

The composition of upstream open reading frames (uORF) in four genes from *Trypanosoma cruzi* typical strains

Lauren Hubert Jaeger · Adelton Brandão

Received: 23 December 2010 / Accepted: 5 April 2011 / Published online: 3 May 2011
© Springer-Verlag 2011

Abstract Upstream open reading frames (uORF) are small open reading frames located in the 5' untranslated region (5' utr) of a mature mRNA. We analysed in four strains representing the *Trypanosoma cruzi* groups Tc I, Tc II, Tc IV and Tc VI the uORF present in 5' utr sequences of four genes: P-type H⁺-ATPase 1, DEAD/H RNA helicase, casein kinase 1.1 and ferredoxin–NADP⁺ reductase. A segment in the 5' utr at each of these genes encompassing one or more uORF was PCR amplified and sequenced. An analysis of these sequences reveals that the uORF in *T. cruzi* show minor variations; however, these nucleotide substitutions mirror the divergence of *T. cruzi* strains into major groups.

Introduction

Trypanosoma cruzi is the causative agent of Chagas disease in humans, and affects about ten million of people in Latin American countries (WHO 2010). It has a widespread host range, shows genetic and biological diversity and induces a variety of clinical presentations in patients (Coura and Viñas 2010). Molecular analysis allowed the clustering of *T. cruzi* strains and isolates into two major phylogenetic lineages, Tc I and Tc II, through the use of markers such as 24S α rDNA (Souto et al. 1996) and mini-exon gene (Fernandes et al. 1998). Recently, a haplotype analysis postulated the existence of three *T. cruzi* ancestral genomes,

leading to the proposition of a third lineage, named Tc III (Pena et al. 2009). Most of the molecular markers used for the *T. cruzi* strain diversity analysis are based on intergenic sequences or repetitive segments. However, *T. cruzi* genes like all of eukaryotes possess transcribed and non-translated segments that present interesting properties (Brandão 2006). One of these segments is the 5' untranslated region (5' utr), which in some mRNA exhibits short open reading frames denominated upstream open reading frames (uORF) (Calvo et al. 2009). To date, the occurrence of uORF in *T. cruzi* has been reported out for only one gene (Teixeira et al. 1999), and the sequence composition of uORF in *T. cruzi* populations remains unknown. In order to investigate uORF composition in *T. cruzi*, we analysed the nucleotide sequences of uORF from selected genes in four strains, which are representative of three ancestral genomes—Tc I, Tc II and Tc III (Pena et al. 2009). We chose the strains CL Brener (Tc VI), Y (Tc II), Dm28c (Tc I) and INPA4167 (Tc IV). These strains were previously typed as Tc I, Tc II and Tc zymodeme 3 (Tc Z3) by the mini-exon multiplex PCR (Fernandes et al. 2001) and re-allocated into the one of six DTU (Tc I–VI) after the new consensus (Zingales et al. 2009). All strains were maintained in BHI medium supplemented with foetal bovine serum at 28°C. Epimastigote cultures at the end of the log phase were centrifuged at 800 g for 10 min. The cells were resuspended in 1 ml of phosphate-buffered saline (pH 7.0) and centrifuged as above. Afterwards, the DNA was extracted with DNAzol Reagent according to the manufacturer's instruction. We searched for and selected from the GenBank, NCBI, *T. cruzi* genes which are single copy or two to three copies in the genome. Additionally, these genes should have both the 5' utr experimentally determined and one or more uORF. Four genes fulfilled these criteria and are listed in Table 1. The primer sequences were based on the *T. cruzi* CL Brener

L. H. Jaeger · A. Brandão (✉)
Laboratory of Interdisciplinary Medical Researches,
Instituto Oswaldo Cruz,
Fiocruz Av. Brasil 4365,
21040-360 Rio de Janeiro, Brazil
e-mail: abrandao@fiocruz.br

Table 1 *T. cruzi* genes, accession number, and primers designed to amplify the uORF-containing 5' utr

