The EMBO Journal vol.10 no.12 pp.3877 - 3885, 1991

Requirement of a polypyrimidine tract for *trans*-splicing in trypanosomes: discriminating the PARP promoter from the immediately adjacent 3′ splice acceptor site

Jin Huang and Lex H.T.Van der Ploeg[1]

Department of Genetics and Development, College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA

Communicated by B.Pernis

[1]Corresponding author

We studied sequence requirements for *trans*-splicing at the 3′ splice acceptor site of a procyclic acidic repetitive protein (PARP) coding gene in trypanosomes. In transient CAT transfection assays with linker scanning (LS) mutants in a PARP promoter−3′ splice acceptor site−CAT construct, minor differences in the sequence composition of the polypyrimidine tract (nt −36 to −5 with respect to the 3′ splice acceptor site) severely affected the CAT activity. Analysis of steady-state CAT RNA in stably transformed trypanosomes revealed that the LS mutations had indeed affected the pre-mRNA splicing efficiency. The data indicate that mini-exon addition is not required simply for maturation of polycistronic pre-mRNA but is also essential for the generation of functional mRNA from monocistronic genes, since unspliced monocistronic pre-mRNA did not accumulate or allow synthesis of CAT. We postulate that mini-exon addition at polycistronically transcribed genes, which can have drastically different polypyrimidine tracts at each of their 3′ splice acceptor sites, can occur with different efficiencies for each gene of the array thus affecting mRNA abundance.

Key words: PARP promoter/polypyrimidine tract/mRNA/*trans*-splicing/trypanosomes

Introduction

Every trypanosome mRNA is believed to consist of two exons: a 5′ capped, non-translated, 39 nt mini-exon or spliced leader and the main coding exon, which are joined by *trans*-splicing (Van der Ploeg, 1986; Agabian, 1990). The mini-exon is derived from the 5′ end of a 140 nt mini-exon donor or spliced leader RNA (medRNA). The mechanism of *trans*-splicing is similar to *cis*-splicing: (i) the 5′ and 3′ boundaries of the donor−intron and acceptor−intron sequences encode the canonical, GT and AG dinucleotides also found at the intron−exon boundaries of *cis*-spliced introns; (ii) Y-structured, branched intermediates, analogous to the lariats in *cis*-splicing have been identified (Murphy *et al.*, 1986; Sutton and Boothroyd, 1986; Laird *et al.*, 1987; Ralph *et al.*, 1988) and the branch sites have been shown to involve a 2′−5′ phosphodiester bond at adenosine residues in the acceptor intron (Sutton and Boothroyd, 1988). However, instead of only one unique branch site, several potential branch sites were identified in the region between nt position −56 and −46, upstream of the 3′ splice acceptor sites of the α and β tubulin genes in trypanosomes (Patzelt *et al.*, 1989); and (iii) *trans*-splicing requires several small nuclear RNAs (snRNAs) including U2, U4 and U6 snRNA (Tschudi *et al.*, 1986; Mottram *et al.*, 1989; Tschudi and Ullu, 1990) although the U5 snRNA may be absent (Agabian, 1990; Mottram *et al.*, 1989); the medRNA has been proposed to be the functional equivalent of the U1 snRNA (Bruzik *et al.*, 1988; Bruzik and Steitz, 1990).

A polypyrimidine tract is required for *cis*-splicing, though it may play a more essential role in mammalian than in yeast pre-mRNA splicing (Green, 1986; Padgett *et al.*, 1986). The significance of the polypyrimidine tract in *trans*-splicing is unclear (Laird *et al.*, 1989; Patzelt *et al.*, 1989) and indeed the nucleotide sequences required for *trans*-splicing have not been determined. Recently established techniques for transient and stable transfection of African trypanosomes and other kinetoplastid protozoa (Bellofatto and Cross, 1989; Clayton *et al.*, 1990; Cruz and Beverley, 1990; Kapler *et al.*, 1990; Laban *et al.*, 1990; Lee and Van der Ploeg, 1990; Rudenko *et al.*, 1990; ten Aasbroek *et al.*, 1990; Eid and Sollner-Webb, 1991) allow a detailed analysis of the mechanism of *trans*-splicing. We have characterized the sequence elements at the 3′ splice acceptor sites that are required for *trans*-splicing. Their identification may shed light on the mechanism of *trans*-splicing, the proteins and snRNPs involved and the role of *trans*-splicing in mRNA production.

Results

Linker scanning analysis of sequences upstream of the 3′ splice acceptor site

The PARP (procyclic acidic repetitive protein) coding genes are found at four loci with two (α and β) PARP genes arranged in a polycistronic array with a promoter immediately upstream of the first α-PARP gene of each array juxtaposed to a 3′ splice acceptor site (Clayton *et al.*, 1990; Pays *et al.*, 1990; Rudenko *et al.*, 1990). Since the regulatory sequences for transcription initiation and *trans*-splicing are found in close proximity, a careful dissection of both types of regulatory elements is required to understand their individual roles in regulating mRNA abundance. In contrast to protein coding genes transcribed by RNA polymerase II, the PARP promoter controls α-amanitin resistant transcription. We had previously proposed that this transcription could be mediated by RNA polymerase I (pol I) and that *trans*-splicing might add a cap to pol I derived protein coding transcripts (Shea *et al.*, 1987; Rudenko *et al.*, 1989, 1991; Van der Ploeg, 1990). Deletion analysis of the PARP promoter has revealed that the minimal promoter element is confined to a region between nt position −400 and −86 (the transcription initiation site is located at about −86) upstream of the 3′ splice acceptor site (Pays *et al.*, 1990; Rudenko *et al.*, 1990; this paper and S.Brown and L.H.T. Van der Ploeg, unpublished). We now wished to determine whether mutations at the 3′ splice acceptor site,

© Oxford University Press

that did not affect PARP promoter function, would affect production of mRNA. In addition, we determined whether the synthesis of unspliced monocistronic mRNA, which theoretically is translatable, can substitute for the synthesis of *trans*-spliced mRNA.

We focused on the region between the transcription initiation site (at nt −86; Pays *et al.*, 1990) and nt position +7 downstream of the 3′ splice acceptor site (nt position 0) of one of the B locus α-PARP genes. The 5′ boundary was defined by the location of the transcription initiation site. The 3′ boundary of the sequences tested for *trans*-splicing was determined arbitrarily by comparing the expression levels of two constructs that extended from nt position −820 upstream, to nt positions +3 and +7 downstream of the 3′ splice acceptor site, placed in front of the chloramphenicol acetyl transferase (CAT) gene. Since each of these constructs gave identical CAT efficiencies we confined our analysis to the region extending upstream from nt position +7. We have not yet analyzed the significance of sequences further downstream (beyond nt +7) for *trans*-splicing.

