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 *c*onserved *i*ntercoding *s*equences (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