

Members of a Large Retroposon Family Are Determinants of Post-Transcriptional Gene Expression in *Leishmania*

Frédéric Bringaud^{1,2}*, Michaela Müller^{3,4}, Gustavo Coutinho Cerqueira^{5,6}, Martin Smith^{3,4}, Annie Rochette^{3,4}, Najib M. A. El-Sayed^{5,7,8}, Barbara Papadopoulou^{3,4*}, Elodie Ghedin^{5,9*}

1 Laboratoire de Génomique Fonctionnelle des Trypanosomatides, Université Victor Segalen Bordeaux 2, Bordeaux, France, **2** UMR-5234 CNRS, Bordeaux, France, **3** Infectious Diseases Research Center, Centre Hospitalier de l'Université Laval Research Center, Quebec, Canada, **4** Department of Medical Biology, Faculty of Medicine, Laval University, Quebec, Canada, **5** The Institute for Genomic Research, Rockville, Maryland, United States of America, **6** Departamento de Bioquímica e Imunológica, Universidade Federal de Minas Gerais, Minas Gerais, Brazil, **7** Department of Cell Biology and Molecular Genetics, The University of Maryland, College Park, Maryland, United States of America, **8** Center for Bioinformatics and Computational Biology, The University of Maryland, College Park, Maryland, United States of America, **9** Division of Infectious Diseases, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

Trypanosomatids are unicellular protists that include the human pathogens *Leishmania* spp. (leishmaniasis), *Trypanosoma brucei* (sleeping sickness), and *Trypanosoma cruzi* (Chagas disease). Analysis of their recently completed genomes confirmed the presence of non-long-terminal repeat retrotransposons, also called retroposons. Using the 79-bp signature sequence common to all trypanosomatid retroposons as bait, we identified in the *Leishmania major* genome two new large families of small elements—LmSIDER1 (785 copies) and LmSIDER2 (1,073 copies)—that fulfill all the characteristics of extinct trypanosomatid retroposons. LmSIDERs are ~70 times more abundant in *L. major* compared to *T. brucei* and are found almost exclusively within the 3'-untranslated regions (3' UTRs) of *L. major* mRNAs. We provide experimental evidence that LmSIDER2 act as mRNA instability elements and that LmSIDER2-containing mRNAs are generally expressed at lower levels compared to the non-LmSIDER2 mRNAs. The considerable expansion of LmSIDERs within 3'UTRs in an organism lacking transcriptional control and their role in regulating mRNA stability indicate that *Leishmania* have probably recycled these short retroposons to globally modulate the expression of a number of genes. To our knowledge, this is the first example in eukaryotes of the domestication and expansion of a family of mobile elements that have evolved to fulfill a critical cellular function.

Citation: Bringaud F, Müller M, Coutinho Cerqueira G, Smith M, Rochette A, et al. (2007) Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. PLoS Pathog 3(9): e136. doi:10.1371/journal.ppat.0030136

Introduction

Trypanosomatids are members of the kinetoplastid family of unicellular protists, which includes human pathogens responsible for Chagas disease (*Trypanosoma cruzi*), African sleeping sickness (*Trypanosoma brucei*), and leishmaniasis (*Leishmania* spp.). *T. brucei* and *T. cruzi* belong to the monophyletic *Trypanosoma* group, which is distantly related to all the other trypanosomatids, including *Leishmania* spp. [1]. Kinetoplastid protein-coding genes are often organized as large directional gene clusters (DGCs) that form polycistronic units [2–4]. Individual mRNAs with a 39-nt 5' capped spliced leader sequence and 3' poly(A) tail are generated from the polycistronic pre-mRNAs via 5' *trans*-splicing and 3' cleavage-polyadenylation reactions [5]. Several lines of evidence raise the intriguing possibility that in trypanosomatids poly(A) addition is coupled to *trans*-splicing of the downstream gene [6,7]. *trans*-splicing signals are often U-rich polypyrimidine tracts, which precede AG acceptor sites on average 50–100 nt upstream of the translational start site. There is no consensus polyadenylation signal in trypanosomatid mRNA, and evidence obtained from a small number of loci suggests that polyadenylation occurs within a short region 100–400 nt upstream of the next polypyrimidine *trans*-splicing signal [7,8]. It was recently reported that in 89% of all available cDNA sequences from *T. brucei*, polyadenylation usually occurs at an A residue located between 80 and 300 nt from

a downstream polypyrimidine tract [9]. The aforementioned polycistronic transcription, and the absence of pol II promoters in all known protein-coding genes, necessitate that gene expression be controlled post-transcriptionally. Indeed, numerous examples in kinetoplastids, including *Leishmania*, show that sequences predominantly located in the 3'-untranslated regions (3'UTRs) control mRNA stability and translation [10–18].

Transposable elements (TEs) are DNA sequences capable of moving from one chromosomal region to another. They are classified into two major groups based on the mechanisms used for their transposition. Class I TEs, or retroelements,

Editor: Peter Myler, Seattle Biomedical Research Institute, United States of America

Received March 7, 2007; **Accepted** July 27, 2007; **Published** September 28, 2007

Copyright: © 2007 Bringaud et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: DGC, directional gene cluster; DIRE, degenerated *ingi*-related element; LINE, long interspersed element; LTR, long-terminal repeat; LUC, luciferase; RIME, ribosomal mobile element; SIDER, short interspersed degenerated retroposon; SINE, short interspersed element; TE, transposable element; TSD, target site duplication; UTR, untranslated region

* To whom correspondence should be addressed. E-mail: bringaud@u-bordeaux2.fr (FB); barbara.papadopoulou@crchul.ulaval.ca (BP); ghedine@dom.pitt.edu (EG)

© These authors contributed equally to this work.

Author Summary

Transposable elements (TEs) are DNA sequences capable of moving from one chromosomal region to another. A considerable fraction of higher eukaryote genomes is comprised of TEs, as exemplified in human (over 40% of the genome) and maize (over 50% of the genome). There is now a growing body of evidence to suggest that TEs can be functionally important and not just “junk,” “selfish,” or “parasitic” DNA sequences that make as many copies of themselves as possible. Indeed, during the past ten years, a considerable number of TE copies have been described as domesticated or exapted elements playing a cellular function, such as transcriptional regulation and contribution to protein-coding regions. TE domestication has been described for only a few copies of TE families, and exaptation of a whole TE family has not been reported so far. We provide evidence that *Leishmania* spp., unicellular protists responsible for human diseases, have recycled and expanded a whole family of short and extinct TEs (retroposons) that have evolved to fulfill an important biological pathway, i.e., regulation of gene expression. We also observed that *Trypanosoma brucei* (a close relative of *Leishmania* spp.) developed other approaches to maintain the same cellular function.

transpose via reverse transcription of an RNA intermediate and are further divided into the long-terminal repeat (LTR) retrotransposons with LTRs and the non-LTR retrotransposons, also called retroposons. Class II TEs, or DNA transposons, move strictly through a DNA intermediate. A considerable fraction of higher eukaryote genomes comprises TEs, as exemplified in human (over 40% of the genome) [19] and maize (over 50% of the genome) [20]. There is now a growing body of evidence to suggest that TEs can be functionally important and not just “junk,” “selfish,” or “parasitic” DNA sequences that make as many copies of themselves as possible [21–23]. For example, there is a considerable number of domesticated TE copies that act as transcriptional regulatory elements or contribute to protein-coding regions of cellular genes (for review see [24–26]).

The recent completion of the *Trypanosoma* genome projects confirmed the presence of LTR retrotransposons and non-LTR retrotransposons (transposons) but no DNA transposons [2–4]. Retroposons constitute the most abundant TEs described in the genome of *T. cruzi* and *T. brucei* (~3% of nuclear genome), while no potentially active TEs have been characterized to date in *L. major* [3]. The most abundant retrotransposons, *ingi* and ribosomal mobile element (RIME) in *T. brucei* [27–29] and L1Tc and NARTc in *T. cruzi* [30,31], are distributed across their respective genomes, although they do show a relative site specificity for insertion [32,33]. The *T. brucei* RIME (0.5 kb) appears as a truncated version of the *T. brucei ingi* (5.25 kb), in which the central 4.7 kb fragment has been deleted (Figure 1). Similarly, the *T. cruzi* NARTc (0.25 kb) element was derived from L1Tc (4.9 kb) by a 3' deletion [30]. The potentially functional *ingi* and L1Tc each encode a large single multifunctional protein that is probably responsible for their retrotransposition and that of the short non-autonomous RIME and NARTc, respectively [32,33]. Consequently, *ingi*/RIME and L1Tc/NARTc are considered as pairs of retroposons, as previously described for the human long interspersed element 1 (LINE1)/Alu, the eel UnaL2/UnaSINE1, and the plant LINE/S1 pairs [34–37]. Until now, potentially active or short non-autonomous TEs have not

been detected in the *L. major* genome [3,4]. However, the genome does contain degenerated retroelements (*L. major* degenerated *ingi*/L1Tc-related elements [LmDIREs]) corresponding to remnants of extinct *ingi*/L1Tc-like retroposons [38]. Interestingly, the *ingi*/RIME and L1Tc/NARTc pairs and DIREs share the first 78–79 nucleotides even though they are otherwise unrelated to each other [30,38] (Figure 1). This “79 bp signature,” therefore, constitutes the hallmark of trypanosomatid retroposons.

Using the “79 bp signature” for BLASTN searches, we identified in the *L. major* genome 1,858 short (~550 bp), noncoding and degenerated retroposons that belong to two new large families of relatively conserved repetitive DNA elements (*L. major* short interspersed degenerated retroposon 1 [LmSIDER1] and LmSIDER2), which display all the hallmarks of trypanosomatid retroposons. LmSIDER1 and LmSIDER2 are predominantly located in the 3'UTR of *L. major* mRNAs and represent the most abundant TEs now characterized in trypanosomatid genomes. Considering that regulation of gene expression in *Leishmania* is mediated almost exclusively by sequences within 3'UTRs, we hypothesized that LmSIDERs may play a role in the regulation of gene expression. In the present study, we provide experimental evidence that members of the second retroposon subfamily in *L. major*, LmSIDER2, promote mRNA destabilization. We conclude that *Leishmania* spp., but not trypanosomes, have recycled and probably expanded an extinct family of short retroposons that participate in the maintenance of an essential cellular function, i.e., the regulation of gene expression.

Results

Characterization of Short Degenerated Retroposons in *L. major* and *T. brucei* Genomes

All *ingi*/RIME, L1Tc/NARTc, and DIRE present in the *T. brucei*, *T. cruzi*, and *L. major* genomes have been identified and annotated [3]. These different retroposon families contain at their 5'-extremity a 79-bp conserved motif (called “79 bp signature”), which constitutes the hallmark of trypanosomatid retroposons [38]. In order to identify other repeated sequences containing the “79 bp signature,” we surveyed the *L. major* and *T. brucei* genomes for the presence of the first 79 bp of *ingi* and 78 bp of L1Tc. BLASTN searches initially detected 108 significant matches in the *L. major* genome, in addition to identifying the LmDIRE sequences. Comparison of the sequences located downstream of these 108 “79 bp signature” matches revealed two heterogeneous groups of sequences, which we named LmSIDER1 and LmSIDER2. After several rounds of BLASTN searches with complete LmSIDER1 and LmSIDER2 sequences, we identified 1,858 related sequences (785 LmSIDER1 and 1,073 LmSIDER2) in the *L. major* genome (Figure 1). Coordinates for these elements on each *L. major* chromosome are listed in Table S1. A phylogenetic analysis of 789 LmSIDER sequences confirmed their division into two distinct subfamilies, LmSIDER1 and LmSIDER2 (Figure 2). A similar BLASTN analysis of the *T. brucei* genome revealed 22 sequences forming two groups of relatively conserved sequences ranging from 558 to 587 bp, named *T. brucei* short interspersed degenerated retroposon 1 (TbSIDER1) (ten sequences) and TbSIDER2 (12 sequences) (Figure 1).


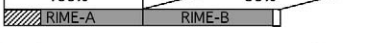



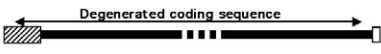

Name	Size (Kb)	Structure	Copy number per haploid genome			References
			Tb	Tc	Lm	
<u>ingi</u>	5.2		115 (3)	-	-	[28-29]
RIME	0.5		86	-	-	[27]
TbSIDER	0.57		26	-	-	This work
<u>L1Tc</u>	4.9		-	320 (15)	-	[31]
NARTc	0.25		-	133	-	[30]
<u>DIRE</u>	~5		73 (0)	257 (0)	52 (0)	[38]
LmSIDER	~0.55		-	-	1858	This work

Figure 1. Description and Copy Number of the Trypanosomatid Retroposons

Retroelement names and sizes are indicated in the left margin. The names of coding or potentially coding retroposons (including the few retrotransposition-competent *ingi* and L1Tc elements) are underlined and boldfaced; the other elements are short non-autonomous retroposons (RIME, NARTc, TbSIDER, and LmSIDER). The central panel ("Structure") represents the schematic map of the retroelements highlighting nucleotide sequence conservation, such as for the *T. brucei ingi*/RIME and *T. cruzi* L1Tc/NARTc pairs. The grey boxes represent conserved sequences between autonomous and non-autonomous members of a pair (the percentage of identity is indicated below), the hatched boxes represent the 79-bp sequence conserved at the 5'-extremity ("79 bp signature"), and the white boxes indicate the adenosine-rich stretch terminal sequences. The right panel indicates the number of each retroelement per haploid genome, including minichromosomes for *T. brucei* (dashes indicate the absence of elements in the corresponding genome). For autonomous elements, the value in brackets indicates the number (per haploid genome) of potentially functional elements, which may code for their own retrotransposition.

