Nucleic Acids Research, 2005, Vol. 33, No. 13 ***4235–4242***
doi:10.1093/nar/gki742

Assigning functions to genes: identification of S-phase expressed genes in *Leishmania major* based on post-transcriptional control elements

Aviad Zick[1], Itay Onn[1,2], Rachel Bezalel[1], Hanah Margalit[2] and Joseph Shlomai[1,*]

[1]Department of Parasitology, The Kuvin Center for the Study of Infectious and Tropical Diseases and [2]Department of Molecular Genetics and Biotechnology, The Hebrew University-Hadassah Medical School, Jerusalem 91120, Israel

Received May 10, 2005; Revised and Accepted July 11, 2005

ABSTRACT

Assigning functions to genes is one of the major challenges of the post-genomic era. Usually, functions are assigned based on similarity of the coding sequences to sequences of known genes, or by identification of transcriptional *cis*-regulatory elements that are known to be associated with specific pathways or conditions. In trypanosomatids, where regulation of gene expression takes place mainly at the post-transcriptional level, new approaches for function assignment are needed. Here we demonstrate the identification of novel S-phase expressed genes in *Leishmania major*, based on a post-transcriptional control element that was recognized in *Crithidia fasciculata* as involved in the cell cycle-dependent expression of several nuclear and mitochondrial S-phase expressed genes. Hypothesizing that a similar regulatory mechanism is manifested in *L.major*, we have applied a computational search for similar control elements in the genome of *L.major*. Our computational scan yielded 132 genes, of which 33% are homologues of known DNA metabolism genes and 63% lack any annotation. Experimental testing of seven of these genes revealed that their mRNAs cycle throughout the cell cycle, reaching a maximum level during S-phase or just prior to it. It is suggested that screening for post-transcriptional control elements associated with a specific function provides an efficient method for assigning functions to trypanosomatid genes.

INTRODUCTION

Trypanosomatids are eukaryotic protozoa that diverged early from the main lineage of eukaryotes (1). Their genes are transcribed forming polycistronic messages that mature into individual mRNAs by a process of *trans*-splicing, which involves the addition of a spliced leader, a capped RNA of ~40 nt, to the 5′-end, and polyadenylation at the 3′-end [reviewed in (2)]. The lack of transcriptional control in trypanosomatids suggests that other control mechanisms may have evolved in order to regulate mRNA levels. Indeed, it was shown that S-phase expressed genes in the trypanosomatid *Crithidia fasciculata* are controlled by a special, cell cycle-dependent, post-transcriptional regulation (3–6). It was demonstrated that mRNA transcripts of four genes accumulate at the beginning of S-phase, and degrade rapidly after DNA replication is completed. All the genes showing this post-transcriptional cell cycle-dependent regulation encode for proteins that participate in DNA metabolism. These include the genes encoding the large subunit of replication protein A (*RPA1*), Dihydrofolate reductase-thymidylate synthetase (*DHFR-TS*), the kinetoplast type II DNA topoisomerase (*TOP2*) and the histone H1-like kinetoplast associated protein 3 (*KAP3*) (3–6). It was shown that cycling is controlled prior to mRNA maturation (5), and that an octamer sequence [(C/A)AUAGAA(G/A)], located at either the 5′ or the 3′ untranslated regions (UTRs) of the mRNAs, was involved in the regulation of the cell cycle-dependent accumulation of these mRNAs. Six copies of the octamer sequence, cloned into the 5′ UTR of the gene encoding calcium binding protein (CaBP), expressed constitutively in *C.fasciculata* cells, conferred strong cycling to the CaBP transcript, while a single base change in this sequence was sufficient to abolish the cycling of a reporter gene *in vivo* (3). Two protein complexes, designated CSBP I and CSBP II, were

*To whom correspondence should be addressed. Tel: 972 2 6758089; Fax: 972 2 6757425; E-mail: Shlomai@cc.huji.ac.il
Present address:
Itay Onn, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

shown to bind specifically to this sequence. A single base change of the octamer consensus sequence was sufficient to prevent binding of the protein complexes to the mRNA (6). In addition, it has been recently shown that in *Leishmania infantum* mRNA transcripts of the kinetoplast topoisomerase II accumulate during the S-phase of the cell cycle. Examination of the *L.infantum* topo II gene's flanking sequences revealed the presence of a sequence element which is similar to the *C.fasciculata* octamer associated with mRNA cycling (7).

As the genome of the trypanosomatid *Leishmania major* is extensively sequenced, a current major challenge is to assign functions to the novel genes that are identified. Usually, such functions are assigned by sequence similarity of the coding sequences to known genes, or by identification of transcriptional *cis*-regulatory elements that are known to be involved in regulation of a specific pathway. In trypanosomatids it is conceivable that functions can be assigned by identification of specific post-transcriptional control elements. Here we speculate that similarly to *C.fasciculata*, the *L.major* S-phase expressed genes may be regulated by a cycling control element present in their UTRs, which fulfills a similar biological function. Thus, we attempt to identify novel *L.major* S-phase expressed genes, through the identification of a similar cycling control element in their UTRs. Indeed, searching *L.major* genome for cycling control elements that exhibit the same characteristics of the *C.fasciculata* control elements yielded 132 genes, of which 33% encode for proteins with known functions during S-phase. 63% of the identified genes lacked any annotation and were predicted as putative S-phase expressed genes. Experimental testing of seven of the genes identified in the screen showed that they reach a maximum level in S-phase or just prior to it, suggesting that identification of the cycling control element is valuable for determining new *L.major* genes that are expressed during the S-phase of the cell cycle.

