Identification of the regulatory elements controlling the transmission-stage specific gene expression of PAD1 in *Trypanosoma brucei.*

Paula MacGregor and Keith Matthews

Supplementary data

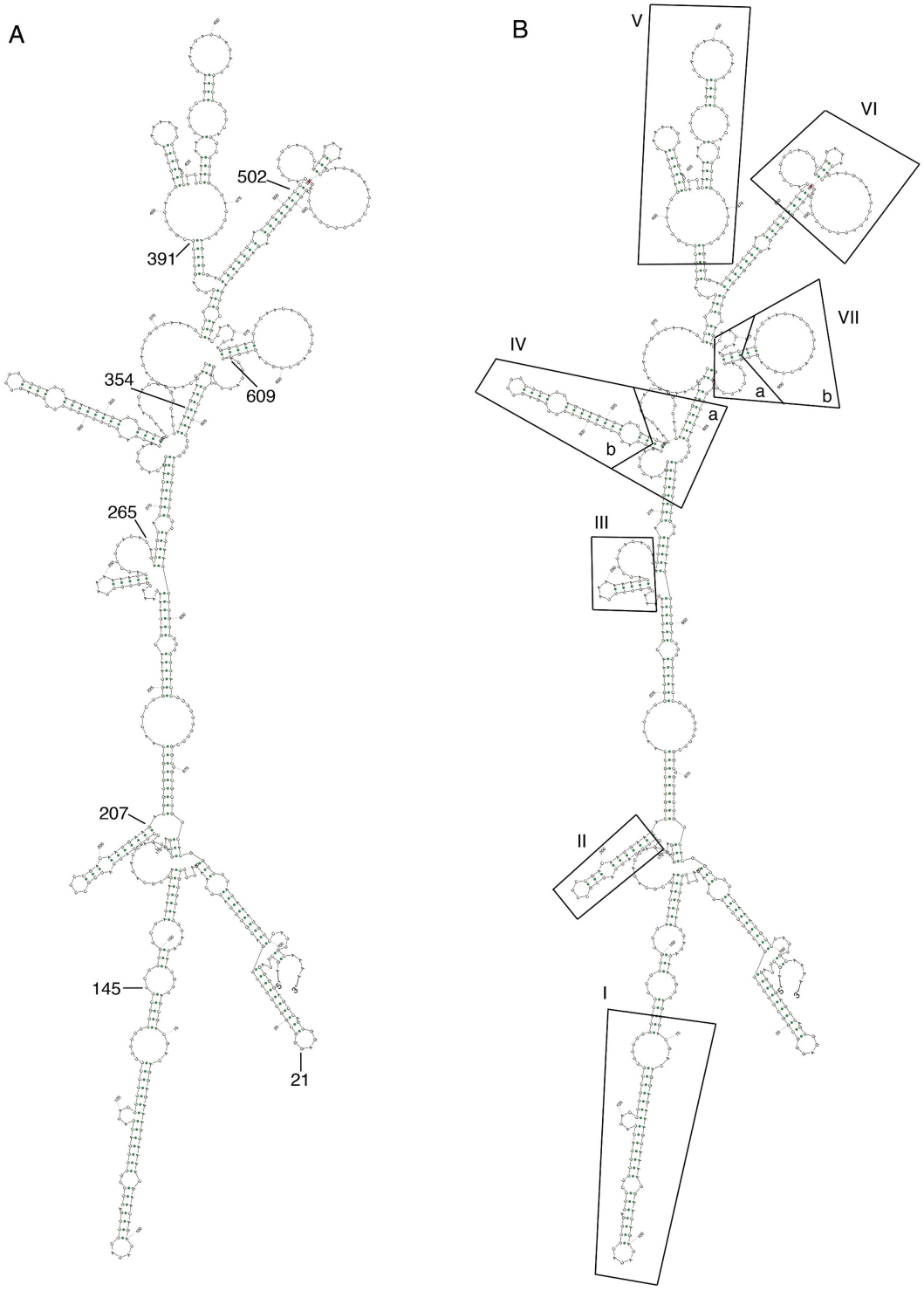
Supplementary Figure 1.