Lm, *L. major*; Tb, *T. brucei*; Tc, *T. cruzi*.
doi:10.1371/journal.ppat.0030136.g001

The work reported will hereafter primarily focus on the LmSIDER2 family. One thousand thirteen LmSIDER2s were aligned with the inclusion of numerous gaps to maximize the alignments (Figure S1). The aligned LmSIDER2 sequences ranged between 178 bp and 702 bp, with a mean of 545 bp. Although LmSIDER2 sequences are highly heterogeneous in composition and size (the alignment comprises 1,612 positions), a conserved core sequence was identified (538 bp) by

removing insertions (1,074 positions) (Figure S2). The removed positions (66.6% of the positions in the original alignment) account for 14.5% of the aligned nucleotides. The defined core sequence was used to perform all the subsequent bioinformatics analyses.

To determine whether the LmSIDER2 sequences are significantly conserved, we performed a chi-square (χ^2) test on the LmSIDER2 core and the flanking sequences (200 bp

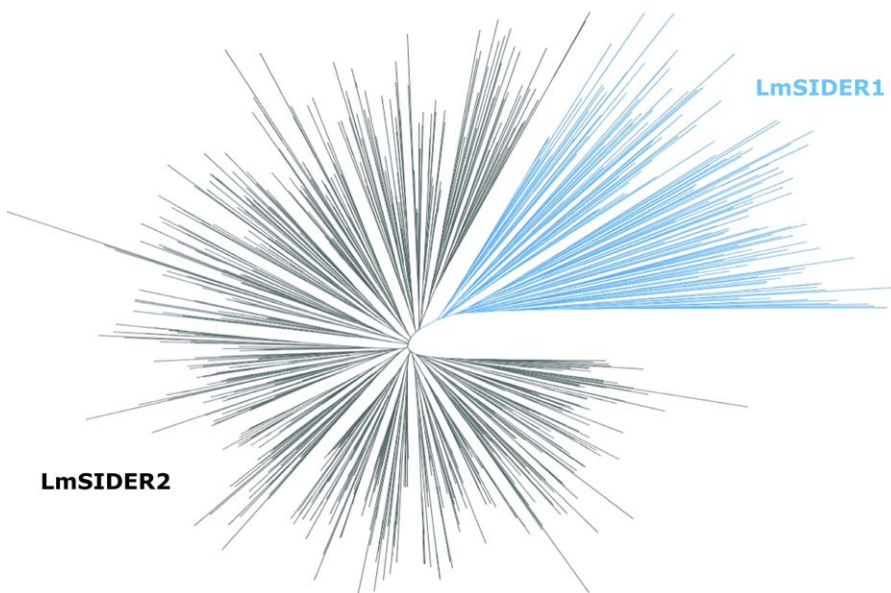


Figure 2. Minimum Evolution Phylogenetic Tree of 785 LmSIDER Sequences

Only the LmSIDER sequences longer than 400 bp and smaller than 700 bp were considered to produce an alignment as described in Materials and Methods. The unrooted phylogenetic tree displays 140 LmSIDER1 (blue cluster) and 645 LmSIDER2 (black cluster) sequences.

doi:10.1371/journal.ppat.0030136.g002

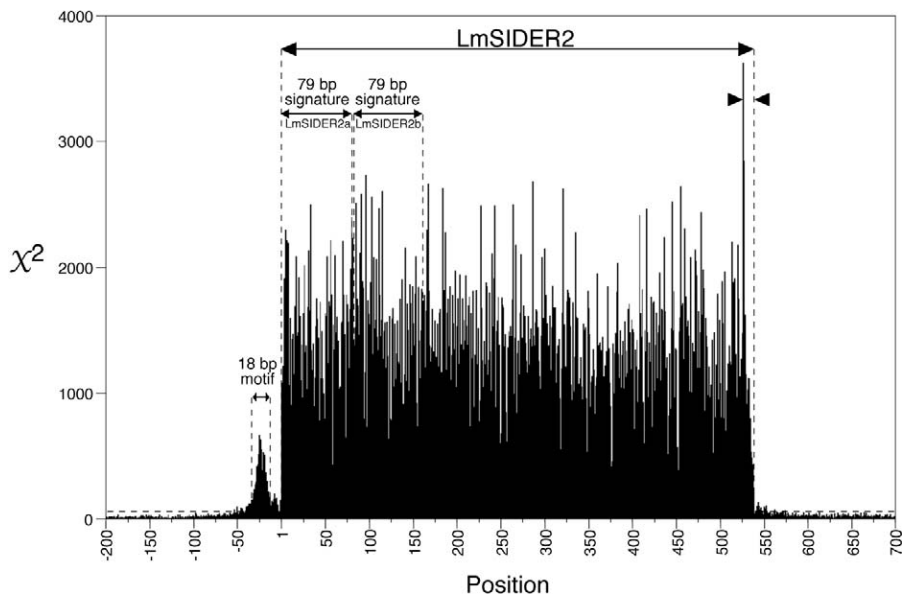


Figure 3. The χ^2 Values for Individual Positions of the LmSIDER2 Core Sequence (538 bp) and Adjacent Regions (200 bp Upstream and 160 bp Downstream)

The χ^2 values were calculated as described in Materials and Methods from the set of 1,013 aligned LmSIDER2. The base composition of the whole *L. major* genome sequence was used to determine the background base distribution. The χ^2 values above the broken horizontal line correspond to significance levels of $p < 0.001$ for three degrees of freedom. Both “79 bp signatures,” called LmSIDER2a and LmSIDER2b, are positioned. The adenosine-rich stretch and the 18-bp thymidine-rich motif located at the 3'-extremity and upstream of the LmSIDER2, respectively, are indicated between arrowheads.

doi:10.1371/journal.ppat.0030136.g003

upstream and 160 bp downstream) (Figure 3). All positions of the LmSIDER2 core show a χ^2 score far above the threshold line corresponding to significant levels (using three degrees of freedom, a χ^2 value of 16.3 corresponds to a significance level of $p < 0.001$), indicating that the LmSIDER2 core is conserved. The flanking regions are not conserved, except for a thymidine-rich stretch (18 residues) starting at 15 bp upstream from the LmSIDER2 (unpublished data).

LmSIDER and TbSIDER Sequences Contain All Hallmarks of Trypanosomatid Retroposons

Several lines of evidence demonstrate that members of LmSIDER2 are clearly related to retroposons identified in trypanosomes (*ingi*/RIME, L1Tc/NARTc, and DIRE). (i) Two tandemly arranged “79 bp signatures” are found at the 5'-extremity of the LmSIDER2 core (Figures 3 and 4). These are 68% and 62% identical with the first 79 bp residues of the *T. brucei ingi*/RIME. (ii) The 3'-extremity of the LmSIDER2 core sequence is composed of an adenosine-rich stretch, which is a hallmark of retroelements due to the requirement of an RNA intermediate during retrotransposition [39] (Figures 3 and 5). (iii) The LmSIDER2 sequences show a high GC content (65.3%), similar to the one seen in LmDIRE (64.5%), as compared to the rest of the *L. major* genome (59.7%) (Table 1). The GC content is also higher for the *T. brucei* RIMEs (53.8%), *ingis* (52.3%), and TbDIREs (48.7%), as compared to the rest of the *T. brucei* genome (41%). The relative lower GC content of the TbDIREs compared to the *ingi*/RIME sequences is probably due to the accumulation of point mutations in TbDIREs, as previously observed for extinct retroposons [24]. This interpretation may also explain the relative lower GC content bias observed in the degenerated LmSIDER2 and LmDIRE sequences, compared to the potentially active *ingi*

and RIME elements. (iv) As previously observed for the *T. brucei ingi*/RIME and *T. cruzi* L1Tc/NARTc retroposons, an 18-bp thymidine-rich motif is conserved upstream of LmSIDER2 (Figures 3 and 5). According to the current model of retrotransposition, this sequence motif corresponds probably to the recognition site of the endonuclease encoded by *ingi* L1Tc-related elements [32,33]. (v) During retrotransposition, the retroposon-encoded endonuclease performs two asymmetrical single-strand cleavages, leading to a duplication of the residues between both cleavages. The duplicated motif, flanking the newly inserted retroposons, is called target site duplication (TSD) (Figure 5). One hundred ninety-one LmSIDER2 sequences (18.9% of the aligned LmSIDER2) are flanked by a conserved motif (>75% identity) ranging from 11 bp to 19 bp (69 of them being 13 bp long), which resemble vestiges of TSDs. For three of them, the 11–13-bp TSD is conserved without mismatch (Figure 5A). Interestingly, the size of TSD flanking LmSIDER2 and the *ingi*/RIME/L1Tc/NARTc elements is similar (~13 bp versus 12 bp) [32,33]. (vi) 90% of the identified LmDIRE sequences (47 out of 52) overlap with a LmSIDER2 sequence at their 5'- and/or 3'-extremities (unpublished data), suggesting that LmDIRE (previously characterized as retroelement vestiges related to *ingi* and L1Tc [38]) and LmSIDER are related. This last observation suggests that LmSIDER was derived from LmDIRE by deletion, as observed for the *T. brucei ingi*/RIME and *T. cruzi* (L1Tc/NARTc) autonomous/non-autonomous pairs of retroposons [30] (Figure 1).

Similarly, both TbSIDER groups show hallmarks of trypanosomatid retroposons, including the presence of the “79 bp signature” (Figure 4) and an adenosine-rich stretch (Figure 5B and 5C) at their 5'- and 3'-extremity, respectively. In addition, one member each of the TbSIDER1 and

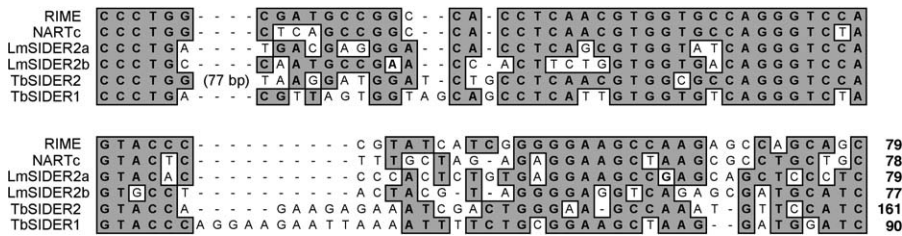


Figure 4. Comparison of the “79 bp Signature” Consensus Sequences between Different Trypanosomatid Retroposons

The first 79 bp of RIME, 78 bp of NARTc, 161 bp of TbSIDER2, 90 bp of TbSIDER1, and both “79 bp signatures” located at the 5'-extremity of the LmSIDER core sequence (LmSIDER2a and LmSIDER2b) were aligned, with the introduction of gaps (-) to maximize the alignments. Identical residues are shaded in grey.

doi:10.1371/journal.ppat.0030136.g004

TbSIDER2 groups is flanked by a degenerated TSD sequence (Figure 5B and 5C).

SIDERS Are Extinct Retroposons

The difficulties encountered in performing the LmSIDER2 alignment reflect the high level of divergence of this TE family. To study the extent of this divergence and gain better insight into the evolutionary dynamics of the LmSIDER2 family, we calculated the percentage of divergence between the consensus LmSIDER2 core sequence deduced from the alignment and each LmSIDER2 core sequence. Since the consensus sequence is assumed to approximate the element's original sequence at the time of insertion, the percentage of substitution from the consensus sequence is correlated to the age of a given element (the age corresponds to the time of

retrotransposition). The divergence ranged between 12% and 40%, with median and mean values of 20% and 17%, respectively (Figure 6). The high level of divergence between the consensus and the most conserved LmSIDER2 sequence (12%) implies that LmSIDER became extinct a long time ago.