MATERIALS AND METHODS

Sequence extraction

Sequences of the *C.fasciculata* genes encoding for the large subunit of replication protein A (*RPA1*), Dihydrofolate reductase-thymidylate synthetase (*DHFR-TS*), the kinetoplast DNA topoisomerase II (*TOP2*) and the histone H1-like kinetoplast associated protein 3 (*KAP3*) were obtained from GenBank (www.ncbi.nlm.nih.gov, accession no. Z23163, M22852, X59623 and AY143553, respectively). Search for homologous protein sequences and motifs, was conducted using Blastp on NCBI non-redundant database (www.ncbi.nlm.nih.gov/BLAST/). For *L.major* genomic data and sequence search we used the geneDB database at the Sanger institute UK (www.genedb.org).

Identifying common motifs in unaligned sequences

To identify the common motifs in unaligned sequences, we used the MIsmatch TRee Algorithm (MITRA) (8) for finding regulatory elements in DNA sequences (http://www1.cs.columbia.edu/compbio/mitra/).

Growth of *L.major* cell cultures

L.major strain MHOM/IL/1986/BLUM (LRC-L509) promastigotes were grown at 28°C in M199 with Hank's salt and L-glutamine w/o sodium bicarbonate, 25 mM HEPES (Sigma), 8 mM sodium bicarbonate adjusted to pH 6.9 with NaOH supplemented with 0.1 mM adenosine (Sigma), 23 μM folic acid (Sigma), 35 μM xanthine (Sigma), 8 μM hemin (Sigma), 10% heat-inactivated fetal calf serum (Beit-HaEmek, Israel), BME vitamin solution (Beit-HaEmek, Israel), 100 U/ml penicillin, 0.1 mg/ml streptomycin (Beit-HaEmek, Israel) and 2 mM L-glutamine (Beit-HaEmek, Israel).

Synchronization of *L.major* cell cultures

L.major promastigotes cultures of 1.5×10^7 cells/ml, were incubated for 16 h in growth medium containing 400 μg/ml hydroxyurea (Sigma). Cells were then harvested by centrifugation at 1600 *g* for 16 min at 4°C, washed three times with phosphate-buffered saline (PBS) and resuspended in twice the volume of fresh medium with no hydroxyurea. The cells remained synchronous for about 14 h. Synchronization was validated by fluorescence microscopy of DAPI stained cells, which was described previously (9). Samples of cells, withdrawn at intervals of 1 h throughout the cell cycle were stained with propidium iodide and submitted to fluorescence activated cell sorter (FACS) analysis (see below). Samples of 20 μg of RNA were prepared for Northern blot hybridization analysis (see below).

Flow cytometery

All procedures were conducted at 4°C, unless otherwise indicated. Samples of $0.7–3 \times 10^7$ *L.major* cells were centrifuged at 1600 *g* for 8 min. Each pellet was resuspended in 200 μl of PBS, fixed by adding 10 ml of ice cold methanol and stored at −20°C. The fixed cells were centrifuged at 1600 *g* for 8 min and resuspended in 200 μl PBS, containing 50 μg/ml RNase A (Sigma) and incubated for 30 min at 4°C. The cells were filtered through a silk mesh and stained with 25 μg/ml propidium iodide (Sigma). The cell suspension was then analyzed by a flow cytometer (FACS; Becton Dickinson), using Cellquest analysis software (Version 2.0.2, Becton Dickinson).

PCR

L.major genomic DNA was prepared using ZR Genomic DNA Kit™ (Zymo Research) or DNeasy Tissue kit (Qiagen), according to manufacturers' recommended protocol. DNA probes, 400–500 bp long, for the *L.major* genes *LmjF06.0860*, *LmjF21.1210*, *LmjF15.1450*, *LmjF12.1220*, *LmjF31.0070*, *LmjF12.0510* and *LmjF33.0650*, were designed based on the coding sequence in the *L.major* geneDB database (The Sanger institute UK, www.genedb.org). The gene sequences were analyzed using the Primer 3 software on the WWW (10). The following oligonucleotides were used as PCR primers:

LmjF06.0860: Upper primer AGCTCATCGTGGAGACCATC &
Lower primer ACCTCCATGTCCGTCAACTC product size 414 bp;
LmjF21.1210: Upper primer AGCACAACGTTGCTTCACAC &

Lower primer CTTCTCCACCTCCTTGATCG product size 457 bp;
LmjF15.1450: Upper primer GAGCGCAACATAATCCTTGG &
Lower primer GAGATAGGCTCGTGCTCGTC product size 460 bp;
LmjF12.1220: Upper primer AATTCGCACTATTCCGTTGG &
Lower primer ACCTGGGTCTTTGTGATTGG product size 419 bp;
LmjF31.0070: Upper primer TGGAGATACCGGACCTCAAC &
Lower primer TAGCGGTTGCACTCGTACAC product size 459 bp;
LmjF12.0510: Upper primer GTGAGCTCCTCAAAACGGAG &
Lower primer ACCACCAGCCACTACCTTTG product size 445 bp;
LmjF33.0650: Upper primer ACTCTATGGGGTTGCCACAG &
Lower primer CTGACTCCTCCAGCAAGGAC product size 467 bp.
α-TUBULIN: Upper primer ATGCGTGAGGCTATCTGCATCCACAT &
Lower primer TAGTGGCCACGAGCGTAGTTGTTCG product size 323 bp.