Gene	GenBank accession no.	5' utr length	uORF position	Amino acids	Primer forward/reverse
P-type H ⁺ -ATPase 1	AF254412	312	44...85	13	5' CACTTTCCTTGACGTCCCAT 3'
			148...219	23	5' TTGGCAAGGAGCTCTTCCGC 3'
Ferredoxin–NADP ⁺ reductase	AY206009	130	81...116	11	5' CATGCGCTTTGCATGGCTGT 3'
					5' CGACGGCGATTTGTACGCGT 3'
DEAD/H RNA helicase	AF117891	116	26...49	7	5' TGCCCCCCTACATTTCCT 3'
					5' TGCGAAGATATGCATCGCGG 3'
Casein kinase 1.1	AF274060	216	4...51	15	5' CAAGGGGATAAAATGACCGC 3'
					5' AAGATCTCGCCGAACGACCC 3'

genomic sequences, available at GeneDB (www.genedb.org). They were designed to amplify a segment that includes the 5' utr and the first 20–30 bases of the main orf in each selected gene, as shown in Fig. 1. PCR amplification was performed with non-Hot Start Taq polymerase (Invitrogen) using the following thermal profile: one cycle at 94°C, 5 min; 30 cycles at 96°C, 10 s; 55°C, 30 s; 72°C, 30 s and one cycle at 72°C, 10 min. The amplification fragments were electrophoresed on 1.5% agarose gel followed by ethidium bromide staining and UV light visualization. The fragments were ligated to plasmid vector with the TOPO-TA ligation kit (Invitrogen). Recombinant plasmids were transferred to *Escherichia coli* TOP10 competent cells (Invitrogen) grown in LB broth containing 100 µg/ml of ampicillin and purified by Wizard Plus SV Minipreps kit (Promega) according to the manufacturer's specifications. An aliquot of this solution (10–50 ng) was used for cycle sequencing with a BigDye terminator kit (Applied Biosystems) with primers forward T7 (5' TAATAC GACTCACTATAGGGCGA 3') and reverse SP6 (5' TTCTA TAGTGTCACCTAAAT 3'). The sequencing fragments were electrophoresed in an ABI3730 capillary DNA sequencer. The obtained sequences were edited and aligned by using the MEGA 4.1 software (Kumar et al. 2008). DNA sequences obtained in this work were submitted to GenBank under accession no. GU784884, GU784880, GU784886,

GU784892, GU784893, GU784894, GU784910, GU784911, GU784912, GU784913, GU784914. We obtained fragments with expected lengths from the selected genes in all strains except the casein kinase 1.1 fragment in strain INPA4167, whose amplification failed to yield a fragment. Though the number of uORF analysed in this work is insufficient to allow generalizations at genomic scale, mutations in uORF were evaluated assuming the population structure of *T. cruzi* as composed of three major lineages according to both mini-exon gene typing and the proposed three ancestral genomes (Pena et al. 2009). In the strains analysed here, we observed that most of the uORF exhibit few sequence variations. A graphical representation of the uORF and their corresponding alignment in *T. cruzi* strains are shown in Figs. 2 and 3. The detailed observations with respect to uORF composition for each of the analysed gene are described below. For P-type H⁺-ATPase 1 and ferredoxin–NADP⁺ reductase genes, their uORF do not vary if we take into account the presence of initiation codon and amino acid content. In the casein kinase 1.1. gene, uORF1 showed variation in its length. In the DEAD/H RNA helicase gene, the uORF initially observed in the Tulahuén strain did not appear in the other analysed strains.

P-type H⁺-ATPase 1

Two copies in *T. cruzi* genome and for the strains analysed here, four uORF are observed. The first uORF is present in all strains and shows a change in amino acid composition (INPA4157) as a result of a transition (A→G) at the second codon position. The second uORF is present in all strains, and one amino acid is altered (strains Dm28c and INPA4167) as a result of transition (A→G) in the second codon position. In one event exclusive to strain Dm28c, two amino acids are altered as a result of one transversion in the first codon position (C→G) and a transition in the first codon position (A→G). The third uORF overlaps to the main orf. In strain INPA4167, one amino acid is altered as a result of transversion in the second codon position

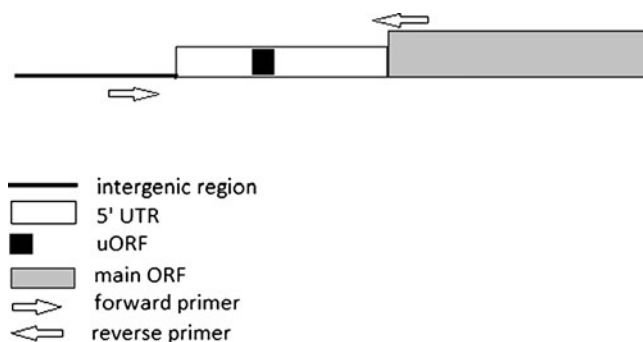


Fig. 1 uORF amplification. Schematics for amplification of uORF-containing 5' utr. Arrowheads indicate primer hybridization sites