The same analysis was carried out on the *T. brucei* RIME/TbSIDER and *T. cruzi* NARTc elements, which are the only short retroposons characterized so far in the trypanosome genomes [40]. For TbSIDERs, the percentage of divergence from the consensus TbSIDER1 and TbSIDER2 core sequences ranged between 11.6% and 18% and 8% and 13.7%, with median values of 16% and 11%, respectively (Figure 6). This indicates that TbSIDERs are also extinct TEs, as observed for LmSIDERs. In contrast, RIME and NARTc are far more

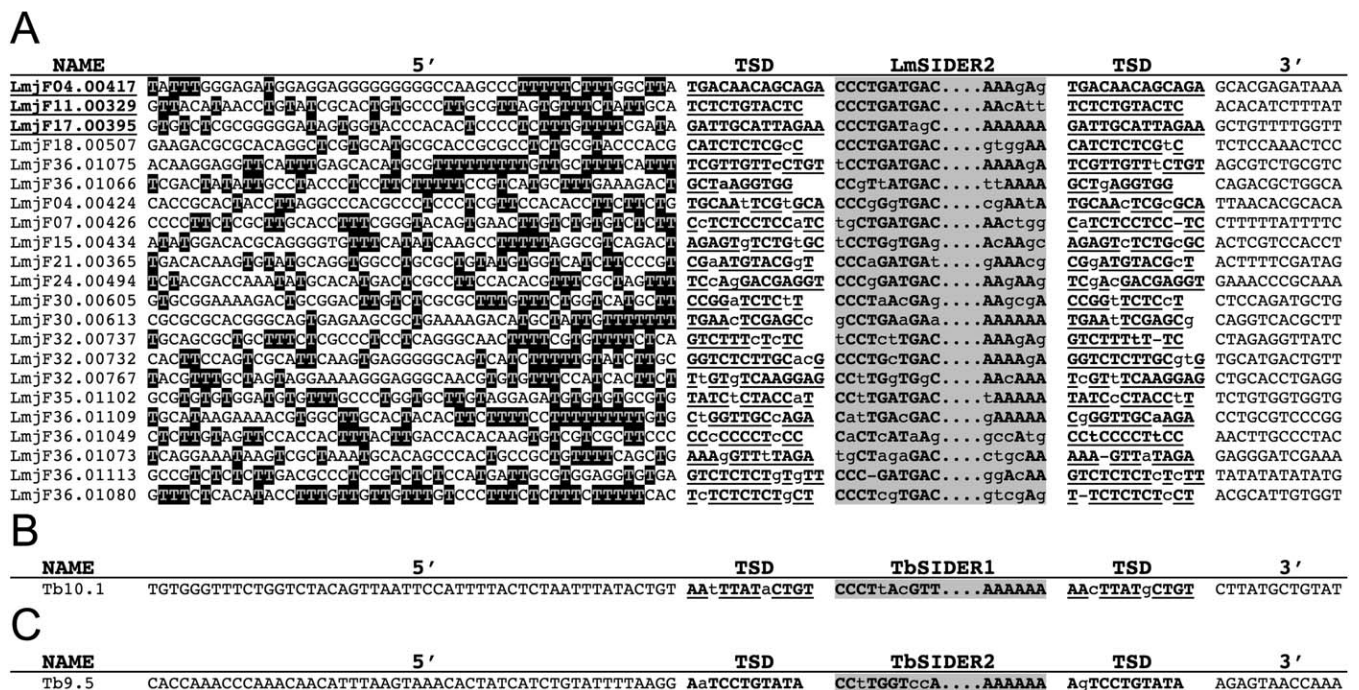


Figure 5. Comparison of the Target Site Duplication (TSD) Flanking LmSIDER2 (A), TbSIDER1 (B), and TbSIDER2 (C) Sequences

In the left margin, the name of the element is indicated (the chromosome number is followed by the locus number). In this figure, only the LmSIDER/TbSIDER elements flanked by a TSD presenting two mismatches at the most are shown. The underlined names mean that the corresponding element is flanked by conserved TSD. The alignment of all the selected sequences was based on the retroelement sequences (grey column headed “LmSIDER2/TbSIDER1/TbSIDER2”) from which only the first 10 bp and the last 6 bp are shown (the conserved residues of the retroposon are boldfaced and capital characters). The TSD flanking the retroelements is indicated by boldfaced and underlined capital characters for the conserved residues. Lowercase characters in the TSD column correspond to nonconserved residues. In (A), T residues within the 5'-flanking sequences (called “5'”) that are abundant upstream of the TSD are indicated with white characters on a black background.

doi:10.1371/journal.ppat.0030136.g005

Table 1. GC Percentage of Trypanosomatid Retroposons

Category	<i>T. brucei</i>	<i>L. major</i>
Whole genome	41.0	59.7
<i>ingi</i>	52.3	—
RIME	53.8	—
TbSIDER	48.9	—
TbDIRE / LmDIRE	48.7	64.5
LmSIDER	—	65.3

doi:10.1371/journal.ppat.0030136.t001

conserved compared to SIDERs (median divergence value of 4% and 2%, respectively) (Figure 6). In addition, 13.8% and 22.5% of the analyzed RIME and NARTc sequences are over 99% identical with the consensus sequence, respectively, indicating recent retrotransposition activities in the trypanosome genomes.

SIDER Distribution in the *L. major* and *T. brucei* Genomes

The *T. brucei* and *L. major* genomes are highly syntenic, with approximately 70% of all genes remaining in the same genomic context [40]. This large-scale synteny enables a comparative analysis of TE distribution in these two completed trypanosomatid genomes. The trypanosomatid genomes are characterized by their unique arrangement of DGCs, which are separated by short (0.9–14 kb) divergent or convergent strand-switch regions. For example, the *L. major* genome (32.6 Mb) has 36 pairs of chromosomes (0.25–2.7 Mb) that are organized into 133 DGCs of tens to hundreds of protein-coding genes (up to 1.26 Mb per DGC) [4]. The *T. brucei* genome is more compact (26 Mb), with 11 pairs of megachromosomes (1.1–5.5 Mb) containing subtelomeric genes at both extremities, which account for ~20% of the genome (~5.2 Mb) [2], while *L. major* chromosomes do not contain large subtelomeric regions [4].

Interestingly, retroposons do not show the same distribution in the *L. major* and *T. brucei* genomes (Tables 2 and 3). Indeed, almost all of LmSIDERs and LmDIREs in *L. major* are located in DGCs (95.4% of the TE), while the *ingi*, RIME, TbDIRE, and TbSIDER retroposons in *T. brucei* are primarily located in subtelomeric regions (60.1% of the TE). Tables 2 and 3 also show that strand-switch regions display the highest TE richness in both *T. brucei* and *L. major*, i.e., over 110 TE per Mb, which corresponds to 23.4% (54 TE) and 4.6% (88 TE) of the retroposons, respectively. The most striking observation is that retroposons are ~50 times more abundant in *L. major* DGCs compared to *T. brucei* DGCs (1,821 versus 38), despite the high level of synteny observed between these regions, which contain an equivalent number of protein-coding genes [40]. This extraordinary difference is the consequence of the unusual distribution and high copy number of LmSIDERs, as exemplified by the comparative analysis of *T. brucei* Chromosome 6 and *L. major* Chromosome 30, which are almost completely syntenic (Figure 7) (see Figures S3 and S4 for the other chromosomes).

LmSIDER2 Are Located within the 3'UTR of mRNAs

Since most LmSIDERs are present in the intergenic regions of DGCs, it was important to determine where they are located in regards to the pre-mRNA processing sites. Individual mature mRNAs in trypanosomatids are generated

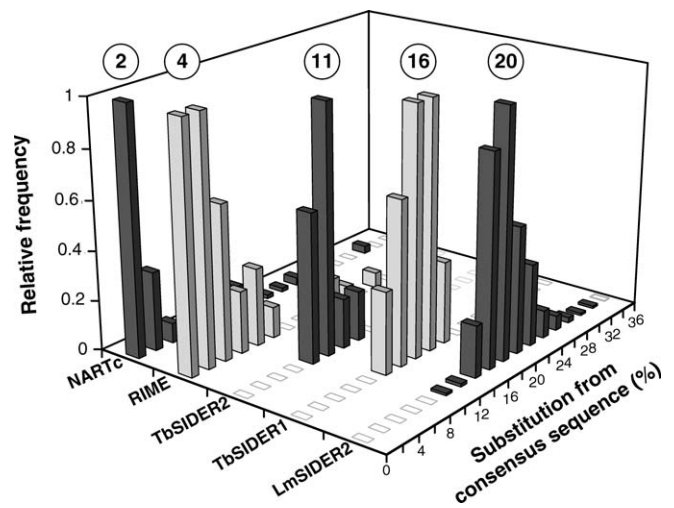


Figure 6. Divergence between Members of LmSIDER2 (1,013 Copies), TbSIDER1 (10 Copies), TbSIDER2 (12 Copies), RIME (70 Copies), and NARTc (115 Copies)

Bases covered by the *L. major* (LmSIDER2), *T. cruzi* (NARTc), and *T. brucei* (RIME, TbSIDER1, and TbSIDER2) short retroposons were sorted by their divergence from their consensus sequence. The consensus sequences, determined from the alignment of the core sequence of all the analyzed retroposons, approximate the element's original sequence at the time of insertion. The number of retroposons per fraction of 2% divergence is expressed as a fraction of the highest value, for which an arbitrary value of 1 has been assigned. The percentage of divergence was calculated using the matching region of the consensus sequences. The circled numbers on the top indicate the median value of each graph.

doi:10.1371/journal.ppat.0030136.g006

from polycistronic precursors by 5' *trans*-splicing of a 39-nt capped leader RNA and 3' polyadenylation [41]. To determine the putative position of polyadenylation sites in *L. major*, we used the prediction algorithm previously developed for trypanosome mRNA processing sites [9]. There are 8,162 genes annotated in version 4.0 of the *L. major* genome. The algorithm could predict the vast majority of the 5'UTRs and 3'UTRs of those genes with the exception of 121 5'UTRs (1.5%) and 569 3'UTRs (7%).

Of the 1,858 LmSIDERs characterized in the *L. major* genome, 1,356 were found to overlap with a 3'UTR, and 494 have at least one 3'UTR upstream, including 85 LmSIDERs found in strand-switch regions. Conversely, 1,852 have at least one 5'UTR downstream, including 50 LmSIDERs overlapping with the 5'UTR of a gene. Because 73% of the LmSIDERs are found within 3'UTRs, we calculated the median distance of these elements to the upstream stop codon (680 bp) and the downstream ATG (978 bp), as well as the distances from the polypyrimidine tract (833 bp) and putative polyadenylation site (734 bp) (Figure 8). The average location of LmSIDER2s is in the middle of the *in silico*-predicted 3'UTRs (at almost equal distance from the upstream stop codon and the downstream polyadenylation site), which clearly demonstrates that most LmSIDER2s are located in the 3'UTR of mRNAs.

LmSIDER2-Containing Transcripts Are on Average Expressed at Lower Levels Relative to Transcripts Lacking LmSIDER2

3'UTRs are known to play a key role in regulating gene expression in *Leishmania* [13,15,18,42–46]. The widespread

Table 2. Retroposon Distribution in the *T. brucei* Genome

Location of Retroposons	Size ^a (Mb)	Percentage				Total	Density ^b (TE per Mb)
		TbSIDER	RIME	<i>ingi</i>	DIRE		
DGC	20.8	4.5	24.3	11.3	22.5	16.5	1.8
Coding strand	20.8	0	11.4	3.1	5.0	5.6	0.6
Noncoding strand	20.8	4.5	12.9	8.2	17.5	10.8	1.2
Strand-switch	0.48 ^c	9.1	31.4	14.4	40.0	23.4	113
Convergent	0.23	0	15.7	8.2	25	12.6	126
Divergent	0.25	9.1	15.7	6.2	15	10.8	100
Subtelomere	5.2	86.4	44.3	76.3	37.5	60.1	26.7

Number of identified TE copies per haploid genome (only megachromosomes are considered [2]): TbSIDER, 22; RIME, 70; *ingi*, 99; DIRE, 40; total, 231.

^aSize of the considered region of the genome.

^bNumber of TE copies per Mb.

^cThe average size of the 95 strand-switch regions is estimated as 5 kb.

doi:10.1371/journal.ppat.0030136.t002

distribution of LmSIDER elements within the *Leishmania* genome and their predominant localization in 3'UTRs, therefore, support the hypothesis that LmSIDER2 may contribute to the regulation of gene expression in this organism. To test this hypothesis, we used custom-designed low density DNA oligonucleotide microarrays to determine expression profiles of LmSIDER2-containing mRNAs in *L. major* promastigotes and *L. major* lesion amastigotes isolated from BALB/c mice. Oligonucleotide microarrays were designed to represent 154 *L. major* genes, from which only 38 bear LmSIDER2 in their 3'UTR. Four independent hybridization experiments were scanned and analyzed using recommended statistic parameters for low spot density arrays in the GeneSpring software. The overall pattern of gene expression for *L. major* promastigotes and amastigotes is shown in the scatterplot of normalized data in Figure 9A. Approximately 50% of the LmSIDER2-containing transcripts are developmentally regulated in either *L. major* promastigotes or amastigotes, without any bias towards a particular life stage (24% amastigotes versus 26% promastigotes) and with the majority of genes being constitutively expressed (Figure 9A; Table S2). Interestingly, from these LmSIDER2-containing transcripts, more than 75% have signal intensities that are lower than the mean intensity of all the spots, as

compared to 40% for the non-LmSIDER2 transcripts (Figure 9A). The minority of LmSIDER2-containing more abundant transcripts (~25%) may be explained by a higher degeneracy of LmSIDER2 that results in a nonfunctional element or by the presence of additional elements within the 3'UTR.

To gain independent evidence for the relatively lower expression of LmSIDER2 mRNAs, a randomly selected number of *L. major* transcripts containing or lacking LmSIDER2 that are most likely clustered within the same transcription unit on three distinct chromosomes were analyzed by quantitative northern blotting. LmjF13.0440, LmjF24.1260, LmjF24.1360, and LmjF36.3810 transcripts harbor LmSIDER2 in their 3'UTR, whereas LmjF13.0430, LmjF24.1250, LmjF24.1280, and LmjF36.3910 do not. LmjF13.0430/LmjF13.0440 and LmjF24.1250/LmjF24.1260 are tandemly linked, whereas LmjF24.1280/LmjF24.1360 and LmjF36.3810/LmjF36.3910 are part of the same transcription unit but are separated by seven to eight genes (Figure 9B). Figure 9B demonstrates that LmSIDER2-containing mRNAs are systematically expressed at much lower levels compared to their co-transcribed genes lacking LmSIDER2. Taken together, these results argue for a more general role of LmSIDER2 in downregulating mRNA expression.

Table 3. Retroposon Distribution in the *L. major* Genome

Location of Retroposons	Size ^a (Mb)	Percentage			Total	Density ^b (TE per Mb)
		LmSIDER1	LmSIDER2	DIRE		
DGC	32.6	97.0	94.3	92.2	95.4	55.9
Coding strand	32.6	79.1	79.9	70.6	79.3	46.5
Noncoding strand	32.6	17.9	14.4	21.6	16.1	9.4
Strand-switch	0.72 ^c	3.0	5.7	7.8	4.6	122
Convergent	0.34	1.5	2.5	5.9	2.2	124
Divergent	0.38	1.5	3.2	1.9	2.4	121
Subtelomere	0	0	0	0	0	0

Number of identified TE copies per haploid genome: LmSIDER1, 785; LmSIDER2, 1,073; DIRE, 52; total, 1,910.

