Molecular & Biochemical Parasitology 183 (2012) 140–150

Contents lists available at SciVerse ScienceDirect

Molecular & Biochemical Parasitology

In silico identification of conserved intercoding sequences in *Leishmania* genomes: Unraveling putative *cis*-regulatory elements

E.J.R. Vasconcelos [a], M.C. Terrão [a], J.C. Ruiz [c], R.Z.N. Vêncio [b], A.K. Cruz [a,*]

[a] Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, Brazil
[b] Departamento de Computação e Matemática, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, Brazil
[c] Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 23 November 2011
Received in revised form 16 February 2012
Accepted 17 February 2012
Available online 25 February 2012

Keywords:
Leishmania
Putative regulatory motifs
UTRs
Post-transcriptional gene regulation
Bioinformatics

ABSTRACT

In silico analyses of *Leishmania* spp. genome data are a powerful resource to improve the understanding of these pathogens' biology. Trypanosomatids such as *Leishmania* spp. have their protein-coding genes grouped in long polycistronic units of functionally unrelated genes. The control of gene expression happens by a variety of posttranscriptional mechanisms. The high degree of synteny among *Leishmania* species is accompanied by highly conserved coding sequences (CDS) and poorly conserved intercoding untranslated sequences. To identify the elements involved in the control of gene expression, we conducted an *in silico* investigation to find conserved intercoding sequences (CICS) in the genomes of *L. major*, *L. infantum*, and *L. braziliensis*.

We used a combination of computational tools, such as Linux-Shell, PERL and R languages, BLAST, MSPcrunch, SSAKE, and Pred-A-Term algorithms to construct a pipeline which was able to: (i) search for conservation in target-regions, (ii) eliminate CICS redundancy and mask repeat elements, (iii) predict the mRNA's extremities, (iv) analyze the distribution of orthologous genes within the generated LeishCICS-*clusters*, (v) assign GO terms to the LeishCICS-*clusters*, and (vi) provide statistical support for the gene-enrichment annotation. We associated the LeishCICS-*cluster* data, generated at the end of the pipeline, with the expression profile of *L. donovani* genes during promastigote–amastigote differentiation, as previously evaluated by others (GEO accession: GSE21936). A Pearson's correlation coefficient greater than 0.5 was observed for 730 LeishCICS-*clusters* containing from 2 to 17 genes. The designed computational pipeline is a useful tool and its application identified potential regulatory *cis* elements and putative regulons in *Leishmania*.

1. Introduction

Genomic sequencing of a wide variety of organisms enables scientists to compare and explore the similarities and peculiarities among different species at the molecular level.

The trypanosomatids are protozoan parasites of medical relevance that have evolved a unique organization of their genetic information. The control of gene expression in these organisms differs from most of the eukaryotes mainly in the lack of transcriptional control, and in the absence of canonical RNA polymerase II promoters and of typical transcription factors. In trypanosomatids, most of the regulatory events occur at the post-transcriptional level. Protein-coding genes are organized as polycistronic units transcribed in a large pre-mRNA strand that are further co-transcriptionally processed by two coupled reactions known as *trans*-splicing and polyadenilation [1,2]. These two events are crucial for the proper maturation of monocistronic mRNA [3]. In addition, 20 species of the trypanosomatid *Leishmania* are the causative agents of diseases affecting 350 million people in 88 countries (www.who.int/leishmaniasis/en). The relevance of these organisms led to a number of genome initiatives on several members of the Trypanosomatidae family.

Among the genomes that have already been sequenced are three species of *Leishmania*. These are *Leishmania Leishmania major*, *Leishmania Leishmania infantum*, and *Leishmania Viannia braziliensis* (www.genedb.org and tritrypdb.org). Comparative analyses of

Abbreviations: CICS, conserved intercoding sequence; UTR, untranslated regions; ORF, open reading frame; CDS, coding DNA sequence; SIDER, short interspersed degenerated retroposon; DIRE, degenerated *ingi*-related element; GO, gene ontology; OCG, orthologous candidates in a set of genes.

* Corresponding author at: Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo/Av. Bandeirantes 3900, 14049-900 Ribeirão Preto, SP, Brazil.
Tel.: +55 16 36023318; fax: +55 16 36331786.
E-mail addresses: eltonjrv@usp.br (E.J.R. Vasconcelos), mcterrao@usp.br (M.C. Terrão), jeronimo@cpqrr.fiocruz.br (J.C. Ruiz), rvencio@usp.br (R.Z.N. Vêncio), akcruz@fmrp.usp.br (A.K. Cruz).

0166-6851/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi: 10.1016/j.molbiopara.2012.02.009

trypanosomatid genomes have revealed a high degree of synteny, which is even higher among *Leishmania* species [4,5]. Despite the estimated 20–100 million years that separates *Leishmania* (*Viannia*) and *Leishmania* (*Leishmania*) subgenera, more than 99% of the genes are syntenic between the three annotated genomes [5]. Nevertheless, conservation is rare among intercoding regions [6].

Because functional genomic elements are under selective pressure, they may acquire mutations at a rate slower than that of non-functional sequences. Thus, the phylogenetic footprinting hypothesis, first described by Tagle et al., in 1988, predicts that sequence conservation among non-coding regions surrounding homologous genes in different species most likely indicates the same functional role [7]. Although not useful on studies of specific genomic regions [10], the phylogenetic footprinting has been successfully used in a wide range of genome scale studies, in the search for regulatory *cis*-elements [7–9]. In fact, a variety of computational tools are currently available for phylogenetic footprinting aiming the discovery of regulatory elements [11–13]. Following this idea, by identifying conserved sequence motifs in a highly divergent genomic landscape, it is possible to discover new functional elements.

Herein we describe the establishment of an *in silico*-based pipeline to search for the conserved intercoding sequences (CICS) which might have functional roles in *Leishmania* genomes, including regulatory events at a post-transcriptional level. We used genomics and transcriptomics databases as tools for comparative analyses inter- and intra-species on the identification and characterization of CICS. Our approach resulted in a novel and rich source that is ready available for the trypanosomatid research community.

2. Methods

2.1. Obtaining Leishmania complete genomes

The *L.* (*L.*) *major* (MHOM/IL/81/Friedlin), *L.* (*L.*) *infantum* (MCAN/ES/98/LLM-877/JPCM5), and *L.* (*V.*) *braziliensis* (MHOM/BR/75/M2904) genomes (fasta files and annotation artemis files) are publicly available for download at ftp://ftp.sanger.ac.uk/pub4/pathogens/Leishmania. The annotation versions used in this study were *LmjFwholegenome_20070731_V5.2*, *LinJwholegenome_20080508.v3.0a*, and *LbrM_wholegenome_V2_29072008*, respectively. These versions have gene accession IDs consistent with the TriTrypDB v2.5.

2.2. Extraction of CDS-flanking regions

With *ad hoc* PERL scripts we extracted CDS-flanking sequences by taking an entire fasta file of a chromosome and another file containing the coordinates of each CDS from the same chromosome as its input. We took 2 kb upstream from the ATG and 2 kb downstream from the STOP codon of all CDSs in the three *Leishmania* genomes. These CDS-flanking sequence fasta files of each species are suitable for BLASTn analysis. In the case that the length between two CDSs was less than 4 kb the script divided the length value in two and assigned it to the 3′ end-flanking-region of the upstream CDS and 5′ end-flanking-region of the downstream CDS.

2.3. BLASTn analysis

To search for similarities in the *Leishmania* spp 2 kb CDS-flanking sequences we used the BLASTn [14] algorithm with the most sensitive word size (-W7), the low complexity regions filter turned on (-F T), and a 10^{-3} *E-Value* threshold. We first compared LmjF-2Kb-flanking-CDS against LbrM-2Kb-flanking-CDS and extracted the conserved sequences using the MSPcrunch algorithm [15], with the options -w -H. These LmjF-LbrM conserved products served as input for a second round of BLASTn against LinJ-2Kb-flanking-CDS. By re-running the MSPcrunch algorithm we obtained the conserved intercoding sequences (CICS) between the three *Leishmania* species in a redundant database that we called CICS_DB. In none of the described steps mutual BLAST was used.