We constructed linker scanning (LS) mutants (McKnight and Kingsbury, 1982) in the 3′ splice acceptor site region between nt position −83 and −1 by replacing 8 nt sequences with a synthetic *Sac*II linker (5′-GCCGCGGC; Figure 1A; two linker scanner mutants replace longer regions between nt positions −83 and −72, and between −66 and −55). The potential polypyrimidine stretch that precedes the AG dinucleotide of the 3′ splice acceptor site has a strong bias for uridines; we therefore replaced these with cytidines and guanosines. Furthermore, we chose a synthetic oligonucleotide without adenosines, to ensure that new cryptic adenosine branch sites would not be encoded in the linker

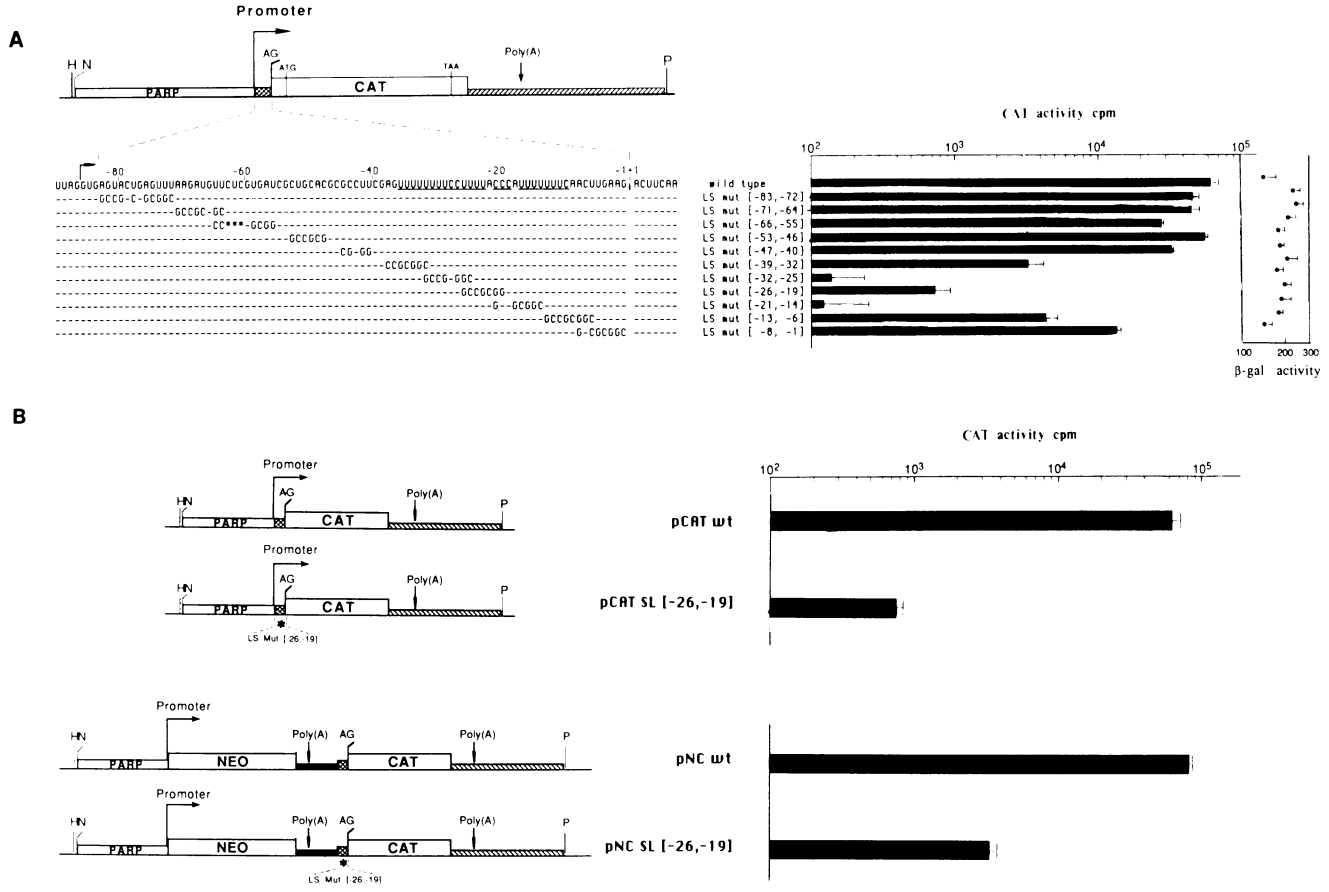


Fig. 1. Requirement of the 3′ splice acceptor site sequence elements for efficient *trans*-splicing. (**A**) Linker scanning (LS) mutant constructs and their *trans*-splicing efficiencies. The nucleotide sequences of all LS mutants were confirmed by dideoxyribonucleotide sequencing and are shown aligned with the wild type 3′ splice region sequence. The polypyrimidine tracts are underlined. The arrow at nucleotide −86 indicates the transcription initiation site. LS mutagenesis was conducted in the 83 nt region (dotted box) upstream of the 3′ splice acceptor site (indicated by a vertical arrow between the −1 and +1 nucleotide positions). The general structure of the plasmid constructs is depicted schematically in the physical map shown at the top. It is composed of the promoter (box labeled PARP) and 3′ splice acceptor site of the PARP α gene (dotted box; B locus derived) and the CAT gene followed by a PARP gene fragment containing the B1 PARP intergenic region (hatched box; Rudenko *et al.*, 1990). Only modified nucleotides are shown; asterisks denote missing nucleotides in the linker scanner mutant [−66, −55]; unchanged nucleotides are indicated with hyphens, mutated nucleotides are shown in each of the LS mutants. The right panel shows the average CAT activities after transfecting each of the constructs into procyclic *T.brucei*. The transformation experiment was performed three times with triplicate samples. Each value represents the average of triplicate samples, from a single experiment, with the standard deviation shown on top of the bar. A construct lacking the PARP promoter, but containing the PARP 3′ splice site, the CAT gene and the PARP intergenic region, served as the negative control. The background (1000 c.p.m.) was subtracted and the absolute counts and their standard deviation are shown on a logarithmic scale in the right panel. The CAT activity for each transfectant was determined as described (Rudenko *et al.*, 1990). The corresponding β-galactosidase activity was measured according to Nolan *et al.* (1988) and the values are also shown on a logarithmic scale in the most right hand panel. (**B**) To prove that the LS mutagenesis did not affect the activity of the PARP gene promoter, two polycistronic constructs, pNC wt and pNC LS[−26, −19], were made. In these constructs, the promoter element was separated from the 3′ splice acceptor site region (dotted box) by the neomycin phosphotransferase coding sequence and part of the intergenic region of the β,α tubulin genes (black bar). [H, *Hin*dIII; N, *Nde*I; P, *Pst*I; AG, dinucleotide at the 3′ splice acceptor site; poly(A), putative polyadenylation sites].

sequence. The different LS mutants were placed downstream of the PARP promoter and immediately upstream of the CAT reporter gene (Figure 1A). The entire sequence upstream of nt position +7, with the exception of the LS sequence itself, is thus identical to the wild type PARP promoter and its 3′ splice acceptor site. The different constructs were transfected into culture form (procyclic) trypanosomes (stock 427-60).