^aSize of the considered region of the genome.

^bNumber of TE copies per Mb.

^cThe average size of the 96 strand-switch regions is estimated as 7.5 kb.

doi:10.1371/journal.ppat.0030136.t003

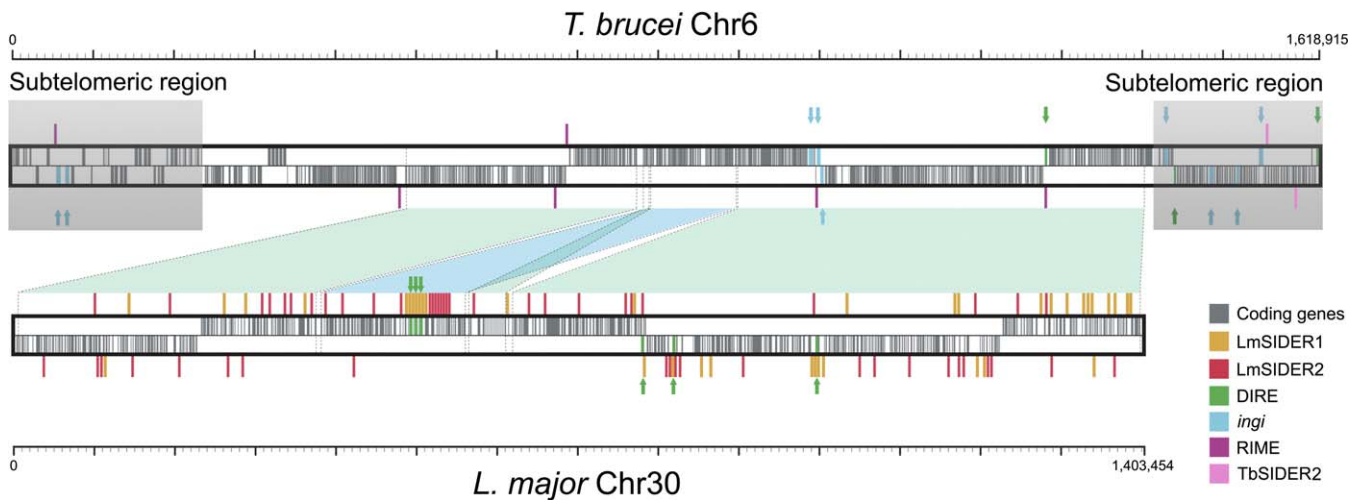


Figure 7. Comparative Analysis of TbChr6 and LmChr30 Syntenic Chromosomes

The syntenic regions between the *L. major* Chromosome 30 (LmChr30) and the *T. brucei* Chromosome 6 (TbChr6) are represented by blue diamonds. The grey shaded extremities of TbChr6 represent subtelomeric regions primarily composed of pseudogenes. The position of protein-coding genes and retroposons in each chromosome is indicated by vertical bars with the color code displayed on the right margin. Protein-encoding genes and *ingi* and DIRE retroposons are shown on the upper or lower part of the schematic chromosomes, depending on their strand location. Above or below the schematic chromosomes, the other retroposons (LmSIDER1, LmSIDER2, TbSIDER2, and RIME) are indicated, as are the blue and green arrows, which show the position of *ingi* and DIRE retroposons, respectively. The size (bp) of the chromosomes is indicated by the scale bars.
doi:10.1371/journal.ppat.0030136.g007

Mutational Analysis of LmSIDER2 mRNAs Shows That LmSIDER2 Downregulate mRNA Expression Levels

We have recently identified conserved regulatory elements within the 3'UTR of a large set of developmentally regulated transcripts in *Leishmania* and showed that these elements operate principally at the translational level [16,17]. While characterizing the LmSIDER families, we found that these regulatory elements are part of the LmSIDER1 subfamily.

We next wanted to obtain direct evidence for the role of LmSIDER2 elements in the regulation of gene expression using luciferase (LUC) as a reporter mRNA. For this, two members of the LmSIDER2 subfamily were selected for further analysis. LmjF08.1270 encodes a hypothetical protein of unknown function [47] and LmjF36.3810 encodes an

aminomethyltransferase. Both harbor LmSIDER2 in their 3'UTR. The LmSIDER2 in the LmjF08.1270 transcript (LmSIDER2-1270) is 563 nt long and is located at the end of a 1,531-nt-long 3'UTR (53 nt upstream from the mapped polyadenylation site, unpublished data). In the case of LmjF36.3810, LmSIDER2 (LmSIDER2-3810) is 610 nt long and is located within a 1,831-nt 3'UTR, at 534 nt from the 3'end of the mRNA (see Figure 10A). The sequence identity between the two LmSIDER2 is 60%. The full-length 3'UTR of either LmjF08.1270 or LmjF36.3810 mRNAs was cloned downstream of the LUC reporter gene. LUC reporter constructs with the whole 3'UTR lacking LmSIDER2 or the LmSIDER2 alone were also made (Figure 10A). Each construct was transfected into *L. major* promastigotes, and

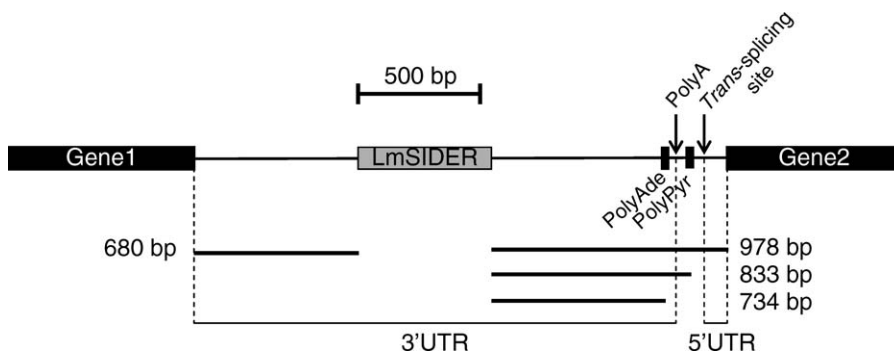


Figure 8. Predominant Localization of LmSIDERs in 3'UTRs

Within the intergenic region, *trans*-splicing generally occurs at an AG dinucleotide (*trans*-splicing site) downstream of a long polypyrimidine tract (PolyPyr). Polyadenylation (PolyA) of the upstream cistron takes place possibly as part of a coupled process together with *trans*-splicing. Consequently, the 3'UTR of the *gene1* mRNA (upstream) ends at the putative polyadenylation site (PolyAde), and the 5'UTR of the *gene2* mRNA (downstream) starts at the *trans*-splicing site. This figure shows the average relative position of the polyadenylation sites and the polypyrimidine tracts estimated with a previously developed algorithm [9], as well as the position of the LmSIDERs. The median size between LmSIDERs and the stop codon of the upstream gene (*gene1*), PolyAde, PolyPyr, or the start codon of the downstream gene (*gene2*) are also indicated.
doi:10.1371/journal.ppat.0030136.g008

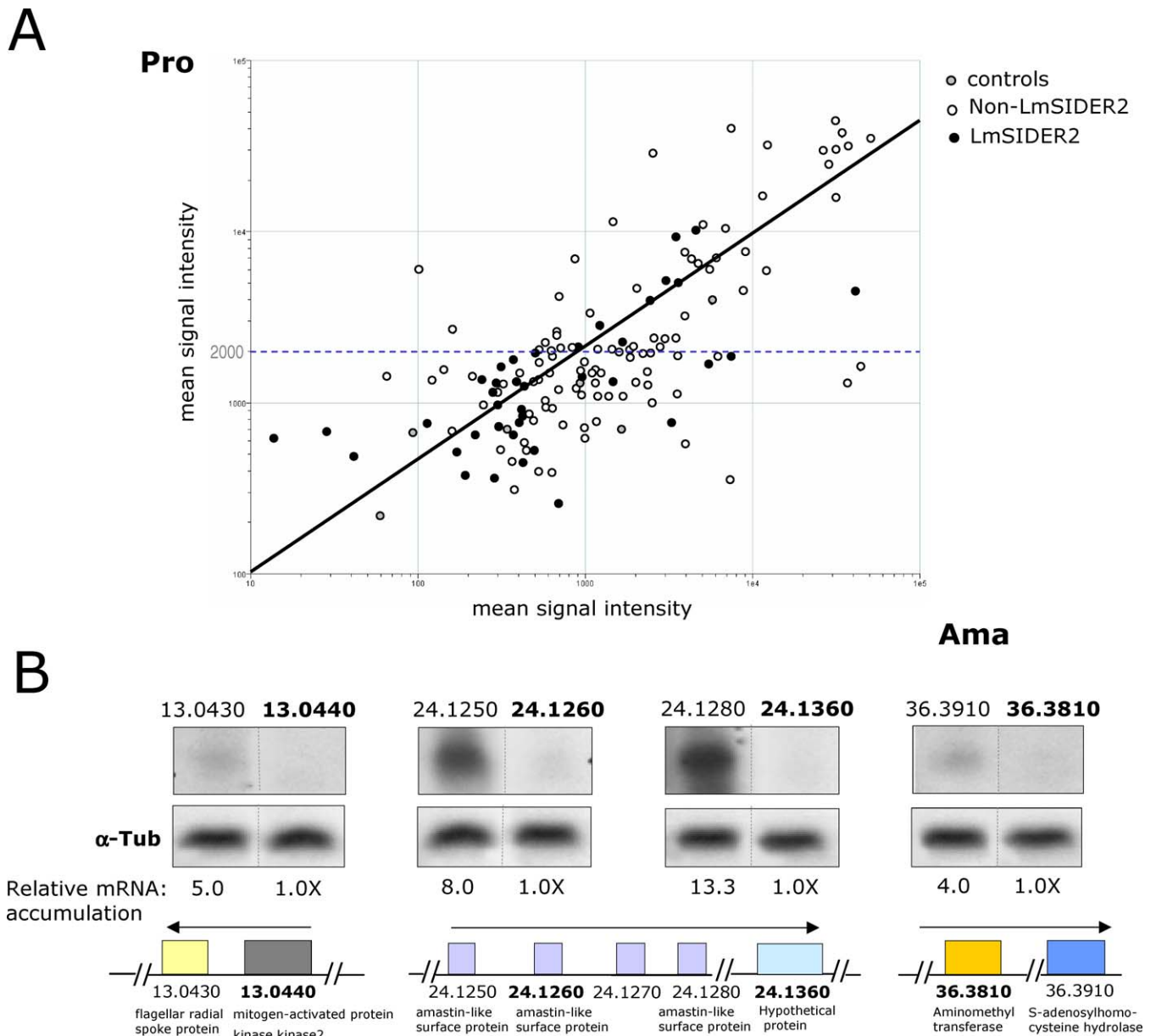


Figure 9. LmSIDER2-Containing mRNAs Are Expressed for the Most Part at Lower Levels Relative to Transcripts Lacking LmSIDER2

(A) Log-scale scatter plot comparing ratios of normalized hybridization intensities between fluorescently labeled *L. major* promastigotes (Cy3) and amastigotes isolated from mice lesions (Cy5) RNA samples. A custom-designed low density DNA oligonucleotide-based microarray comprising 154 *L. major* genes, from which only 38 are predicted to harbor LmSIDER2 in their 3'UTR, was used for this study. Total RNA was purified from *L. major* promastigotes (Pro) grown to mid-log phase and from *L. major* lesion amastigotes (Ama). Probes were synthesized from total RNA and hybridized to microarrays in quadruplicate. Hybridization experiments were scanned and analyzed using recommended statistic parameters for low spot density arrays in the GeneSpring software. Mean signal intensities of all spots corresponding to one gene were background subtracted and normalized with the mean spot intensity of the alien RNA NAC1. Genes were considered as differentially regulated when their expression ratios satisfied a *p*-value below 0.05. The mean signal intensity within the array was calculated to be ~2,000 (horizontal dotted line). 50% of the genes were identified as significantly differentially expressed. 75% of the LmSIDER2-containing transcripts showed signal intensity lower than the mean intensity of all the spots as compared to 40% for the non-SIDER2 transcripts.