2.4. Clustering CICS with SSAKE algorithm

To eliminate sequence redundancy on the CICS_DB fasta file and to facilitate the detection of how many, and which, CDSs share conserved intercoding sequences inter- and intra-species, we ran a novel clustering tool called SSAKE [16]. This open source algorithm has been developed to assemble millions of short sequences (20–30 bp) produced in the next-generation sequencing methods. Therefore, the CICS_DB.fasta was used as an entry file to SSAKE as if it was an output file from a deep-sequencing project. The algorithm was executed successfully with no warning messages and the parameters used were the following: -p 0 -c 1 -m 16 -z 20. Besides generating a fasta file with all the clustered sequences, SSAKE also provides a very informative file (*.readposition*) that shows how many times identical sequences overlap and their coordinates along the clustered sequence's length (personal communication with SSAKE's author). The clustering process allowed us to generate what we called SSAKE clustered CICS (sc-CICS), from which we generated a group of genes bearing a common CICS, which were named LeishCICS-*clusters*. The common sequence from a LeishCICS-*cluster* was named LeishCICS (see Section 3).

2.5. Filtering unclassified and simple repeats, low complexity regions, and non-coding RNA sequences

In spite of running BLASTn with the filter for low complexity regions turned on (-F T), we could not remove all repeats from our sc-CICS databases. Therefore, we used a specialized algorithm called RepeatMasker to run an additional filtering step (http://www.repeatmasker.org). We ran this program with the options -s and -specie "leishmania" and were able to identify unclassified and simple repeats, and low complexity regions as well. We used the RepeatMasker Library release 20090604. To mask non-coding RNAs we used a fasta file with 980 ncRNA sequences extracted from the tritrypDB downloadable file (LmajorAnnotatedTranscripts_TriTrypDB-2.5.fasta) as library.

2.6. Mapping previously characterized extinct retroelements (SIDERs and DIREs)

We obtained SIDER (short interspersed degenerated retroposon) sequences coordinates annotated in the *L.* (*L.*) *major* genome from supplementary table of Bringaud et al. [21], and DIREs (degenerated *ingi*-related elements) from the annotation files .artemis of *L.* (*L.*) *major* chromosomes downloaded from ftp://ftp.sanger.ac.uk/pub4/pathogens/Leishmania/major/CHROMOSOMES. We performed a BLASTn search (-FF -W7 -e1e-3) of sc-CICS against the whole *L.* (*L.*) *major* genome to identify SIDERs and DIREs in our dataset. With Shell/Unix and PERL language it was possible to take the BLAST best hit of each sc-CICS and to compare the BLAST subject coordinates with those from SIDERs and DIREs. The sc-CICS BLAST best hits that fell into the SIDER or DIRE coordinates, were considered as "within SIDER region" or "within DIRE region," respectively.

2.7. In silico prediction of the mRNA 5′- and 3′-UTRs

The Pred-A-Term algorithm [17] was used to predict the potential mRNA processing sites in the *Leishmania* chromosomes [*trans-splicing* acceptor sites (SAS) and polyadenylation sites]. We

wrote PERL scripts to automate the generation of fasta files which worked as input for Pred-A-Term and to evaluate whether LeishCICS could be located within a putative mRNA UTR, using a BLASTn threshold of 80% identity and 70% of alignment coverage. Distance (nt) of LeishCICS from the annotated start and stop codons and its orientation in the mRNA molecule were also assessed.

2.8. Orthology analysis of genes bearing LeishCICS

After the generation of the LeishCICS-*clusters* dataset, it was possible to search for orthologous candidates in the group of genes bearing a common LeishCICS (OCG) within their UTRs. We obtained a complete list of the orthologous candidates from LmjF-*vs*-LinJ, LmjF-*vs*-LbrM and LbrM-*vs*-LinJ after a BLASTp comparison (-FF -e1e-3 -v1 -b1). A PERL script was written to detect the presence of orthologous candidates between the LeishCICS-*clusters* of each species. R codes were used to construct the Venn diagram.

2.9. Assignment and GO enrichment analysis

It was also possible to assign functional themes to the set of genes within a LeishCICS-*cluster*. The association was made using Shell/Unix tools and *ad-hoc* PERL scripts, taking as a template the annotation files .gff (http://tritrypdb.org/common/downloads/release-2.5/Lmajor/Lmajor_TriTrypDB-2.5.gff, Lbraziliensis_TriTrypDB-2.5.gff, and Linfantum_TriTrypDB-2.5.gff), which contain the gene ontology (GO) classification for genes of each species studied here. Using R tools we applied the Fisher's exact test for the GO enrichment analysis of LeishCICS-*clusters* [18].

2.10. In silico validation of putative regulatory elements using whole transcriptome data

Recently, the largest time series analysis of gene expression during differentiation of promastigotes to axenic amastigotes was published for *Leishmania*. The authors used a microarray chip with *L.* (*L.*) *infantum* probes and evaluated the *L.* (*L.*) *donovani* transcript levels [19]. This data is available at the Gene Expression Omnibus database from NCBI (http://www.ncbi.nlm.nih.gov/geo) with the GSE21936 accession number. We performed a BLASTn (-FF -W7 -e1e-5) of all CDS probes from GPL10446 (see Lahav et al. [19]) against all *L.* (*L.*) *infantum* annotated CDSs (TriTrypDB-v2.5), taking only those probes presenting a single CDS perfect match (100% identity with the entire probe sequence), to allow the association between probe IDs and CDS IDs. For our Pearson's correlation analysis on genes from LeishCICS-*clusters*, the database of transcript expression values contained only those probes related to the 5619 genes filtered on the basis of statistical significance [19]. Scripts in R language were written for this purpose.

2.11. Discovering motifs in LeishCICS-clusters training set

The MEME algorithm version 4.5 [20], specialized in the discovery of motifs in a set of biological sequences, was used to strengthen LeishCICS as putative regulatory motifs. As the training set for the algorithm, we took a sample of 100 LeishCICS-*clusters* using their 2 kb CDS-flanking sequences as input. The parameters used were: -dna -minw 20 -maxw 200 -revcomp -nmotifs 10. We used the get_fasta_stats.pl script (http://www.genome.ou.edu/informatics.html) and Unix/Shell tools to generate our background Markov model file specified in the –bfile option.

3. Results and discussion

The main result of this work was the establishment of a novel information resource useful for the search of functional and regulatory elements. This resource is publicly available to the trypanosomatids' research community at: http://labpib.fmrp.usp.br/LeishCICS/ (Supplementary data 1 and 2). The LeishCICS resource comprises a sequence database (LeishCICS.fasta) and the set of genes bearing the same conserved sequences in their CDS-flanking regions (LeishCICS-*clusters*). The usefulness of the resource was validated by indicating regulatory motifs common to a group of co-regulated genes using one of the most comprehensive system-level time-series transcriptome data publicly available for *Leishmania* (GSE21936: NCBI-GEO accession number).

To achieve that we used a combination of computational tools, such as Linux-Shell, PERL and R languages, BLAST, MSPcrunch, SSAKE, and Pred-A-Term algorithms to construct a pipeline. Combination of these algorithms permitted us to (a) format data files, (b) search for conservation in target-regions and create a redundant fasta file with millions of CICS, (c) cluster this fasta file to eliminate redundancy and to generate LeishCICS-*clusters*, (d) predict the mRNA's extremities and keep only those CICS which were within 5′- or 3′-UTRs, (e) search for orthologous genes within each LeishCICS-*cluster* dataset, (f) assign GO terms to the LeishCICS-*clusters* and (g) give statistical support for the gene-enrichment annotation.

3.1. Generation of clustered CICS

All the sequences that flank annotated CDSs from the genomes of *L.* (*V.*) *braziliensis*, *L.* (*L.*) *infantum*, and *L.* (*L.*) *major* were extracted from the whole genome database of each organism using *ad hoc* PERL scripts (see Section 2 and Fig. 1). Based on BLASTn similarity searches, the pipeline was designed to identify the conserved sequences (greater than 75% identity) and to map them on the nearest neighbor CDS of each species. The rescued CICS were organized in an intermediate redundant data bank (redundant CICS_DB) which has 12,894,423 million sequences (Fig. 1).

CICS_DB is a redundant fasta file with more than 12 million sequences varying from 20 to 700 nt in length. Of these 12 million sequences, 99.3% are less than 100 nt long and approximately 5.9 million are short sequences (20–30 nt). Therefore, we used a special algorithm, SSAKE [16], to cluster these sequences and to eliminate redundancy and group identical sequences (Fig. 1). The algorithm originally developed to assemble short sequences produced by the next-generation sequencing strategies was useful to assemble CICS. This analysis generated 16,226 SSAKE-clustered CICS, the so-called sc-CICS.

3.2. CICS classification: SIDERs, DIREs, and novel putative elements

To organize the generated database we performed a computational analysis to seek characterized sequences and unravel novel elements with putative functional roles.

From 16,226 sc-CICS sequences, we identified several classes of repeats and two well characterized retroelements, the SIDERs and the DIREs, providing support that the applied approach is robust.