>Pad1-2
agcttaggggagccagtggagggcggcttcctgacgtttattcaacattttgtgtattta
tataatgttagcgacgacaggagaaagtaaaatgaacagatatactttgattccttttta
ctatactttccttttttttgcttctttgtctatacatatcgcataccattagtgagtt
207⇓
cctttattatcgcctccgacaatgaagacggtgcgccaactctggtaaacgataaataaa

cggttttaaaaaatcatctttacacatattgtacgagcagtccatgcattattattcttt
354⇓
ttttttgagttttctcatcagaagagatcgatacaacaacaccagtcactatatgaacac
accaacttatatcgttaaattcttcacttagctctcgtttatgatcgataaaagttcgaa
ttcacctcaaaaaatttccagaattaatattatattttctttccattctttgctattttc
502⇓
agagcgatctccttactctcaccCACTCTTCACCCGTGGAAACACCATGTAATCCTTTTC
TTCTTGCGGTGGGAGAGAGGAGAGTGCAGACCAACACAGGCACCCCAACACCGTTGATCC
GTAGTGCTACCTATCTCGTCTTCATATCTTTACTGCTTGTTATAATTTTGTCTTTTTTAC
CTTTTTTTTTCGGCTGCGCTGTCACGGTTTTATAAATGTCCATTGC**AAAA**TAGTTGCTCT
CAAGACACACATCAATAGCATTATATCCGCCCAATACCTCTTCTTTCATTCTCCCAAAAC
TTAATGCTTTCATTTCTCCCTTTTTGACTTATTGAACCTTTGAACAATACCAC**AG**ACTAA
CAGCAATAAACAAGCACGCAGTCACGTAAAGGAACCTAACATTTTAGGGAAACAAAAATT
ATCTGGCAGCACACTGGGCGAAACATCAGGAGGAATACAACTGAGCGTACTGTACAAAAC
ATTGCC

Annotated intergenic sequence between PAD1 and PAD2.
The polyadenylation site for PAD1 is indicated in red, the splice leader addition site for PAD2 (ag) is underlined. The black region represents the sequences unique to the PAD1-2 intergenic region. The green, cyan and blue regions represent sequences shared with other PAD gene family intergenic regions: green (PAD 5-6 and PAD 7-8 intergenic region), cyan (PAD1-2, PAD2-3, PAD5-6, PAD7-8 intergenic regions), blue (all PAD intergenic regions). Arrows indicate selected positions of interest, delimiting particular deletion boundaries. The sequence shown is derived from the *T. brucei* reference genome (TREU927/4). The gene code identifiers are: PAD1 (Tb927.7.5930), PAD2 (Tb927.5.5940).

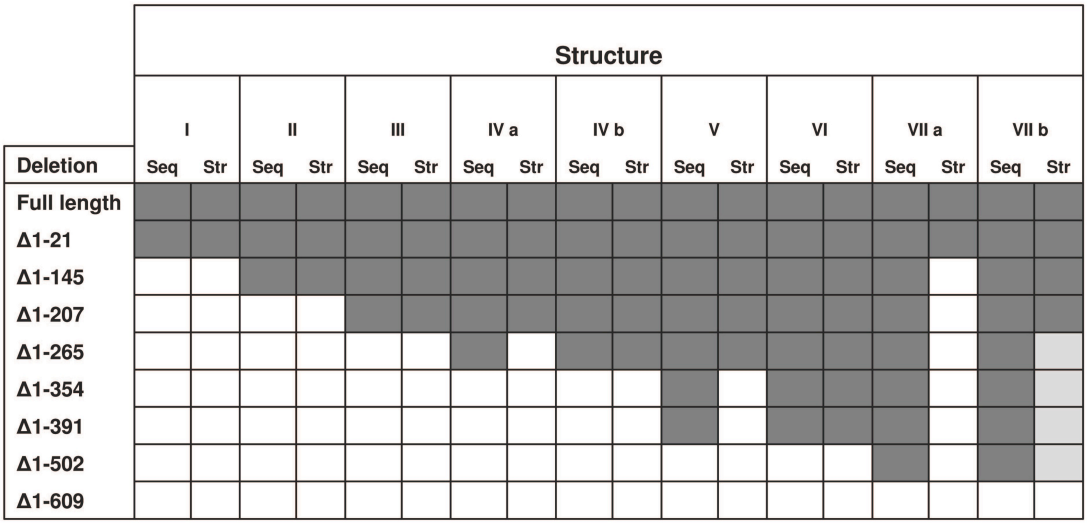
Supplementary Figure 2.



