4516–4523 *Nucleic Acids Research, 1998, Vol. 26, No. 19*

In vivo selection for intronic splicing signals from a randomized pool

John Bouck[1,2], Samuel Litwin[1], Anna Marie Skalka[1,2] and Richard A. Katz[1,*]

[1]Institute for Cancer Research, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA 19111, USA and [2]The Cell and Molecular Biology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA

Received April 13, 1998; Revised and Accepted August 11, 1998

ABSTRACT

Retroviruses utilize balanced splicing to express multiple proteins from a single primary transcript. A number of *cis*-acting signals help maintain this balance, including the branch point sequence (BPS), polypyrimidine tract (PPyT) and sequences within the downstream exon. In general, regulated splicing requires weak splicing signals and we have previously shown the same requirement for the simple retrovirus, avian sarcoma virus (ASV). Here we take advantage of the requirement for balanced splicing in retroviral replication to examine the sequence constraints of an intronic splicing element. Selection for replication competence makes it possible to amplify and identify functional sequences from a pool of all possible sequences. In this report we examine the role of pyrimidines within the PPyT. Our results provide *in vivo* confirmation that the functional strength of a PPyT is related to its length and uridine content and that the PPyT plays a role in the second step of the splicing reaction. We also show that the minimal distance between the 3′-splice site and the BPS in this system is 16 nt. With modification, the selection system described here can be used to examine the sequence constraints of other exonic or intronic splicing elements *in vivo*.

INTRODUCTION

Splicing is the process by which intervening sequences are removed from precursor messenger RNAs (pre-mRNAs). The reaction takes place within a large complex called the spliceosome and requires the participation of five small nuclear ribonucleoprotein particles (snRNPs) as well as many non-snRNP proteins (for reviews see 1–5). Pre-mRNA splicing occurs by a two-step mechanism (Fig. 1C). In the first step, the donor 5′-splice site (5′-ss) is cleaved via a nucleophilic attack by a 2′-OH from an adenosine within the intronic branch point sequence (BPS) near the acceptor 3′-splice site (3′-ss). This creates two intermediates, a lariat intron still linked to the downstream exon and a free upstream exon. The second step comprises an attack on the 3′-ss by the 3′-OH of the newly liberated upstream exon. This joins the two exons and releases the lariat intron.

In vertebrates, *cis*-acting RNA signals that direct intron removal are only loosely conserved (6). RNA elements that participate include the 5′-ss, 3′-ss, BPS and a pyrimidine-rich stretch downstream of the BPS, the polypyrimidine tract (PPyT) (Fig. 1A). Additional elements within the exon and intron may contribute to the splicing of specific pre-mRNAs (7–10). snRNPs recognize the 5′-ss, 3′-ss and BPS through limited base pairing between the RNA component of the snRNP and the pre-mRNA (11–13). A subset of the protein components of the spliceosome act to assist the interaction between snRNAs and their target sequences (14,15). Other proteins involved in splicing participate in the selection of exons and in the organization of snRNPs during splicing (for a review see 16).

The PPyT does not interact directly with snRNPs but, rather, is recognized by one or more protein splicing factors. Furthermore, this element appears to have multiple roles in the splicing reaction. Early in the reaction the splicing factor U2AF binds to the PPyT and promotes the association of U2 snRNP with the adjacent BPS (15,17). A late function for the PPyT is implied by the observation that certain mutations within this region allow the first step to proceed but inhibit the second step (18,19). For example, Smith *et al*. (19) have shown that insertion of a stem–loop into the PPyT of an efficiently spliced pre-mRNA results in a specific block after the first step. Sequence requirements of the PPyT have been examined by a number of groups. Mutational analysis has demonstrated that insertion of purines into the PPyT generally has a negative effect on splicing; long stretches of U residues activate splicing more than other combinations of pyrimidines and a continuous stretch of five U residues is important for efficient splicing (20–22).

Retroviruses employ balanced splicing, amongst other mechanisms, to express multiple proteins from a single primary transcript (23). The unspliced mRNA encodes the structural and enzymatic proteins while a spliced mRNA encodes the envelope glycoproteins. In addition to providing essential proteins, the unspliced message is also packaged into the virion to serve as the viral genome. Simple retroviruses, such as avian sarcoma virus (ASV), do not encode proteins that regulate splicing and, therefore, this incomplete or ‘balanced’ splicing is mediated through interactions between the viral RNA and the host cell splicing machinery. We have found that this regulation can be manifested as a restriction or ‘partial block’ at either step in the splicing pathway (18). The determinants include sub-optimal signals (the BPS and PPyT), which in turn are responsive to the ASV *env* exonic splicing enhancer (ESE) (18,24,25).

The requirement for balanced splicing imposes strong selective pressure on retroviral splicing signals. Because it is relatively easy to manipulate retroviral DNA, large numbers of test

*To whom correspondence should be addressed. Tel: +1 215 728 3668; Fax: +1 215 728 2778; Email: r_katz@fccc.edu

sequences can be examined for functionality by selection for replication competence. We have previously demonstrated that the ratio of spliced and unspliced retroviral mRNAs produced in ASV-infected cells can be dramatically affected by U→C transitions that are outside a five U stretch within the ASV *env* PPyT. Here we employ the retroviral system to further examine the requirements of the PPyT, starting from DNA populations that include all possible combinations of pyrimidine sequences of a defined length. The method we describe for examining the PPyT could be modified to examine other splicing elements within either the intron or exon.

MATERIALS AND METHODS

PCR cloning and *in vivo* selection

The method of cloning the randomized PPyT into the ASV genome is described in the text and Figure 2. The *in vivo* selection scheme for viruses containing competent PPyTs is outlined in Figure 1B and described in the text. The oligonucleotide primers used in the first step of PCR cloning were: upstream primer, 5′-GTTATTCGCTTAAGCCTAGAG$(Y)_{8,10,12}$GCAGGCAGT-TCTGACTGGATA-3′ (the BPS and AG intron border are underlined); downstream primer, 5′-GACAGCTTATCATCGATA-3′. The second PCR step used the same downstream primer and a separate upstream primer: 5′-GGAAGCCGTCATAAAGG-3′.

In vivo analysis

Assays for the presence of reverse transcriptase (RT) activity in the supernatant of transfected cells were performed essentially as described (18). An aliquot (3 µl) of supernatant was mixed with poly(C) template, oligo(dG) primer and [α-^{32}P]dGTP and incubated at 37°C for 90 min. The reactions were spotted onto DEAE–Dextran paper and washed three times with 1× SSC and twice with 95% ethanol. The filter was then dried and subjected to autoradiography. S1 nuclease analysis and primer extension were performed as described previously (18).

Statistical analysis

A test of binomial proportions was used to compare the frequency of U residues at each position in either a 10mer or 12mer to the overall fraction of U residues in the collection. The binomial used is an approximation to the hypergeometric distribution, suitable for large numbers.