The primers were used in the PCR, as follows: 50 µl reaction mixture included 0.625 U of FailSafe™ PCR Enzyme Mix (EpiCenter), 10 ng of *L.major* genomic DNA, 0.5 µM of Upper and Lower primers, and FailSafe™ PreMix F (EpiCenter). The reaction mixture was incubated for 2 min at 95°C, for 34 cycles of: 30 s at 94°C, 30 s at 58°C, 1 min at 72°C followed by 10 min at 72°C. DNA probes were labeled using [α-^{32}P]-dCTP, using random primers (Amersham Bioscience) and DNA polymerase I Klenow fragment (Fermentas), following the manufacturer's recommended protocol.

Northern blot hybridization analysis

Total cell RNA was isolated from samples of synchronous *L.major* cells at 1 h intervals after the removal of hydroxyurea, using TRI REAGENT™ (Molecular Research Center Inc.), following the manufacturer's procedure at a ratio of 2.5–5 × 10^8 *L.major* cells/ml TRI REAGENT™. Samples of 20 µg RNA were loaded onto the 18% formaldehyde, 1% agarose gel, at 1 V/cm, for 3 h, at room temperature, then transferred onto the nylon membranes (Sartorius) and hybridized with the radioactively-labeled PCR probe, as described previously (11). Quantification was conducted by exposure to an imaging plate, and analysis using PhosphorImager [Bio Imaging Analyzer (BAS1000; Fuji)]. Each membrane was stripped by incubation for 3 h, at 80°C in 50% formamide, 5% SDS, 50 mM Tris–HCl (pH 7.5) and re-probed.

RESULTS AND DISCUSSION

Characterizing the cycling control element in known S-phase expressed genes of *C.fasciculata*

Examination of the cycling control elements based on the available experimental data in *C.fasciculata* indicated that in most cases they appear more than once in the UTR. We have used the known occurrences of the cycling control element in *C.fasciculata* to characterize it by the number of repeats and their location relative to the gene's ORFs. This analysis revealed several characteristics: (i) The element's core contained the consensus sequence ATAGAA; (ii) When two copies of the motif were present within the same UTR they were separated by a maximum of 500 nt; (iii) The motif appeared on the same strand as the coding sequence (CDS); (iv) At least one of the identified motifs was positioned at a distance not greater than 500 nt from the beginning or the end of the CDS; (v) The motif was located either at the 5′ or the 3′ UTR of the gene, but not within the CDS. These characteristics can be used as a set of rules which can be applied in the genome-wide search for S-phase expressed genes that are supposed to be regulated by the cycling control element.

Conservation of the cycling control element in *L.major*

To study the conservation of the octamer sequence in *L.major* we compared the coding sequences and UTRs of *C.fasciculata* genes that were shown to contain the octamer sequence (and demonstrated the consequent mRNA cycling) with their *L.major* orthologues. These orthologues (as determined by a BLAST search in *L.major* genome) are: *RPA1*, encoding the 51 kDa *RPA1* subunit (geneDB accession no. LmjF28.1820), *TOP2*, encoding the DNA topoisomerase II (geneDB accession no. LmjF15.1290), *KAP3*, encoding the kinetoplast associated protein 3 (geneDB accession number LmjF32.3780) and *DHFR-TS*, encoding the dihydrofolate reductase-thymidylate synthetase (geneDB accession number LmjF06.0860). Comparison of *L.major* and *C.fasciculata* sequences revealed high similarity within the CDS (Figure 1): identity of 67% in *KAP3*, 79.2% in *DHFR-TS*, 85% in the *TOP2* and 91.6% in *RPA1*. However, the similarity decreases significantly in the flanking regions. A search for the octameric

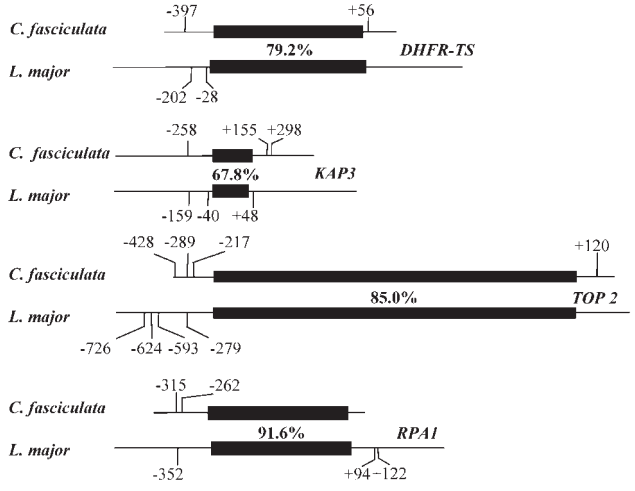

Figure 1. Location of the cycling control sequences in *C.fasciculata* genes and in their orthologues in *L.major*. The sequences of *C.fasciculata* mRNAs, previously shown to specifically accumulate in S-phase, are shown. The heavy lines represent the CDS. The numbers indicate the location of the *cis*-acting cycling elements in respect to the CDS. *L.major* orthologous genes, including an additional 1 kb at each end are shown. The location of the CATAGA sequence and its distance from the CDS are indicated. The identity percentage of the coding sequences is shown between the CDS.