Structural features of the *PAD1* 3'UTR as predicted by Sfold. (A) The predicted secondary structure of the full length *PAD1* 3'UTR using position 710nt as the polyadenylation site. The nucleotide positions indicated correlate with the position of

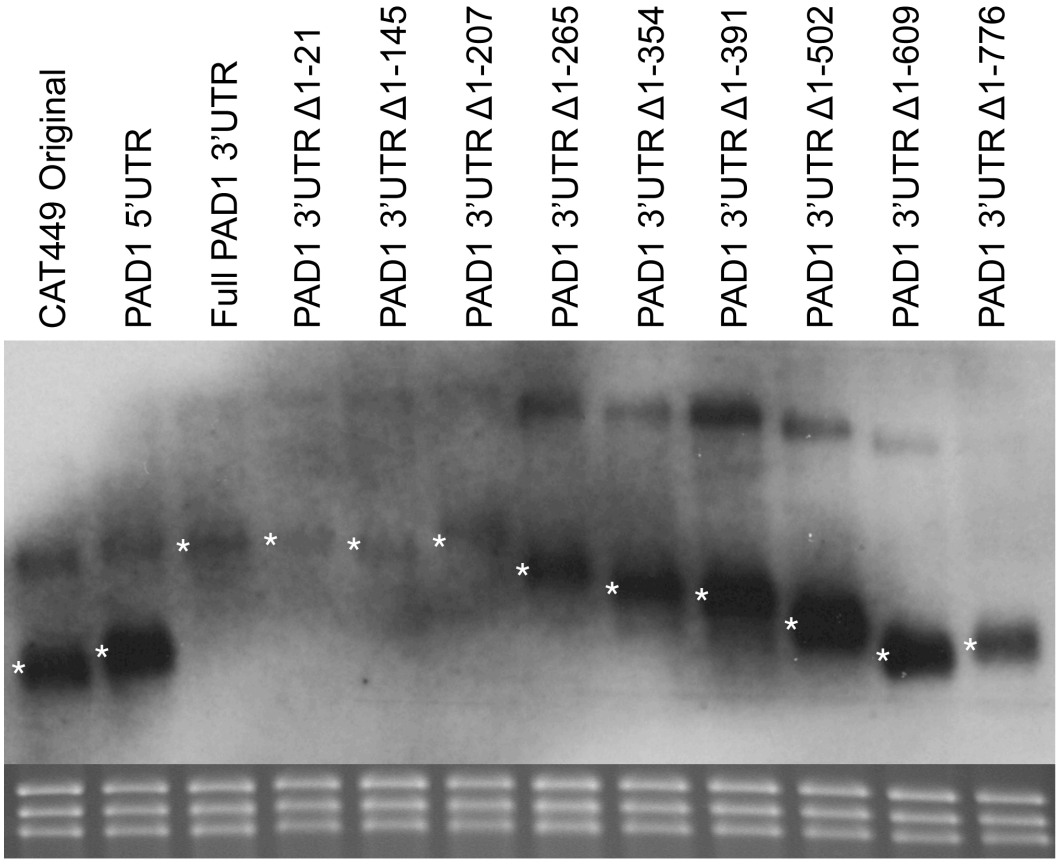
the deletions made in the *PAD1* 3'UTR deletion series. (B) There are various stem-loops and bulges in the *PAD1* 3'UTR, some of which have been indicated by numbered boxes. This is not an exhaustive classification of the observed structural features and do not relate to known functional features. The extent to which individual predicted structures are preserved in each deletion of the PAD1 3'UTR is summarised in Supplementary Figure 3. Sfold analysis was carried out using standard folding parameters (with a folding temperature of 37°C) and in each case the consensus Ensemble centroid structure is shown. For the full length PAD1 3'UTR structure, ΔG°_{37} = -145.86. (1-3)

Supplementary Figure 3



The features predicted to be present or absent in the *PAD1* 3'UTR deletion series. The presence (white squares) or absence (dark grey squares) of the sequence (Seq) and Sfold-predicted secondary structure (Str) of each of the structural features highlighted in Supplementary Figure 2, in each of the 3'UTR deletions, are indicated. For four of the deletion constructs structure VIIb was still present but in a somewhat disrupted form and this is indicated by light grey squares in the table.