The relative CAT activity (always measured in the linear range of the assay) was affected most severely in the LS mutants with replacements between nt position −39 and −1 (Figure 1A). This region encodes the AG dinucleotide at the 3′ splice acceptor site and a polypyrimidine stretch, the longest uninterrupted stretch measuring 14 nt (interrupted by LS mut[−32, −25] which almost entirely abolished CAT activity). Interestingly, merely a replacement of the sequences between nt positions −13 and −6, changing four uridines within the first 7 nt of the polypyrimidine tract reduced the CAT activity to <10% of the wild type. Deletion of the AG dinucleotide at the 3′ splice acceptor site in LS mut[−8 to −1] led to the use of a newly generated 3′ splice acceptor site at nucleotide position −7 (Figure 1A and next sections), although at only 20% of wild type CAT activity.

The β-galactosidase (β-gal) activity from a co-transfected β-gal reporter gene with a (wild type) 3′ splice acceptor site was a control to show that alterations in CAT activity did not result from a varying transfection efficiency. Please note that, though the absolute β-gal activity in our assay is low, the β-gal control is sufficiently accurate to reveal, for instance, that the CAT activity of LS mut[−21, −14] does not result from an overall poor transfection efficiency (the β-gal controls show comparable activity in the transfection with this LS mutant and the transfection with the CAT wild type plasmid).

Discriminating promoter mutations from splice acceptor site mutations

The close proximity of the PARP promoter and its 3′ splice acceptor sequences made it essential to ensure that all the mutations analyzed affected the efficiency of *trans*-splicing only and had not affected the promoter function or the location of the transcription initiation site. We performed two experiments to verify this.

First, we constructed an array of protein coding genes in which the promoter (nt position −820 to nt −80) and the 3′ splice acceptor site (nt position −80 to nt position +7) were separated by ~1.3 kb of DNA encoding the neomycin phosphotransferase gene and the polyadenylation site of α and β tubulin (see physical map in Figure 1B). The efficiency

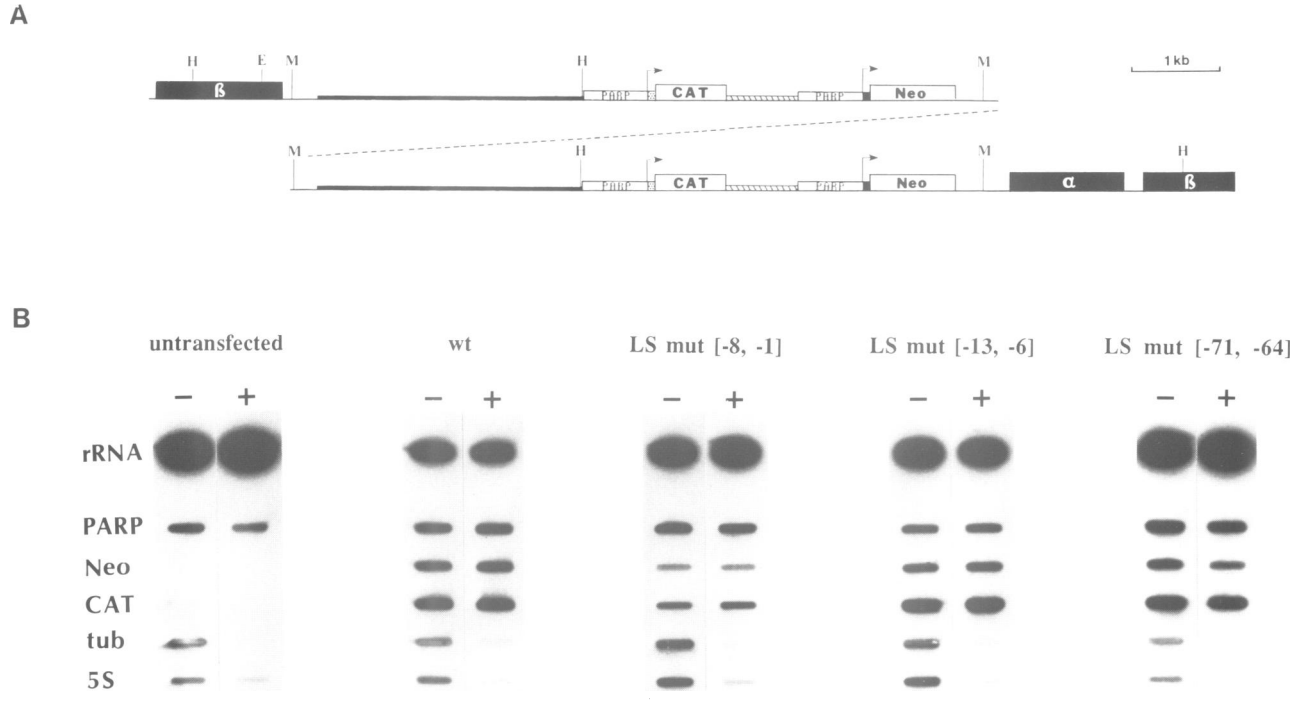


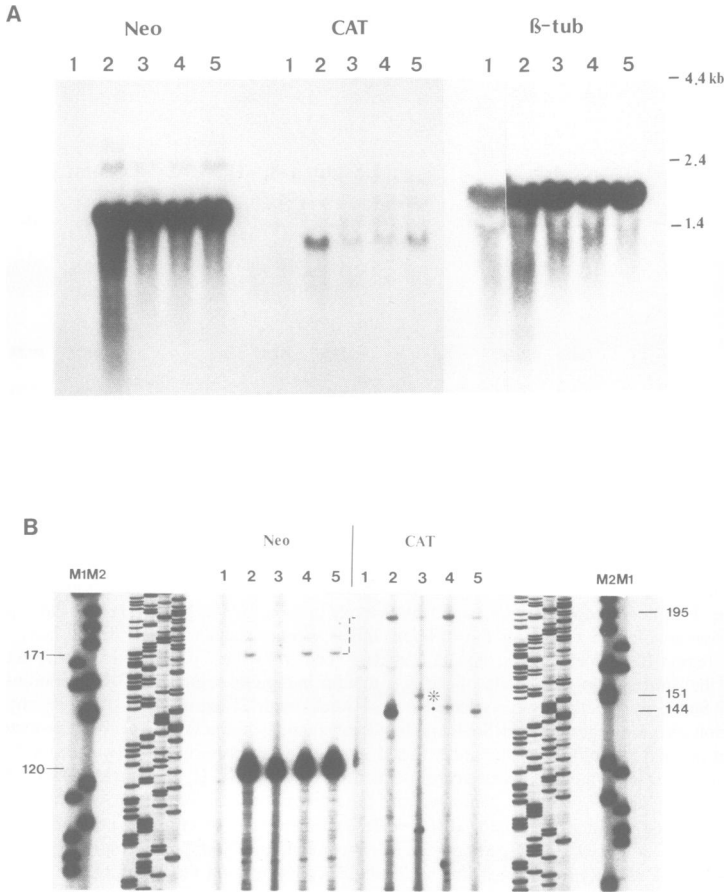
Fig. 2. (**A**) Integration of the LS mutant constructs into the tubulin locus of *T.brucei* by homologous recombination. Each gene is controlled by its own PARP promoter (arrow). The 3′ splice acceptor region (small black box) in front of the Neo^r gene is wild type, while the splice site of the CAT gene in different constructs is variable and represents the different linker scanner mutants (small dotted box). The Neo^r gene is flanked at its 3′ end by the β,α tubulin intergenic region (thin line), and is separated at its 5′ end from the CAT gene by the PARP intergenic region (hatched box). The constructs were linearized at the *Mlu*I site in the middle of the β,α tubulin intergenic region, and electroporated into procyclic *T.brucei* as described (Rudenko *et al.*, 1990). Stable transfectants were selected over 4 weeks with 25 μg/ml of G418. The physical map of the integrated gene copies was determined by restriction enzyme digestion and Southern blot analysis of genomic DNA of the transfected cell lines (data not shown). The copy number of the construct in each homologous integration varied among the different transfectants, from 1 to 5 copies (as indicated by the dotted line; see text). The thick line represents the Bluescript vector sequences. H, *Hin*dIII; E, *Eco*RI; M, *Mlu*I. (**B**) Analysis of nascent RNA derived from LS mutants. The signs '+' and '−' in each panel indicate the presence and absence of α-amanitin (final concentration 1 mg/ml). ^{32}P-Labelled nascent RNA was hybridized to filters containing different DNA fragments: (from top to bottom) a *Bgl*II fragment encoding the rDNA repeat (PR4; White *et al.*, 1986); a 0.8 kb *Eco*RI PARP coding sequence fragment; the Neo^r coding region (~1 kb); the coding sequence of the CAT gene (~0.8 kb); a 0.75 kb *Hin*dIII−*Eco*RI fragment from the β tubulin coding region; and a 5S rRNA genomic fragment (~0.8 kb). Hybridization signals were quantified with a Betagen Betascope 603 Blot Analyzer.