(B) The steady-state levels of four pairs of transcripts that are part of the same transcription unit on three different *L. major* chromosomes were estimated by quantitative northern blotting. LmjF13.0440, LmjF24.1260, LmjF24.1360, and LmjF36.3810 transcripts (in bold) harbor LmSIDER2 in their 3'UTR, whereas LmjF13.0430, LmjF24.1250, LmjF24.1280, and LmjF36.3910 do not. LmjF13.0430 and LmjF13.0440 genes are tandemly linked on Chromosome 13, LmjF24.1250 and LmjF24.1260 are tandemly linked on Chromosome 24, and LmjF24.1280 and LmjF24.1360 genes (chr 24) and LmjF36.3810 and LmjF36.3910 genes (chr 36) are separated by seven and eight genes, respectively. The predicted putative function of the above protein-coding genes is indicated at the lower panel. Equal amounts (~20 μ g) of total RNA were loaded on agarose gel prior to transfer onto a nylon membrane and hybridized with gene-specific probes that were of the same length, GC content, and labeling activity. mRNA levels were quantitated with respect to the amount of total RNA loaded on the gel as verified by hybridization using the alpha-tubulin (α -Tub) gene-specific probe. The normalized numbers are indicated below the blots.

doi:10.1371/journal.ppat.0030136.g009

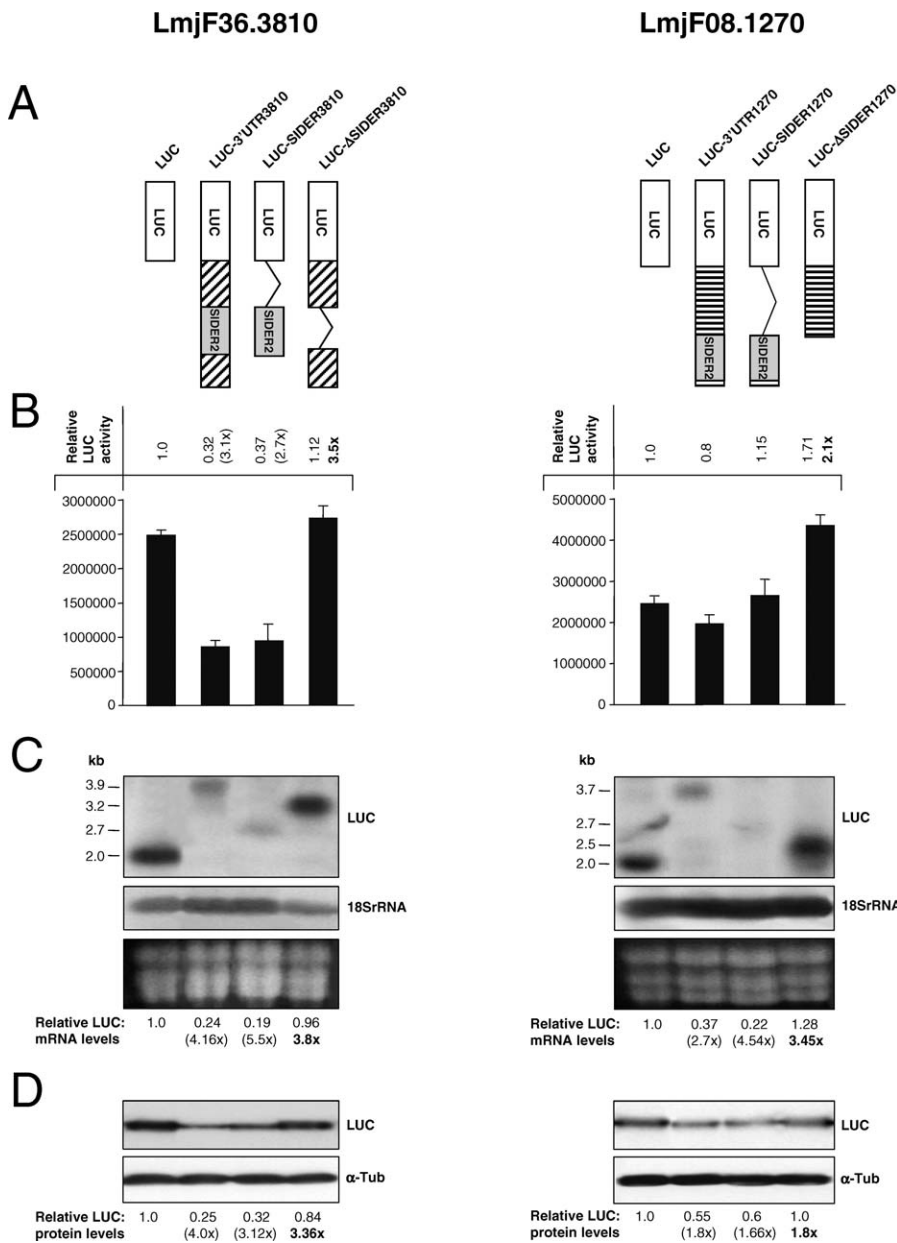


Figure 10. LmsIDER2 Promotes mRNA Downregulation in *L. major*

The potential role of LmsIDER2-containing LmjF36.3810 and LmjF08.1270 3'UTRs in regulating either mRNA or protein levels was evaluated in *L. major* recombinant parasites grown as promastigotes.

(A) Schematic representation of the different LUC chimeric constructs used in this study with the corresponding name indicated on the top. The cross-hatched boxes indicate 3'UTR sequences other than the SIDER sequence. In all the LUC-expressing vectors, the LUC transcript is processed at the 5'-end by sequences from the alpha-tubulin intercistronic region that provides signals for *trans*-splicing (see Materials and Methods). A series of 3'UTR sequences comprising either the full-length 3'UTR (3'UTR-3810, 3'UTR-1270) from the LmjF36.3810 (3,810) and LmjF08.1270 (1,270) transcripts, or the defined LmsIDER2 sequence alone (SIDER-3810, SIDER-1270), or the 3'UTR lacking the LmsIDER2 element (Δ SIDER-3810, Δ SIDER-1270), were cloned downstream of the reporter gene firefly luciferase (LUC). The presence and location of LmsIDER2 within the 3'UTR of LmjF08.1270 and LmjF36.3810 transcripts was verified by 3'UTR mapping. The pSPYNEO α LUC vector (LUC) lacking any regulatory 3'UTR region [16] was used here as a control. The above LUC constructs were introduced by electroporation into the *L. major* LV39 strain to obtain stable recombinant parasites. The copy number of the different LUC vectors per cell was similar as estimated by Southern blot hybridization (unpublished data).

(B) LUC activity was measured as indicated in Materials and Methods. The data were normalized relatively to the control transfectant (LUC). Numbers in parentheses correspond to fold differences in LUC activity with respect to the control LUC strain. The number in bold represents the fold difference in LUC activity compared to the full-length 3'UTR-3810 or 3'UTR-1270. Values are mean + standard error of four independent experiments.

(C) Northern blot analysis of total RNA extracted from *L. major* promastigotes expressing the different LUC-chimeric constructs described in (A). RNA loading on the gel was monitored by hybridization to the 18S rRNA-specific probe. A section of the ethidium-stained gel containing the three ribosomal RNAs is also shown to demonstrate loading. Northern blot hybridization was repeated at least two times and similar results were obtained.

(D) Western blot analysis of total protein lysates from *L. major*-LUC recombinant promastigotes using the anti-LUC antibody. Membranes were stripped and reacted with an anti-alpha-tubulin (α -Tub) antibody to verify protein loading. Western blot analyses were carried out with three different cultures for each transfectant and similar results were obtained. Numbers in parentheses correspond to fold differences relative to the LUC control protein steady-state levels.

doi:10.1371/journal.ppat.0030136.g010

stable recombinant parasites were analyzed for LUC activity. Relative LUC activity was calculated by comparing the values obtained with either *SIDER2*-expressing or *SIDER2*-lacking recombinant parasites to the LUC control [16]. Figure 9B demonstrates that the LmjF36.3810 3'UTR (LUC-3'UTR3810) results in a 3.1-fold decrease in LUC activity in comparison to the LUC control. A similar decrease (2.7-fold) was obtained with the LmjF36.3810 *LmSIDER2* alone (LUC-*SIDER2*3810). Contrasting with this, deletion of *SIDER2*3810 in *L. major* LUC- Δ *SIDER2*3810 promastigotes caused a 3.5-fold increase in LUC activity with respect to the LUC-3'UTR3810 and LUC-*SIDER2*3810 recombinant parasites. In the case of LUC-3'UTR1270 and LUC-*SIDER2*1270 promastigote cultures, the presence of *LmSIDER2* had only a slight effect on LUC activity; however, the deletion of *LmSIDER2* in LUC- Δ *SIDER2*1270 resulted in a 2.1-fold increase in LUC activity (Figure 10B), which is consistent with a putative role of *LmSIDER2* in regulating LmjF08.1270 gene expression.

To investigate the basis of the differences observed in LUC activity between *LmSIDER2*-bearing and *LmSIDER2*-lacking LUC chimeric constructs, we first tested the effect of *LmSIDER2* on LUC mRNA abundance by northern blotting. RNA loading on the gel was monitored by hybridization to the 18S rRNA-specific probe. The LmjF36.3810 or LmjF08.1270 *LmSIDER2* reduces the levels of LUC chimeric mRNAs by an average of 5-fold with respect to the LUC control mRNA levels (Figure 10C). In contrast to this, deletion of *LmSIDER2*-3810 or *LmSIDER2*-1270 retroposons causes a marked increase in LUC mRNA accumulation (3.45- to 3.8-fold). These findings indicate that *LmSIDER2* could downregulate mRNA abundance.

To determine the relative contribution of mRNA abundance to the observed LUC activity, we evaluated the level of LUC protein expression derived from the *LmSIDER2*-containing 3'UTRs by western blotting (Figure 10D). In the case of LUC-3'UTR3810 and LUC-*SIDER2*3810 transfectants, the amount of LUC mRNA dictates the amount of LUC protein. A linear correlation was also observed between LUC- Δ *SIDER2*3810 mRNA accumulation and LUC- Δ *SIDER2*3810 protein levels (Figure 10C and 10D). These findings establish that *LmSIDER2*-3810 does not alter translational regulation in *L. major* promastigotes, but rather confers lower mRNA levels. However, although *LmSIDER2*-1270 clearly contributes to lower steady-state RNA levels, the decrease in mRNA (2.7-fold to 4.54-fold) does not perfectly correlate with LUC protein levels (1.6-fold to 1.8-fold decrease), and LUC activity remained practically unchanged between LUC-3'UTR1270 and LUC-*SIDER2*1270 recombinant parasites in comparison to the LUC control (Figure 10B–10D). These data suggest that in the context of LmjF08.1270, other sequences might compensate for the downregulation effect of *LmSIDER2* on mRNA abundance, probably by increasing translation rates.

LmSIDER2 Are Involved in mRNA Destabilization

As regulation of gene expression in *Leishmania* is known not to occur at the transcriptional level, and as there is virtually no evidence for differential splicing [10], the most likely mechanism for lower abundance of *LmSIDER2* mRNAs is through altered mRNA stability. To examine whether lower accumulation of *LmSIDER2*-3810- and *LmSIDER2*-1270-containing LUC chimeric transcripts in *L. major* promastigotes could be due to mRNA destabilization, we measured

half-lives of the LUC transcripts that bear or lack *LmSIDER2* using actinomycin D treatment to block de novo transcription and northern blot hybridization to visualize mRNAs. Analysis of the data revealed that LUC-3'UTR1270 and LUC-3'UTR3810 transcripts have half-lives of 45 min and 80 min, respectively (Figure 11A and 11B). *LmSIDER2* deletion resulted in a marked increase of the half-life of the LUC transcript by 3.0- to 5.5-fold, respectively (Figure 11A and 11B). We also evaluated the half-lives of the single copy endogenous LmjF36.3810 and LmjF08.1270 mRNAs, which are very short (~16 and 14 min, respectively) (Figure 11C and 11D). The differences in the half-lives observed between the endogenous and the episomal *LmSIDER2*-containing transcripts can be explained by the higher copy number (~35) of the latter compared to that of the former.

Discussion

The discovery of transposable elements in trypanosomatid genomes was recently advanced by the completion of the *T. brucei*, *T. cruzi*, and *L. major* genomic sequence [2–4]. Here, we describe a newly discovered family of extinct retroposons in *Leishmania*, named *LmSIDER2* (1,073 copies), that are predominantly located in the 3'UTR of mRNAs, and show that members within this family play a role in the regulation of gene expression.

Evolution of Autonomous/Non-Autonomous Retroposon Pairs in Trypanosomatids

The genomes of higher eukaryotes contain pairs of autonomous/non-autonomous retroposons composed of small noncoding elements, which use for their own mobility the retrotransposition machinery encoded by autonomous elements (for review see [48]). This is exemplified by the retroposon pairs described in human (LINE1/Alu, LINE2/MIR, and LINE2/Ther-1) [24,34,37], fish (UnaL2/UnaSINE1) [36], reptiles (CR1-like LINE/SINE) [49], and plants (Bali1/S1) [35]. In these examples, the small noncoding elements, called small interspersed elements (SINEs), are tRNA-, 5S RNA-, or 7SL RNA-related sequences [50–52]. In contrast, the small noncoding partners (RIME and NARTc) of the trypanosome *ingi*RIME (*T. brucei*) and L1Tc/NARTc (*T. cruzi*) pairs are derived from the autonomous retroposons (*ingi* and L1Tc) by deletion of the coding sequence [27–31]. The truncated RIME and NARTc elements became fixed in the trypanosome genome with copy numbers equivalent to that of the autonomous *ingi* and L1Tc retroposons (see Table 1) [32,33]. In addition, all trypanosomatid genomes analyzed so far contain degenerated retroposons related to *ingi* and L1Tc (DIRE) [38]. The majority of *LmDIRE* sequences identified in the *L. major* genome (90%) overlaps with a subset of *LmSIDER2*, suggesting that the latter are derived from the former by deletion. This indicates the existence of an *LmDIRE*/*LmSIDER2* pair comparable to the trypanosome *ingi*RIME and L1Tc/NARTc pairs.