Fisher's exact test (FET) was used to compare proportions of pyrimidines in several contexts. The number of U residues in the selected 10 nt PPyTs was compared with that in the selected 12 nt PPyTs using FET, as well as comparison of the proportion of U residues in one position of the sequences to the same proportion in the rest of the positions.

Two sample T tests were used to compare the longest runs of C residues (or U residues) in one set of sequences with those in another set. This test compared the mean maximum run length in one group with that in the other. The Wilcox test on two samples was also used, with adjustments for ties, for the same purpose.

A permutation test was constructed to compare the longest runs of C residues (or U residues) in any viral sequence set with the lengths of such runs if the letters in each sequence were placed randomly. First, the average length of such runs was determined for a sequence set. Next, each sequence in the set was randomly permuted, the longest sequence of C residues (or U residues) was found within each sequence and their overall average length determined. The average length of the permuted sequences was compared with that of the non-permuted sequences. If the average length of the permuted sequences equaled or exceeded the average length of the non-permuted sequences in 50 (of 1000) or fewer cases then we concluded that the average run lengths of C residues (or U residues) was longer than expected, with a P value ≤5%.

We estimated the number of sequences that would function given the presence of a duplicate as described previously (26).

RESULTS

Construction of viral DNAs containing a randomized PPyT

We have previously reported that transitions within the *env* PPyT of ASV severely affected viral replication and altered the balance of RNA splicing (18). To examine the contribution of each pyrimidine within the PPyT, we established an *in vivo* system that selects PPyTs from a randomized pool based upon their ability to support balanced splicing and concomitant replication. The first step (Fig. 1A) was to use a PCR-based approach to incorporate randomized PPyTs into the proviral genome (Fig. 2). An oligodeoxynucleotide corresponding to the 3′-ss region, including the BPS and 3′-ss, was synthesized. To randomize the PPyT, equal amounts of U and C nucleotides were added during synthesis at each PPyT position. In order to test the effects of length as well as pyrimidine composition, three different oligonucleotides were synthesized that would create randomized PPyTs of 8, 10 and 12 nt in length. These oligodeoxynucleotides were incorporated into the ASV genome through the two-step PCR procedure described in Figure 2. The backbone viral DNA clone used to accept the randomized segment is a mutant version of the pLD6 ASV clone denoted IS1. The virus encoded by this clone contains a non-coding window spanning the PPyT, which allows the selection to be independent of any viral reading frame constraints (27). Because the final selection is for viable viruses, deleterious mutations that may accumulate during the PCR or cloning steps will not appear in the final population.

Bacteria were transformed with the library of viral plasmids containing the randomized PPyTs and a sample of the primary transformants was counted to determine the number of individual clones that made up the starting pool. Taking into account the number of unique PPyTs of each length (2^8, 2^{10} and 2^{12}), the likelihood that all possible combinations of pyrimidines within the PPyT exists in the starting pool can be calculated (Table 1). The results indicate that the starting pools should be sufficiently complex to test all PPyTs.

To estimate the randomness of the three pools, transformants representing libraries of DNA clones from each length of PPyT were pooled and the plasmid DNAs extracted. Each pool of DNA was subjected to automated DNA sequencing and the results showed that the sequences flanking the PPyTs were unique, while the PPyT was randomized as expected (data not shown). Within the PPyT, the peak traces showed a qualitative bias towards C residues, although the degree of bias is unclear because comparison of fluorescence intensities is not quantitative. To estimate the degree of randomness of the 12 and 10 nt PPyTs more accurately, bacteria were transformed with an aliquot of DNA from these two pools, 16 and 9 individual clones were isolated, respectively, and the region of the PPyTs in each clone was sequenced. The average percent of U residues present in the PPyTs of these clones is

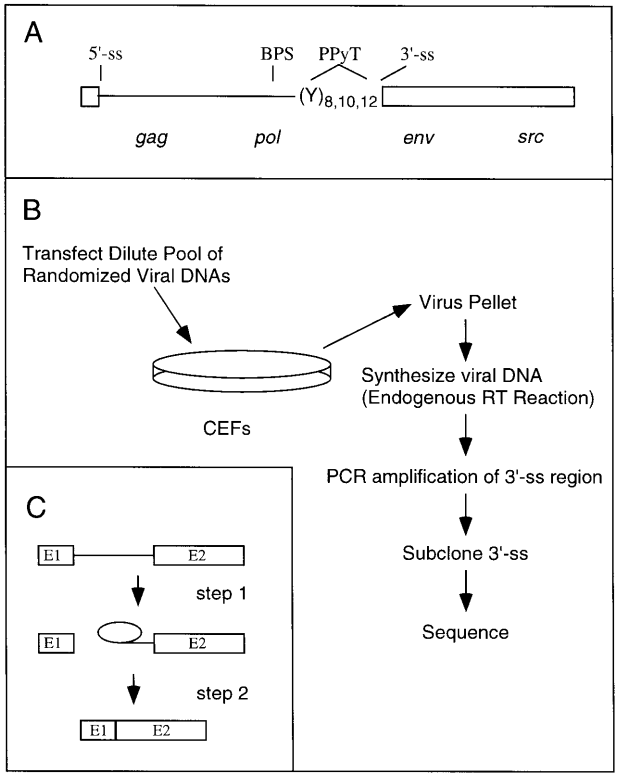


Figure 1. Schematic diagram of the ASV pre-*env* mRNA transcript and the selection protocol employed. (**A**) The intron and exon structure of the *env* transcript is depicted. PPyT that was subjected to randomization is indicated by $(Y)_{8,10,12}$. The viral genes, *gag*, *pol*, *env* and *src* are indicated below the genome and the locations of the splicing signals are listed above, including the 5′-splice site (5′-ss), branch point sequence (BPS), polypyrimidine tract (PPyT) and 3′-splice site (3′-ss). (**B**) A flow chart of the selection protocol is shown (see text for details). (**C**) Diagram of the two steps of the splicing reaction.

