2426–2432 *Nucleic Acids Research, 1998, Vol. 26, No. 10*

Analysis of sequences and predicted structures required for viral satellite RNA accumulation by *in vivo* genetic selection

Clifford D. Carpenter and Anne E. Simon*

Department of Biochemistry and Molecular Biology, University of Massachusetts, Amherst, MA 01003, USA

Received January 13, 1998; Revised and Accepted March 23, 1998

ABSTRACT

***In vivo* genetic selection was used to study the sequences and structures required for accumulation of subviral sat-RNA C associated with turnip crinkle virus (TCV). This technique is advantageous over site-specific mutagenesis by allowing side-by-side selection from numerous sequence possibilities as well as sequence evolution. A 22 base hairpin and 6 base single-stranded tail located at the 3′-terminus of sat-RNA C were previously identified as the promoter for minus strand synthesis. Approximately 50% of plants co-inoculated with TCV and sat-RNA C containing randomized sequence in place of the 22 base hairpin accumulated sat-RNA in uninoculated leaves. The 22 base region differed in sat-RNA accumulating in all infected plants, but nearly all were predicted to fold into a hairpin structure that maintained the 6 base tail as a single-stranded sequence. Two additional rounds of sat-RNA amplification led to four sequence family 'winners', with three families containing multiple variants, indicating that evolution of these sequences was occurring in plants. Three of the four sequence family winners had the same 3 bp at the base of the stem as wild-type sat-RNA C. Two of the winners shared 15 of 22 identical bases, including the entire stem region and extending two bases into the loop. These results demonstrate the utility of the *in vivo* selection approach by showing that both sequence and structure contribute to a more active 3′-end region for accumulation of sat-RNA C.**

INTRODUCTION

Plus (+) strand RNA virus replication by RNA-dependent RNA polymerases (RdRp) proceeds through a complementary minus (–) strand intermediate followed by synthesis of a copy of the (+) strand. This process requires promoters on the (+) and (–) strand RNAs that allow the RdRp to selectively amplify its cognate RNA. In addition to promoters responsible for full-length (–) strand synthesis, internal RdRp promoters located on the (–) strand intermediate direct the synthesis of 3′-co-terminal subgenomic RNAs that serve as mRNAs for downstream open reading frames.

Promoters for RdRp have been localized by deletion analysis and structural and sequence determinants analyzed by 'reverse genetics'. In this process, mutations that disrupt sequence or structural elements are generated in full-length transcripts and biological activity of mutant templates are assessed either in the whole organism, in cell culture or *in vitro*. The high error rate of RdRp, estimated at 10^{-4} (1), can lead to additional alterations or reversions that increase the biological fitness of weak mutated promoters (2–4). Using such techniques a wide variety of single and multiple hairpins have been identified as important promoter elements (5). In addition, tertiary structural interactions that help maintain tRNA-like structures (6), promote interactions ('kissing') between separated hairpins (2) and form elements such as pseudoknots (7,8) have been identified as important features of some RdRp promoters. Short (11–20 base) primary sequences without obvious secondary structures formed by canonical base pairing have also been identified on (–) strands as promoters for subgenomic RNA synthesis of brome mosaic virus (9) and full-length synthesis of a subviral RNA of turnip crinkle virus (TCV; 10).

TCV, with its associated subviral satellite RNAs (sat-RNAs), has proven to be a useful model for studying promoters required for amplification of RNA (3,10–12) and subgenomic RNA synthesis (13). TCV is a single component, (+) stranded RNA virus of 4054 bases (14,15) that is associated with sat-RNAs of 194 (sat-RNA D) and 356 bases (sat-RNA C). Sat-RNA C is a chimeric RNA composed of nearly full-length sat-RNA D at the 5′-end and two regions of TCV genomic RNA at the 3′-end (16) and requires the helper virus for amplification in host cells. A combination of *in vivo* and *in vitro* approaches has led to identification of the 3′-terminal 29 bases of sat-RNA C as the promoter for (–) strand synthesis. This promoter is composed of a 22 base hairpin and a 6 base single-stranded 3′-terminal tail (12). *In vitro* (12) and *in vivo* (3) analyses of the hairpin using site-specific mutagenesis suggested that while a hairpin is required for biological activity, the primary sequence of the loop and stem are of limited importance.