The trypanosome *ingi*RIME and L1Tc/NARTc pairs are considered active, since the very low level of sequence divergence observed is consistent with recent retrotransposition activities. The *T. brucei* and *T. cruzi* genomes contain several potentially active *ingi*/L1Tc, which encode a single long and conserved protein [3,32,33]. This contrasts with the *L. major* *LmDIRE*/*LmSIDER2* pair, which has lost its retro-

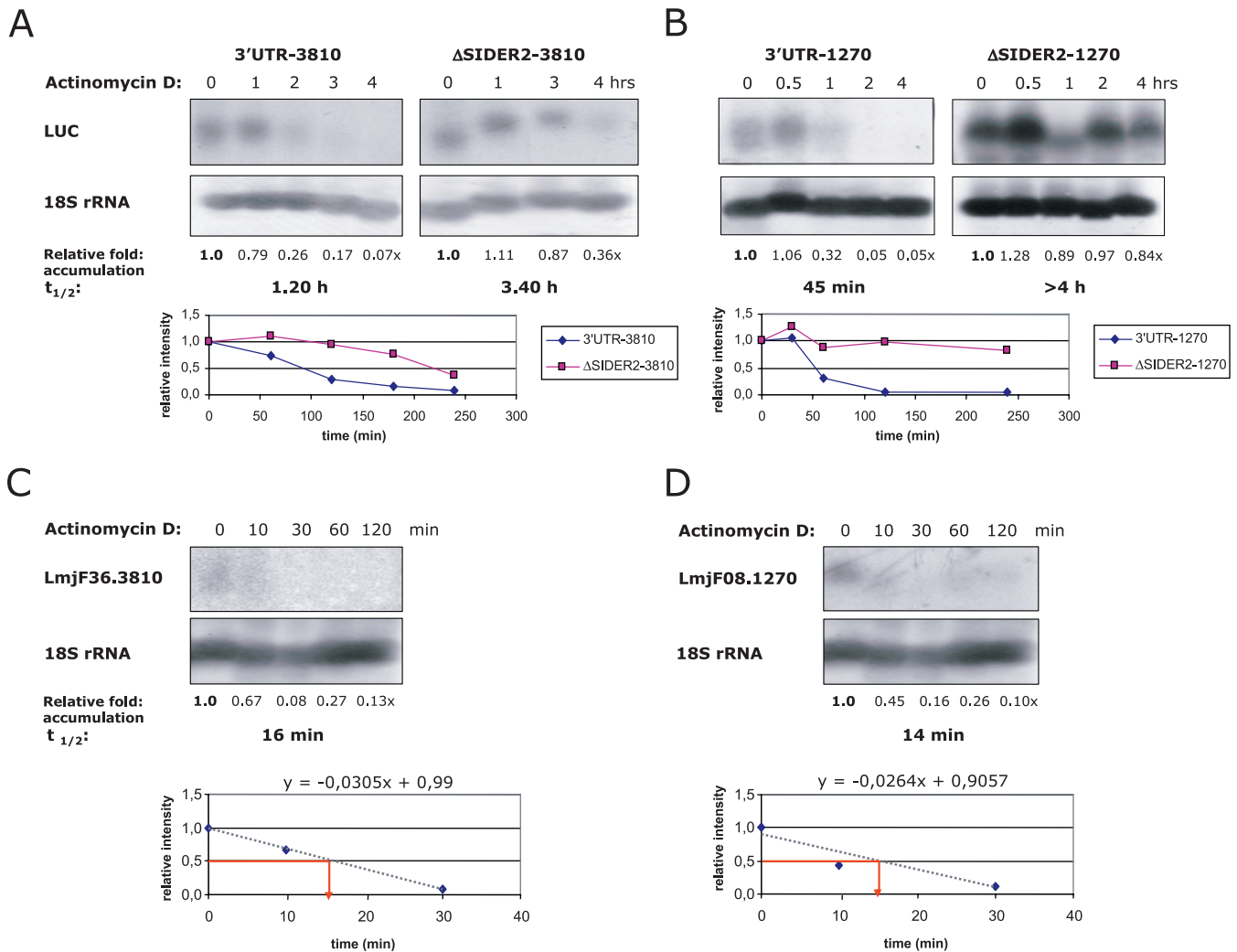


Figure 11. LmSIDER2 Is Involved in mRNA Destabilization

(A and B) Comparison of mRNA half-lives ($t_{1/2}$) in *L. major* promastigotes between LmSIDER2-containing LUC chimeric mRNAs, 3'UTR-3810 (A) and 3'UTR-1270 (B), and LUC mRNAs lacking the LmSIDER2 element, Δ SIDER2-3810 (A) and Δ SIDER2-1270 (B). Approximately 10^7 *L. major* recombinant promastigotes/ml were treated with 10 μ g/ml of actinomycin D, an inhibitor of de novo transcription, and RNA was isolated at the time points shown and analyzed by northern blotting. Transcript levels were normalized with respect to the amount of the rRNA loaded using the 18S rRNA probe as an internal control. The levels of mRNAs were assessed using phosphorimaging. The values shown below the blots represent the LUC mRNA fold accumulation with respect to its abundance prior to the addition of actinomycin D (0). This is a representative experiment out of two that showed very similar results.

(C and D) The decay of the endogenous LmjF36.3810 (C) and LmjF08.1270 (D) transcripts was also assessed in *L. major* promastigotes using actinomycin D (using northern blotting). The half-lives of LmSIDER2-containing LUC chimeric mRNAs and LmjF36.3810 and LmjF08.1270 endogenous transcripts were estimated based on hybridization intensities with gene-specific probes normalized with the 18S rRNA probe with respect to different time points of actinomycin D treatment. Experiments shown here are representative of two that showed very similar results.

doi:10.1371/journal.ppat.0030136.g011

transposition activity. Indeed, the only reverse transcriptase domains identified in the completed *L. major* genome belong to LmDIREs, which have accumulated numerous point mutations after their extinction [38]. In the absence of functional retroposons, the noncoding LmSIDER families can be considered extinct as well, since their members need enzymes produced in *trans* by autonomous retroposons for their mobilization. Consequently, the LmSIDER and LmDIRE families probably became extinct simultaneously, when the last active LmDIRE disappeared from the *L. major* genome, as proposed for the extinct human LINE2/MIR and rodent LINE1/B1 pairs [24,53]. The simultaneous extinction of the human autonomous LINE2 and non-autonomous MIR retro-

posons is illustrated by their similar nucleotide substitution level [24]. However, this comparative analysis cannot be done for the LmDIRE/LmSIDER pair because of the inappropriately low number of LmDIRE sequences available for such a statistical analysis [38]. The high level of divergence (12%) between the consensus and the most conserved LmSIDER2 sequence suggests that LmSIDER became extinct a long time ago. The rise and fall of TE families has been well documented in several genomes [19,24]. For example, it was estimated that the human LINE2 retroposons, which show at least 18% divergence with the consensus LINE2 sequence, became extinct 50–100 million years ago [19]. In the absence of trypanosomatid fossil records and thus of a molecular

clock, the date of LmSIDER extinction cannot be estimated with accuracy. It probably occurred after the speciation of the *Trypanosoma* and *Leishmania* genus 200–500 million years ago [54], since trypanosomes still contain putative active elements [38].

Exaptation of LmSIDERs by *L. major* and Their Role in Modulating Gene Expression

The recent completion and comparative analysis of eukaryotic genomes provides evidence that several superfamilies of short non-autonomous retroposons (e.g., SINE) have been conserved and distributed among a wide range of species [55–57]. These conservations suggest that numerous extinct retroposons were domesticated hundreds of million years ago and are still functional in several species. While superfamilies of retroposons are conserved and shown to be functional, exaptation of a TE family to the extent described here has not been reported so far. Here, we provide evidence that *Leishmania* spp. have recycled a whole family of short retroposons (LmSIDER2), which have evolved to fulfill important biological pathways such as the regulation of gene expression, whereas its close relative *T. brucei* developed other approaches to maintain similar cellular functions. Retroposon-mediated regulation at transcriptional or post-transcriptional levels [23,48,58–61] remains a relatively rare event in other eukaryotes and is not thought to be an intrinsic function of retroposons. Most LmSIDERs (95.4%) are located within intergenic regions of DGCs, mainly in 3'UTRs, while 95.5% of the TbSIDERs are located outside DGCs; the retroposon density in DGCs being ~50 times higher in *L. major* than *T. brucei*. This contrasting SIDER distribution can also be correlated with the difference in the average size of intergenic regions between *L. major* and *T. brucei* (1,432 bp versus 721 bp) [40], in part due to the presence of LmSIDERs in the 3'UTRs.

We have previously identified a conserved 450–550-bp element located in the 3'UTR of several *Leishmania* amastigote-specific transcripts that is implicated in stage-specific translational control [16,17]. Interestingly, this element belongs to the LmSIDER1 subfamily of retroposons, which comprises at least 785 sequences across the *Leishmania* genome (A. Rochette, M. Smith, P. Padmanbhan, B. Papadopoulou, unpublished data). In this study, we presented several lines of evidence showing that LmSIDER2 promotes mRNA destabilization. This conclusion stems from a comprehensive microarray analysis, from northern blotting data, and from a more direct reporter gene analysis of selected mRNAs. The functional distinction between LmSIDER1 and LmSIDER2 is consistent with the way they clustered in a phylogenetic tree.

The ability of LmSIDER2 to destabilize mRNA seems to be intrinsic and context independent, since it can be functional at different distances from the poly(A) tail and even outside the context of the endogenous 3'UTRs (see Figure 10). LmSIDER2-containing mRNAs are generally expressed at lower levels compared to non-SIDER2-bearing transcripts and are short-lived (half-lives of ~15 min). Taken together, these observations suggest that LmSIDER2 are *cis*-acting components of a regulatory pathway that generally down-regulates gene expression to ensure rapid turnover of a specific subset of *Leishmania* mRNAs. Throughout its complex life cycle, *Leishmania* is subjected to a variety of rapidly changing environmental conditions, and rapid mRNA turn-

over can permit the parasite to adapt its pattern of protein synthesis to continuously changing physiological needs. We hypothesize that the mRNA-destabilizing function of LmSIDER2 can be enhanced or blocked as needed due to their particular sequence or structure (LmSIDER2 elements are highly heterogeneous), and/or the presence of other elements in the 3'UTR of *Leishmania* transcripts. This is in agreement with our preliminary results in *L. infantum* amastigotes, where the 36.3810 SIDER2 becomes inactive due to the presence of a downstream element (M. Müller, B. Papadopoulou, unpublished data), and with the observation that none of the highly expressed housekeeping genes harbor LmSIDER2 (unpublished data).

In the case of LmjF36.3810 and LmjF08.1270 transcripts, which are both constitutively expressed in *L. major*, the LmSIDER2 destabilizing element works as efficiently in amastigotes as it does in promastigotes (unpublished data). Our microarray data on 38 LmSIDER2-containing transcripts are also consistent with these observations. However, other stages of the parasite, irrespective of whether they are morphologically distinct (e.g., metacyclics) or not, exist where the function of these elements might be more crucial. Indeed, we also found that several transcripts reported to be upregulated in the metacyclic stage of *L. major* [62] contain LmSIDER2 (unpublished data). Likewise, the role of these elements might be more evident as the parasite experiences a specific environmental challenge, particularly in the rather dynamic ecological niche inside its insect host. Indeed, a number of short-lived mRNAs are known to be responsive to specific extracellular environmental stimuli in other systems where expression is regulated by sequences in 3'UTRs (e.g., the AU-rich elements of inflammatory cytokines and growth factors) [63,64]. Alternatively, the role of LmSIDER2 might be to negatively modulate gene expression and thereby check that mRNAs, stage-specific or constitutively expressed, are maintained at nontoxic levels (for instance, mRNAs encoding structural proteins are generally expected to be more abundant than those encoding regulatory proteins).

Comparison of the *L. major* and *T. brucei* genomes showed that SIDERs are ~70 times more abundant in *L. major* compared to *T. brucei* [38]. Considering that the majority of LmSIDERs is co-transcribed with coding genes and that members of the LmSIDER families are shown to play a role in the regulation of gene expression, whereas most of the very few TbSIDERs are distributed in the relatively silent subtelomeric regions, it is tempting to propose that *Leishmania*, but not trypanosomes, have exapted and expanded the SIDER retroposons. The reasons behind this extraordinary LmSIDER expansion are currently unknown. The widespread genomic distribution of LmSIDER2 and our functional data on both LmSIDER1 and LmSIDER2 members raises the interesting possibility that numerous *Leishmania* transcripts encoding a wide repertoire of functionally diverse proteins may be regulated by a similar mechanism in response to specific environmental stimuli and/or growth conditions. The involvement of TE in coordinated expression of genes was already proposed in the seventies [65].

We propose that *Leishmania*, an organism with no known control at the level of transcription initiation, has acquired the ability to post-transcriptionally coordinate gene regulation via short retroposons (LmSIDERs) in the 3'UTR. This is consistent with the prevailing notion that retroelements

likely emerged as genomic parasites and gradually invaded the genomes of most eukaryotic cells, but later became an integral part of their genome and were used for the benefit of these organisms.

Materials and Methods

Identification of LmSIDERs. A BLASTN search of the *L. major* genome with the first 79 residues of the *T. brucei ingi*/RIME (“79 bp signatures”) revealed 108 homologous sequences, corresponding to the 5′-extremity of degenerated retroposons, subsequently called LmSIDERs. A multiple alignment (ClustalW [66]) of the sequences located downstream from these 108 “79 bp signatures” (1 kb) was then done to define six groups of related but very heterogeneous sequences, ranging from 450 to 790 bp in length. The 3′-extremity of most of these relatively conserved sequences was composed of an adenosine-rich stretch, as generally observed for retroposons. In order to identify other LmSIDER in the *L. major* genome, a second BLASTN search was performed with one representative from each group of sequences. About 1,500 matches were retained. A third BLASTN search conducted with a subset of very divergent LmSIDER identified new sequences. Some of these newly identified LmSIDER were used for a fourth BLASTN search. We stopped this reiterative BLASTN search approach after two additional runs, since no more sequences were detected, with a total of 1,858 identified LmSIDER elements. The BLASTN analysis also revealed that LmSIDER could be separated into two groups composed of 785 (LmSIDER1) and 1,073 (LmSIDER2) sequences (see Figure 2).