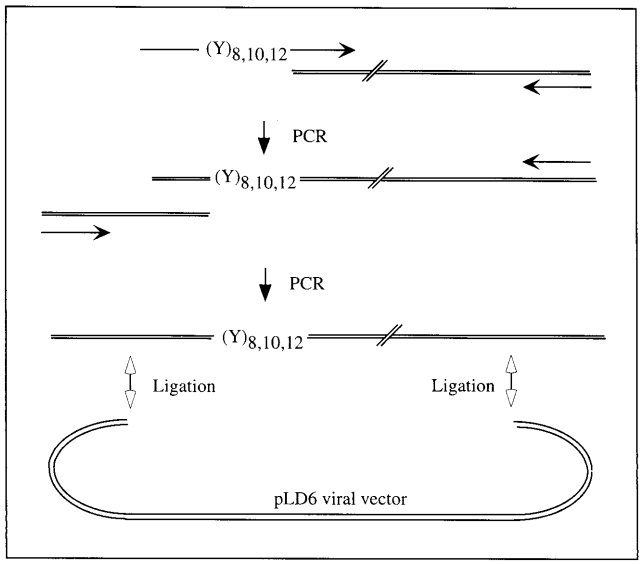


Figure 2. Generation of randomized viruses through a PCR cloning strategy. Small horizontal arrows indicate primers; a region of one primer was randomized and the three different lengths of randomized sequences are indicated by $(Y)_{8,10,12}$. This degenerate oligonucleotide contains unique sequences flanking the randomized region and was used to amplify a fragment of viral DNA corresponding to the downstream exon. A second PCR reaction was used to extend this fragment using sequences upstream of the randomized region. To reduce any potential bias introduced by PCR, the templates for each reaction did not contain sequences corresponding to the randomized region. Lastly, the 1 kb PCR product was ligated into the viral vector IS1 (see text) using unique restriction enzyme sites.

shown in Table 1. If the pools were completely random, then it is expected that the average number of U residues in the PPyTs would be 50%. We note that the starting pools contained ~40% U, suggesting that a bias toward C residues was introduced in our system. The 8 nt pool was not examined in this manner.

Table 1. Characterization of randomized libraries before selection

Length of Randomized PPyT	Distance From BPS to 3'-ss	Number of Clones in Randomized Pool	Predicted Complexity of PPyTs	Fold Over Represen-tation in Random Pool	Average Percent Uridine in Randomized Pool
12	18	16,000	4096	4	38
10	16	16,000	1024	16	41
8	14	16,000	256	64	nt

nt, not tested.

In vivo selection of competent PPyTs

In our selection protocol (Fig. 1B) transfection of chicken embryo fibroblast (CEF) tissue culture cells with DNA initiates the first round of viral replication. If the DNA does not encode a replication-competent virus then the virus will not spread beyond this point, consequently, only replication-competent virus will be amplified. To limit 'helper' effects which might allow more than one virus to be amplified in a single culture dish by functional complementation, we determined empirically the amount of DNA which results in virus appearance in only ~50% of the transfected plates (~10 ng/60 mm plate). Using this strategy, second round infections should be 'one hit' and any defective genomes should be lost, since no complementation will occur.

Following transfection, the cultures were maintained for 10 days to allow any replication-competent viruses to spread. After this period, supernatants were collected and the virus was pelleted and analyzed as outlined in Figure 1B. First, an endogenous reverse transcription reaction was performed to produce viral DNA (18). Next the 3′-ss region, including the BPS, PPyT and exonic splicing enhancer (ESE), was amplified by PCR (18). The amplified fragments were then subcloned and the 3′-ss region was subjected to sequence analysis.

To test for the occurrence of multiple viruses in a single tissue culture plate, several plasmid DNAs from the final subcloning step were isolated and sequenced. Six transfected plates were analyzed in this manner and from three to six clones derived from the same transfection were analyzed. Results from two out of the six transfections indicated that more than one virus was amplified in the same tissue culture plate. In one case, two clones from the same culture differed by a single nucleotide in the PPyT region. A likely interpretation is that a virus introduced from the starting pool had acquired a second mutation during the 10 day passage in tissue culture. In the second case, the two clones differed in sequence significantly, suggesting that two independent viruses were amplified to significant levels in the same culture. These results show that more than one replication-competent virus