of expression of CAT from this construct (pNC wt) and one additional construct with a mutated 3′ splice acceptor site (pNC LS[−26, −19]) positioned upstream of the CAT gene were compared with the CAT efficiencies obtained with the constructs analyzed in Figure 1A. Figure 1B shows that the promoter and 3′ splice acceptor site in these constructs functioned similarly to those of the constructs described in Figure 1A. These data indicate that the linker scanner mutants are likely to have affected the *trans*-splicing efficiency only since their positioning, relative to the promoter or transcription initiation site, did not alter the efficiency of CAT expression.

Secondly, to ensure that the mutations had not affected the promoter and to allow analysis of the CAT pre-mRNA and CAT steady-state mRNA, we analyzed promoter functioning in several stably transfected trypanosome lines. We generated plasmid constructs in which the CAT and Neo genes were placed in a tandem array, each gene being under the control of its own PARP promoter (see Figure 2 for physical map). These constructs were integrated, by homologous recombination, into the α and β tubulin locus of *Trypanosoma brucei* as shown in the physical maps in Figure 2A. Only the construct with the LS mut[−8, −1] was integrated as a single copy. The constructs with the wild type splice site, the LS mut[−13, −6] and the LS mut[−71, −64] were integrated as tandem arrays of three, four and five copies, respectively (data not shown). We isolated nuclei from these stable transfectants and determined the transcriptional efficiencies of the CAT and Neo genes in a nuclear run-on assay with ^{32}P-labeled nascent RNA. The comparison of the efficiencies of transcription of the CAT and Neo genes relative to control genes discriminates between the effects of the mutations on promoter function and *trans*-splicing. In each case we found that the transcription of CAT and Neo genes was, as expected, α-amanitin resistant. Importantly, the CAT gene was always transcribed at twice the efficiency of the Neo gene in every cell line. This indicates that the promoter in front of CAT functioned similarly in every cell line. In addition, the CAT gene is the first gene located downstream of the α-amanitin sensitively transcribed β tubulin gene and the fact that the CAT gene was transcribed more efficiently than the Neo gene may imply that the second (Neo) promoter, in a tandem array of promoters, functions less efficiently and that read through transcription into plasmid sequences must be limited (as shown in Lee and Van der Ploeg, 1990). Finally, the level of transcription of CAT and Neo genes, relative to the control genes (see for instance PARP) roughly correlated with the number of integrated plasmid copies. We therefore conclude that the linker scanner mutants had not significantly affected promoter function. We can also conclude that sequences of significance for promoter function must be located upstream of nt position −80.

Analysis of steady-state RNA in transformants which express CAT genes with mutated 3′ splice acceptor sites

We next analyzed the steady-state RNA from these transfected cell lines and determined whether the LS mutations indeed affected the pre-mRNA *trans*-splicing



efficiency. First, in Northern hybridizations (Figure 3A) each of the LS mut[−71, −64] (lanes 5), [−13, −6] (lanes 4) and [−8, −1] (lanes 3) constructs generates CAT mRNA of ~1.4 kb. The steady-state CAT RNA levels in each cell line varied. The differences in CAT mRNA overall reflected the predicted altered *trans*-splicing efficiencies that we had observed in the transient CAT assays (Figure 1) and the varying CAT gene copy number. For instance, the construct with the wild type splice site (lane 2) present in three copies, produced much more CAT mRNA than LS mut[−13, −6] (lane 4) which is present in four copies (the predicted amount of CAT mRNA, as deduced from the transient CAT activity of this LS mutant corrected for CAT gene copy number was 1.3 × 7% of wild type CAT mRNA). Similarly LS mut[−8, −1] (lane 3) present at one copy produced much less CAT mRNA than the wild type construct (the predicted amount of CAT mRNA as deduced from the transient CAT activity of this LS mutant was 0.3 × 20% of wild type). LS mut[−71, −64] (present in five copies in the genome; lane 5) produced as expected more CAT mRNA than the other two LS mutants. However, its CAT mRNA level should have been 1.6 × 75% of wild type, while it produced less CAT mRNA than wild type. We do not have a good explanation for this discrepancy. As expected, the amount