Identification of TbSIDERs. The first 79 residues of the *T. brucei ingi*/RIME (“79 bp signatures”) were used to perform a BLASTN search of the *T. brucei* genome database (version 3.0 of The Institute for Genomic Research’s [TIGR] *T. brucei* assembly). For this BLAST analysis, the annotated RIME, *ingi*, and DIRE sequences were masked using Repeat Masker (<http://www.repeatmasker.org/>). A multiple sequence alignment (ClustalW [66]) of the regions located downstream of 51 identified “79 bp signatures” (1 kb) and defined two groups of related sequences, named TbSIDER1 (ten sequences) and TbSIDER2 (12 sequences), while the other 29 sequences were unique and appeared not to be related to retroposons.

Multiple alignments and phylogenetic analysis of LmSIDER sequences. We used ClustalW (<http://www.ebi.ac.uk/tools/clustalw/>), MUSCLE (<http://www.drive5.com/muscle/>), and 3DCoffee (<http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee.cgi/index.cgi>) programs to perform a multiple sequence alignment of all (1,073 sequences) or different subsets of (from 50 sequences) LmSIDER2. None of these attempts produced in and of themselves a satisfactory alignment, probably because of the high degree of divergence and size polymorphism. The MUSCLE program produced a workable alignment from a selection of 50 full-length and relatively closely related LmSIDER2 sequences. This multiple alignment was manually refined to generate a framework used to manually align, one by one, the LmSIDER2 sequences. The final alignment contained 1,013 LmSIDER2 sequences (Figure S1) The LmSIDER2 core sequence was generated by deleting all positions showing a gap for at least 50% of the aligned sequences, which represents 66.6% of the positions (1,074 positions out of 1,612 in the original alignment) (Figure S2). The statistical and comparative analyses were performed using this LmSIDER2 core sequence.

LmSIDERs were extracted from the most recent *L. major* genome annotation (<http://www.genedb.org/>) for phylogenetic analysis. We extracted SIDER (formerly named LmRIME) sequence regions between 400 and 700 nucleotides long using Artemis [67]. An automated multiple sequence alignment was generated by comparing individual sequences to a Hidden Markov Model (HMM) using HMMER 1.8.5 (<http://hmmer.janelia.org/>). The HMM profile used to align the LmSIDERs was generated using 15 representative sequences selected from the manual alignment shown in Figure S1 (24.0477, 36.1076, 29.0524, 31.0641, 33.0760, 36.1128, 36.1087, 35.1074, 34.0878, 31.0653, 25.0573, 38.0225, 34.0863, 14.0386, 28.0581). Limiting the amount of sequences in the profile minimizes position-specific base composition bias. To facilitate visualization of the subsequent tree, we removed additional LmSIDERs displaying >95% identity to at least one other aligned sequence using an ad-hoc JAVA script (<http://java.sun.com/>). The final alignment contains 785 LmSIDER sequences (140 LmSIDER1 and 645 LmSIDER2). The 785 resulting LmSIDERs were submitted to a Minimum Evolution phylogenetic analysis based upon the number of differences using the MEGA3 program [68]. Furthermore, only parsimonious informative sites were considered.

The phylogenetic tree was displayed using HyperTree JAVA program [69].

Divergence between members of retroposon families. TbSIDER1 (ten), TbSIDER2 (12), RIME (70), and NARTc (115) sequences were separately aligned using ClustalW (<http://www.ebi.ac.uk/tools/clustalw/>), whereas LmSIDER2 sequences (1,013) were manually aligned as described above. The core sequences, deduced from these alignments, were defined as described above for the LmSIDER2 core sequence. 21, 2, 44, and 19 positions were removed from the original TbSIDER2, TbSIDER1, RIME, and NARTc alignments, which corresponds to 4%, 0.4%, 8.1%, and 6.7% of the positions, respectively. The core consensus sequences were reconstituted by considering the most conserved residue at each position of the alignment. Then, the percentage of substitution from the consensus was determined for each sequence aligned by calculating the sequence identity of each sequence with the consensus. The consensus sequence was created with BioEdit (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>) using a threshold frequency for inclusion of 26%. Gaps were treated like residues.

Statistical analysis. To quantify the degree of conservation at each column in the core sequence multialignments, a chi-square (χ^2) score was computed comparing the observed distribution of ACGTs in the column to the distribution in the entire genome. The background ACGT distribution for the genome was obtained by counting the occurrences of each base in the set of all assembled chromosomes. Then, in each of the four multiple-alignments at each column, the chi-square score was computed as

$$\chi^2 = \sum_{i \in \{A,C,G,T\}} \frac{(o_i - e_i)^2}{e_i} \quad (1)$$

where o_i is the observed number of occurrences of character i in the given column, and e_i is the expected number of occurrences of character i computed as the proportion of character i in all assemblies multiplied by the number of sequences in that column of the multialignment. Using three degrees of freedom, a χ^2 value of 16.3 corresponds to a significance level of $p < 0.001$.

Determination of mRNA processing sites. The chromosomes and genomic coordinates of all *L. major* coding sequences were retrieved from version 4.0 of the assembly and annotation database hosted at TIGR. Using the predictive algorithm developed by Benz et al. [9], we scanned all *L. major* chromosomes to locate the putative polypyrimidine tract and splice acceptor and polyadenylation sites for each gene, thus delimiting the coordinates of the putative 5′UTR and 3′UTR. We selected the splice acceptor signal nearest to the start codon. This choice was based on what was observed in *T. brucei*, where EST mapping validated that 66% of the genes primarily used the closest site [9]. The distance between each LmSIDER element and its closest downstream and upstream gene on each chromosome strand was computed, disregarding the strand on which the element was located. The distance between each LmSIDER and the first methionine codon of the nearest downstream gene was calculated to determine a list of LmSIDERs that overlapped with the in silico-predicted 3′UTRs or 5′UTRs. Then, the distance between 3′UTR overlapping LmSIDERs and polypyrimidine and polyadenylation sites of the overlapping gene was calculated.

Leishmania culture. The *L. major* LV39 strain used in this study was described previously [70]. Promastigotes were cultured at pH 7.0 and 25 °C in SDM-79 medium supplemented with 10% heat-inactivated FCS (Wisent, <http://www.wisent.ca/>) and 5 µg/ml hemin. Intracellular *L. major* amastigotes were isolated from footpad lesions of infected BALB/c mice as previously described [71].

Plasmid construction and transfections. The expression vector pSPYNEO α LUC was described previously [16] and is referred to as LUC-control in the present study. The LUC-chimeric mRNAs transcribed from this vector are processed in *Leishmania* using sequences within the alpha-tubulin intergenic region cloned at the 5′-end. The different LUC-chimeric constructs listed in Figure 10 were made as follows. The full-length 3′UTR of LmjF36.3810 and LmjF08.1270 transcripts from the termination codon to 434 bp beyond the poly(A) site in the case of LmjF36.3810, and to 84 bp beyond the poly(A) site in the case of LmjF08.1270, or the LmSIDER2 element or the 3′UTR lacking LmSIDER2, were amplified by PCR using Taq DNA polymerase (Qiagen, <http://www.qiagen.com/>) and primers with inserted BamHI or PstI restriction sites (see Table S3). PCR products were cloned into vector pCR2.1 (Invitrogen, <http://www.invitrogen.com/>), digested with BamHI or PstI (New England Biolabs, <http://www.neb.com/>) and subcloned into the BamHI site downstream of the LUC gene in vector pSPYNEO α LUC [16]. All constructs have been verified by sequencing. Purified plasmid vector DNA (10–20 µg, Qiagen) were

transfected into *Leishmania* by electroporation as described previously [72]. Stable transfectants were selected with 0.04 mg/ml G-418 (Sigma, <http://www.sigmaaldrich.com>).

LUC assay. The LUC activity of the recombinant parasites was determined as described previously [17]. Briefly, mid-log-phase promastigotes were diluted 1:100 in SDM-79 supplemented with 10% glycerol and counted in a Neubauer counting chamber. Equivalents of 4×10^7 and 2×10^7 parasites were spun, the pellet resuspended in 5× luciferase lysis (Promega, <http://www.promega.com>) buffer and frozen at -80°C . Twenty μl of each lysate was then mixed with an assay buffer (Promega) containing D-luciferin potassium salt, and LUC activity was measured in a luminometer (Dynex MLX, <http://www.dynextechnologies.com>).

RNA and protein manipulations. Total RNA of *L. major* promastigotes was isolated using the TRIzol reagent (Gibco BRL, <http://www.invitrogen.com>) following manufacturer instructions. Northern blot hybridizations were performed following standard procedures [73]. To prepare soluble protein lysates, *Leishmania* cells were harvested by centrifugation, washed with ice-cold phosphate-buffered saline (PBS), resuspended in Laemmli buffer, and syringed with a microsyringe (ten times). Proteins were quantified using Amido Black 10B (Bio-Rad, <http://www.bio-rad.com>), and 50 μg of total protein extracts were loaded onto 10% SDS-PAGE gels. The gels were transferred on a polyvinylidene difluoride membrane (Immobilon-P; Millipore, <http://www.millipore.com>) and the membranes were incubated for 90 min in blocking buffer (PBS with 0.1% Tween 20 and 5% nonfat dry milk). The first antibody, a goat anti-luciferase pAb (Promega) diluted 1:10,000 in blocking buffer, was incubated with the membrane for 90 min with agitation. Following three washes with PBST (PBS supplemented with 0.1% Tween 20), a second antibody, a donkey anti-goat (Santa Cruz Biotechnology, <http://www.scbt.com>) diluted 1:10,000 in blocking buffer, was incubated for 45 min with the membrane. After additional washes, the blot was visualized by chemiluminescence with a Renaissance kit (New Life Science Products, <http://las.perkinelmer.com>). RNA and protein levels were estimated by densitometric analyses using a PhosphorImager with ImageQuant 5.2 software.

RNA stability assays. To determine the half-life of LmSIDER2-containing transcripts, mid-log phase *L. major* promastigote cultures were incubated with 10 $\mu\text{g}/\text{mL}$ of actinomycin D (Sigma), an inhibitor of de novo transcription. At specific times post-addition of the drug, 10-ml culture aliquots were pelleted by centrifugation, washed once with Hepes-NaCl buffer, and lysed in 1 ml TRIzol reagent (Gibco BRL). Total RNA was extracted from these samples and subjected to northern blot hybridization. Quantitation of the different transcripts was done by densitometric analysis using a PhosphorImager with the ImageQuant 5.2 software.

DNA microarray analysis and quantitative real-time RT-PCR. Thirty-eight *L. major* genes predicted to harbor LmSIDER2 in their 3'UTR were chosen for DNA microarray analysis, as part of a previously described 70-mer oligonucleotide array comprising a total of 154 selected genes [47]. Total RNA from *L. major* promastigotes and lesion amastigotes isolated from infected BALB/c mice was prepared using the TRIzol reagent (Gibco BRL) and purified using the RNeasy kit (Qiagen). Quality and quantity of the RNA was assessed by RNA 6000 Nano Assay Chips (Agilent Technologies, <http://www.home.agilent.com>) and a Bioanalyzer (Agilent Technologies). Probes for microarray hybridization were prepared using the indirect Micromax TSA labeling and detection kit (Perkin Elmer, <http://las.perkinelmer.com>). For each labeling reaction, 2 μg of purified RNA was spiked with two exogenous mRNAs (*NAC1* and *CAB1* from *Arabidopsis thaliana* at 2.5 $\text{pg}/\mu\text{l}$; Stratagene, <http://www.stratagene.com>) to adjust for variations in the incorporation efficiency of the modified nucleotides and differences in first-strand cDNA synthesis reactions. Hybridization, washes, and detection of fluorescence were done as described previously [47]. Four independent microarray experiments including dye swapping were scanned, and signal intensities for each spot were exported into GeneSpring software (Agilent) for further analysis. Local background was subtracted from each spot on the array, and intensity-dependent normalization was carried out within arrays. Cy5/Cy3 ratio for each spot was normalized with Cy5/Cy3 ratio for the *A. thaliana* *NAC1* spike. Genes were only considered as statistically different in their expression if they satisfied a *p*-value cutoff of 0.05. Expression ratios of three LmSIDER2-containing genes (LmjF31.1890, LmjF33.2550, LmjF08.1270) and one non-LmSIDER2 gene (LmjF16.1430) were confirmed by quantitative real-time RT-PCR as described previously [47]. These ratios were normalized using the *GAPDH* ratio to give a fold difference of expression. To exclude eventual amplification of mouse transcripts, cDNA from mouse macrophages served as negative control in each experiment.

Supporting Information

Figure S1. Alignment of 1,013 LmSIDER2

The alignment, saved under the Philip format, was performed as described in Materials and Methods with the introduction of gaps (-) to maximize the alignments. The LmSIDER names indicate the chromosomal localization followed by the model number.

Found at doi:10.1371/journal.ppat.0030136.sg001 (1.6 MB DOC).

Figure S2. Alignment of the Core Sequence of 1,013 LmSIDER2

This alignment was generated from the one presented in Figure S1 by deleting all positions showing a gap for at least 50% of the aligned sequences.

Found at doi:10.1371/journal.ppat.0030136.sg002 (553 KB DOC).

Figure S3. Distribution of Genes and Retroposons on the 36 *L. major* Chromosomes

The central scale bars showing the size of the chromosomes (kb) separate features located on different strands. The position of protein-encoding genes and retroposons is indicated by vertical bars with the color code shown on the right margin. Protein-encoding genes and DIREs are shown on both central panels, while the upper or lower part of the schematic chromosomes display the position of LmSIDER1 and LmSIDER2.

Found at doi:10.1371/journal.ppat.0030136.sg003 (888 KB DOC).