Supplementary Figure 4



CAT mRNA abundance increases with sequential deletion of the *PAD1* 3'UTR. A representative northern blot of the CAT449 *PAD1* 3'UTR deletion series is shown, revealing that CAT mRNA abundance is down-regulated by the *PAD1* 3'UTR compared to the CAT449 original construct. This repression of mRNA expression is alleviated with progressive deletions. Upper bands are processing intermediates, often observed when using the CAT449 construct (4). The relevant band in each lane is marked with an asterisk. The predicted size of each band is: CAT449 Original = 862nt; PAD1 5'UTR = 914nt; Full PAD1 3'UTR = 1449nt; PAD1 3'UTR Δ1-21 = 1428nt; PAD1 3'UTR Δ1-145 = 1304nt; PAD1 3'UTR Δ1-207 = 1242nt; PAD1 3'UTR Δ1-265 = 1184nt; PAD1 3'UTR Δ1-354 = 1095nt; PAD1 3'UTR Δ1-391 = 1058nt; PAD1 3'UTR Δ1-502 = 947nt; PAD1 3'UTR Δ1-609 = 840nt. The polyadenylation site used in the PAD1 3'UTR Δ1-776 transcript is not known and as

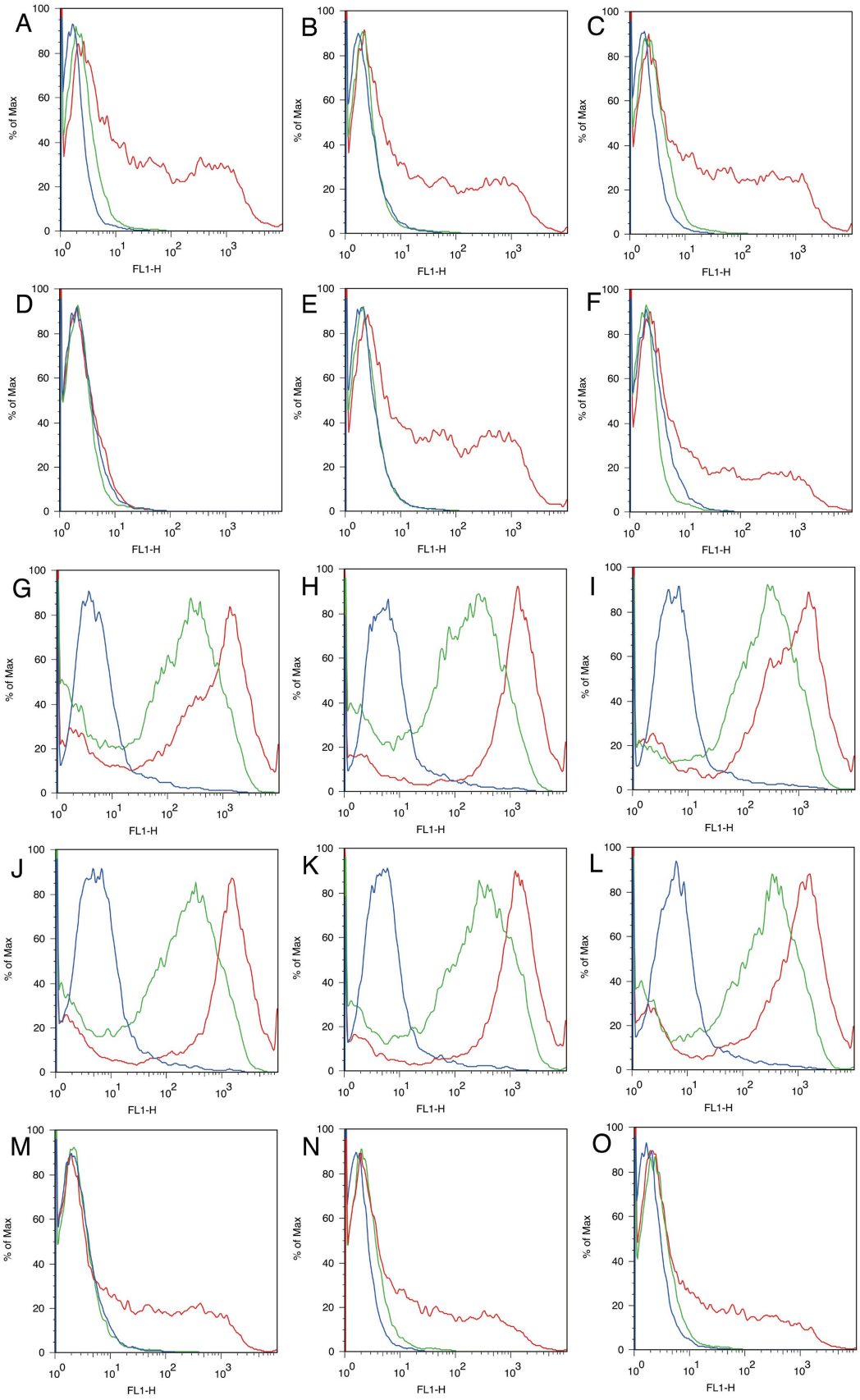
such the transcript length cannot be predicted accurately, however, it would be expected to be between 795nt and 956nt. For construct 3'UTR Δ1-207 the observed band ran higher than anticipated in each of two separate cell lines, perhaps reflecting use of an alternative polyadenylation site. Nonetheless the mRNA and protein abundance for this cell line was similar to its adjacent sequential deletions (see also Figure 2B, C).