Results obtained using site-specific mutagenesis to establish the importance of primary sequence and secondary/tertiary structures in promoter sequences of RNA templates are limited by the difficulty

*To whom correspondence should be addressed. Tel: +1 413 545 0170; Fax: +1 413 545 4529; Email: simon@biochem.umass.edu

of attempting all possible combinations of nucleotides at each position. *In vitro* genetic selection, also known as SELEX (systematic evolution of ligands by exponential enrichment) (17,18), can be used to circumvent such limitations by allowing simultaneous analysis of large numbers of randomized nucleotide combinations that have high affinity for specific nucleic acid binding proteins or other target molecules (19). The complexity of the nucleotide population decreases in each round of selection, with 'winners' emerging in the final round representing an enriched population of molecules that have outperformed competing molecules. *In vitro* SELEX has been used to analyze sequences that bind to alfalfa mosaic virus coat protein (20), which is required for amplification of the virus *in vivo*, and templates that bind to coliphage Qβ replicase and that contain promoters for RNA amplification by the Qβ replicase (21,22). The Qβ system is particularly suited to analysis of sequences/structures required for viral replication by *in vitro* SELEX, since purified highly active Qβ replicase is able to exponentially amplify promoter-containing RNAs (23). Unfortunately, highly efficient exponential amplification of RNA genomes has not been achieved for other RNA viruses, thereby limiting promoter characterization by *in vitro* SELEX.

Recently, randomized selection approaches have been applied to *in vivo* studies to identify RNA–RNA interactions required for splicing in *Saccharomyces cerevisiae* (24,25) and iterative randomized selection (more similar to *in vitro* SELEX with multiple rounds of selection) used to characterize RNA nuclear import signals in *Xenopus laevis* oocytes (26) and exonic splicing enhancers in quail fibroblast cultures (27). We now report that iterative randomized selection combined with natural *in vivo* evolution can be used to analyze sequences required for amplification of viral-associated RNAs. Results obtained using this approach strongly suggest that both sequence and structure of the hairpin promoter at the 3′-end of sat-RNA C (+) strands contribute to efficient RNA amplification of sat-RNA C.

MATERIALS AND METHODS

Generation of randomized templates for *in vitro* transcription

Oligonucleotides T7C5′ (GTAATACGACTCACTATAGGGA-TAACTAAGGG) and $C_{313\text{-}327}$ (TATCTATTGGTTCGG) were used as primers in a polymerase chain reaction (PCR) with pT7C+, a full-length wild-type cDNA clone of sat-RNA C (11), to generate a cDNA product containing a T7 RNA polymerase promoter upstream of a sat-RNA C sequence truncated by 29 bases at the 3′-end. Standard PCR conditions (50 µl volume) contained 1 ng template plasmid, 25 pmol each oligonucleotide, 1 U pyrostase enzyme (Molecular Genetics Resources, Tampa, FL) and buffer supplied by the manufacturer. Optimal conditions were 50 cycles of PCR at 93, 32 and 72°C for 1 min each, with an additional unit of enzyme added after the 25th cycle. After phenol extraction and gel purification the cDNA was used in a second PCR with the T7C5′ oligonucleotide and a 44mer (positions 313–356 of sat-RNA C) containing 22 randomized nucleotides in positions 328–349 (Fig. 1B). cDNA products of the second PCR contained a T7 RNA polymerase promoter upstream of full-length sat-RNA C containing 22 randomized nucleotides between positions 328 and 349.