These degenerated retroelements were recently found in an *in silico* screening of the *L.* (*L.*) *major* genome [21]. These are elements (~500 nt long) that have lost their transposition capability and are located mainly in the Directional Gene Clusters (DGCs) coding strand, more precisely in the 3′-UTRs of some genes. They are either involved in the stage-specific translational control (SIDER1) [22,23] or in the destabilization of the transcript (SIDER2) [21]. Those SIDER elements have also been described for the *L.* (*L.*) *infantum* and *L.*

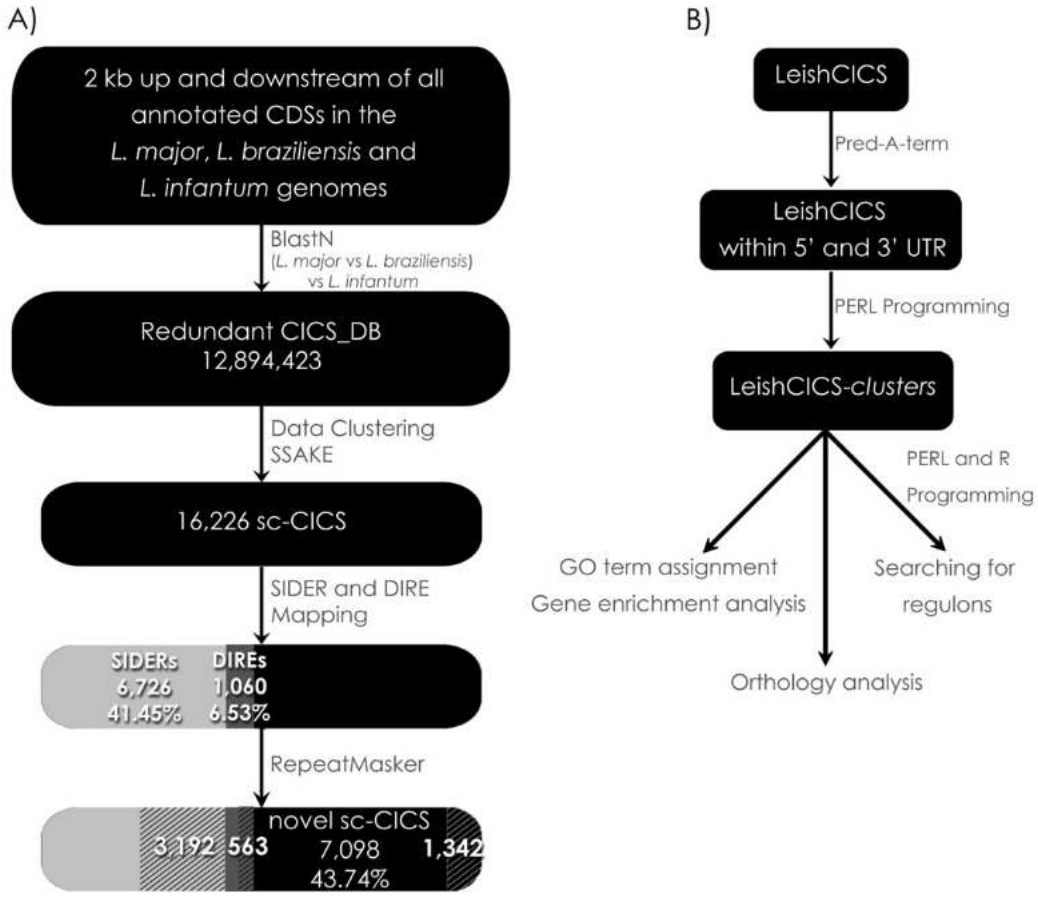

Fig. 1. The computational pipeline based on BLAST similarity searches and SSAKE clustering. (A) We performed a BLASTn of all *L.* (*L.*) *major* CDS-flanking regions against all *L.* (*V.*) *braziliensis* CDS-flanking regions. The conserved products of this first round of BLASTn were then submitted to a second round against all the *L.* (*L*) *infantum* CDS-flanking regions. The redundant CICS_DB contained 12,894,423 conserved sequences between the three species. The CICS_DB were then clustered by SSAKE and 16,226 SSAKE-clustered CICS (sc-CICS) were rescued. The sc-CICS were then mapped within SIDER (6726) and DIRE (1060) regions and subsequently masked by RepeatMasker to filter previously described repeats and known ncRNAs (3192 and 563 unclassified repeats within SIDERs and DIREs, respectively and 1342 from miscellaneous origins). After this procedure we obtained 7098 novel sc-CICS to be further analyzed. (B) After taking the maximum of three multiplets of each selected sc-CICS (see text from results on Section 3.3), we obtained 9225 sequences, the LeishCICS. Subsequently, we performed a series of analysis which involves: (i) the mRNA's UTRs prediction with Pred-A-Term algorithm and the mapping of LeishCICS within these regions; (ii) the generation of LeishCICS-*clusters* (set of genes bearing a common LeishCICS within their UTRs); and (iii) searching for orthology, significant ontologies and regulons on the LeishCICS-*clusters* dataset.

(*V.*) *braziliensis* genomes and an *in silico* analysis of 6802 orthologous genes from *L.* (*L*) *major*, *L.* (*L*) *infantum*, and *L.* (*V.*) *braziliensis* revealed 839 orthologs harboring SIDER elements [24]. Thus, these extinct retroposons may have evolved differently and obtained preferential assimilation or conservation in some species, being involved with genotypic diversity and species-specific patterns of gene expression control. Other *Leishmania* truncated retroposons, derived from the *T. brucei ingi* and ribosomal mobile elements (RIME), named LmDIREs (*L. major* degenerated *ingi*-related elements) [25], were also found in our sc-CICS.

We mapped those sc-CICS located within SIDERs and DIREs in the *L.* (*L*) *major* genome using SIDER and DIRE coordinates from the genome (Bringaud et al. [21]; and artemis files of *L.* (*L.*) *major* chromosomes downloaded from ftp://ftp.sanger.ac.uk/pub4/pathogens/Leishmania/major/CHROMOSOMES, respectively). Because these elements are approximately 0.5 kb long [21] and most of the CICS herein identified were shorter than 100 bp, we termed the sc-CICS located within SIDER or DIRE coordinates as "within SIDER region" or "within DIRE region," respectively (Fig. 2).

To roughly estimate the sensitivity of our pipeline we considered the fraction of existing SIDERs that can be readily identified in our sc-CICS sequence database relative to the number of elements existing simultaneously in all three genomes analyzed. We calculated the number of distinct SIDER coordinates present in the 6726 sc-CICS within SIDER regions (Fig. 2 "first bar"). Using the supplementary table from Bringaud et al. [21], which contains 1851 LmSIDERs, their chromosome location and IDs, it was possible to identify 807 distinct LmSIDER IDs among the 6726 subset of sc-CICS. We achieved 43.6% (807/1851) sensitivity. Considering the sequence heterogeneity of SIDER copies within a single genome and the fact that we have compared genomes from three species, this result indicates that the pipeline was able to identify regulatory *cis*-elements.

The same 16,226 sc-CICS sequences were subjected to the RepeatMasker algorithm (see Section 2.5) to more stringently remove the low complexity sequences and repeats that had not been eliminated by the BLASTn analysis previously conducted. Among simple repeats and low complexity regions, RepeatMasker was also able to detect two types of *Leishmania* unique repeats described in the RepBase repository (www.girinst.org) as "unclassified" repeats: RS3_LM and IR2_LM, both from *L.* (*L.*) *major* [26,27]. In addition, direct and inverted repeats were also found, which are known to facilitate gene amplification, contributing to the parasite's genetic plasticity [28–32]. One hundred and ninety-six RS3_LM and 4262 IR2_LM were masked and included as "unclassified repeats" (Fig. 2, hatched first and second bars and third bar). Simple repeats and some of the annotated *L.* (*L.*) *major* ncRNAs were

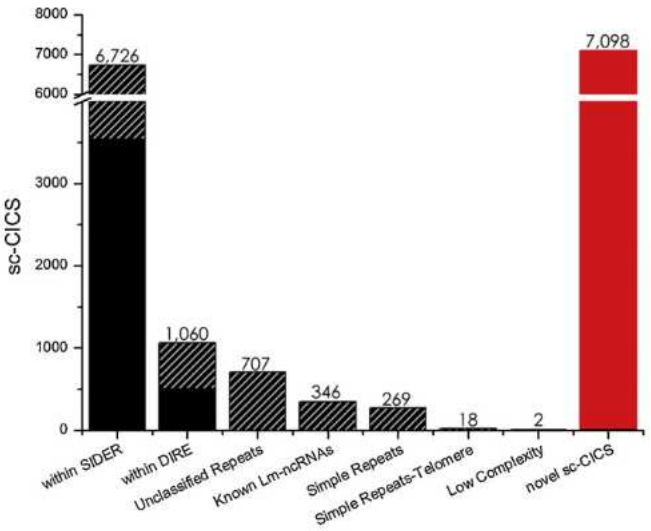

Fig. 2. Conserved InterCoding Sequences classification (total of 9128 sequences). The SSAKE clustered CICS (sc-CICS) were located within SIDER and DIRE regions (6726 and 1060 respectively, summing up 47.9% of the total sc-CICS). The 3192 sc-CICS within SIDER and 563 within DIRE regions were masked as unclassified repeats (hatched regions on corresponding bars). RepeatMasker also detected known *L. (L.) major* non-coding RNA genes, simple repeats, and low complexity repeats (hatched bars). The novel sc-CICS represents 43.7% of the total sc-CICS [16,226 (classified + unclassified) − 9128 (classified) = 7098].

masked; they represent 1.8% (289/16,226) and 2.1% (346/16,226) of the total sc-CICS dataset, respectively (Fig. 2).