Table 2. PPyTs selected *in vivo*

				Number of Uridines
13 nt Tracts				
1	AAGCCUAG:G	CCCUCCUUUCCCU	GCAGG	5
2	AAGCCUAGAG	UUCCCCCCCCCCU	GCAGG	3
3	AAGCCUAGAG	CUCCCCCCCUCCC	GCAGG	2
12 nt Tracts				
4	AAGCCUAGAG	UCCCUUUUUUCU	GCAGG	8
5	AAGCCUAGAG	CCUUUCUUCUCU	GCAGG	7
6	AAGCCUAGAG	UCCUCUUUCCUU	GCAGG	7
7	AAGCCUAGAG	UUCUUCCUCUCU	GCAGG	7
8	AAGCCUAGAG	UUUUCCCCCCUU	GCAGG	6
9	AAGCCUAGAG	UUUUCCCCCUCU	GCAGG	6
10	AAGCCUAGAG	UUUUCCCUCCUC	GCAGG	6
11	AAGCCUAGAG	UUUUCUUCCCCC	GCAGG	6
12	AAGCCUAGAG	UUUCCUCCCCUU	GCAGG	6
13	AAGCCUAGAG	UUCCCCUUCUUC	GCAGG	6
14	AAGCCUAGAG	UCCCCCUUUUUC	GCAGG	6
15	AAGCCUAGAG	UCCUCCCUCUUU	GCAGG	6
16	AAGCCUAGAG	UCCUUCCCUUCU	GCAGG	6
17	AAGCCUAGAG	UCUUCCUCCUCU	GCAGG	6
18	AAGCCUAGAG	CCUUCUUUCUCC	GCAGG	6
19	AAGCCUAGAG	CCUUUCUUCUCC	GCAGG	6
20	AAGCCUAGAG	CCUCCCUUCUUU	GCAGG	6
21	AAGCCUAGAG	UUUUCCCCCCCU	GCAGG	5
22	AAGC:UAGAG	UUCCCCCCUUCU	GCAGG	5
23	AAGCCUAGAG	UUCCCUCCCCUU	GCAGG	5
24	AAGCCUAGAG	UUCCUUCCCCUC	GCAGG	5
25	AAGCCUAGGG	UUCCUCUCCCUC	GCAGG	5
26	AAGCCUAGAG	UCUCUCUCCCCU	GCAGG	5
27	AAGCCUAGAG	UCUCCUCCUUCC	GCAGG	5
28	AAGCCUAGAG	UCUCCUCUCCCU	GCAGG	5
29	AAGCCUAGAG	UCCUCCUCUCCU	GCAGG	5
30	AAGCCUAGAG	UCCUUCCCUUCC	GCAGG	5
31	AAGCCUAGAG	CUUUUCCCCCCU	GCAGG	5
32	AAGCCUAGAG	CUCCUCCCCUUU	GCAGG	5
33	AAGCCUAGAG	CCUUCCCUUCCU	GCAGG	5
34	AAGCCUAGAG	CCUCCUUCCUCU	GCAGG	5
35	AAGCCUAGAG	CCUCCUUUCCUC	GCAGG	5
36	AAGCCUAGAG	CCCUCUCUUCCU	GCAGG	5
37	AAGCCUAGAG	CCCUUUCUUCCC	GCAGG	5
38	AAGCCUAGAG	CCCCUCUCCUUU	GCAGG	5
39	AAGCCUAGAG	UUUCCCCUCCCC	GCAGG	4
40	AAGCCUAGAG	UUCUCCCCCCUC	GCAGG	4
41	AAGCCUAGAG	UUCCUCUCCCCC	GCAGG	4
42	AAGCCUAGAG	UCUCCUCCCUCC	GCAGG	4
43	AAGCCUAGAG	UCUCCCUUCCCC	GCAGG	4
44	AAGCCUAGAG	UCCCUCUUCCCC	GCAGG	4
45	AAGCCUAGAG	UCCCCCCCUCUU	GCAGG	4
46	AAGCCUAGAG	CUUCCCCCCUCU	GCAGG	4
47	AAGCCUAGAG	CUUCCCCCUCUC	GCAGG	4
48	AAGCCUAGAG	CUCCCUCCUCUC	GCAGG	4
49	AAGCCUAGAG	CCUUCCCUUCCC	GCAGG	4
50	AAGCCUAGAG	CCUCCUCUCCCU	GCAGG	4
51	AAGCCUAGAG	CCCCUUCCCUUC	GCAGG	4
52	AAGCCUAGAG	CCCCCUCCUCUU	GCAGG	4
53	AAGCCUAGAG	CCCCCUUUCCCU	GCAGG	4
54	AAGCCUAGAG	UUCCCUCCCCCC	GCAGG	3
55	AAGCCUAGAG	UUCCCCUCCCCC	GCAGG	3
56	AAGCCUAGAG	UCCCCUCUCCCC	GCAGG	3
57	AAGCCUAGAG	UCCCCUCUCCCC	GCAGG	3
58	AAGCCUAGAG	UCCCCCCCUUCC	GCAGG	3
59	AAGCCUAGAG	CUCCCUCCUCCC	GCAGG	3
60	AAGCCUAGAG	CUUCUCCCCCCC	GCAGG	3
61	AAGCCUAGAG	CCCUCCCCCCUU	GCAGG	3
62	AAGCCUAGAG	CCCUCUUCCCCC	GCAGG	3
63	AAGCCUAGAG	CCCCUCUCCUCC	GCAGG	3
64	AAGCCUAGAG	CCCCCUCCUCCU	GCAGG	3
65	AAGCCUAGAG	CCCCCUUUCCCC	GCGGG	3
66	AAGCCUAGAG	CCCCCCUCUCUC	GCAGG	3
67	AAGCCUAGAG	CUCUCCCCCCCC	GCAGG	2
11 nt Tracts				
68	AAGCCUAGAG	CUCUUCCUCCU	GCAGG	5
69	AAGCCUAGAG	CUCUCCCUCCC	GCAGG	3
10 nt Tracts				
70	AGGCCUAGAG	UUUUUCUUCC	GCAGG	7
71	AAGCCUAGAG	UUUUCCUUCC	GCAGG	6
72	AAGCCUAGAG	UUCUUCUUCC	GCAGG	6
73	AAGCCUAGAG	CUCCUUUCUU	GCAGG	6
74	AAGCCUAGAG	CCUUUUUUCC	GCAGG	6
75	AAGCCUAGAG	UUUCCCCCUU	GCAGG	5
76	AAGCCUAGGG	UUCCUUCCCU	GCAGG	5
77	AAGCCUAGAG	CUCUUCCUUC	GCAGG	5
78	AAGCCUAGAG	CUCCUUCUCU	GCAGG	5
79	AAGCCUAGAG	CCUUUCUUCC	GCAGG	5
80	AAGCCUAGAG	UUUCCCCUCC	GCAGG	4
81	AAGCCUAGAG	UUCUCUCCCC	GCAGG	4
82	AAGCCUAGAG	UUCCUUCCCC	GCAGG	4
83	AAGCCUAGAG	UCCCUUCCCU	GCAGG	4
84	AAGCCUAGAG	CUCUCUUCCC	GCAGG	4
85	AAGC:UAGAG	CCCUCUCCUU	GCAGG	4

could be selected in the transfection step under our conditions, but, as expected, this is a relatively rare event.

Characterization of the PPyTs in replication-competent viruses

The selection procedure outlined above was performed several times for each of the viral DNA pools. In total, ~150 100 mm plates were transfected and replication-competent virus appeared in ~80 plates. Sequences of the PPyTs from viruses selected for replication competence are listed in Table 2. We focused on the 12 nt tracts and representative results that were obtained from the 10 and 8 nt tracts. A total of 86 viruses were selected in the three experiments and the 3′-ss of each was sequenced and a summary is presented in Table 3.

Because the selection procedure is dependent upon viral replication and retroviruses have high mutation rates, it is expected that mutations outside the PPyT might be found in the virus populations at the end of the selection procedure. Additional mutations could also be introduced during the various PCR steps.

Table 3. Characteristics of selected PpyTs

Length of Random PPyT	Total Number of PPyTs Selected	Number of Duplicate PPyTs Selected	Number of Aberrant Tracts Selected	Average Percent Uridine After Selection
12	68	1	5	40
10	14	0	3	49
8	4	0	4	nt

nt, not tested.

To estimate the occurrence of mutations outside the PPyT we examined sequences immediately surrounding the PPyT as well as sequences upstream of the BPS and downstream of the 3′-ss region. Analysis of the 10 and 12 nt tracts revealed that 8 out of the 85 viruses selected *in vivo* contained mutations outside the PPyT (Table 3). One secondary mutation that is not shown in Table 2 consists of a single base change in the *pol* gene, 51 nt