cycling control sequence [(C/A)ATAGAA(G/A)] revealed a single copy of the sequence in the *lmKAP3* and the *lmTOP2* flanking sequences. A search for the core sequence, ATAGAA, resulted in the addition of one more site in the *lmKAP3* flanking sequence. The locations of these elements in regard to the ORFs were consistent with the rules determined for the *C.fasciculata* sequences. However, they did not appear in a few copies in the UTRs. A possible explanation for these observations was that the signal in *L.major* is slightly different from the signal in *C.fasciculata*, and by searching for the exact *C.fasciculata* signal we might have missed it. To this end we turned to analyze the flanking regions of *L.major* genes involved in DNA metabolism, attempting to identify in them a common signal that might be the *L.major* cycling control element.

UTRs of *L.major* homologues of DNA metabolism genes contain a CATAGA motif

Genes involved in DNA metabolism in *L.major* were identified by their sequence similarity to sequences of genes encoding known proteins that participate in DNA metabolism in *Homo sapiens*, *Xenopus laevis*, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*, as revealed by BLAST (Supplementary Table 1). The sequences flanking the identified ORFs, a 1000 nt stretch residing 5′ to the ATG or 3′ to the stop codon, were analyzed. In the case of genes that were spaced at a distance shorter than 1000 nt, the intergenic region between the stop codon and the ATG of the next ORF was analyzed. In this set of sequences, containing 164 sequences, the shortest sequence was 292 nt long, the longest sequence was 1000 nt long and the average sequence length was 927.8 nt. The MITRA algorithm for detection of common motifs in unaligned sequences identified a CATAGA sequence as a significant motif in the set of sequences analyzed. This sequence, which comprises the 5′ sequence of the *C.fasciculata* octamer, is proposed by us as a putative cycling control sequence (PCS) in *L.major*. Interestingly, analysis of the genomic context of the PCSs in these genes revealed that they are rich in adenine, while in the whole intergenic regions adenine was the least frequent nucleotide (Figure 2).

As shown in Figure 1, re-analyzing the *RPA1*, *TOP2*, *KAP3* and *DHFR-TS* UTRs by searching for the PCS resulted in a total of 12 hits in all four genes. Ten of the elements were located within less than 500 nt from the coding sequence. Two PCSs were identified in *lmDHFR-TS* upstream region. Four PCSs were identified in the *lmTOP2* upstream region. Three PCSs were found in *lmRPA1*, one upstream and two downstream to the CDS. Three PCSs were found in *lmKAP3*, two upstream and one downstream to the CDS. These locations are consistent with the rules determined above, but differ from the locations of the previously characterized *C.fasciculata* mRNA cycling signal in the gene homologues' UTRs (Figure 1). Given the nucleotide background in the intergenic regions of the 164 genes in our training set, which is 0.19, 0.3, 0.24 and 0.27 for A, C, G and T, respectively, the probability to find the sequence CATAGA is 1.3×10^{-4}, and hence, the probability to find it at least once per 1000 nt in all four sequences is less than 2.4×10^{-4}. The finding of three pairs and a quartet of the CATAGA sequence in the examined sequences is, therefore, highly statistically significant, and suggests that this sequence may have a functional role in *L.major* S-phase expressed genes.

Screening the *L.major* genome for genes containing a CATAGA sequence

To get an idea of the abundance of the PCS in the genome, we first searched the CATAGA hexamer in the whole genome sequence, and found that it occurred 6,324 times in the *L.major* genome. Pairs of PCSs that are located less than 500 nt apart on the same strand, as defined above, were found 554 times. Since the experimental data in *C.fasciculata* suggested that the cycling control elements reside in the UTRs of the genes they regulate, we aimed to identify genes that exhibit the PCS in their UTRs. Thus, we extracted those genes where the PCS pairs resided in a distance less than 500 nt from a CDS, either upstream or downstream. This analysis retrieved 132 CDSs. Examination of the distance between neighboring PCSs revealed that 44% of these sequences are positioned less than 100 nt apart (Figure 3). Extending the distance to

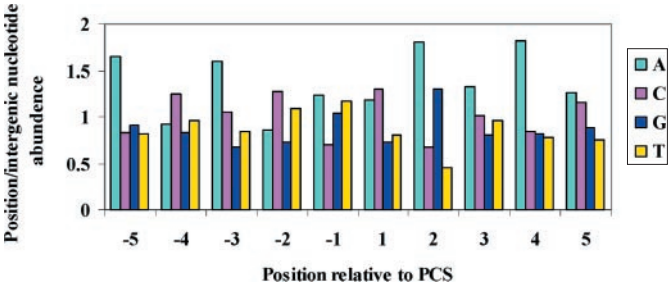

Figure 2. The genomic context of the PCS. The nucleotide frequency in the ±5 positions flanking the PCS was compared to the nucleotide frequency in the intergenic regions of DNA metabolism gene homologues in *L.major*. The ratio of the nucleotide frequency in the positions flanking the PCS to the nucleotide frequency in the intergenic regions is shown.