The results obtained suggest that the designed computational pipeline is able to detect potential regulatory elements, as the sc-CICS within SIDER regions constitute the second major group, representing 41.45% (6726/16,226) of the sc-CICS dataset (Figs. 1 and 2). sc-CICS within DIRE regions were also represented, but at lower levels (6.3% of the total sc-CICS) (Figs. 1 and 2). The remaining 7098 sc-CICS (43.7%), which did not fall into any of the predefined categories, were named novel sc-CICS and were further investigated *in silico* to evaluate their potential as functional *cis*-elements of the *Leishmania* spp genomes (Figs. 1 and 2).

3.3. Generation of LeishCICS-clusters: genes bearing common CICS

After the exclusion of known elements from the 16,226 clustered sequences, the "novel sc-CICS" were used as bait to generate a database of *Leishmania* genes bearing common CICS. To generate such database we used SSAKE, and must emphasize that its output files may contain entirely identical fragments or regions with patches of identity. Given that the conserved sequences assembled do not necessarily represent a continuous region in the analyzed genomes, we arbitrarily extracted from the SSAKE *.readposition* file (see Section 2) those identical fragments more frequently represented in a given sc-CICS, which we named a "multiplet", because it occurs more than once in the analyzed genomes. Within a sc-CICS there may be one to several multiplets (Fig. 3A and B). Novel sc-CICS may also be derived from a single occurrence in each genome, which we named "singlet" (Fig. 3C).

For the investigation further conducted with the sc-CICS database, to the 7098 novel sc-CICS (Fig. 2) we added 343 sc-CICS from the set of non-SIDER and non-DIRE "unclassified repeats" containing multiplets (third bar on Fig. 2 and data not shown). The total was 7441 sc-CICS. An *ad hoc* PERL script was designed to select from each sc-CICS the maximum of three multiplets, the ones most abundantly represented (as exemplified by circles in Fig. 3A). At the end of this process we had 9225 final sequences to work with (4476 derived from multiplets and 4749 derived from singlets), which were named LeishCICS (Supplementary data 1 and 2).

Because known *cis* regulatory elements are located within the mRNA UTRs, it would be of relevance to position LeishCICS relative to all the sites for *trans*-splicing and polyadenylation on the annotated mRNAs from the *Leishmania* genomes. For that, we used the Pred-A-Term algorithm [17] and a series of PERL scripts were written to localize the LeishCICS within the putative UTRs. All the LeishCICS-*clusters*, their genes and predicted UTRs are depicted on the three tabular files (one for each species) provided in Supplementary data 2. Those files contain LeishCICS IDs, gene accession numbers (TriTrypDB v2.5), UTR length and LeishCICS sequences, their distance (nt) from the START or STOP codons and orientation within the mRNA. Noteworthy, LeishCICS within putative 3′-UTRs are much more frequent than within 5′-UTR, about

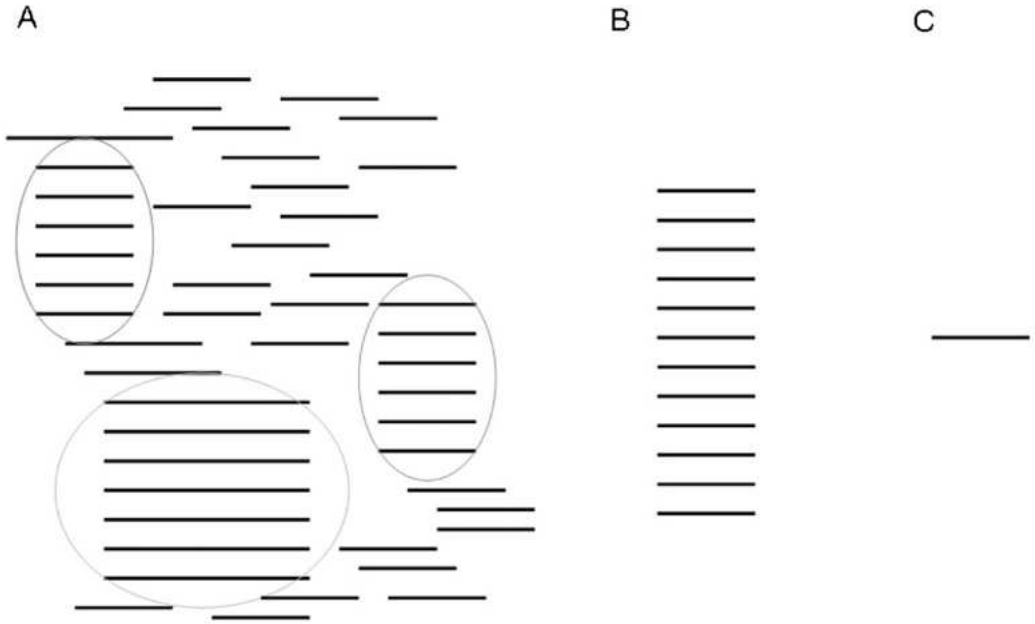

Fig. 3. Types of CICS clustered by SSAKE algorithm (sc-CICS). (A) and (B) are multiplets, identical sequences occurring in many regions of the genomes. (C) Single CICS on the three *Leishmania* genomes.

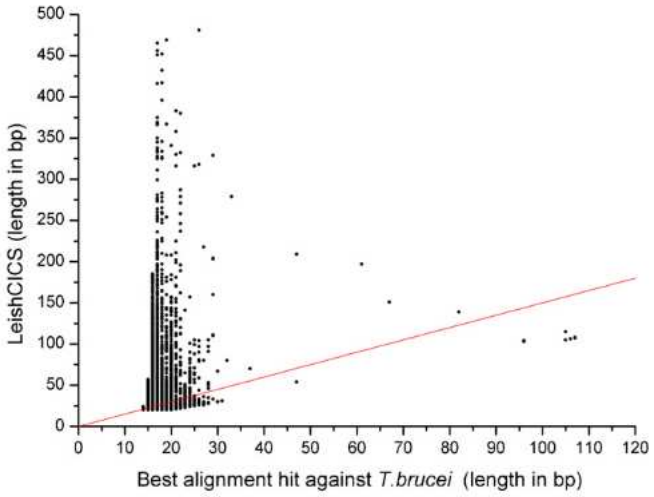

Fig. 4. BLASTn search of the LeishCICS *versus Trypanosoma brucei* whole genome (TbTREU927 from tritrypDB version 2.5). The *Y*-axis is the query sequence length and the *X*-axis represents the best-hit alignment length in a threshold of 80% identity, and *e-value* = 1. Dots under the line represent conserved sequences in more than 2/3 of the query length between LeishCICS and *T. brucei*. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

85% of the cases in each species (see Supplemetary data 2), corroborating previous studies which revealed the participation of *cis* elements mainly within the 3′-UTR [21,23,33,34]. This database proved useful for the subsequent analyses in the search for putative post-transcriptional regulons in *Leishmania*. We named LeishCICS-*clusters* those sets of genes bearing the same conserved sequences in their putative UTRs.