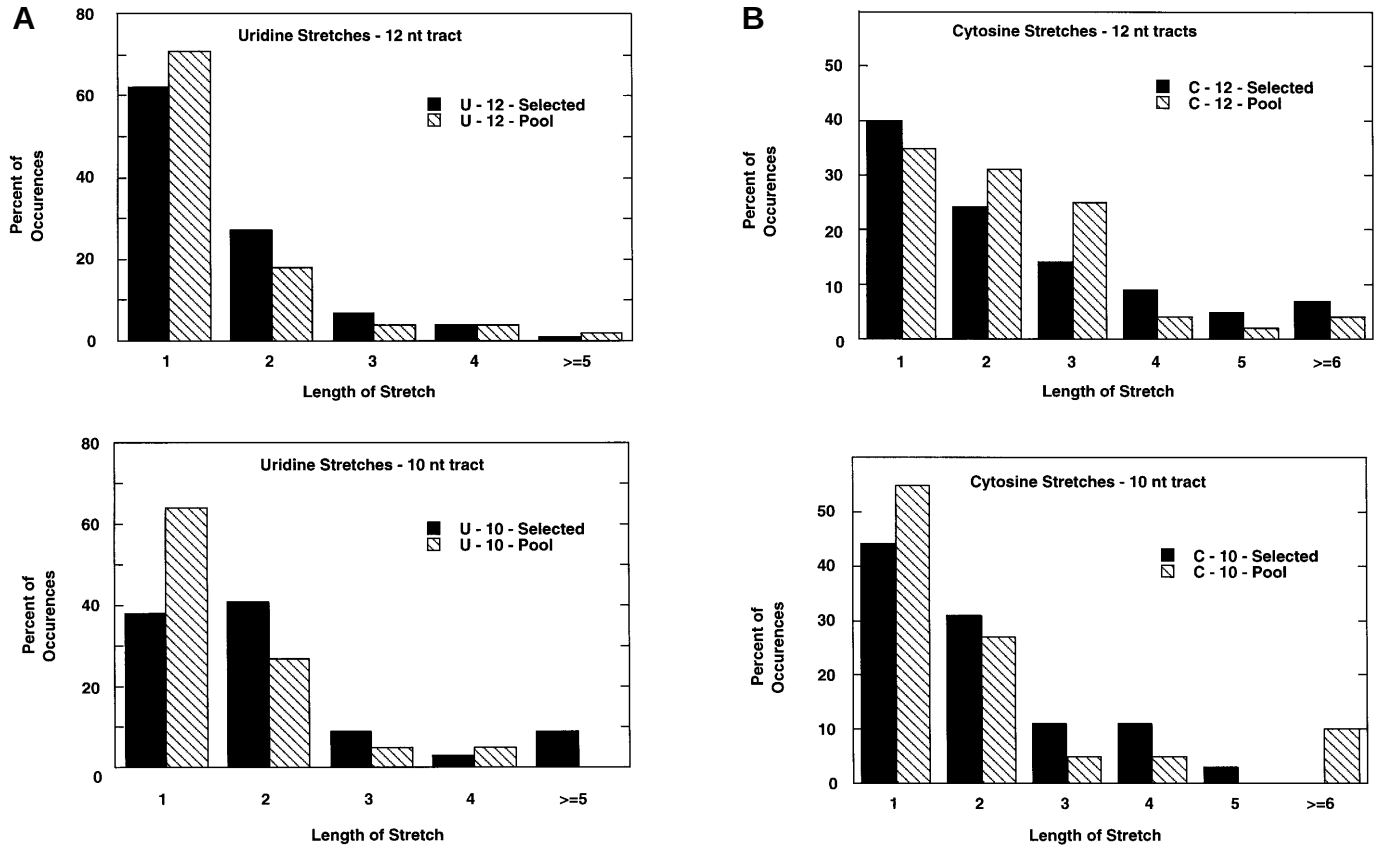


Figure 3. Occurrence of U and C stretches in PPyTs. PPyTs from the *in vivo* selected viruses and the unselected pools of DNA were examined for the presence of groups of nucleotides. (**A**) Groups of U residues are shown for the 10 and 12 nt randomizations. (**B**) Groups of C residues are shown for the 10 and 12 nt randomizations.

upstream of the BPS (tract 46). This change introduces a stop codon into the integrase reading frame and the virus is predicted to express a protein that is truncated at the C-terminus by 16 amino acids. Apparently, this mutation does not significantly affect viral replication (see below). Together, the 10 and 12 nt selection experiments produced 81 tracts, of which 8 differed from the input material, suggesting that the rate of occurrence of secondary mutations is <10%.

Two of the isolated tracts were identical (Table 2, tracts 56 and 57), which allows us to estimate the minimum number of 12 nt tracts that should be capable of supporting balanced splicing and viral growth at the 95% confidence level to be 466 (26). The maximum likelihood estimate is 2134, indicating that about half of all combinations of pyrimidines (4096) are likely to support balanced splicing in this system. Because no duplicate tracts were identified in the 14 PPyTs from *in vivo* selected viruses containing 10 nt PPyTs, no estimate can be made concerning the likely number of competent tracts that exist.

Transfections using the 8 nt PPyT pool never produced replication-competent viruses containing tracts of 8 nt. Of the four viruses that were isolated, two tracts were 10 nt long, one was 11 and one 12. We also note that a significantly larger amount of DNA was needed to initiate viral infection (1 µg/60 mm plate) and that the percentage of plates that produced viruses was also reduced compared with the 10 and 12 nt tracts (~10% versus ~50%). These results suggest that 8 nt is below the minimal distance between the BPS and 3′-ss. Because the 10 nt tracts were the shortest isolated, we conclude that the minimal length of the PPyT in this system is 10 nt and the corresponding minimal distance between the BPS and 3′-ss is 16 nt.

Table 3 shows the composition of pyrimidines within the PPyTs of viruses selected *in vivo*. We have focused on the U composition because previous studies indicated that the number of U residues, as well as U tracts, within the PPyT are important functional determinants (20–22). The average percent U within the PPyTs of viruses selected *in vivo* was compared with the percentage determined for the starting pool of randomized DNA (Table 1). In the 12 nt tracts, the percent of U residues was not significantly different between the selected viruses and the starting pool of DNA. However, for the 10 nt tracts that were selected *in vivo* the percentage of U residues was significantly higher than for the 12 nt tracts that were selected, with a *P* value of 0.011 (see Materials and Methods). This suggests that to achieve approximately the same level of splicing, the shorter tract requires a higher U content. This is consistent with other reports which show that the strength of a PPyT is related to length and U content.

Published reports demonstrated the importance of a five U stretch in the PPyT and thus we analyzed the starting pools of DNA and the selected viruses for such stretches (Fig. 3A). The numbers of four, three, two and single U arrangements were also determined. To determine the significance of these comparisons, a computer simulation was used for the generation of a random sampling (Materials and Methods). Importantly, the array of U and C stretches in the starting pool sequences did not differ significantly from a purely random sample, indicating both that the starting pool contained a wide variety of PPyTs and that a

Name	BPS	PPyT	3'-SS	Uridines
Wild Type	AGGC AGGCGAG	CCCUCUUUUU	GCAG	6
IS1	UAAG CCUAGAG	CCCUCUUUUU	GCAG	6
10Y-6U	UAAG CCUAGAG	CCUUUUUUCC	GCAG	6
10Y-7U	UAGG CCUAGAG	UUUUUCUUCC	GCAG	7
12Y-2U	UAAG CCUAGAG	CUCUCCCCCCCC	GCAG	2
12Y-4U	UAAG CCUAGAG	CUUCCCCCCUCU	GCAG	4
12Y-6U-1	UAAG CCUAGAG	UUUUCCCCCCUU	GCAG	6
12Y-6U-2	UAAG CCUAGAG	UCCCCCUUUUUC	GCAG	6
12Y-7U	UAAG CCUAGAG	CCUUUCUUCUCU	GCAG	7

Figure 4. Description of parental and *in vivo* selected 3′-ss regions examined for functional capability. The BPS, PPyT and 3′-ss are listed for the viruses that we examined in more detail after selection. The sequences of two previously characterized viruses are listed, wild-type and IS1. The viruses derived from the selection procedure are listed in two groups; one group contains PPyTs of 10 nt in length, the other 12 nt. The number of U residues within each tract is listed.

comparison with these sequences is valid. Comparisons between the tracts of selected viruses and the starting pool of DNA (12 nt PPyT) showed no significant difference in the occurrence of U stretches of any length (Fig. 3A, top). Similar examination of the 10 nt tracts revealed that the PPyTs of selected viruses contain more U stretches of ≥5 nt and fewer single U residues than the starting pool. Due to the small sample size of the 10 nt tracts, only the reduction of single U residues is significant (P = 0.054). However, this could be interpreted as a consequence of an increased occurrence of longer U stretches.