Supplementary Figure 5

Predicted structures of the *PAD1* 3'UTR internal deletions. The predicted ensemble secondary structure of the full length *PAD1* 3'UTR (left hand structure) compared to the Δ386-564nt internal *PAD1* 3'UTR deletion (middle structure) and the Δ354-624nt internal *PAD1* 3'UTR deletion (right hand structure) using position 710nt as the polyadenylation site. Boxes on full-length structures indicate the structures that are deleted. Boxes on deletion constructs indicate approximate location of the deleted structures. Constructs were designed so as to minimise disruption to the remainder of the structure. Sfold analysis was carried out using standard folding parameters (with a folding temperature of 37°C) and in each case the consensus Ensemble structure is shown. For the full length PAD1 3'UTR structure, ΔG°_{37} = -

145.86; for the Δ354-624 structure, ΔG°_{37} = -97.99; for the Δ386-564 structure, ΔG°_{37} = -115.60.

Supplementary Figure 6



Flow cytometry profiles of triplicate EP Procyclin expression assays used to derive the EP expression values in Figure 5. After treatment with (A-C) dH_2O (D-F) DMSO (G-I) 100µM pCPT-cAMP (J-L) 10µM 8pCPT-2'-O-Me-cAMP or (M-O) 5µM Troglitazone, cells were induced to differentiate to procyclic forms by addition of 6mM cis aconitate. In all profiles blue represents 0 hours, green represents 6 hours and red represents 24 hours post-induction.

Supplementary Methods

Plasmid Constructions

The pHD617 vector was described (Reference 32) and a variant with the hygromycin resistance gene replaced with a puromycin resistance gene had been created previously. The pHD617 CAT-*PAD1* 3'UTR puroR vector for use in monomorphic bloodstream forms was produced by excising the Actin 3'UTR with BamHI and XhoI and replacing it with a PCR amplicon of the *PAD1* 3'UTR generated with oligonucleotides with added BamHI and XhoI restriction sites (5' TTA GGA TCC GCT TAG GGG AGC CAG TGG AGG GC 3' and 5' TTA CTC GAG GGC AAT GTT TTG TAC AGT ACG CTC AG 3'). The pHD617 CAT-*PAD1* 3'UTR puroR vector for use in pleomorphic bloodstream forms was modified further by replacing the existing Procyclin 5'UTR containing a tetracycline repressor binding site with one not containing this site by excising the original 5'UTR with KpnI and HindIII and inserting a PCR amplicon of the Procyclin 5'UTR generated with oligonucleotides with added KpnI and HindIII restriction sites (5' TAA GGT ACC GTC ATT GGG GTT AAG CGG AAA GGT G 3' and 5' TTA AAG CTT GTG AAT TTT ACT TTT TGG TGT AAT TGA AG 3'). The pHD617 GUS-Actin 3'UTR hygroR vector was produced by excising the CAT reporter gene ORF with HindIII and BamHI and replacing it with a PCR amplicon of the GUS ORF generated with oligonucleotides with added HindIII and BglII restriction sites (5' TTA AAG CTT ATG TTA CGT CCT GTA GAA ACC CCA AC 3' and 5' TTA AGA TCT TCA TTG TTT GCC TCC CTG CTG CGG 3').