3.4. Searching LeishCICS in Trypanosoma brucei genome

To investigate conservation of LeishCICS in another trypanosomatid, we performed a BLASTn comparison of the 9225 LeishCICS against the whole *Trypanosoma brucei* genome (TREU927 from tritrypDB v2.5). Using a threshold of identity >80% and e-value < 1, we obtained hits for 7145 LeishCICS, but only for 2996 did the alignment extension cover more than 2/3 of the query length (Fig. 4, dots under red line). As shown in Fig. 4, the vast majority of larger LeishCICS (>100 nt) did not present reliable hits in *T. brucei*. The four largest LeishCICS (varying in size from 582 to 696 nt) did not match any *T. brucei* sequence. They either flanked ORFs of *Leishmania*-only hypothetical conserved proteins or the first CDS of two transcription units; one from chromosome 32 and another from chromosome 09 (data not shown).

All the LeishCICS bearing a *T. brucei* homologue sequence longer than 90 bp (LeishCICS-s4915, LeishCICS-s11049, LeishCICS-s4152, LeishCICS-s3222, LeishCICS-s6171, LeishCICS-s3287, LeishCICS-s9083; Fig. 4, dots under red line and greater than 90 bp) matched a single genomic region, within the annotated 5′-UTR of a hypothetical *T. brucei* protein (Tb927.8.6040). Nevertheless the Pred-A-Term prediction positioned these LeishCICS within the putative 3′-UTR of the *Leishmania* 60S ribosomal L12 protein gene (LmjF24.2210, LinJ24_V3.2300, and LbrM24_V2.2290). Curiously, this gene is just upstream of the Tb927.8.6040 *Leishmania* ortholog, and the predicted 5′UTR of Tb927.8.6040 is unusually long (~1.7 kb), as determined by Digital Gene Expression analysis [35]. Interestingly, recent RNA-seq analyses of the *L. (L.) major* and *L. (L.) infantum* transcriptomes confirmed this long 5′-UTR for this hypothetical protein gene on those species (tritrypdb.org). Therefore, it is possible to assume that the *in silico* predicted lengths of the *Leishmania* UTRs in this region are incorrect, and the conserved elements are part of the 5′-UTR of the LmjF24.2200, LinJ24_V3.2290, and LbrM24_V2.2280 genes. Considering the evolutionary distance between the genera these LeishCICS may turn out to be a novel 5′ *cis* regulatory element.

Another LeishCICS conserved in *T. brucei* (LeishCICS-s7888, Fig. 4, dot under red line with 47 bp on *X* axis) caught our attention. It is a sequence 54 nucleotides long (*Y* axis) that might be part of the 3′-UTR of orthologous genes coding for the putative replication factor C, subunit 5 (Tb927.10.7990, LmjF36.6710, LinJ36_V3.7020, and LbrM35_V2.7060). Most of the LeishCICS (81%) are about 20–60 nt long, and the great majority of sequences within this class (2983 LeishCICS) have also been maintained in *T. brucei* (Fig. 4, dots under red line between 15 and 30 bp on *X* axis).

The detection of LeishCICS across distinct genera of trypanosomatids suggests the existence of a strong selective pressure. Considering the estimated time of divergence between *Trypanosoma* and *Leishmania* genera, predating the emergence of mammals (~165 million years) [36], we hypothesize that the common TrypCICS may represent important functional elements for these parasites' biology. Further comparative analyses implemented with other trypanosomatid genera need to be conducted for a better identification and characterization of TrypCICS.

3.5. LeishCICS bears short motifs previously described

Recently, in a high-throughput study of the *T. brucei* cell cycle-regulated transcriptome, Archer and co-workers identified short motifs (5–8 nt) that are present in the UTRs of co-regulated genes during the cell cycle [37]. We searched for these short motifs using regular expression, and considered both their forward and reverse orientation. We found four LeishCICS bearing the 8-mer and five harboring the 9-mer motifs possibly located within the 3′-UTRs of mRNAs (Table 1). The octameric cycling control sequence [(C/A)ATAGAA(G/A)] has been previously described in *Crithidia fasciculata* [38] and *L. (L.) major* [39]. The 9-mer [ATGTAnAGT] motif has been identified in the 3′UTR of the *L. (L.) mexicana* paraflagellar rod gene (PRE, paraflagellar rod regulatory element) [33,40]. None of the LeishCICS presenting the octameric cycling control motif were found flanking those genes mentioned by Zick et al. [39] (*RPA1*, *TOP2*, *KPA3*, and *DHFR-TS*). Therefore, we believe that this 8-mer is not common to the flanking regions of the four aforementioned genes in the three species studied here. On the other hand, we found a different group of genes bearing that 8-mer motif: cytochrome c oxidase subunit V, and three hypothetical protein genes (Table 1). Our investigation showed that the PRE 9-mer motif is also found in the orthologous genes of the *Leishmania* species analyzed (Table 1).

3.6. Using MEME algorithm to reinforce LeishCICS as motifs

To verify if our conservation-based computational method was able to discover putative regulatory motifs, we ran MEME version 4.5, an algorithm specialized for the elicitation of motifs in a user pre-defined set of biological sequences [20]. Despite its great qualities, MEME takes too long to run and needs a training set of input sequences that can not be as large as whole genomes or even the original conserved intercoding sequences DB (redundant CICS_DB). Therefore, we needed a smaller trained sample. Thus, to confirm LeishCICS as putative regulatory motifs, we took a sample of 100 LeishCICS-*clusters* using their 2 kb CDS-flanking sequences as the training set for MEME. MEME identified as conserved motifs 85 out of the 100 LeishCICS, being 90% of them significant (e-values $< 5 \times 10^{-5}$, data not shown).

3.7. Distribution of orthologous genes within LeishCICS-clusters

Because it would be expected that orthologous genes might keep the same *cis*-regulatory elements in different species, we found it valuable to explore the LeishCICS-*cluster* database (9225

Table 1
LeishCICS possessing previously described short regulatory motifs located at the 3′UTR of the indicated *Leishmania* genes.

Regulatory motif	LbrM gene ID Tritryp-v3.1	LmjF gene ID Tritryp-v3.1	LinJ gene ID Tritryp-v3.1	Product description	LeishCICS ID
8-mer [(C/A)ATAGAA(G/A)]	LbrM.14.0430	LmjF.14.0420	LinJ.14.0430	Hypothetical protein, conserved	LeishCICS-s6297
8-mer [(C/A)ATAGAA(G/A)]	LbrM.22.1500	LmjF.22.1590	LinJ.22.1440	Hypothetical protein, conserved	LeishCICS-s7302
8-mer [(C/A)ATAGAA(G/A)]	LbrM.23.1970	LmjF.26.1710	LinJ.26.1690	Cytochrome c oxidase subunit V	LeishCICS-s8858
8-mer [(C/A)ATAGAA(G/A)]	LbrM.27.2490	LmjF.27.2305	LinJ.27.2230	Hypothetical protein, conserved	LeishCICS-s10140
9-mer [ATGTAnAGT]	LbrM.16.1490	LmjF.16.1425 and LmjF.16.1425	LinJ.16.1510 and LinJ.16.1520	Paraflagellar rod protein 2C	LeishCICS-s5364
9-mer [ATGTAnAGT]	LbrM.25.0910	LmjF.25.1020	LinJ.25.1060	Dehydrogenase-like protein	LeishCICS-s9909
9-mer [ATGTAnAGT]	LbrM.30.0110	LmjF.30.0100	LinJ.30.0100	Hypothetical protein, conserved	LeishCICS-s9836
9-mer [ATGTAnAGT]	LbrM.31.0160	LmjF.29.1760 and LmjF.29.1770	LinJ.29.1880 and LinJ.29.1890	Paraflagellar rod protein 1D	LeishCICS-s4814
9-mer [ATGTAnAGT]	LbrM.32.0490	LmjF.32.0420	LinJ.32.0430	Hypothetical protein, conserved[a]	LeishCICS-s6352

[a] GO process: microtubule-based movement process.

LeishCICS) regarding the presence of orthologous genes within each cluster. The analysis revealed that 61% of the LeishCICS-*clusters* (5665/9225) held orthologous candidates in their group of genes (OCG). In addition, consistently with species evolutionary distances, we observed that most of the OCG sharing a common LeishCICS came from *L.* (*L.*) *major* and *L.* (*L.*) *infantum* (94% of the LeishCICS-*clusters* containing OCG, 5332/5665, Fig. 5 and Supplementary data 5), while fewer OCG from *L.* (*V.*) *braziliensis* were found; 50% of the LeishCICS-*clusters* containing OCG have *L.* (*V.*) *braziliensis* orthologues (2857/5665, Fig. 5 and Supplementary data 5). We have also found that 59% of the LeishCICS-*clusters* containing the three possible orthologous pairs (1159/1962) are present in only one gene per genome, and the three of them are orthologues (Supplementary data 5).