Identical comparisons were made for C stretches in Figure 3B. In this case we find that the selected 12 nt tracts contain an increased length of C stretches compared with the starting pool (P = 0.064). The 10 nt tracts show a shift away from C stretches of >5 nt, although the significance of this is low. Together these data suggest that stretches of five or more U residues may be important for balanced splicing in the 10 nt tracts and that longer stretches of C residues are important for balanced splicing in the 12 nt tracts.

Each position in the tracts was also examined for U or C occurrence. Comparisons of 10 nt tracts from selected viruses with the starting pools of DNA revealed a preference for U at the second position downstream of the BPS (P = 0.028). The viruses containing 12 nt PPyTs demonstrated a preference for U at the first position downstream of the BPS (P = 0.013).

In vivo analysis of viruses containing selected PPyTs

Because the selection procedure produced a wide variety of PPyTs, we decided to examine a representative subset more closely. A total of 10 different PPyTs were chosen based upon their U content and the 3′-ss regions of these isolates were re-inserted into the original infectious viral DNA clone. Three of these reconstructed viral DNAs either failed to produce virus or acquired additional mutations within the PPyT during further passage (data not shown). This suggests that ~30% of the selected viruses are unstable or contain additional mutations outside the 3′-ss region that allowed them to grow in tissue culture. We have not investigated further the viruses that failed to replicate.

Sequences in the 3′-ss region of the seven reconstructed clones that produced virus are shown in Figure 4. The sequences of the wild-type virus as well as the parent IS1 are included for comparison. Two of the reconstructed viruses contained alterations outside the PPyT. The 10Y-7U virus contains an A→G transition at the first position upstream of the BPS and the 12Y-4U virus contains a single base change in the *pol* gene, as discussed earlier.

PPyTs selected *in vivo* allow efficient viral replication

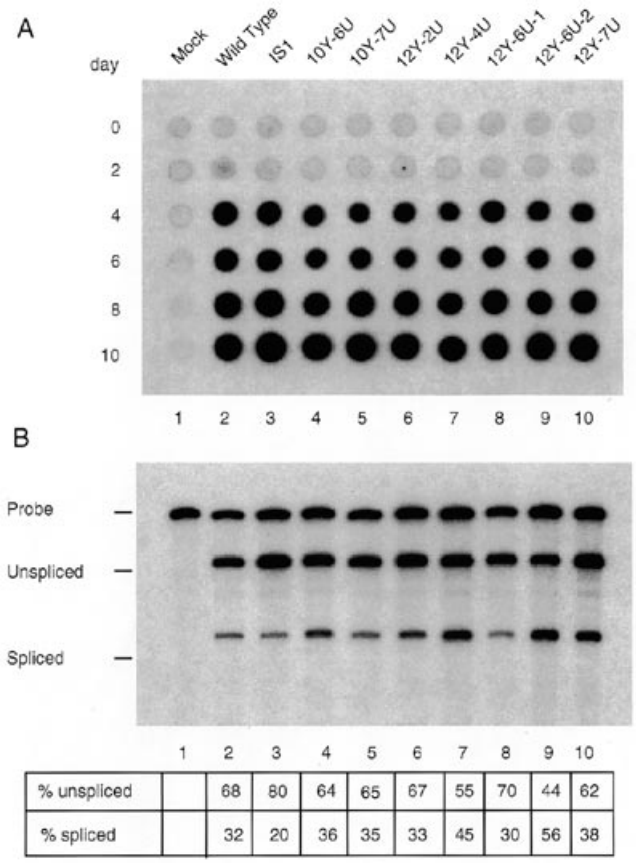
Figure 5A shows the appearance of progeny virus after transfection with each of the reconstructed viral DNA clones as determined by a standard RT assay. The detection of RT activity requires virus spread by re-infection and the level of RT production reflects the efficiency of replication. No RT activity was detected in a mock-transfected tissue culture plate (lane 1). All of the DNA clones tested gave rise to progeny virus at a similar rate to the wild-type. At the end of the 10 day passage in tissue culture, the sequences of the viral populations were determined. The original sequences of the 3′-ss of all of these viruses were maintained during passage (data not shown).

Previously we have shown a correlation between balanced splicing and viral growth. To examine the balance of splicing in the *in vivo* selected viruses, we isolated RNA from infected cells after the 10 day passage shown in Figure 5A. The relative amounts of spliced and unspliced viral transcripts was determined by S1 analysis. By using a probe that hybridizes to the 5′-ss, the relative amount of spliced and unspliced RNA can be determined (Fig. 5B). Radioactivity was quantitated by phosphorimaging and the percentages of spliced and unspliced message detected are indicated beneath each lane. The wild-type virus demonstrates the characteristic pattern; approximately one-third of the viral transcripts are spliced. As expected, IS1 shows slightly reduced splicing as compared with the wild-type (25). The *in vivo* selected viruses display a narrow range of splicing percentages. Most are very similar to the wild-type, with approximately one-third of the RNA spliced. Two viruses, 12Y-4U and 12Y-6U-2 (lanes 7 and 9), show significantly more splicing (~50%). These splicing efficiencies are within the acceptable window for viability as determined previously.

PPyTs selected *in vivo* affect splicing at either step 1 or step 2

The control of balanced *env* splicing in the ASV wild-type and variants has been shown to involve either step 1 or step 2 of the splicing reaction (18). To examine potential differential effects on the two steps by the selected PPyTs, we assayed for the presence of lariat–exon 2 intermediates *in vivo* by primer extension using RNA extracted from infected cells (Fig. 6). The results show that no extension products are produced with RNA from mock-infected cells, whereas a number of extension products are produced with RNA from wild-type-infected cells. These extension products are derived either by run-off transcription to the 5′-end of the unspliced viral transcripts, by premature termination due to pausing of RT or from partially degraded RNA templates. RNA from cells infected with the IS1 virus gives rise to the same extension products as the wild-type and, in addition, a unique product that corresponds to a lariat–exon 2 intermediate. An exon primer was used to distinguish between free lariats and lariat–exon 2 intermediates (data not shown). We have previously shown that IS1 accumulates lariat–exon 2 intermediates, which is characteristic of regulation at step 2 (18,24).