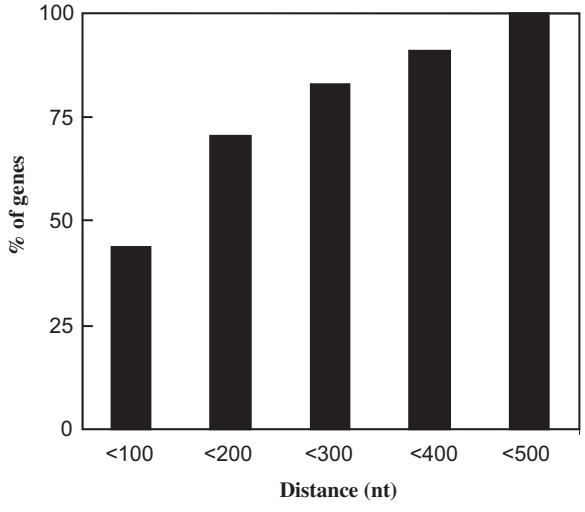

Figure 3. Distance between neighboring CATAGA sequences. The number of genes containing a CATAGA sequence pair within 500 nt of their UTR, either upstream or downstream the CDS, clustered by distance between neighboring CATAGA sequences.

Table 1. Classification of annotated genes which contain at least a PCS pair within 500 nt of their CDS

Assigned role	Number of genes	% of genes
Replication machinery	15	30.6
Cell proliferation	4	8.2
Chromosome structural maintenance	17	34.7
Nucleotide metabolism	2	4.1
Replication associated domain	5	10.2
other	6	12.2
Total number of annotated genes	49	100

200 nt covers 71% of the data and about 10% is added to each 100 nt addition, up to 500 nt (Figure 3).

PCS is abundant within genes encoding for DNA metabolism proteins

In order to determine the possible cellular roles of the 132 genes selected by the above screen, we submitted each of these genes as a query for a Blastx search against the NCBI non-redundant database and for motif identification (Table 1). A biological function could be assigned to 49 genes, accounting for 37% of the dataset, based on either significant similarity to other genes of known function, or by the identification of motifs associated with specific biological functions. The other 83 genes were either orphan sequences or conserved hypothetical genes. Out of the 49 genes for which a function could be assigned, 43 genes are homologous to DNA metabolism genes or have motifs that suggest such a role. This result is highly statistically significant ($P < 10^{-41}$ by a hypergeometric test). Within the group, 30.6% of the genes encode for replication proteins, such as DNA primase, DNA polymerases, PCNA, topoisomerase II and *RPA1* (both 51 and 28 kDa subunits). Some of the identified genes have a known function in nuclear DNA replication, while some of them have previously been associated with kDNA replication in *C.fasciculata* (e.g. the gene encoding the structure-specific endonuclease 1, SSE1). The next class, about 34.7% of the group, contains genes encoding proteins with known function in chromosome maintenance. Genes in this group play a major role in chromosome organization and condensation. Another gene, possibly a thymidine kinase involved in nucleotide metabolism, has been identified in addition to *DHFR-TS*, which was part of our training set. 8.2% of the genes are related to cell proliferation, while another group, encompassing 10.2% of the annotated genes, contain motifs related to DNA replication, such as probable DEAE helicases and nuclease sequence signature. The last group, which contains 12.2% of the annotated genes, involves genes that do not have a known role in DNA metabolism. Supplementary Table 2 shows the complete list of annotated genes, identified by this screen. While the rest of the sequences found using the computational search (63%) have no known function, they are especially intriguing, since they may represent genes, yet unknown, that are involved in *L.major* nuclear and kinetoplast DNA replication. Supplementary Table 1 shows the complete list of these genes.

mRNA transcripts of candidate genes peak at S-phase

To test whether the genes selected by our computational screen are expressed during S-phase, we monitored the mRNA