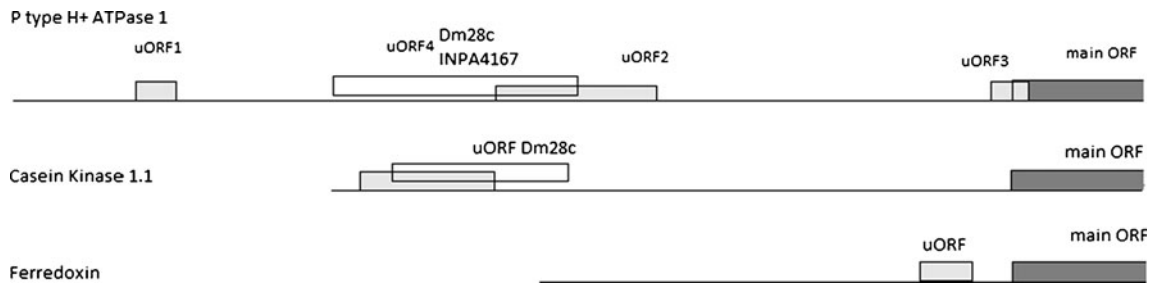


Fig. 2 uORF position. The position of upstream orf in 5' utr of selected genes from *T. cruzi* typical strains. Unfilled boxes indicate the uORF whose position is different or has been created by mutations

(T→A). The fourth uORF is exclusive to strains INPA4167 and Dm28c. It is the result of a transversion (G→T, second codon position), giving rise to a start codon. This uORF has a length of 64 bases (21 amino acids), and it overlaps the second uORF which is common to all strains. Two mutations are observed in strain INPA4167 (A→T) and (G→A). Out of the three uORF, the first one (near the trans-splicing site) is the less variable while the largest one (the second uORF) shows more nucleotide substitutions.

Ferredoxin–NADP+ reductase

A single copy is in the *T. cruzi* genome. It possesses one uORF. For all strains analysed here, only one amino acid is different due to the transversion occurring in the sequence from Dm28c (G→T). One synonymous mutation (at the

third codon position, G→T) is observed in Dm28c. The uORF position in mRNA is near the start codon in the second half of the mRNA. This uORF accounts for 26% of the 5' utr length (26/138).

DEAD/H RNA helicase

There are two mRNA sequences in GenBank for this gene: one isolated from Tulahuén strain (accession no. AF117891) and another one from the CL Brener clone (accession no. AF228510). The mRNA sequence from the Tulahuén strain contains an initiator codon (AUG) in the 5' utr resulting in an upstream open reading frame. However, upon amplification in three other strains (Y, INPA4167, Dm28c) and comparison to genome reference sequence (CL Brener clone), we noticed this initiator codon is not

P-type ATPase	
<u>uorf1</u>	
CLBrener	ATG CAA AAA GGG GGA GTT GGC CTG TTT TTC AAG TTT ATT TAA
Y
Dm28c
4167G.
<u>uorf2</u>	
CLBrener	ATG GCG ACA AGA CCT AAA ATA AAA CGG GCA AGG CCC TCT CAA CGA AGC CAC GGA GGC GGT CTC CAA CAG TGA
Y
Dm28cG. G.. .G.
4167AG.
<u>uorf3</u>	
CLBrener	ATG AAG TGC CAA TGC TAC CGC CGT CCA AGG GTC TGA
Y
Dm28cA. ...
4167G.
<u>uorf4</u>	
CL_Brener	AGG AAA AGC TGG AGC GCG TCC GTG ACC GAA TTT TGT CAA ACG AGC AAA AAT GGC GAC AAG ACC TAA
YA. ...
Dm28c	.T.
4167	.T.C.T.
Casein Kinase1	
CL_Brener	ATG AAA TTT GAC TAT AAC AAC AAC AAC AAA TAC CAA CAA AAG ACG GGG -TA A-- --- --- ---
Y
Dm28c C..G. --AT G.. .AA TAA ACA TTT AA-
Ferredoxin	
CL_Brener	ATG AGC GTG TAC AAG AGT TGT AGC TTT TTT TAT TAG
Y
Dm28cTT ...
inpa4167
Helicase	
CL_Brener	GTT GTG CCT CTC TCC GTG TAT GTG AA
Y
Dm28c	A.. ...
inpa_4167	A.. ...
AF117891_Tulahuén	A-. ...