The original CAT449 reporter construct was described in Reference 31. To create the CAT449 *PAD1* 5'UTR construct the Aldolase 5'UTR was excised with XhoI and

HindIII and replaced by a PCR amplicon of the *PAD1* 5'UTR with added XhoI and HindIII restriction sites (5' TTA CTC GAG AAA CAT GGA CAG TCA ACA TCT CCA TAT G 3' and 5' TTA AAG CTT GGC AAT GTT TTG TAC AGT ACG CTC AG 3'). To create the series of CAT449 vectors with variants of the *PAD1* 3'UTR, the truncated aldolase 3'UTR was excised from the vector by digesting with BamHI and BbsI, this was replaced with a full length or partially deleted *PAD1* 3'UTR PCR amplicon, generated with oligonucleotides with a 5' BamHI and 3'BsmAI site (Full length 5' TTA GGA TCC GCT TAG GGG AGC CAG TGG AGG GC 3', Δ1-21 5' TTA GGA TCC GGC GGC TTC CTG ACG TTT ATT CAA C 3', Δ1-175 5' TTA GGA TCC CTT TGT CTA TAC ATA TCG CAT ACC ATT AG 3', Δ1-207 5' TTA GGA TCC ACG GTG CGC CAA CTC TGG TAA ACG 3', Δ1-265 5' TTA GGA TCC ATA TTG TAC GAG CAG TCC ATG CAT TA 3', Δ1-354 5' TTA GGA TCC GAA CAC ACC AAC TTA TAT CGT TAA ATT C 3', Δ1-391 5' TTA GGA TCC CTC TCG TTT ATC GAT AAA AGT TCG 3', Δ1-502 5' TTA GGA TCC CAC TCT TCA CCC GTG GAA ACA CC 3', Δ1-609 5' TTA GGA TCC CCT ATC TCG TCT TCA TAT CTT TAC TGC 3', Δ1-776 5' TTA GGA TCC AAA CTT AAT GCT TTC ATT TCT CCC TTT TTG 3', All reverse 5' TAA TGC TTG AGA CGG CAA TGT TTT GTA CAG TAC GCT CAG 3').

To delete internal sequences of the *PAD1* 3'UTR the remaining 5' and 3' sections of the 3'UTR were PCR amplified with oligonucleotides with identical restriction sites at the 3' and 5' end, respectively (Δ386-564 (HindIII) 5' TAA AAG CTT TGA AGA ATT TAA CGA TAT AAG TTG GTG TG 3' and 5' TAA AAG CTT TGC AGA CCA ACA CAG GCA CCC CAA 3', Δ354-624 (BamHI) 5' TAA CTC GAG ATA TAG TGA CTG GTG TTG TTG TAT CG 3' and 5' TAA CTC GAG TAT CTT TAC

TGC TTG TTA TAA TTT TGT C 3'). These remaining sections were ligated together, PCR amplified (5' TTA GGA TCC GCT TAG GGG AGC CAG TGG AGG GC 3' and 5' TAA TGC TTG AGA CGG CAA TGT TTT GTA CAG TAC GCT CAG 3') and inserted into the CAT449 vector as described above. To insert sequences of the *PAD1* 3'UTR into the truncated aldolase 3'UTR the CAT449 vector was linearised with BamHI. The 1-354nt or 354-624nt region of the *PAD1* 3'UTR was PCR amplified using both 5' and 3' oligonucleotides with BamHI sites (Insert 1-354nt 5' TTA GGA TCC GCT TAG GGG AGC CAG TGG AGG GC 3' and 5' TTA GGA TCC ATA TAG TGA CTG GTG TTG TTG TAT CG 3', Insert 354-624nt 5' TTA GGA TCC TGA AGA CGA GAT AGG TAG CAC TAC G 3' and 5' TTA GGA TCC GAA CAC ACC AAC TTA TAT CGT TAA ATT C 3') and inserted into the linearised vector. All resulting constructs were sequenced to ensure correct insertion and orientation.

Supplementary References

1. Ding, Y., Chan, C.Y. and Lawrence, C.E. (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res*, **32**, W135-141.
2. Ding, Y., Chan, C.Y. and Lawrence, C.E. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna*, **11**, 1157-1166.
3. Ding, Y. and Lawrence, C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, **31**, 7280-7301.
4. Mayho, M., Fenn, K., Craddy, P., Crosthwaite, S. and Matthews, K. (2006) Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in Trypanosoma brucei: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements. *Nucleic Acids Res*, **34**, 5312-5324.