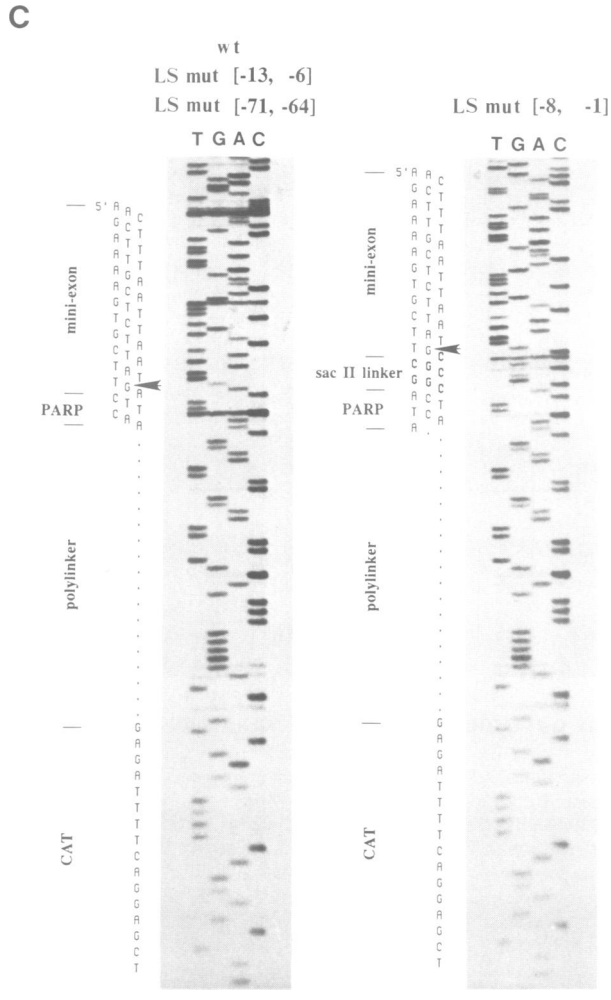


Fig. 3. (**A**) Northern blot analysis of mRNA derived from stable transfectants with the different LS 3′ splice acceptor mutants. The lanes contain: 1, RNA from untransfected trypanosomes; 2, pCAT wt; 3, LS[−8, −1]; 4, LS[−13, −6]; and 5, LS[−71, −64]. Three panels were each hybridized with different ^{32}P-labeled probes: Neo, the coding sequence of the Neo^r gene; CAT, a CAT gene coding sequence probe; and β-Tub, the β tubulin gene probe (see Figure 2B for probe description). Post-hybridizational washes were carried out at 0.1 × SSC at 65°C. (**B**) Primer extension analysis in steady-state RNA to characterize the 5′ ends of mRNA derived from the LS mutants. Two antisense 20mer oligonucleotides were used. One is a Neo^r gene specific primer (5′-CCATCTTGTTCAATCATGCG) complementary to the 5′ end of Neo^r coding sequence, and the other is a CAT gene specific primer (5′-CAACGGTGGTATATCCAGTG) complementary to the 5′ end of the CAT coding sequence. cDNA was synthesized from different RNA samples using a Neo primer (panel Neo) or CAT primer (panel CAT). The different lanes contain: 1, RNA from untransfected trypanosomes; 2, pCAT wt; 3, LS[−8, −1]; 4, LS[−13, −6]; and 5, LS[−71, −64]. The size of major extended products is indicated. An aberrant species in CAT lane 3 is marked with an asterisk. M1 and M2 are end-labeled size standards derived from M13/*Hae*III and PBR322/*Hpa*II, respectively. For accurate sizing the primer extended products are flanked by M13 sequencing ladders. (**C**) DNA sequence of the cDNAs from LS mutant CAT mRNAs. The nucleotide sequence of the CAT cDNAs from pCAT wt, the mutants, LS[−13, −6] and LS[−71, −64] transfectants was identical (left panel; only the sequence from LS mut[−13, −6] is shown), showing mini-exon addition to the correct 3′ splice acceptor site (marked with an arrowhead). The sequence of CAT cDNA from the LS mut[−8, −1] transfectant (right panel) shows a cryptic 3′ splice acceptor site (arrowhead) generated by the *Sac*II linker placed downstream of an adenosine residue, replacing the wild type 3′ splice acceptor site. Horizontal bars indicate the junctions between different sequence elements used to construct the linker scanner mutants. The sequence of the polylinker region that connects PARP-derived sequence and CAT-derived sequence is represented by a dotted line. Sense strand sequence is read (rather than antisense) because the lanes are marked with the complementary bases.

A

```
-80      -70       -60       -50       -40       -30       -20       -10       -1+1      +10
UCCCAGAAGU AUUAUGUUUA AACACGCAUG UGCGUCGGUA GCAAGCCUGG UAAAAUUUGU CUGUACCAUU UCCAAUGAAG CGGUAGGGAA  ODC          Cons.
AAUAUACCUU AGUCAUUCCA UUUCCUAUUG CGGGUACAAC GAUAACGGUG GUAAAACCGU CGGCGUUUUU UUUUUCUAAG UAAUCGAAAC  PGK A       Cons.
ACGGCGCAGU CAGCGACACG CCAGCGUGUG CGCCACCUAA UGUACGAUAU UCUAUUUCUC AUUUUCCAUG ACACGCACAG AGGCGCACAA  VSG 118     BF
GAGAGGGCAA UGCAACGCUG GUUAUCCCAG UUGAUGUGCG CAUUCAUAUA UACCAUUUCC CCAUUUUUUU CAACUUGAAG ACUUCAAUUA  PARP Bβ     PF
ACUGAGUUUA AGAUGUUCUC GUGAUCGCUG CACGCGCCUU CGAGUUUUUU UUCCUUUUAC CCAUUUUUUU CAACUUGAAG ACUUCAAUUA  PARP Bα     PF
UUUUUGAUCG UUCUACUGUG UAACUUUGUG UUUAUCUUUG UUACUUCACU CUUUUUUCAC UCAAAUCGUU UGGGCUGCAG GCUCAGCUCU  PGK C      BF
AGUCUUUUGA UUCAAAAAUU UAAACCGAAU UCCCUUCCCC AAGUCUCGCA GUCACUUCUU UUCAACGUUU UCUCACUUAG CCAACAUACU  PGK B      PF
CAUCGUCCUC CCAUACUUGC ACUCCUGCUC UUCCUUCCUA UUUGCAUUUG CCCACGUGCC CCACGCGUGC AUCACAUAAG CAUCAAGGAA  Hsp 70     Cons.
AUUCGCUUCU UCCUGUCGUG UCACACCUCU UUCUCCCUCU CCCUUUCCGC CUUUUCUUUC AAUCUUGUUU UCUCGACCAG CCCUACUAGA  Tubβ       Cons.
```

B

Fig. 4. (**A**) Examples of 3′ splice acceptor site sequences in some of the trypanosome genes used in the comparison. The sequences are shown as RNA sequences extending from 80 nt upstream (−80) to 10 nt downstream (+10) of the 3′ splice acceptor site (nt position 0). The AG dinucleotides at the 3′ splice acceptor−intron boundaries as well as the AG dinucleotides at alternative 3′ splice acceptor sites located further downstream are underlined. Polypyrimidine tracts are underlined: we defined polypyrimidine tracts located directly upstream of the 3′ splice acceptor site, between nt position −80 and −1, as having at least one sequence with a minimum of five consecutive pyrimidines and interrupted by no more than 22% purines. BF, bloodstream form, specifically expressed; CF, insect form expressed; Cons, constitutively expressed. (**B**) The pyrimidine content of 29 different 3′ splice acceptor sites was calculated (average per 10 nt column presented in A). Sequences were used from the 3′ splice acceptor sites for the following genes: PARP, Procyclic Acidic Repetitive Protein genes (α and β; Rudenko *et al.*, 1990); VSG (117, 118, 221, AnTat 1.1) Variant Surface Glycoprotein gene (Boothroyd and Cross, 1982; Van der Ploeg *et al.*, 1982); ESAG (ESAGs 1−7), Expression Site Associated Genes (Cully *et al.*, 1985; Pays *et al.*, 1989; Berberof *et al.*, 1991); Tub (α and β) Tubulin genes (Sather and Agabian, 1985); Cal (1−3), Calmodulin genes (Tschudi *et al.*, 1985); PGK (A, B, C), Phosphoglycerate Kinase genes (Gibson *et al.*, 1988); PolII (A, B) RNA polymerase II genes (Evers *et al.*, 1989; Smith *et al.*, 1989); TIM, Triosephosphate isomerase gene (Swinkels *et al.*, 1986); Hsp70, Heat Shock Protein 70 kDa gene (Glass *et al.*, 1986); ALD, Aldolase gene (Clayton, 1985); U-ALD, Upstream gene of Aldolase (Vijayasarathy, 1990); I-ALD, Intergenic gene of Aldolase (Vijaysarathy *et al.*, 1990); ODC, Ornithine Decarboxylase gene (Phillips *et al.*, 1987); PP (1, 2), Protein Phosphatase gene (Evers and Cornelissen, 1990).

