Pawar H, Pai K, Patole MS. A novel protein coding potential of long intergenic non-coding RNAs (lincRNAs) in the kinetoplastid protozoan parasite leishmania major. Acta Trop. 2017;167:21–5.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.