RNAs from cultures infected with viruses containing *in vivo* selected PPyTs were analyzed in a similar manner. Some showed a pattern similar to the wild-type while others showed accumulation of lariat–exon 2 intermediates. Both of these types of splicing regulation are observed with the 10 and 12 nt tracts tested. We note that the relative amount of lariat–exon 2 detected varies among viruses. For example, 12Y-7U (lane 10) reproducibly shows greater accumulation of lariat–exon 2 intermediate relative to unspliced than the IS1 parental virus. Conversely, other viruses show an



% unspliced		68	80	64	65	67	55	70	44	62
% spliced		32	20	36	35	33	45	30	56	38

Figure 5. Growth rate and splicing pattern of selected viruses *in vivo*. (**A**) The appearance of RT activity in the supernatant of transfected cells was monitored as described (Materials and Methods). 'Mock' indicates cells transfected with vector DNA containing no viral sequences. (**B**) Total cellular RNA from chronically infected cells was hybridized with a DNA probe spanning the 5′-ss and then subjected to S1 nuclease digestion. The bands corresponding to spliced and unspliced RNA were quantitated by phosphorimaging and the percentage of each species is listed below the lanes. The 'mock' lane indicates RNA isolated from a mock transfection.

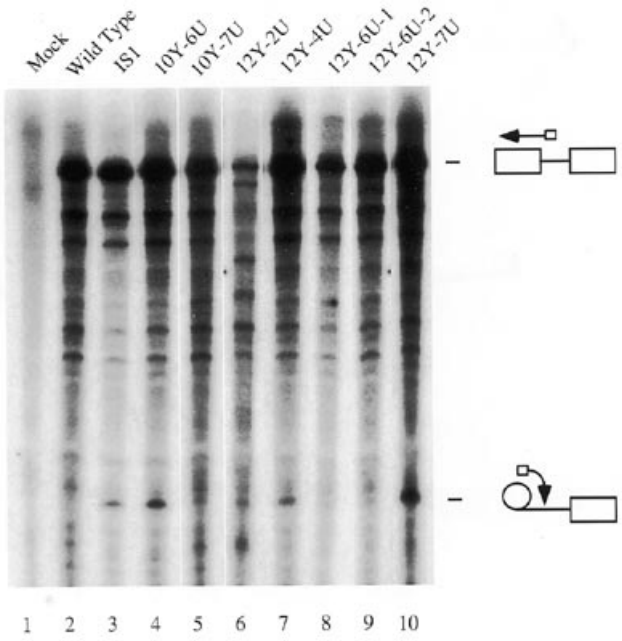


Figure 6. Detection of lariat–exon 2 intermediates *in vivo*. Total cellular RNA from infected cells was examined by primer extension assay using a labeled primer that hybridizes to intronic sequences as indicated in the diagrams to the right. 'Mock' indicates RNA from mock transfected cells.

intermediate level of lariat–exon 2 intermediates (less than IS1, but more than wild-type). The significance of this variation is unclear.

DISCUSSION

In this report we describe the use of a retroviral system to selectively amplify sequences that support balanced splicing from a pool of randomized sequences. This system selects for splice sites that allow weak but functional splicing. As weak splice sites are also a prerequisite for alternative splicing of cellular pre-mRNAs, the results of our studies are applicable to regulation of host pre-mRNA splicing. The experiments described here focus on analysis of the PPyT. The PPyT at the *env* 3′-ss in ASV was altered in length, randomized with respect to C and U residues and replication-competent isolates were obtained. Previously, we found that this PPyT was sensitive to subtle changes; addition of two pyrimidines activated splicing and abrogated viral growth, whereas a single transition within this extended PPyT reduced the amount of spliced product and restored viral growth (18). Thus, the results from our randomization experiments which show that many different PPyTs can support viral growth were unexpected. The diversity of the PPyTs of selected viruses may result from a number of participating factors. Several proteins are predicted to interact with the PPyT and the binding preferences of these proteins is likely to be degenerate (28–32). The wide variety of PPyTs selected in our experiments could reflect different balances in such interactions. Another source of PPyT diversity may result from the fact that different residues in the tract contribute to different steps of the splicing reaction. For example, we have shown that the PPyT can play a role in either the first or second step of the splicing reaction (18).

A close examination of the functional PPyTs selected *in vivo* revealed some trends. For example, shorter PPyTs contained more U residues on average than the longer ones. *In vitro* experiments have demonstrated that increasing the U content of a PPyT will increase usage of the corresponding 3′-ss. This has also been addressed *in vivo* through analysis of two competing 3′-splice sites (33). It was noted that the proximal PPyT was shorter and contained fewer U residues than the distal PPyT. Interchanging the two PPyTs or increasing the length and U content of the proximal PPyT resulted in its activation. In our system, the competition is a splice/no splice choice. In addition, all of the selected PPyTs produced similar ratios of spliced and unspliced products, which allows us to make comparisons between different lengths of PPyTs with all else being equal.

Because we were unable to isolate viruses containing shorter PPyTs, we conclude that the minimal length of a PPyT in our system is 10 nt. One interpretation of this result is that when the BPS and 3′-ss are in close proximity, the splicing machinery is unable to identify them and the splice site is skipped. The 10 nt tract corresponds to a distance between the branch point adenosine and the 3′-ss of 16 nt. Reed (22) has suggested the minimal distance between BPS and 3′-ss to be 16 nt based upon sequence comparisons. Our results provide confirmation of this suggestion.

Comparison of the stretches of U or C residues in the PPyTs of viruses selected *in vivo* reveals that the 10 nt tracts contained more long U stretches than the starting pool and the 12 nt tracts contained longer C stretches. We conclude that the selective pressure with respect to the shorter tracts was different from that on the longer tracts. One interpretation of the data is that the selection in the 10 nt tracts was based upon the presence of longer U stretches and the selection in the 12 nt tracts was for longer C stretches. Because the selection requires maintenance of balanced splicing, the presence of a higher U content in the 10 nt tracts may suggest that recruitment of positive factors can compensate for the short distance between the BPS and 3′-ss. The effect of longer C stretches within the PPyT has not been reported previously and the role is unclear.

Further analysis demonstrated a preference for U at the second position downstream of the BPS for the 10 nt tract and at the first position downstream of the BPS for the 12 nt tract. The meaning of these preferences is not clear, but it may indicate a requirement for U residues at the 5′-end of the PPyT. This has been suggested by others through *in vitro* mutagenesis studies and our data may provide *in vivo* confirmation of these results (20).

In earlier studies, we described the control of balanced splicing in ASV as a partial block in the splicing reaction. This block allows some unspliced transcripts to escape splicing to the cytoplasm, but sufficient RNA is spliced to allow appropriate production of *env* mRNA (18,24). We have shown that a block may occur at two different stages within the splicing pathway, either before step 1 or during step 2. The step at which the block occurs in infected cells is reflected in the presence or absence of lariat–exon 2 intermediates; presence of these intermediates is taken as indicative of a block at step 2, while absence indicates a block at step 1. When we analyzed a subset of the PPyTs from selected viruses for the presence of lariat–exon 2 intermediates, we find that some display this intermediate *in vivo*, while others do not. Because these viruses differ primarily in the composition of pyrimidines within their PPyTs, this supports a role for the PPyT in the second, as well as the first, step of splicing.