of Neo mRNA more closely reflected the *neo* gene copy number in all cell lines, while the tubulin hybridization control showed that identical amounts of RNA were loaded in lanes 2−5. The fact that the CAT mRNA is much less abundant than the Neo mRNA, even though the CAT genes are transcribed at a higher efficiency, presumably reflects an overall lower CAT mRNA stability.

We next performed primer extensions on steady-state RNA with ^{32}P-end-labeled oligonucleotides specific for the Neo and CAT coding sequences allowing a quantitative comparison of CAT pre-mRNA and CAT mRNA levels in a single cell. These quantifications will reveal the inhibitory effects of the LS mutations on *trans*-splicing more accurately since they will determine the ratio of CAT pre-mRNA and mRNA in a single transfectant. In addition, the 5′ ends of the different mRNA molecules were analyzed to address whether the RNAs of the different LS mutants had been processed at all. The primer extension products with the Neo oligonucleotide (Figure 3B) shows a product at 120 nt, which we assume is specific for the 5′ end of mini-exon containing Neo mRNA (see next section) and a second strong stop at 171 nt, assumed to map to the transcription initiation site

of the PARP promoter (Pays *et al.*, 1990; and data not shown). Primer extension products using the CAT oligonucleotide also gave specific strong stops (Figure 3B): the band at 144 nt is specific for the 5′ end of the spliced CAT mRNA (see next section); the second specific strong stop at 195 nt locates the transcription initiation site (Pays *et al.*, 1990). However, in the LS mut[−8, −1] construct, in which the regular 3′ AG splice acceptor site had been deleted, a longer (151 nt) extended product could be detected (band marked with an asterisk) indicating *trans*-splicing at an alternative 3′ splice acceptor site. The significance of the slightly shorter extended product in this primer extension (band marked with a black dot) is unclear and we assume that it represents a non-specific strong stop. In these experiments the ratio of the intensity of the CAT pre-mRNA (band at 195) and CAT mRNA signal (band at 144) in each lane clearly reflects the predicted differences in the *trans*-splicing efficiencies for *each* of the LS mutants. We conclude that the 3′ splice acceptor site mutations had indeed affected the *trans*-splicing efficiency.

The amount of pre-mRNA appeared overall to be unaffected when compared with the wild type, LS mut[−8, −1] and LS mut[−13, −6] cell lines [as identified by the intensity of the band with the strong stop mapping to the initiation site at nt position −86 nt; compare the control Neo signal (at 171 nt) and the CAT signal (at nt 195) intensities for each cell line; minor signal intensity differences reflect CAT, Neo gene copy number differences and experimental variation affecting the CAT and Neo pre-mRNAs comparably]. However, please note that again the amount of CAT pre-mRNA in LS mut[−71, −64] appears low while the ratio of its CAT pre-mRNA to mRNA shows that its 3′ splice site is, as predicted, more efficient than that of the other LS mutants. We conclude that since the CAT pre-mRNA did not accumulate when *trans*-splicing was impaired it must be rapidly turned over.

DNA nucleotide sequence analysis of cDNAs from these four different primer extension products revealed that the mini-exon had been added at the correct 3′ splice acceptor site, except for the cDNAs from LS mut[−8, −1] in which a newly generated cryptic 3′ splice acceptor site at nt position −7 was used (Figure 3C).

Discussion

Our data show that the polypyrimidine tract of the PARP 3′ splice acceptor site, located between nt position −36 and −5, has an essential role in *trans*-splicing. For instance, LS mut[−32, −25] (Figure 1) replacing the sequence UUUUCCUU with the sequence GCCGCGGC almost reduced CAT activity to background levels. To address the significance of polypyrimidine tracts at 3′ splice acceptor sites at other trypanosome genes we compared the nucleotide sequences at 29 different 3′ splice acceptor sites (Figure 4). The distribution of pyrimidines per 10 nucleotide column revealed a significant increase in the number of pyrimidines between nt positions −40 and −10 (Figure 4B). This distribution of pyrimidines correlates with the area where the LS replacements near the PARP 3′ splice acceptor site had the most dramatic effects on the *trans*-splicing efficiency. This suggests that the requirement of a polypyrimidine tract is not just an oddity of the PARP 3′ splice acceptor site. A bias for the use of uridine over cytidine is also observed in this region, with uridines occurring most frequently in the region between nt position −20 and −10. Even though an occasional 3′ splice acceptor site can be preceded by a longer polypyrimidine tract, extending up to nt position −80 (see for instance the β tubulin 3′ splice acceptor site; Figure 4A) this appears to be an exception. A comparison of insect form and bloodstream form specific mRNAs (compare PGK B [insect form (PF)] and PGK C [bloodstream (BF)] did not reveal obvious nucleotide sequence preferences in the polypyrimidine tract, making it unclear whether differential *trans*-splicing, controlled by the 3′ splice acceptor site, can regulate the differential expression of constitutively transcribed genes. The examples in Figure 4A represent 3′ splice acceptor site sequences with the most diverged polypyrimidine tracts (compare for instance the α-PARP, β tubulin, phosphoglycerate-kinase A and ornithine decarboxylase 3′ splice acceptor site sequences). These examples were chosen to highlight the range of polypyrimidine tract length that exists. This nucleotide sequence variation and the varying splicing efficiencies that we observed in our linker scanner mutant analysis predict that these different 3′ splice acceptor sites should function with widely varying efficiencies. In comparing the CAT expression levels in constructs that contained 3′ splice acceptor sites from different genes this variability was indeed observed (unpublished data). An intrinsic variability in the *trans*-splicing efficiency at individual genes from a polycistronic array could therefore represent one of the regulatory mechanisms by which the mRNA abundance of genes from a polycistronically transcribed locus is regulated.