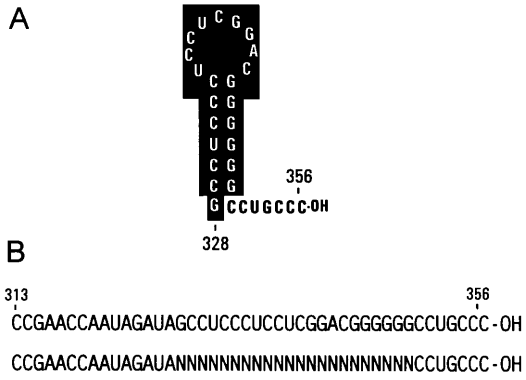

Figure 1. (**A**) The hairpin promoter at the 3′-end of sat-RNA C. The structure of the promoter was determined by chemical and enzymatic probing (12). Bases randomized and subjected to selection are in white. The Δ*G* of the hairpin is predicted by the RNAFOLD program (Genetics Computer Group, University of Wisconsin) to be –10.7 kcal/mol. (**B**) Randomized bases (denoted by N) and a portion of the upstream non-randomized sequence of sat-RNA C are shown. The 7 nt 3′-tail was not randomized, as mutations in this sequence are repaired to the wild-type sequence by TCV RdRp using primers generated from the identical sequence in the TCV genomic RNA by abortive cycling (30).

In vitro transcription and inoculation

The products of the second PCR described above (1/25th of the total) were subjected to transcription using T7 RNA polymerase as previously described (11). The synthesized RNA was divided equally into 15 portions (~5–6 µg/plant) and used to inoculate 15 2-week-old turnip seedlings along with 5 µg helper virus inoculum (HVI) per plant, as previously described (28). HVI is total RNA isolated from turnip plants infected with the genomic TCV RNA and its associated sat-RNA D. At various times post-inoculation RNA was prepared from uninoculated leaves and assayed for the presence of sat-RNA C-sized molecules by gel electrophoresis followed by staining with ethidium bromide (the level of wild-type sat-RNA C that accumulates in plant cells is similar to the level of 5S rRNA; data not shown). Cloning and sequencing of sat-RNA C-sized molecules using a 3′-RACE (rapid amplification of cDNA ends) PCR cloning and sequencing procedure has been previously described (29).

For the second round inoculations, total RNA isolated at 14 days post-inoculation (d.p.i.) from 16 plants containing sat-RNA C-sized species as visualized on polyacrylamide gels was pooled and re-inoculated onto six turnip seedlings (5 µg/plant). Sat-RNA C-sized species accumulating in plants at various d.p.i. were assayed as described above. For third round inoculations, total RNA isolated at 14 d.p.i. from the six second round plants was pooled and re-inoculated onto six turnip seedlings (5 µg/plant).

For competition experiments HVI (10 µg/plant) and equal amounts of wild-type sat-RNA C and/or round three winner sat-RNA C transcripts were inoculated into individual turnip seedlings. RNA was extracted 19 days later and sat-RNA was cloned as described above.

Generation of biologically active sat-RNA C from third round winners

Construction of full-length cDNAs of selected third round sat-RNA C 'winners' for use in competition assays between

selected species and between selected species and wild-type sat-RNA C required the removal of the poly(A) tails added during cloning. To clone the new sat-RNA C species, 19 base oligonucleotide primers complementary to the 3′-terminal bases of the selected sat-RNA C molecules were used in a 30 cycle PCR in the presence of oligonucleotide pT7C5′. The product of the PCR (1/25th of the total) was subjected to transcription *in vitro* using T7 RNA polymerase. Approximately 1 µg of each transcript, as assayed by agarose gel electrophoresis, was combined with 10 µg HVI and inoculated onto individual turnip seedlings, as described above.

RNA gel blot analysis

RNA gel blots were performed as previously described (30). The probe for TCV was complementary to positions 3892–3912 and the probe for sat-RNA C was complementary to positions 175–199.

RESULTS AND DISCUSSION

In vivo selection of RNA promoter sequences

To determine if iterative randomized selection is applicable to analysis of sequences/structures important for viral RNA amplification we chose to analyze the promoter for (–) strand synthesis at the 3′-end of sat-RNA C for the following reasons: (i) the promoter is one of the smallest and simplest RNA promoters and has been extensively characterized (3,12,31); (ii) the promoter sequence is not thought to be involved in any other viral function, such as encapsidation (32) or gene expression, since sat-RNA C is not a template for protein synthesis; (iii) while TCV does not require sat-RNA for infectivity, TCV in the presence of sat-RNA C is at a selective advantage for movement through plants (Q.Kong and A.E.Simon, unpublished observations) due to an undetermined mechanism. Therefore, TCV that is associated with biologically active sat-RNA C is more highly represented in uninoculated leaves of infected plants.