There is a relatively high percentage (38.6%) of LeishCICS-*clusters* (3560/9225 CICS) that do not contain annotated orthologous genes. Interestingly, the LeishCICSs in 52% of these clusters (1857/3560) are not within the predicted UTRs of any gene in, at least, two species (data not shown). We must emphasize that Pred-A-Term algorithm may have failed to predict or mispredicted some of the UTR lengths.

Although orthologous genes from the three *Leishmania* species could be expected to share the same *cis*-regulatory elements, it is possible to assume that the different species may have evolved distinct forms of control of a given gene (or groups of genes). In this case, the *cis*-elements of orthologous genes could have diverged. This scenario is possible and could, in fact, be the explanation for some physiological or host-parasite interaction differences. Noteworthy, the *Leishmania* species investigated are biologically diverse and cause distinct diseases, but their genomes are highly syntenic and their protein coding genes highly conserved. Several features that distinguish these species could result from different processes and elements involved in the regulation of gene expression.

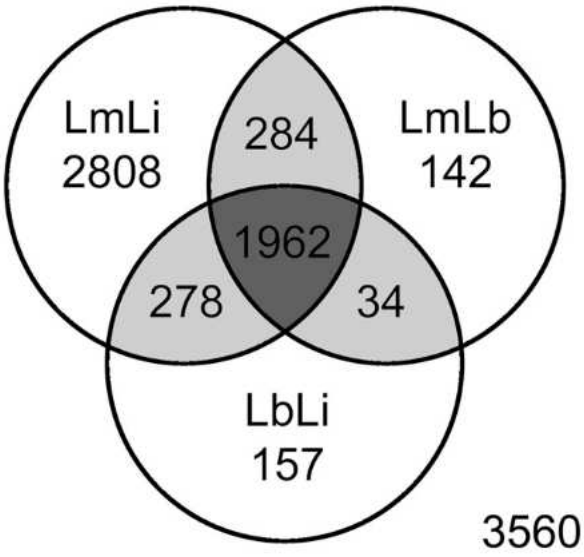

Fig. 5. Distribution of orthologous genes within LeishCICS-*clusters*. The LeishCICS-*clusters* dataset was searched for the presence of orthologous candidates in the group of genes (OCG) bearing a common LeishCICS within their UTRs. The Venn diagram shows the total number of LeishCICS-*clusters* (5665) presenting at least a pair of orthologous candidates. The combination of orthologous pairs are depicted by LmLi, LmLb, LbLi (for pairs of *L.* (*L.*) *major* and *L.* (*L.*) *infantum*, *L.* (*L.*) *major* and *L.* (*V.*) *braziliensis* or *L.* (*V.*) *braziliensis* and *L.* (*L.*) *infantum*, respectively) along with the corresponding number of LeishCICS-*clusters*, in each circle of the Venn diagram. The number 3560, placed on the right inferior corner of the figure, refers to the number of LeishCICS-*clusters* that do not contain orthologous genes.

3.8. Searching for *Leishmania* putative RNA regulons based on LeishCICS

The lack of transcriptional control of gene expression in trypanosomatids led us to hypothesize that the presence of *cis*-elements common to a group of genes across species is an indication of co-regulation. These commonly regulated genes would be part of a "post-transcriptional regulon".

A recently conducted large-scale analysis of gene expression patterns in different life stages of *T. brucei* revealed the existence of post-transcriptional regulons and led authors to speculate that

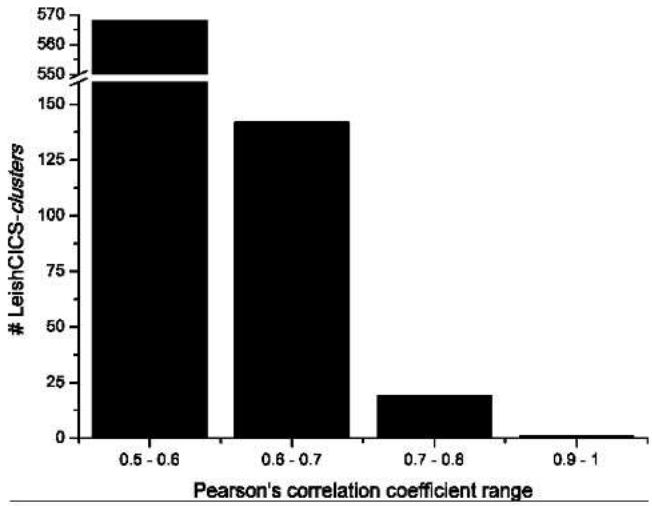

Fig. 6. LeishCICS-*clusters* grouped according to the Pearson's correlation coefficient range. The *L.* (*L.*) *infantum* transcriptome data (GSE-21936) representing the profile of gene expression throughout the promastigote to amastigote differentiation period was the input data for the conducted Pearson's correlation analysis (see text for details).

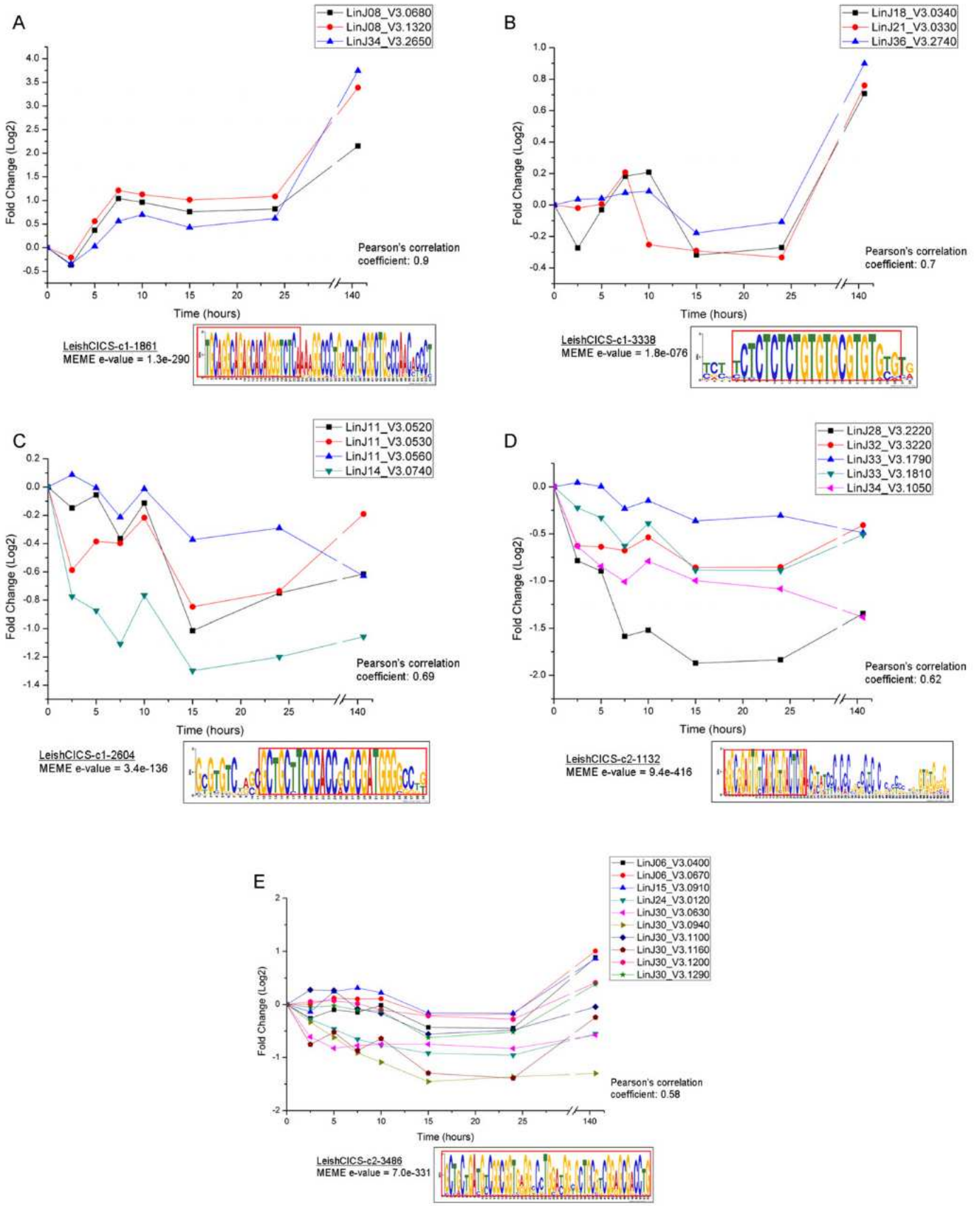

Fig. 7. Co-regulated mRNA transcripts that bear common LeishCICS within their putative 3′-UTRs. The gene expression values were measured at eight time points (0 h, 2.5 h, 5 h, 7.5 h, 10 h, 15 h, 24 h, and 144 h) of differentiation between promastigote and amastigote stages *L. (L.) donovani*. The LeishCICS were *in silico* validated by MEME algorithm (red border boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

genes linked to a regulon might share conserved sequences in their UTRs [41]. The same group identified regulatory motifs in sets of cell-cycle co-regulated genes [37]. Therefore, the LeishCICS-*cluster* dataset constitutes a robust *in silico* source to investigate a possible co-regulated group of genes in *Leishmania*.