Our analysis did not reveal the location of potential branch sites. By analogy to *cis*-splicing LS mutants that affect the adenosine residue at the branch site should abolish *trans*-splicing at the 3′ splice acceptor site. Branch sites for *trans*-splicing had been located between nt −42 and −58 of the trypanosome α and β tubulin genes (Patzelt *et al.*, 1989). However, *trans*-splicing efficiencies of LS mutants in the region between nt position −83 and −40 were hardly affected, ranging from 93% (LS mut[−53, −46]) to 46% (LS mut[−66, −55]) of wild type. The overlapping series of linker scanner mutants between nt position −83 and −40, which did not affect the polypyrimidine tract, replaced eight of the 15 adenosine residues in the 3′ acceptor intron of the pre-mRNA. Their positioning between nt −83 and −40 predicted that one of these should serve as a branch site (Patzelt *et al.*, 1989). However, since none of these linker scanner mutants had markedly reduced CAT activity, any unique adenosine residue in this region is not essential in branch site formation. It is possible that, in contrast to the primary branch sites located between nt position −56 and −46 upstream of the α and β tubulin genes (Patzelt *et al.*, 1989), those at the PARP 3′ splice acceptor sites are located closer to the 3′ splice acceptor site (between nt position −39 and −2). Alternatively, as proposed previously, cryptic branch sites may exist and a single specific branch site may be absent from trypanosome pre-mRNA (Agabian, 1990; Hartshorne and Agabian, 1990). Of the seven adenosine residues between nt position −39 and −2, only one linker scanner mutant, disrupting the adenosine at −18, has a very low CAT activity of 0.2% of wild type, making this a possible candidate for a unique branch site. However, the CAT activity levels of the linker scanner mutants, that replaced adenosine residues in this region, cannot be

interpreted straightforwardly, since their effects are either due to interference with the functioning of the polypyrimidine tract or of the branch site. Since the remaining CAT activity in several of these LS mutants was significantly over background (minimally 1.3% of wild type activity) we can conclude that none of these sequences encoded unique branch sites.

We have not included the first three nt (GTG, −86 to −84) of the pre-mRNA in the LS analysis. We omitted this region, since mutations at the transcription initiation site may affect PARP promoter activity, thereby obscuring its significance for *trans*-splicing. Since adenosine residues are absent in this region it is unlikely to function in branch site formation.

Finally, pre-mRNA that was not *trans*-spliced and which had a 5′ end that located to the transcription initiation site at nt position −86 did not substitute for *trans*-spliced mRNA in CAT production. Hence we can tentatively conclude that *trans*-splicing is an absolutely essential step in the production of functional CAT mRNA since it is even required for the synthesis of functional mRNA from monocistronic genes. Mini-exon addition might affect mRNA stability since unspliced pre-mRNA did not accumulate or allow synthesis of CAT enzyme.

Materials and methods

Construction of linker scanning mutants

A wild type construct, pCAT wt, was first made from plasmid BNspCAT-1 (Rudenko *et al.*, 1990) by adding an 850 bp PARP intergenic region-containing fragment, to the 3′ end of the CAT gene, and by cloning this fragment into the *Hin*dIII and *Pst*I sites of pBluescript SK$^-$. LS mutagenesis (McKnight and Kingsbury, 1982) was conducted in the 83 nt region (Figure 1, dotted box) upstream of the 3′ splice acceptor site (indicated by a vertical arrow between the −1 and +1 nucleotide positions). We made the 5′ and 3′ exoIII deletion series from *Hin*dIII digested, linearized plasmid BSspCat (Rudenko *et al.*, 1990) and from *Bam*HI digested plasmid BNspCAT-1. After ligation of *Sac*II linkers (5′-GCCGCGGC-3′; NE Biolab) to each of the exoIII deletion series, the 5′ exoIII series was digested with the restriction enzymes *Sac*II and *Nco*I, while the 3′ deletion series was cut with *Sac*II−*Hin*dIII. Matching deletion mutants (e.g. 3′Δ-1 and 5′Δ-8) were ligated into a *Hin*dIII and *Nco*I-digested pCAT wt vector to generate an LS mutant construct. Nine out of the 11 LS mutants were constructed according to this protocol. The other two mutants, LS[−83, −72] and LS[−66, −55], were generated following additional manipulation of the *Sac*II linkers [(i.e. after the first linker ligation, the *Sac*II site was made blunt with T4 DNA polymerase and subjected to a second round of linker ligation, *Sac*II digestion and T4 DNA polymerase to make a blunt *Sac*II site. This resulted in sequence replacements of larger stretches, covering 12 bp (5′-GCCGCCGGCGGC)]. Polycistronic constructs pNC wt and pNC LS [−26, −19] were made from the pCAT construct described above and plasmid BNspNeo-T (Lee and Van der Ploeg, 1990). After a multi-step cloning procedure, an *Eco*RI−*Mlu*I fragment from BNspNeo-T was inserted into the *Sca*I site (located 80 bp upstream of the 3′ splice acceptor site) of plasmid pCAT wt and plasmid pCAT LS[−26, −19], respectively. The nucleotide sequences at the ligation junctions were confirmed by dideoxy-ribonucleotide sequencing.

The constructs pCAT wt, LS mut[−8, −1], LS mut[−13, −16] and LS mut[−71, −64] were linearized at the *Xba*I site. The *Xba*I site was made blunt-ended and ligated to a blunt *Hin*dIII−*Bam*HI fragment from plasmid BNspNeo-T. A construct encoding CAT and a Neo^r gene in the same orientation was selected.

Transformations

To test the effect of the different LS mutants on *trans*-splicing efficiency, 10 μg of each LS mutant construct was electroporated into procyclic *T.brucei* (Rudenko *et al.*, 1990), together with 10 μg of the control PARP promoter−β-gal gene construct. The methods to test the β-gal activity are as described by Nolan *et al.* (1988). Stable transformations were performed with *Mlu*I linearized plasmids as described by Lee and Van der Ploeg (1990).

RNA analysis

Steady-state RNA was prepared from stable lines of transfected trypanosomes according to Maniatis *et al.* (1982). RNA was fractionated in a 0.7% formaldehyde gel with 10 μg RNA loaded per lane. Following size fractionation the RNA was transferred to a nitrocellulose filter.

Nuclei preparation and nuclear run-off transcription were performed as described previously (Kooter *et al.*, 1987).

Nucleotide sequence analysis of cDNAs

The primer extension products of CAT mRNAs (Figure 3B) were directly used as templates for PCR amplification (Saiki *et al.*, 1986) with the same antisense CAT primer and a sense oligonucleotide of the mini-exon sequence (5′-AACGCTATTATTAGAGCGGT). Amplified cDNAs were cloned into the *Sma*I site of plasmid Sp64 and their DNA nucleotide sequence determined according to USB DNA sequencing protocol with sequenase enzyme.

Acknowledgements

We thank Stanley Korman, Keith Gottesdiener, Sylvie Le Blancq, Mary Lee, Gloria Rudenko, Demetrios Vassilatis and Scott Zeitlin for critical reading of the manuscript and Gloria Rudenko and Steven Brown for constructing the β-gal plasmid and for help with setting up the β-gal assay. This work was supported by NIH grant AI 21784 to L.H.T.V.d.P. and by a grant from the John D. and Catherine D.MacArthur Foundation. L.H.T.V.d.P. is a Burroughs Wellcome Scholar in Molecular Parasitology.