Figure S4. Distribution of Genes and Retroposons on the 11 *T. brucei* Megachromosomes

The central scale bars showing the size of the chromosomes (kb) separate features located on different strands. The position of protein-encoding genes and retroposons is indicated by vertical bars with the color code shown in the right margin. Protein-encoding genes and *ingi* and DIRE retroposons are shown in both central panels, while the upper or lower part of the schematic chromosomes indicate the position of RIME and TbSIDER retroposons.

Found at doi:10.1371/journal.ppat.0030136.sg004 (856 KB DOC).

Table S1. LmSIDER Sequences Annotated in the *L. major* Genome

The chromosome localization (“chr”), genomic coordinates (“start” and “end”), strand localization (“str”), family (“fam”), and name (“name”) of the annotated LmSIDERs are indicated. The first column (“ID”) shows the name of each LmSIDER annotated in the database (version 4.0 of the assembly) hosted at The Institute for Genomic Research. The last column (“chr_size”) indicates the size of the chromosomes.

Found at doi:10.1371/journal.ppat.0030136.st001 (194 KB PDF).

Table S2. Differential Gene Expression of *L. major* SIDER2-Containing Transcripts Analyzed by DNA Microarrays

Found at doi:10.1371/journal.ppat.0030136.st002 (77 KB PDF).

Table S3. Primers Used for the Generation of the LUC-Expressing Vectors

Found at doi:10.1371/journal.ppat.0030136.st003 (59 KB PDF).

Acknowledgments

We thank Al Delcher for performing the chi square analysis, Daniel Nilsson for providing us with the prediction algorithm previously developed for trypanosome mRNA processing sites [9], Jacques Nicolas and colleagues for useful discussions, and Marc Ouellette and Simon Haile for critical reading.

Author contributions. FB, GCC, MS, NMAES, and EG performed the bioinformatic analyses. MM, AR, and BP conceived and designed the biological experiments.

Funding. FB was supported by the Centre National de Recherche Scientifique (CNRS), the Conseil Régional d'Aquitaine, and the Ministère de l'Éducation Nationale de la Recherche et de la Technologie. This work was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP-12182) awarded to BP. MM is supported by a scholarship from Laval University and by the Deutscher Akademischer Austausch Dienst (DAAD). AR is a fellow of a Fonds de Recherche en Santé de Québec (FRSQ). BP is a Burroughs Wellcome Fund New Investigator in Molecular Parasitology and a member of a CIHR Group on Host-Pathogen Interactions and of a Fonds Québécois de la Recherche sur

la Nature et les Technologies (FQRNT) Center for Host-Parasite Interactions.

References

- Haag J, O'Huigin C, Overath P (1998) The molecular phylogeny of trypanosomes: Evidence for an early divergence of the Salivaria. *Mol Biochem Parasitol* 91: 37–49.
- Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309: 416–422.
- El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, et al. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309: 409–415.
- Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309: 436–442.
- Liang XH, Haritan A, Uliel S, Michaeli S (2003) *trans* and *cis* splicing in trypanosomatids: Mechanism, factors, and regulation. *Eukaryot Cell* 2: 830–840.
- Hug M, Hotz HR, Hartmann C, Clayton C (1994) Hierarchies of RNA-processing signals in a trypanosome surface antigen mRNA precursor. *Mol Cell Biol* 14: 7428–7435.
- Vassella E, Braun R, Roditi I (1994) Control of polyadenylation and alternative splicing of transcripts from adjacent genes in a procyclin expression site: A dual role for polypyrimidine tracts in trypanosomes? *Nucleic Acids Res* 22: 1359–1364.
- Schurch N, Hehl A, Vassella E, Braun R, Roditi I (1994) Accurate polyadenylation of procyclin mRNAs in *Trypanosoma brucei* is determined by pyrimidine-rich elements in the intergenic regions. *Mol Cell Biol* 14: 3668–3675.
- Benz C, Nilsson D, Andersson B, Clayton C, Guilbride DL (2005) Messenger RNA processing sites in *Trypanosoma brucei*. *Mol Biochem Parasitol* 143: 125–134.
- Clayton CE (2002) Life without transcriptional control? From fly to man and back again. *EMBO J* 21: 1881–1888.
- Argaman M, Aly R, Shapira M (1994) Expression of heat shock protein 83 in *Leishmania* is regulated post-transcriptionally. *Mol Biochem Parasitol* 64: 95–110.
- Quijada L, Soto M, Alonso C, Requena JM (2000) Identification of a putative regulatory element in the 3'-untranslated region that controls expression of HSP70 in *Leishmania infantum*. *Mol Biochem Parasitol* 110: 79–91.
- Charest H, Zhang WW, Matlashewski G (1996) The developmental expression of *Leishmania donovani* A2 amastigote-specific genes is post-transcriptionally mediated and involves elements located in the 3'-untranslated region. *J Biol Chem* 271: 17081–17090.
- Wu Y, El Fakhry Y, Sereno D, Tamar S, Papadopoulou B (2000) A new developmentally regulated gene family in *Leishmania amastigotes* encoding a homolog of amastin surface proteins. *Mol Biochem Parasitol* 110: 345–357.
- Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M (2001) Developmental regulation of heat shock protein 83 in *Leishmania*. 3' processing and mRNA stability control transcript abundance, and translation is directed by a determinant in the 3'-untranslated region. *J Biol Chem* 276: 47922–47929.
- Boucher N, Wu Y, Dumas C, Dube M, Sereno D, et al. (2002) A common mechanism of stage-regulated gene expression in *Leishmania* mediated by a conserved 3'-untranslated region element. *J Biol Chem* 277: 19511–19520.
- McNicoll F, Muller M, Cloutier S, Boilard N, Rochette A, et al. (2005) Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*. *J Biol Chem* 280: 35238–35246.
- Folgueira C, Quijada L, Soto M, Abanades DR, Alonso C, et al. (2005) The translational efficiencies of the two *Leishmania infantum* HSP70 mRNAs, differing in their 3'-untranslated regions, are affected by shifts in the temperature of growth through different mechanisms. *J Biol Chem* 280: 35172–35183.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274: 765–768.
- Orgel LE, Crick FH (1980) Selfish DNA: The ultimate parasite. *Nature* 284: 604–607.
- Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601–603.
- Biemont C, Vieira C (2006) Genetics: Junk DNA as an evolutionary force. *Nature* 443: 521–524.
- Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 9: 657–663.
- Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* 19: 68–72.
- Shapiro JA (2005) Retrotransposons and regulatory suites. *Bioessays* 27: 122–125.
- Hasan G, Turner MJ, Cordingley JS (1984) Complete nucleotide sequence of an unusual mobile element from *Trypanosoma brucei*. *Cell* 37: 333–341.
- Kimmel BE, Ole-Moiyoi OK, Young JR (1987) *ingi*, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINES. *Mol Cell Biol* 7: 1465–1475.
- Murphy NB, Pays A, Tebabi P, Coquelet H, Guyaux M, et al. (1987) *Trypanosoma brucei* repeated element with unusual structural and transcriptional properties. *J Mol Biol* 195: 855–871.
- Bringaud F, García-Pérez JL, Heras SR, Ghedin E, El-Sayed NM, et al. (2002) Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. *Mol Biochem Parasitol* 124: 73–78.
- Martin F, Maranon C, Olivares M, Alonso C, Lopez MC (1995) Characterization of a non-long terminal repeat retrotransposon cDNA (LITc) from *Trypanosoma cruzi*: Homology of the first ORF with the ape family of DNA repair enzymes. *J Mol Biol* 247: 49–59.
- Bringaud F, Biteau N, Zuiderwijk E, Berriman M, El-Sayed NM, et al. (2004) The *ingi* and RIME non-LTR retrotransposons are not randomly distributed in the genome of *Trypanosoma brucei*. *Mol Biol Evol* 21: 520–528.
- Bringaud F, Bartholomeu DC, Blandin G, Delcher A, Baltz T, et al. (2006) The *Trypanosoma cruzi* LITc and NARTc non-LTR retrotransposons show relative site-specificity for insertion. *Mol Biol Evol* 23: 411–420.
- Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* 94: 1872–1877.
- Tatout C, Lavie L, Deragon JM (1998) Similar target site selection occurs in integration of plant and mammalian retrotransposons. *J Mol Evol* 47: 463–470.
- Kajikawa M, Okada N (2002) LINES mobilize SINEs in the eel through a shared 3' sequence. *Cell* 111: 433–444.
- Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked *Alu* sequences. *Nat Genet* 35: 41–48.
- Bringaud F, Ghedin E, Blandin G, Bartholomeu DC, Caler E, et al. (2006) Evolution of non-LTR retrotransposons in the trypanosomatid genomes: *Leishmania major* has lost the active elements. *Mol Biochem Parasitol* 145: 158–170.
- Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell* 72: 595–605.
- El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309: 404–409.
- Agabian N (1990) Trans splicing of nuclear pre-mRNAs. *Cell* 61: 1157–1160.
- Aly R, Argaman M, Halman S, Shapira M (1994) A regulatory role for the 5' and 3' untranslated regions in differential expression of hsp83 in *Leishmania*. *Nucleic Acids Res* 22: 2922–2929.
- Brittingham A, Mosser DM (1996) Exploitation of the complement system by *Leishmania* promastigotes. *Parasitol Today* 12: 444–447.
- Larreta R, Soto M, Quijada L, Folgueira C, Abanades DR, et al. (2004) The expression of HSP83 genes in *Leishmania infantum* is affected by temperature and by stage-differentiation and is regulated at the levels of mRNA stability and translation. *BMC Mol Biol* 5: 3.
- Purdy JE, Donelson JE, Wilson ME (2005) Regulation of genes encoding the major surface protease of *Leishmania chagasi* via mRNA stability. *Mol Biochem Parasitol* 142: 88–97.
- Mishra KK, Holzer TR, Moore LL, LeBowitz JH (2003) A negative regulatory element controls mRNA abundance of the *Leishmania mexicana* Paraflagellar rod gene PFR2. *Eukaryot Cell* 2: 1009–1017.
- McNicoll F, Drummelsmith J, Muller M, Madore E, Boilard N, et al. (2006) A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. *Proteomics* 6: 3567–3581.
- Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. *Int Rev Cytol* 247: 165–221.
- Ohshima K, Hamada M, Terai Y, Okada N (1996) The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. *Mol Cell Biol* 16: 3756–3764.
- Ullu E, Tschudi C (1984) *Alu* sequences are processed 7SL RNA genes. *Nature* 312: 171–172.
- Daniels GR, Deininger PL (1985) Repeat sequence families derived from mammalian tRNA genes. *Nature* 317: 819–822.
- Kapitonov VV, Jurka J (2003) A novel class of SINE elements derived from 5S rRNA. *Mol Biol Evol* 20: 694–702.
- Rinehart TA, Grah RA, Wichman HA (2005) SINE extinction preceded LINE extinction in sigmodontine rodents: Implications for retrotranspositional dynamics and mechanisms. *Cytogenet Genome Res* 110: 416–425.
- Overath P, Haag J, Lischke A, O'Huigin C (2001) The surface structure of trypanosomes in relation to their molecular phylogeny. *Int J Parasitol* 31: 468–471.
- Gilbert N, Labuda D (1999) CORE-SINEs: Eukaryotic short interspersed

- retroposing elements with common sequence motifs. *Proc Natl Acad Sci U S A* 96: 2869–2874.
56. Ogiwara I, Miya M, Ohshima K, Okada N (2002) V-SINEs: A new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. *Genome Res* 12: 316–324.
 57. Nishihara H, Smit AF, Okada N (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* 16: 864–874.
 58. Britten RJ (1997) Mobile elements inserted in the distant past have taken on important functions. *Gene* 205: 177–182.
 59. Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 238: 115–134.
 60. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, et al. (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* 7: 597–606.
 61. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441: 87–90.
 62. Saxena A, Worthey EA, Yan S, Leland A, Stuart KD, et al. (2003) Evaluation of differential gene expression in *Leishmania major* Friedlin procyclics and metacyclics using DNA microarray analysis. *Mol Biochem Parasitol* 129: 103–114.
 63. Pieczyk M, Wax S, Beck AR, Kedersha N, Gupta M, et al. (2000) TIA-1 is a translational silencer that selectively regulates the expression of TNF- α . *EMBO J* 19: 4154–4163.
 64. Yaman I, Fernandez J, Sarkar B, Schneider RJ, Snider MD, et al. (2002) Nutritional control of mRNA stability is mediated by a conserved AU-rich element that binds the cytoplasmic shuttling protein HuR. *J Biol Chem* 277: 41539–41546.
 65. Britten RJ, Davidson EH (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* 46: 111–138.
 66. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25: 4876–4882.
 67. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: Sequence visualization and annotation. *Bioinformatics* 16: 944–945.
 68. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5: 150–163.
 69. Bingham J, Sudarsanam S (2000) Visualizing large hierarchical clusters in hyperbolic space. *Bioinformatics* 16: 660–661.
 70. Roy G, Kundig C, Olivier M, Papadopoulou B, Ouellette M (2001) Adaptation of *Leishmania* cells to in vitro culture results in a more efficient reduction and transport of biopterin. *Exp Parasitol* 97: 161–168.
 71. Muyombwe A, Olivier M, Harvie P, Bergeron MG, Ouellette M, et al. (1998) Protection against *Leishmania major* challenge infection in mice vaccinated with live recombinant parasites expressing a cytotoxic gene. *J Infect Dis* 177: 188–195.
 72. Papadopoulou B, Roy G, Ouellette M (1992) A novel antifolate resistance gene on the amplified H circle of *Leishmania*. *EMBO J* 11: 3601–3608.
 73. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: A laboratory manual*. 2nd edition. New York: Cold Spring Harbor Laboratory Press.