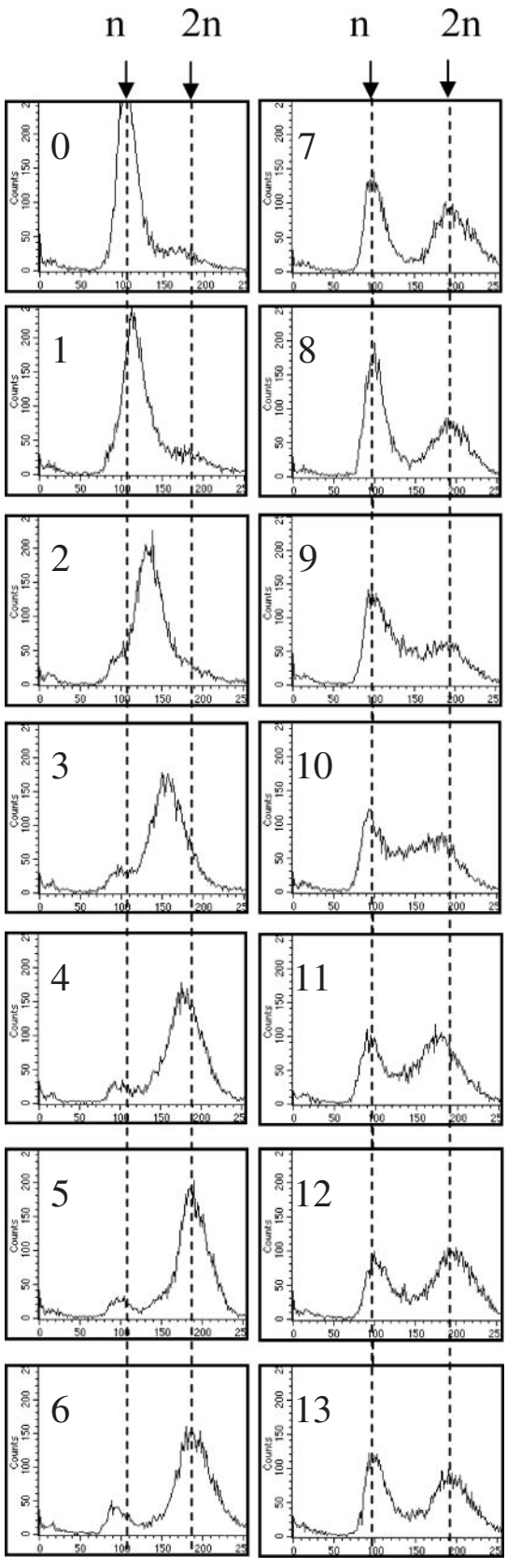

Figure 4. FACS analysis of synchronized *L.major* cells. Cell samples were withdrawn at 1 h intervals after release of hydroxyurea arrest for 13 h, stained with propidium idodide and submitted to FACS analysis, as described in the Materials and Methods.