Fig. 3 uORF sequence. uORF sequences of the selected genes. Dots represent identical nucleotides and underline the deletions. Note that ATPase 1 uORF4 exists only in strains Dm28c and INPA4167; the

initiator codon shows a mutation in the first position in strains Y and CL Brener. ATPase uORF4 is overlapping uORF2

present in these strains. It is not clear whether this is a particular mutation for the Tulahuen strain or a sequencing error. With respect to CL Brener clone, upon comparing two publicly available sequences—an mRNA sequence in GenBank that has been deposited prior to the finishing of *T. cruzi* genome project (accession no. AF228510) and the genome sequence itself—we noticed these sequences differ by six nucleotide positions with respect to the sequence upstream in the main orf. The sequence derived from the Tulahuen strain also displays a second ATG in 5' utr. However, this initiator codon does not result in a typical uORF because it is immediately followed by a terminator codon (TGA). Thus, unless the first initiation codon observed in the Tulahuen strain is not the result of a sequencing error, we conclude there is no uORF in the DEAD/H RNA helicase gene. Notwithstanding, it is worth noting the presence of the palindromic sequence 5' TGTGCCTCTCTCCGTGT 3' embedded into the once thought uORF. A basic local alignment search tool (BLAST) using this palindrome as query sequence matches genes in some organisms showing this palindrome in the vicinity of the 5' utr. This is clearly observed in one putative protein kinase from *Leishmania infantum* chromosome 25 (accession AM502243, region 616,153 to 616,137).

Casein kinase 1.1

Two copies are in the *T. cruzi* genome (two isoforms) (Spadafora et al. 2002) and one in uORF. Only the casein kinase 1.1 copy was analysed here. No amplification was obtained for strain INPA4167. Two insertions in the sequence from Dm28c have caused a frame shift in the terminator codon, giving rise to longer uORF (20 amino acids). This means five additional amino acids in comparison to the CL Brener clone and Y strain, which have uORF with 15 amino acids.

It is worth mentioning that the genes we analysed here are relatively conserved among the trypanosomatids. However, a BLAST search did not detect any similarity to sequences from other trypanosomatids. In conclusion, in typical *T. cruzi* strains representing the major lineages, most of the uORF exhibit few sequence variation and the nucleotide composition of uORF appears to be species

specific: no sequence identity for the genes analysed is observed among trypanosomatid species.

Acknowledgements This work received financial support from Fundação Oswaldo Cruz.

References

- Brandão A (2006) The untranslated regions from genes from *Trypanosoma cruzi*: perspectives for functional characterization of strains and isolates. Mem Inst Oswaldo Cruz 101:775–777
- Calvo SE, Pagliarini DJ, Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci USA 106:7507–7512
- Coura JR, Viñas PA (2010) Chagas disease: a new worldwide challenge. Nature 465:S6–S7
- Fernandes O, Sturm NR, Derré R, Campbell DA (1998) The mini-exon gene: a genetic marker for zymodeme III of *Trypanosoma cruzi*. Mol Biochem Parasitol 95:129–133
- Fernandes O, Santos SS, Cupolillo E, Mendonça B, Derre R, Junqueira AC, Santos LC, Sturm NR, Naiff RD, Barret TV, Campbell DA, Coura JR (2001) A mini-exon multiplex polymerase chain reaction to distinguish the major groups of *Trypanosoma cruzi* and *T. rangeli* in the Brazilian Amazon. Trans R Soc Trop Med Hyg 95:97–99
- Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299–306
- Pena SD, Machado CR, Macedo AM (2009) *Trypanosoma cruzi*: ancestral genomes and population structure. Mem Inst Oswaldo Cruz 104:108–114
- Souto RP, Fernandes O, Macedo AM, Campbell DA, Zingales B (1996) DNA markers define two major phylogenetic lineages of *Trypanosoma cruzi*. Mol Biochem Parasitol 83:141–152
- Spadafora C, Repetto Y, Torres C, Pino L, Robello C, Morello A, Gamarro F, Castanys S (2002) Two casein kinase 1 isoforms are differentially expressed in *Trypanosoma cruzi*. Mol Biochem Parasitol 124:23–36
- Teixeira SM, Kirchoff LV, Donelson JE (1999) *Trypanosoma cruzi*: suppression of tuzin gene expression by its 5'-UTR and spliced leader addition site. Exp Parasitol 93:143–151
- WHO (2010) Chagas disease (American trypanosomiasis) fact sheet no. 340 June 2010. <http://www.who.int/mediacentre/factsheets/fs340/en/index.html/>. Accessed 21 Dec 2010
- Zingales B, Andrade SG, Briones MRS, Campbell DA, Chiari E, Fernandes O, Guhl F, Lages-Silva E, Macedo AM, Machado CR, Miles MA, Romanha AJ, Tibayrenc M, Schijman AG (2009) A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Mem Inst Oswaldo Cruz 104:1051–1054