The promoter for (–) strand synthesis of sat-RNA C is contained within the 3′-terminal 29 bases (Fig. 1A). This sequence can function as an independent promoter to drive transcription of a non-template RNA in *in vitro* reactions containing partially purified TCV RdRp (12). Attempts to study the importance of the 3′-terminal 7 nt by deleting the sequence results in repair of the wild-type sequence using the identical 3′-terminal sequence from the genomic TCV RNA as a template for the repaired segment (31). For this reason only the 22 bases in the hairpin portion of the sat-RNA C promoter were randomized for subjection to *in vivo* selection (Fig. 1B).

Transcripts of sat-RNA C containing 22 randomized bases in place of the 3′-terminal hairpin were either co-inoculated onto 30 turnip seedlings in the presence of the TCV helper virus or inoculated onto 15 seedling leaves 5 days after inoculation of the same leaves with just the helper virus. At 14 d.p.i. RNA was extracted from uninoculated leaves and analyzed by gel electrophoresis. None of the sequentially inoculated plants had detectable sat-RNA C in uninoculated leaves (data not shown). However, 16 of the 30 plants simultaneously inoculated with helper virus and sat-RNA transcripts had sat-RNA C-sized species in uninoculated leaves, with levels varying from ~50% of wild-type to approximately wild-type levels (Fig. 2). The sequences of 2–18 sat-RNA C clones from each of nine randomly

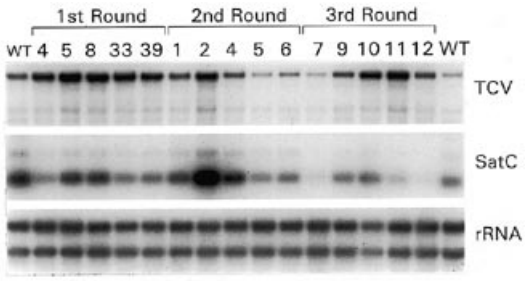

Figure 2. RNA gel blot analysis of total RNA isolated from uninoculated leaves at 14 d.p.i. The blot was sequentially hybridized with probes for TCV (upper), sat-RNA C (middle) and rRNA (lower) as a loading control. The top band in the upper panel is TCV genomic RNA and the lower two bands are the two subgenomic RNAs. In the middle panel the lower band is monomeric sat-RNA C and the upper band is sat-RNA C dimers. Numbers above the lanes refer to specific plants listed in Tables 1 (first number in the name of the sequence), 2 and 3. Only selected plants from the three rounds are shown. WT, plants were inoculated with TCV and wild-type sat-RNA C.

Table 1. First round *in vivo* selection

Name	Sequence[a]	% of clones recovered[b]
WT	GCCUCCCUCCUCGGACGGGGGG	
4-1	AGCCUCCUAAUACCAAUGGAAG	100 (9)
5-1	GCCGGGCAGCAUAUACUCCUGG	25 (8)
5-2	GACACAUGUACACGACAUGCGU	63
5-3	GCCCUCCACACUGCUGAGAAGG	12
7-1	UCUAGGCGCUUCCUAUUGACGC	100 (8)
28-1	GAUUAUCGUCUCAGACUGUAAU	100 (5)
29-1	AAGCCAAACCACGACUCUUUGG	100 (3)
30-1	CGCGGAACACAGACAAAUUCCG	100 (2)
31-1	GACCAGCAUCUCAACCGCUCGU	56 (18)
31-2	CCCGGGGUGUUACACAAUACCU	28
31-3a	CAGCCUUCCAUCUUGGUAAAGG	5
31-3b	CAGCCUUuCAUCUUGGUAAAGG	5
31-3c	CAGCCUUCuAUCUUGGUAAAGG	5
33-1	UGCGCGUCCCACCUGAGGACCG	54 (13)
33-2	UGGGCACUAACCUUAAGGUACU	23
33-3	UGGCGCACUGUUAUCAGACCGC	23
39-1a	GGACCAGCUGAAAUAAACUGUC	46 (13)
39-1b	GGACCAGCUGAAAUAAgCUGUC	64

[a]Clones within divisions originated from the same plant. lower case letters denote differences from an arbitrarily selected 'parental' clone.
[b]Numbers in parentheses reflect the number of clones sequenced from a given plant

selected plants that were accumulating sat-RNA C-sized species are shown in Table 1. None of the sat-RNA contained the wild-type 22 base sequence and no two plants produced clones with the same sequence. Only a single species was isolated from five plants (plants 4, 7 and 28–30), while sequence variants that differed by only a single position were found in two plants (plants 31 and 39). The latter result suggests that in addition to selection from the original population of sequences capable of forming viable promoters, evolution of these sequences was occurring from the multiple rounds of replication necessary for the presence of sat-RNA C in uninoculated leaves.