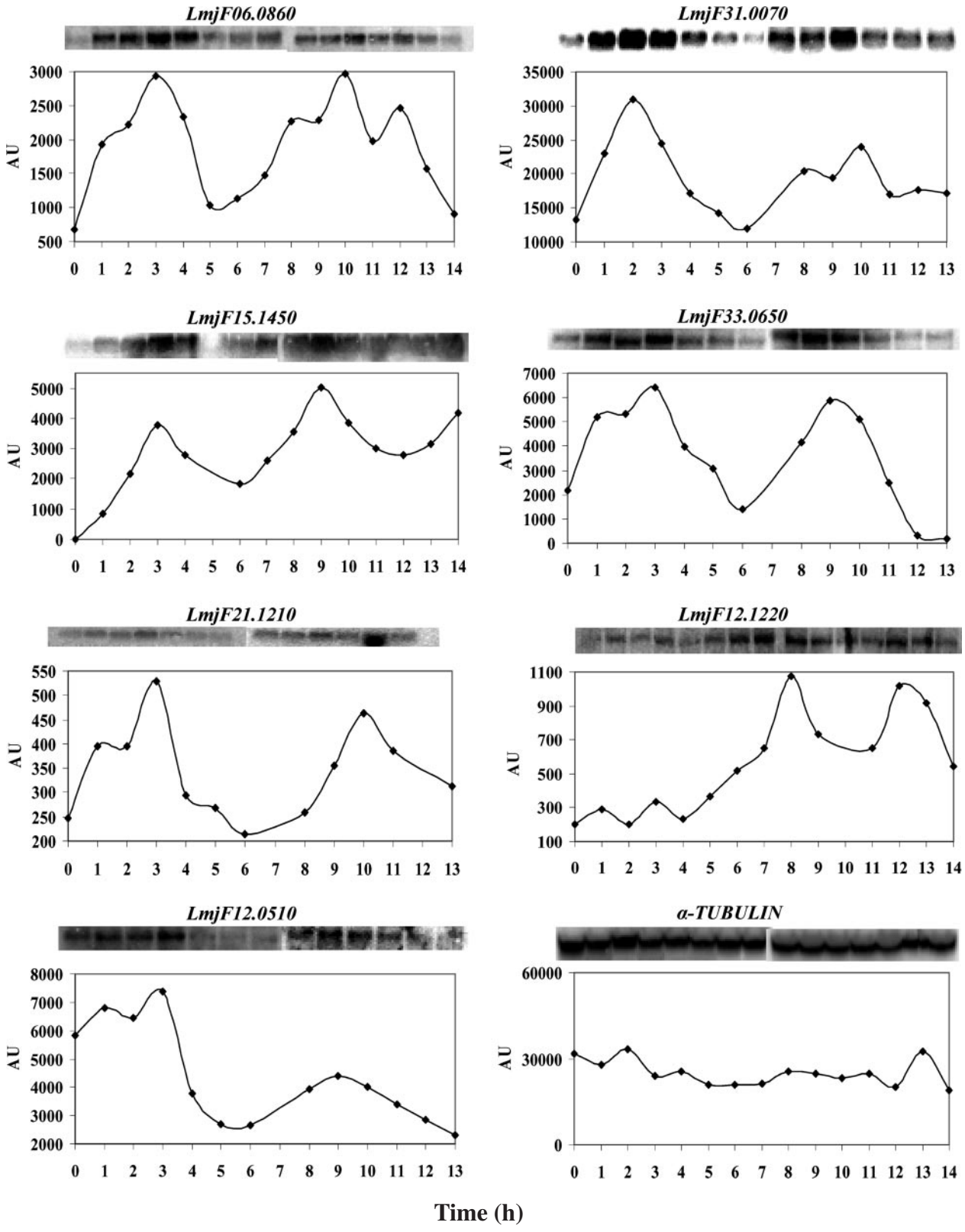

Figure 5. Cycling of mRNA transcripts of putative S-phase genes throughout the *L.major* cell cycle. Total cell RNA was extracted from samples of synchronized *L.major* cells, withdrawn at 1 h intervals after the release of hydroxyurea arrest. RNA samples (20 μg) were gel electrophoresed, blotted onto nylon membranes and submitted to northern blot hybridization analysis using radioactively-labeled probes, as described in the Materials and Methods. The level of mRNA was quantified using phosphorImaging. *LmjF06.0860*—homologue of *CfDHFR-TS*; *LmjF15.1450*—homologue of *PCNA*; *LmjF21.1210*—homologue of thymidine kinase gene; *LmjF12.0510*—homologue of *HsCep164* gene, encoding the Cep164 protein, which was putatively localized to the centromer (15); *LmjF31.0070*—homologue of African swine fever virus helicase gene, *NP_042734*; *LmjF33.0650*—a conserved hypothetical gene, with a predicted mitochondrial signal peptide; LmjF12.1220—homologue of *X.laevis AND-1* gene, encoding a nuclear DNA binding protein with HMG and WD domains (12). α-tubulin gene was used as a control.

levels of several candidates throughout the cell cycle, in hydroxyurea-synchronized *L.major* cultures. We followed the progress through two rounds of the cell cycle (~14 h) by fluorescence microscopy of DAPI stained cells, monitoring the changes in cells morphology, and by FACS analyses of propidium idodide-stained cells, following the changes in their DNA content (Figure 4).

Abundance of specific mRNA transcripts of seven of the genes found in our screen was followed during the progress in the *L.major* cell cycle, using northern blot hybridization analysis of cell samples, withdrawn at 1 h intervals after the removal of hydroxyurea. The gene encoding α-tubulin, which does not show the PCSs in either of its UTRs, was used as a control. The seven genes that were analyzed vary in their putative relevance to S-phase: *LmjF06.0860* is the gene homologue of *CfDHFR-TS*, which contains two cycling control elements (Figure 1) and was shown to cycle in an S-phase dependent manner (12,13). *LmjF21.1210* and *LmjF15.1450* are homologues of the thymidine kinase and PCNA encoding genes, respectively, both of which encode for proteins involved in DNA metabolism in other eukaryotes, and are expected to be expressed during S-phase. The next three genes have homologues with unassigned function in other organisms, but there are some hints that may suggest that they are involved in DNA metabolism: The *LmjF12.1220* gene is similar (*E*-value 2×10^{-19}) to the *X.laevis AND-1* gene, X98884, encoding a nuclear DNA binding protein with HMG and WD domains (14). The *LmjF12.0510* gene is similar (*E*-value 5×10^{-18}) to the *H.sapiens Cep164* gene, NM_014956, encoding the Cep164 protein, which was putatively localized to the centromer (15). The *LmjF31.0070* gene contains an SSL2 domain (*E*-value 4×10^{-24}) and is highly similar (*E*-value 10^{-150}) to the African swine fever virus helicase gene, *NP_042734*. The last gene, *LmjF33.0650*, is a conserved hypothetical gene, with a predicted mitochondrial signal peptide, which contains seven PCSs within its flanking sequence.

The steady state levels of mRNA transcripts of the *LmjF06.0860*, *LmjF21.1210*, *LmjF15.1450*, *LmjF12.0510*, *LmjF31.0070* and *LmjF33.0650* genes were found to vary during cell cycle progression. Two peaks of mRNA could be detected during two consecutive cell cycles. The first peak occurred at 2 or 3 h after the release from hydroxyurea arrest, during the first cell cycle, and the second peak occurred at 9–10 h during the second cell cycle. Both peaks occurred within the respective S-phases of the two consecutive cell cycles monitored (Figure 4). The levels of RNA transcripts were 3- to 7-fold higher during S-phase for most of these genes (Figure 5). The *LmjF12.0510* had an additional prominent transcript of a smaller than expected size. This transcript reached a maximum level that was 2-fold higher at the second hour after the removal of hydroxyurea, relative to its abundance during the rest of the cell cycle. The *LmjF12.1220* transcript was expressed periodically, but reached maximum levels at 8 h and subsequently at 12 h post hydroxyurea arrest release, just prior to S-phase (Figure 5). The α-tubulin transcript, used as a control, was found to be expressed at a constant level during cell cycle progression (Figure 5).

Our results show that genes in *L.major* that contain the PCSs in their flanking regions cycle throughout the cell cycle, reaching their peak levels during S-phase or at the G_1/S boundary. Three of the genes, *LmjF06.0860*, *LmjF21.1210* and *LmjF15.1450*, were selected for the experimental testing because their homologues are known to play a role in DNA metabolism, and our results support the possible involvement of these genes in DNA metabolism in *L.major*. The four other genes have no homologues with known function. Our results suggest that the cellular function of these genes may be related to the S-phase, and furthermore, it may allude to the function of their homologues in other organisms. Thus, it is conceivable that the gene products of *Cep164*, the human homologue of *LmjF12.0510* and helicase *NP_042734*, the African swine fever virus homologue of *LmjF31.0070*, play a role in DNA metabolism during the S-phase. Likewise, it is possible that *AND-1*, *LmjF12.1220* homologue in *X.laevis*, plays a role at the G_1/S boundary. Finally, *LmjF33.0650* which is a hypothetical gene conserved only within the group of Trypanosomatidae, is also implied to have a role in DNA metabolism, and thus, could be a potential drug target. There are other genes among the 132 candidates with the PCSs in their UTRs that are specific to trypanosomatids, and therefore may also be useful in future development of anti-trypanosomatid drugs (16).