Previously, we have shown that in our system, *in vivo* selected suppressor mutations in the BPS and PPyT affect *in vitro* splicing as well as UV crosslinking of two splicing factors, SAP49 and U2AF^{65} (34). Similar analysis was carried out with the selected PPyTs described here. Preliminary results indicate that the crosslinking patterns of these two factors are also affected when the various selected PPyTs are compared (data not shown).

Other groups have used *in vivo* selection procedures to examine elements important for splicing. Studies by Chen and Chasin examined exon skipping in the dihydrofolate reductase gene (35). This negative selection procedure uncovered inactivating mutations in the PPyT, BPS and 3′-ss. Coulter *et al*. (36) examined exonic sequences which could activate splicing, starting with a pool of DNA that was randomized at a short stretch within the downstream exon. Through iterative rounds of transfection, amplification of spliced product and reconstruction of unspliced template, a number of sequences that promote splicing at the upstream intron were selected. This procedure is limited to examination of sequences within the exon. The retroviral system described here is capable of identifying weak but functional splicing elements that reside within an intron. With minor modifications, this same retroviral system procedure can be used to examine any element important for splicing.

ACKNOWLEDGEMENTS

We would like to thank Warren Kruger and John Taylor for critical comments on the manuscript. We also thank the Fox Chase Cancer Center Fannie E. Ripple Foundation Biotechnology Core Facility for preparation of DNA primers. This work was supported by National Institutes of Health grants CA71515, AI40385 and AI40721, institutional grant CA06927 from the National Institutes of Health and also by an appropriation from the Commonwealth of Pennsylvania. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute or any other sponsoring organization.

REFERENCES

1 Moore,M.J., Query,C.C. and Sharp,P.A. (1993) In Gesteland,R.F. and Atkins,J.F. (eds), *The RNA World*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2 Inoue,K., Ohno,M. and Shimura,Y. (1995) *Gene Exp.*, **4**, 177–182.
3 Rio,D.C. (1993) *Curr. Opin. Genet. Dev.*, **3**, 574–584.
4 Adams,M.D., Rudner,D.Z. and Rio,D.C. (1996) *Curr. Opin. Cell Biol.*, **8**, 331–339.
5 Chabot,B. (1996) *Trends Genet.*, **12**, 472–478.
6 Senapathy,P., Shapiro,M.B. and Harris,N.L. (1990) *Methods Enzymol.*, **183**, 252–278.
7 Carlo,T., Sterner,D.A. and Berget,S.M. (1996) *RNA*, **2**, 342–353.
8 Arrigo,S., Yun,M. and Beemon,K. (1987) *Mol. Cell. Biol.*, **7**, 388–397.
9 Amendt,B.A., Hesslein,D., Chang,L.J. and Stoltzfus,C.M. (1994) *Mol. Cell. Biol.*, **14**, 3960–3970.
10 Stoltzfus,C.M. and Fogarty,S.J. (1989) *J. Virol.*, **63**, 1669–1676.
11 Mount,S.M., Pettersson,I., Hinterberger,M., Karmas,A. and Steitz,J.A. (1983) *Cell*, **33**, 509–518.
12 Zhuang,Y. and Weiner,A.M. (1989) *Genes Dev.*, **3**, 1545–1552.
13 Zhuang,Y. and Weiner,A.M. (1986) *Cell*, **46**, 827–835.
14 Staknis,D. and Reed,R. (1994) *Mol. Cell. Biol.*, **14**, 2994–3005.
15 Valcarcel,J., Gaur,R.K., Singh,R. and Green,M.R. (1996) *Science*, **273**, 1706–1709.
16 Kramer,A. (1996) *Annu. Rev. Biochem.*, **65**, 367–409.
17 Lee,C.G., Zamore,P.D., Green,M.R. and Hurwitz,J. (1993) *J. Biol. Chem.*, **268**, 13472–13478.
18 Bouck,J., Fu,X.-D., Skalka,A.M. and Katz,R.A. (1995) *Mol. Cell. Biol.*, **15**, 2663–2671.
19 Smith,C.W., Porro,E.B., Patton,J.G. and Nadal-Ginard,B. (1989) *Nature*, **342**, 243–247.
20 Coolidge,C.J., Seely,R.J. and Patton,J.G. (1997) *Nucleic Acids Res.*, **25**, 888–896.
21 Roscigno,R.F., Weiner,M. and Garcia,B.M. (1993) *J. Biol. Chem.*, **268**, 11222–11229.
22 Reed,R. (1989) *Genes Dev.*, **2**, 2113–2123.
23 Coffin,J.M. (1996) In Fields,B.N., Knipe,P.M. and Howley,P.M. (eds), *Fields Virology*, 3rd Edn. Lippincott-Raven, Philadelphia, PA, Vol. 2, pp. 1767–1847.
24 Fu,X.-D., Katz,R.A., Skalka,A.M. and Maniatis,T. (1991) *Genes Dev.*, **5**, 211–220.
25 Katz,R.A. and Skalka,A.M. (1990) *Mol. Cell. Biol.*, **10**, 696–704.
26 Litwin,S. and Shlochik,M. (1990) *J. Exp. Med.*, **171**, 293–297.
27 Katz,R.A., Kotler,M. and Skalka,A.M. (1988) *J. Virol.*, **62**, 2686–2695.
28 Ruskin,B., Zamore,P.D. and Green,M.R. (1988) *Cell*, **52**, 207–219.
29 Patton,J.G., Mayer,S.A., Tempst,P. and Nadal-Ginard,B. (1991) *Genes Dev.*, **5**, 1237–1251.
30 Patton,J.G., Porro,E.B., Galceran,J., Tempst,P. and Nadal,G.B. (1993) *Genes Dev.*, **7**, 393–406.
31 Gil,A., Sharp,P.A., Jamison,S.F. and Garcia,B.M. (1991) *Genes Dev.*, **5**, 1224–1236.
32 Singh,R., Valcarcel,J. and Green,M.R. (1995) *Science*, **268**, 1173–1176.
33 Dominski,Z. and Kole,R. (1991) *Mol. Cell. Biol.*, **11**, 6075–6083.
34 Bouck,J., Fu,X.-D., Skalka,A.M. and Katz,R.A. (1998) *J. Biol. Chem.*, **273**, 15169–15176.
35 Chen,I.T. and Chasin,L.A. (1993) *Mol. Cell. Biol.*, **13**, 289–300.
36 Coulter,L.R., Landree,M.A. and Cooper,T.A. (1997) *Mol. Cell. Biol.*, **17**, 2143–2150.