To evaluate whether *Leishmania* genes sharing a LeishCICS constitute a regulon, we used the resource generated by a high-throughput analysis of *L.* (*L.*) *donovani* gene expression recently published [19]. These authors analyzed the transcriptome changes during differentiation from promastigote to amastigote stages using a microarray chip of *L.* (*L.*) *infantum* genome (available at the NCBI-GEO database, GSE21936). Using the GSE21936 dataset we selected the 5619 genes that have been previously filtered based on their Pearson's correlation (>0.4 within replicates). This number of genes represents 68.2% of the whole *L.* (*L.*) *infantum* protein-coding genes. We used this set of filtered *L.* (*L.*) *infantum* genes to search for a possible correlation between their expression patterns with the presence of LeishCICS. Our analysis revealed that 730 LeishCICS-*clusters* (16% of the 4476 LeishCICS derived from multiplets) presented a Pearson's correlation coefficient of >0.5 (Fig. 6 and Supplementary data 3). Given the filtered genes in Lahav's work, the LeishCICS-*clusters* may contain genes presenting

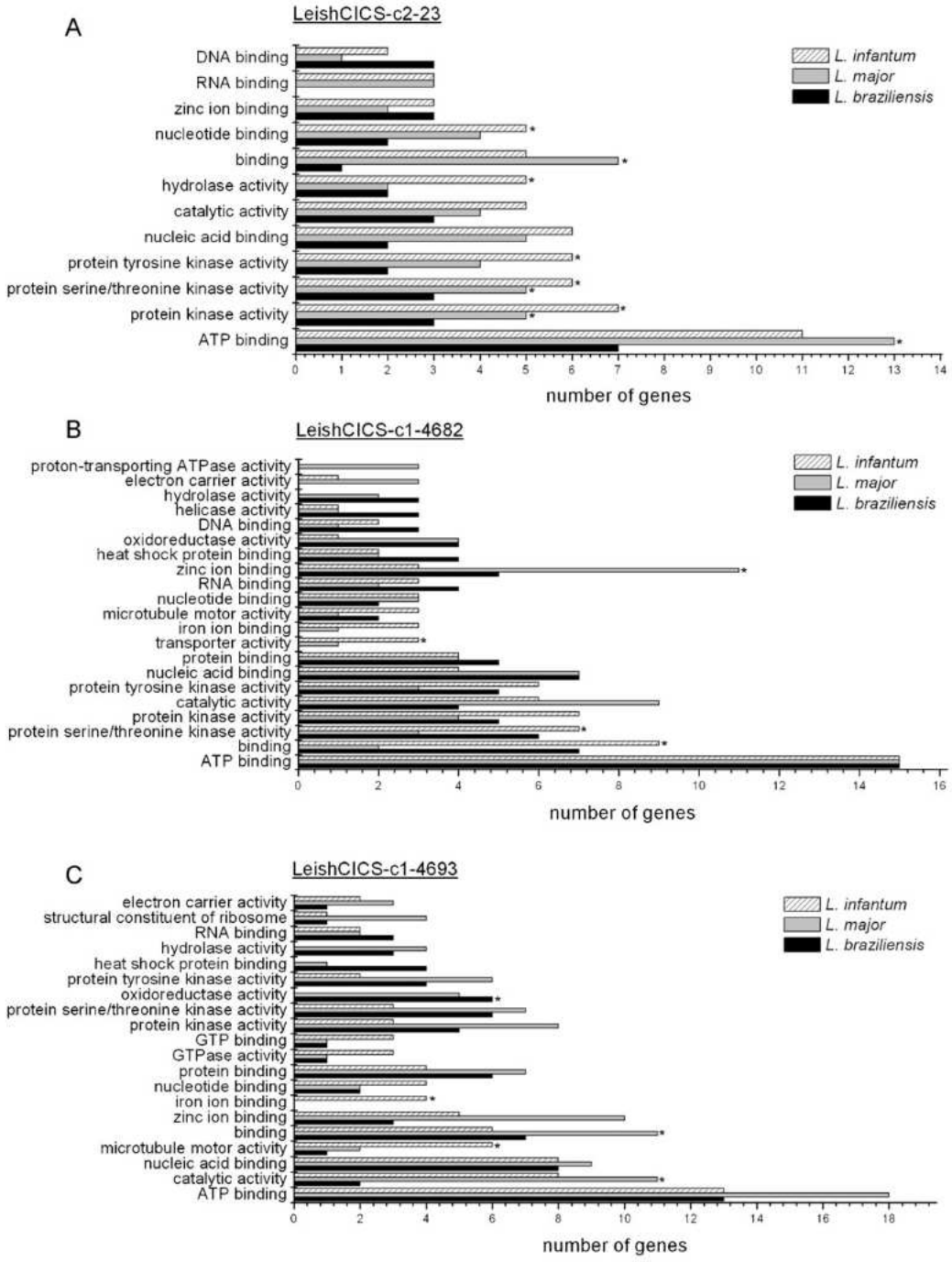

Fig. 8. Clustering genes by functional themes. Three of the most representative LeishCICS-*clusters* (clusters with an elevated number of genes in each species). LeishCICS-c2-23 (A), LeishCICS-c1-4682 (B), and LeishCICS-c1-4693 (C) had their genes grouped by molecular function (GO terms). Gene enrichment analysis was conducted using Fisher's Exact Test and the asterisks (*) indicate that the set of genes is specifically associated (enriched) to the given term (Fisher Test *p-value* < category frequency on gene's list for each species in a cluster).

non-reliable expression values, which were excluded from our Pearson's correlation analyses.

A mean of 4.4 genes per LeishCICS-*cluster* was found in those groups of genes with a Pearson's correlation coefficient ranging from 0.5 to 0.6 (varying from 2 to 17 genes per LeishCICS-*cluster*). More stringent Pearson's correlation values (from 0.6 to 0.7 and 0.7 to 0.8) rescued LeishCICS-*clusters* with a smaller number of genes (from 2 to 9 and 2 to 3 genes, respectively). Only one LeishCICS-*cluster* (LeishCICS-c1-1861) presented a coefficient above 0.9 (Fig. 6); it includes three amastin-like protein genes filtered by Lahav's threshold (Fig. 7A and Supplementary data 3). It is worth mentioning that the LeishCICS-c1-1861 element is within the first 1100 nucleotides of the long *L.* (*L.*) *infantum* amastin 3′-UTR transcript (Supplementary data 2 – LeishCICS-cluster-LinJ.tab). It has been shown previously that two different elements within this UTR play a role in the control of amastin expression. Nevertheless, the sub-region of LeishCICS-c1-1861 has not been explored yet [23]. It is also present within the 3′UTR of 13 *L.* (*L.*) *major* amastin-like transcripts (11 in chromosome 8 and 2 in chromosome 34) and may be an extra regulatory element of the UTR.

LeishCICS-c1-3338 flanked three genes filtered by Lahav's threshold presenting higher expression levels in amastigotes (Fig. 7B and Supplementary data 3). We observed a high number (~50%) of thymidine residues in the motif (Fig. 7B) and it is known that U-rich elements (URE) in 3′UTRs may be involved in the kinetics control of mRNA decay in a deadenylation-independent manner [34]. Therefore, the mRNA decay kinetics of genes bearing LeishCICS-c1-3338 should be investigated.

Other examples of groups of genes with common LeishCICS displaying a notable synchronicity of mRNA variation levels throughout the differentiation period are depicted (Fig. 7C–E).

The highlighted LeishCICS, as well as others that have been rescued with the pipeline, are potential regulatory motifs. These sequences may act as binding sites for regulatory proteins, playing a key role in post-transcriptional mechanisms of gene expression control. Nevertheless, experimental validation to confirm the profile of mRNA levels during differentiation and correlation of the identified CICS with such patterns must be conducted.