Obviously, not all S-phase expressed genes in *L.major* show the cycling control element as defined by our rules. As can be seen in Supplementary Table 1, some DNA metabolism gene homologues in *L.major* do not show the PCS at all, and some show it only once, and therefore were not included in the group of 132 candidates. It is possible that with more experimental knowledge on DNA metabolism genes in *L.major*, the rules determined in this study will be refined, or other predictors of S-phase expressed genes may be found. It is also possible that other groups of functionally related genes in *L.major* are regulated by other specific post-transcriptional control signals. The approach described here is applicable to such groups of genes as well. By compiling a database of known genes related to a specific function, a common element can be detected in their UTRs, and used, in turn, for detecting additional genes that contain the signal and may play a similar role. Finally, this approach can be used more widely for the annotation of genes in other trypanosomatids, with possible important implications to the assignment of function to their homologues in the genomes of other organisms.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

NOTE ADDED IN PROOF

Our analysis has been performed prior to the recent completion of the sequence of the *L. major* genome [Ivens,A.C. *et al.* (2005) Science, **309**, 436–442].

ACKNOWLEDGEMENTS

This study was supported, in parts, by grant No. 623 from the Israel Science Foundation (ISF) and by grant No. 2001006 from

the United State-Israel Binational Science Foundation (BSF), Jerusalem, Israel. I.O. was supported by a Yeshaya Horowitz Fellowship. Funding to pay the Open Access publication charges for this article was provided by grant No. 623 from the Israel Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Baldauf,S.L., Roger,A.J., Wenk-Siefert,I. and Doolittle,W.F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, **290**, 972–977.
2. Clayton,C.E. (2002) Life without transcriptional control? From fly to man and back again *EMBO J*., **21**, 1881–1888.
3. Mahmood,R., Hines,J.C. and Ray,D.S. (1999) Identification of cis and trans elements involved in the cell cycle regulation of multiple genes in *Crithidia fasciculata*. *Mol. Cell Biol*., **19**, 6174–6182.
4. Mahmood,R., Mittra,B., Hines,J.C. and Ray,D.S. (2001) Characterization of the *Crithidia fasciculata* mRNA cycling sequence binding proteins. *Mol. Cell Biol*., **21**, 4453–4459.
5. Avliyakulov,N.K., Hines,J.C. and Ray,D.S. (2003) Sequence elements in both the intergenic space and the 3′ untranslated region of the *Crithidia fasciculata* KAP3 gene are required for cell cycle regulation of KAP3 mRNA. *Eukaryot. Cell*, **2**, 671–677.
6. Mittra,B., Sinha,K.M., Hines,J.C. and Ray,D.S. (2003) Presence of multiple mRNA cycling sequence element-binding proteins in *Crithidia fasciculata*. *J. Biol. Chem*., **278**, 26564–26571.
7. Hanke,T., Ramiro,M.J., Trigueros,S., Roca,J. and Larraga,V. (2003) Cloning, functional analysis and post-transcriptional regulation of a type II DNA topoisomerase from Leishmania infantum. A new potential target for anti-parasite drugs. *Nucleic Acids Res*., **31**, 4917–4928.
8. Eskin,E. and Pevzner,P.A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, **18**, S354–363.
9. Abu-Elneel,K., Robinson,D.R., Drew,M.E., Englund,P.T. and Shlomai,J. (2001) Intramitochondrial localization of universal minicircle sequence-binding protein, a trypanosomatid protein that binds kinetoplast minicircle replication origins. *J. Cell Biol*., **153**, 725–734.
10. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.
11. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. 2nd edn.. Cold Spring Harbor press, Cold Spring Harbor, NY.
12. Pasion,S.G., Brown,G.W., Brown,L.M. and Ray,D.S. (1994) Periodic expression of nuclear and mitochondrial DNA replication genes during the trypanosomatid cell cycle. *J. Cell Sci*., **107**, 3515–3520.
13. Pasion,S.G., Hines,J.C., Ou,X., Mahmood,R. and Ray,D.S. (1996) Sequences within the 5′ untranslated region regulate the levels of a kinetoplast DNA topoisomerase mRNA during the cell cycle. *Mol. Cell Biol*., **16**, 6724–6735.
14. Kohler,A., Schmidt-Zachmann,M.S. and Franke,W.W. (1997) AND-1, a natural chimeric DNA-binding protein, combines an HMG-box with regulatory WD-repeats. *J. Cell Sci*., **110**(9), 1051–1062.
15. Andersen,J.S., Wilkinson,C.J., Mayor,T., Mortensen,P., Nigg,E.A. and Mann,M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, **426**, 570–574.
16. Choe,Y., Brinen,L.S., Price,M.S., Engel,J.C., Lange,M., Grisostomi,C., Weston,S.G., Pallai,P.V., Cheng,H., Hardy,L.W. *et al*. (2005) Development of alpha-keto-based inhibitors of cruzain, a cysteine protease implicated in Chagas disease. *Bioorg. Med. Chem*., **13**, 2141–2156.