References

Agabian,N. (1990) *Cell*, **61**, 1157−1160.
Bellofatto,V. and Cross,G.A.M. (1989) *Science*, **244**, 1167−1169.
Berberof,M., Pays,A. and Pays,E. (1991) *Mol. Cell. Biol.*, **11**, 1473−1479.
Boothroyd,J.C. and Cross,G.A.M. (1982) *Gene*, **20**, 279−287.
Bruzik,J.P. and Steitz,J.A. (1990) *Cell*, **62**, 889−899.
Bruzik,J.P., Van Doren,K., Hirsch,D. and Steitz,J.A. (1988) *Nature*, **335**, 559−562.
Clayton,C.E. (1985) *EMBO J.*, **4**, 2997−3003.
Clayton,C., Fueri,J.P., Itzhaki,J.E., Bellofatto,V., Sherman,D.R., Wisdom,G.S., Vijayasarathy,S. and Mowatt,M.R. (1990) *Mol. Cell. Biol.*, **10**, 3036−3047.
Cruz,A. and Beverley,S.M. (1990) *Nature*, **348**, 171−173.
Cully,D.F., Ip,H. and Cross,G.A.M. (1985) *Cell*, **42**, 173−182.
Eid,J. and Sollner-Webb,B. (1981) *Proc. Natl. Acad. Sci. USA*, **88**, 2118−2121.
Evers,R. and Cornelissen,A.W.C.A. (1990) *Nucleic Acids Res.*, **18**, 5089−5095.
Evers,R., Hammer,A., Kock,J., Jess,W., Borst,P., Memet,S. and Cornelissen,A.W.C.A. (1989) *Cell*, **56**, 585−597.
Gibson,W.C., Swinkels,B.W. and Borst,P. (1988) *J. Mol. Biol.*, **201**, 315−325.
Glass,D.J., Polvere,R.I. and Van der Ploeg,L.H.T. (1986) *Mol. Cell. Biol.*, **6**, 4657−4666.
Green,M.R. (1986) *Annu. Rev. Genet.*, **20**, 617−708.
Hartshorne,T. and Agabian,N. (1990) *Genes Dev.*, **4**, 2121−2131.
Kapler,G.M., Coburn,C.M. and Beverley,S.M. (1990) *Mol. Cell. Biol.*, **10**, 1084−1094.
Kooter,J.M., Van der Spek,H.J., Wagter,R., d'Oliveira,C.E., Van der Hoeven,F., Johnnson,P.J. and Borst,P. (1987) *Cell*, **51**, 261−272.
Laban,A. and Wirth,D. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9119−9123.
Laban,A., Tobin,J.F., Curotto de Lafaille,M.A. and Wirth,D.F. (1990) *Nature*, **343**, 572−574.
Laird,P.W. (1989) *Trends Genet.*, **11**, 204−208.
Laird,P.W., Zomerdijk,J.C.B.M., De Korte,D. and Borst,P. (1987) *EMBO J.*, **6**, 1055−1062.
Layden,R.E. and Eisen,H. (1988) *Mol. Cell. Biol.*, **8**, 1352−1360.
Lee,M.G.-S. and Van der Ploeg,L.H.T. (1990) *Science*, **250**, 1583−1589.
Maniatis,T., Fritsch,E. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
McKnight,S.L. and Kingsbury,R. (1982) *Science*, **217**, 316−324.
Mottram,J., Perry,K.L., Lizardi,P.M., Luhrmann,R., Agabian,N. and Nelson,R.G. (1989) *Mol. Cell. Biol.*, **9**, 1212−1223.
Murphy,W.J., Watkins,K.P. and Agabian,N. (1986) *Cell*, **47**, 517−525.
Nolan,G., Fiering,S., Nicolas,J.-F. and Herzenberg,L.A. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2603−2607.

Padgett,R.A., Grabowski,P.J., Konarska,M.M. and Sharp,P.A. (1986) *Annu. Rev. Biochem.*, **55**, 1119–1150.
Patzelt,E., Perry,K.L. and Agabian,N. (1989) *Mol. Cell. Biol.*, **9**, 4291–4297.
Pays,E., Tebabi,P., Pays,A., Coquelet,H., Revelard,P., Salmon,D. and Steinert,M. (1989) *Cell*, **57**, 835–845.
Pays,E., Coquelet,H., Tebabi,P. and Steinert,M. (1990) *EMBO J.*, **9**, 3145–3151.
Phillips,M.A., Coffino,P. and Wang,C.C. (1987) *J. Biol. Chem.*, **262**, 8721–8727.
Ralph,D., Huang,J. and Van der Ploeg,L.H.T. (1988) *EMBO J.*, **7**, 2539–2545.
Rudenko,G., Bishop,D., Gottesdiener,K. and Van der Ploeg,L.H.T. (1989) *EMBO J.*, **13**, 4259–4263.
Rudenko,G., Le Blancq,S., Smith,J., Lee,M.G.-S., Rattray,A. and Van der Ploeg,L.H.T. (1990) *Mol. Cell. Biol.*, **10**, 3492–3504.
Rudenko,G., Chung,H.M., Vinh,P.P. and Van der Ploeg,L.H.T. (1991) *EMBO J.*, **10**, 3387–3397.
Saiki,R.K., Bugawan,T.L., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1986) *Nature*, **324**, 163–166.
Sather,S. and Agabian,N. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 5695–5699.
Shea,C., Lee,M.G.-S. and Van der Ploeg,L.H.T. (1987) *Cell*, **50**, 603–612.
Smith,J.L., Levin,J.R., Ingles,J.C. and Agabian,N. (1989) *Cell*, **56**, 815–827.
Sutton,R.E. and Boothroyd,J.C. (1986) *Cell*, **47**, 527–535.
Sutton,R.E. and Boothroyd,J.C. (1988) *EMBO J.*, **7**, 1431–1437.
Swinkels,B.W., Gibson,W.C., Osinga,K.A., Kramer,R., Veeneman,G.H., Van Boom,J.H. and Borst,P. (1986) *EMBO J.*, **5**, 1291–1298.
ten Asbroek,A.L.M.A., Quellette,M. and Borst,P. (1990) *Nature*, **348**, 174–175.
Tschudi,C. and Ullu,E. (1990) *Cell*, **61**, 459–466.
Tschudi,C., Richards,F. and Ullu,E. (1986) *Nucleic Acids Res.*, **14**, 8893–8903.
Tschudi,C., Young,A.S., Ruben,L., Patton,C.L. and Richards,F.F. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 3998–4002.
Van der Ploeg,L.H.T. (1986) *Cell*, **47**, 479–480.
Van der Ploeg,L.H.T. (1990) In Hames,B.D. and Glover,D. (eds), *Frontiers in Molecular Biology. Gene Rearrangements*. IRL Press, Oxford, pp. 51–97.
Van der Ploeg,L.H.T.,Liu,A.Y.C., Michels,P.A.M., DeLange,T., Borst,P., Majumder,K., Weber,H. and Veeneman,G.H. (1982) *Nucleic Acids Res.*, **10**, 3591–3604.
Vijayasarathy,S., Ernest,I., Itzhaki,J.E., Sherman,D., Mowatt,M.R., Michels,P. and Clayton,C.E. (1990) *Nucleic Acids Res.*, **18**, 2967–2975.
White,T.C., Rudenko,G. and Borst,P. (1986) *Nucleic Acids Res.*, **14**, 9471–9489.

Received on August 16, 1991; revised on September 12, 1991