3.9. Gene enrichment by the assignment of GO categories to LeishCICS-cluster genes

The assignment of GO categories to genes sharing a CICS on their flanking regions may add relevant information to associate those genes with their putative functions to infer the existence of a possible regulon. This approach has been already used to group those genes bearing SIDERs in *L.* (*L.*) *major*, *L.* (*L.*) *infantum*, and *L.* (*V.*) *braziliensis* genomes as a mean of unraveling shared biological functions or processes [24]. The main focus of this kind of analysis is to act as a tool to cluster genes by themes after a high-throughput experiment.

Those groups of functionally related genes in different LeishCICS-*clusters* are depicted in Fig. 8 and Supplementary data 4. We have used the GO molecular function terms assigned to the *Leishmania* genes in the GFF files downloaded at tritrypdb.org (see Section 2). After the gene set enrichment analyses some categories have gained prominence, indicating that the list of genes in a LeishCICS-*cluster* is specifically associated (enriched) to those terms (asterisks in Fig. 8). We may speculate that conserved motifs within UTRs, common to a group of genes included in a GO term, might be required in regulatory events shared by the functionally related mRNAs.

4. Conclusion

We generated a computational pipeline to identify, map, and characterize conserved intercoding sequences (CICS) in the *Leishmania* genomes, aiming to provide publicly available data on conserved sequences to be further tested as *cis*-elements involved in gene expression regulation. Our results indicate that the developed strategy is useful to identify conserved sequences that might have regulatory roles in *Leishmania* species. In fact, we rescued several regulatory *cis*-elements previously described by others. We demonstrated that several of the novel conserved *cis*-elements are placed within untranslated regions of *Leishmania* transcripts that share a similar pattern of expression. This indicates that those CICSs may play a functional role on the control of gene expression.

Authors' contributions

EJRV performed all the *in silico* analyses, idealized the computational pipeline and drafted the manuscript; MCT contributed to the design of the study; JCR conceived the original project and helped in some computational analyses; RZNV performed the statistical analysis; AKC contributed to the conception and coordination of the study; EJRV, MCT, RZNV and AKC wrote the final manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by AKC FAPESP grant 06/50323–7 and CNPq. EJRV and MCT were supported by FAPESP fellowships (2008/53929–9 and 2008/04969–8).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molbiopara.2012.02.009.

References

[1] 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.
[2] Vanhamme L, Pays E. Control of gene expression in trypanosomes. Microbiol Rev 1995;59:223–40.
[3] 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.
[4] El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science 2005;309:404–9.
[5] 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.
[6] Laurentino EC, Ruiz JC, Fazelinia G, Myler PJ, Degrave W, Alves-Ferreira M, et al. A survey of *Leishmania braziliensis* genome by shotgun sequencing. Mol Biochem Parasitol 2004;137:81–6.
[7] Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 1988;203:439–55.
[8] Leung JY, McKenzie FE, Uglialoro AM, Flores-Villanueva PO, Sorkin BC, Yunis EJ, et al. Identification of phylogenetic footprints in primate tumor necrosis factor-alpha promoters. Proc Natl Acad Sci USA 2000;97:6614–8.
[9] Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 2000;288:136–40.
[10] Holtom B. A Paralogy Based Strategy for Identifying Regulatory Elements in Mammalian Genomes. PhD Thesis. Wolfson College, University of Oxford; 2008.
[11] Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res 2002;12:739–48.
[12] Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. J Biol 2003;2:13.
[13] Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res 2002;12:832–9.

[14] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[15] Sonnhammer EL, Durbin R. A workbench for large-scale sequence homology analysis. Comput Appl Biosci 1994;10:301–7.
[16] Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 2007;23:500–1.
[17] Smith M, Blanchette M, Papadopoulou B. Improving the prediction of mRNA extremities in the parasitic protozoan *Leishmania*. BMC Bioinformatics 2008;9:158.
[18] Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2007;23:401–7.
[19] Lahav T, Sivam D, Volpin H, Ronen M, Tsigankov P, Green A, et al. Multiple levels of gene regulation mediate differentiation of the intracellular pathogen *Leishmania*. Faseb J 2011;25:515–25.
[20] Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994;2:28–36.
[21] Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM, et al. Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. PLoS Pathog 2007;3:1291–307.
[22] Rochette A, McNicoll F, Girard J, Breton M, Leblanc E, Bergeron MG, et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in *Leishmania* spp. Mol Biochem Parasitol 2005;140:205–20.
[23] McNicoll F, Muller M, Cloutier S, Boilard N, Rochette A, Dube M, et al. Distinct 3′-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*. J Biol Chem 2005;280:35238–46.
[24] Smith M, Bringaud F, Papadopoulou B. Organization and evolution of two SIDER retroposon subfamilies and their impact on the *Leishmania* genome. BMC Genomics 2009;10:240.
[25] Bringaud F, Ghedin E, Blandin G, Bartholomeu DC, Caler E, Levin MJ, et al. Evolution of non-LTR retrotransposons in the trypanosomatid genomes: *Leishmania major* has lost the active elements. Mol Biochem Parasitol 2006;145:158–70.
[26] Ortiz G, Segovia M. Characterisation of the novel junctions of two minichromosomes of *Leishmania major*. Mol Biochem Parasitol 1996;82:137–44.
[27] Ivens AC, Lewis SM, Bagherzadeh A, Zhang L, Chan HM, Smith DF. A physical map of the *Leishmania major* Friedlin genome. Genome Res 1998;8:135–45.
[28] Coderre JA, Beverley SM, Schimke RT, Santi DV. Overproduction of a bifunctional thymidylate synthetase-dihydrofolate reductase and DNA amplification in methotrexate-resistant *Leishmania tropica*. Proc Natl Acad Sci USA 1983;80:2132–6.
[29] White TC, Fase-Fowler F, van Luenen H, Calafat J, Borst P. The H circles of *Leishmania tarentolae* are a unique amplifiable system of oligomeric DNAs associated with drug resistance. J Biol Chem 1988;263:16977–83.
[30] Grondin K, Roy G, Ouellette M. Formation of extrachromosomal circular amplicons with direct or inverted duplications in drug-resistant *Leishmania tarentolae*. Mol Cell Biol 1996;16:3587–95.
[31] Genest PA, ter Riet B, Dumas C, Papadopoulou B, van Luenen HG, Borst P. Formation of linear inverted repeat amplicons following targeting of an essential gene in *Leishmania*. Nucleic Acids Res 2005;33:1699–709.
[32] Dias FC, Ruiz JC, Lopes WC, Squina FM, Renzi A, Cruz AK, et al. Organization of H locus conserved repeats in *Leishmania* (*Viannia*) *braziliensis* correlates with lack of gene amplification and drug resistance. Parasitol Res 2007;101:667–76.
[33] Mishra KK, Holzer TR, Moore LL, LeBowitz JH. A negative regulatory element controls mRNA abundance of the *Leishmania mexicana* Paraflagellar rod gene PFR2. Eukaryot Cell 2003;2:1009–17.
[34] Haile S, Dupe A, Papadopoulou B. Deadenylation-independent stage-specific mRNA degradation in *Leishmania*. Nucleic Acids Res 2008;36:1634–44.
[35] Veitch NJ, Johnson PC, Trivedi U, Terry S, Wildridge D, MacLeod A. Digital gene expression analysis of two life cycle stages of the human-infective parasite, *Trypanosoma brucei gambiense* reveals differentially expressed clusters of co-regulated genes. BMC Genomics 2010;11:124.
[36] Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, et al. Gene synteny and evolution of genome architecture in trypanosomatids. Mol Biochem Parasitol 2004;134:183–91.
[37] Archer SK, Inchaustegui D, Queiroz R, Clayton C. The cell cycle regulated transcriptome of *Trypanosoma brucei*. PLoS One 2011;6:e18425.
[38] Mahmood R, Hines JC, Ray DS. Identification of cis and trans elements involved in the cell cycle regulation of multiple genes in *Crithidia fasciculata*. Mol Cell Biol 1999;19:6174–82.
[39] Zick A, Onn I, Bezalel R, Margalit H, Shlomai J. Assigning functions to genes: identification of S-phase expressed genes in *Leishmania major* based on post-transcriptional control elements. Nucleic Acids Res 2005;33:4235–42.
[40] Holzer TR, Mishra KK, LeBowitz JH, Forney JD. Coordinate regulation of a family of promastigote-enriched mRNAs by the 3′UTR PRE element in *Leishmania mexicana*. Mol Biochem Parasitol 2008;157:54–64.
[41] Queiroz R, Benz C, Fellenberg K, Hoheisel JD, Clayton C. Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. BMC Genomics 2009;10:495.