Computer generated secondary structures for the sat-RNA C cloned in the first round are shown in Figure 3. All sequences could be folded into hairpins ranging in stability from –2.8 to –10.5 kcal/mol (the wild-type hairpin is –10.7 kcal/mol). Thirteen

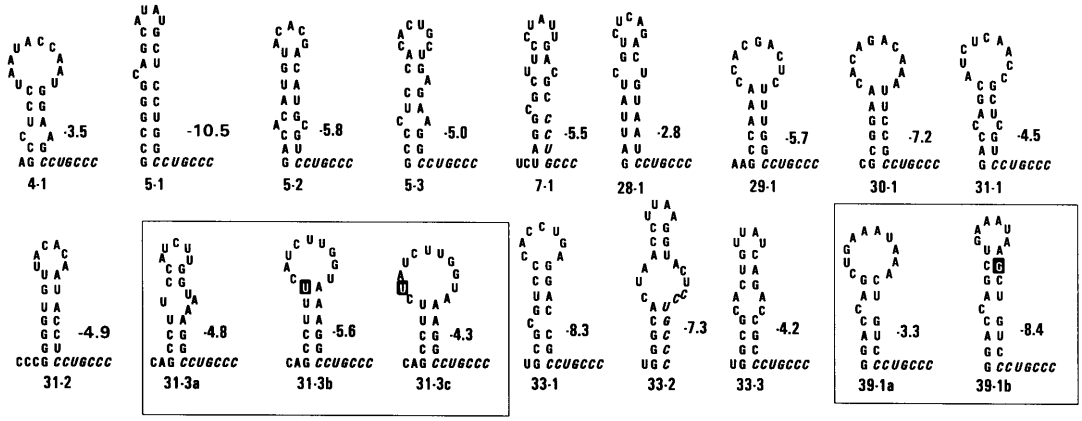

Figure 3. Computer derived secondary structures for first round selected sat-RNA C. The 22 base randomized sequence and 7 base non-randomized 3′-tails (in italics) are shown. The numbers below the hairpins denote the clone numbers from Table 1. Numbers to the right of the hairpins are the computer-derived Δ*G* values for stability of the hairpins. Boxed hairpins are variants of a single sequence and all boxed members were found in a single plant. Boxed nucleotides denote base alterations from an arbitrarily selected 'parental' sequence.

of the 15 clones (variants of a single sequence are counted once) have the non-randomized 3′-terminal 6 nt as a single-stranded tail, as did wild-type sat-RNA C. The remaining clones have the 3′-terminal 6 nt as part of the hairpin stem. Loop sequences incapable of forming canonical base pairs ranged from 4 to 12 bases compared with the wild-type 9 bases. Three of the clones (5-1, 5-3 and 29-1) contained the identical 3 bp at the base of the hairpin as wild-type sat-RNA C. Together these results confirm earlier studies indicating that the TCV RdRp is capable of utilizing a variety of hairpins as promoters for (–) strand synthesis *in vivo* (3).

Both sequence and structure contribute to the activity of the promoter sequence at the 3′-end of sat-RNA C

To enrich for more fit promoter sequences, RNA from the 16 plants accumulating 'first round' sat-RNA was pooled and used to inoculate six turnip seedlings. The amount of sat-RNA accumulating at 14 d.p.i. ranged from ~50 to 200% of wild-type levels (Fig. 2). Sat-RNA C clones were analyzed from two plants at 14 and 34 d.p.i. (to determine if the sat-RNA populations changed over time within a plant) and four additional plants at 34 d.p.i. (Table 2). Since the majority of clones were present as variants of sequences found in the first round, clones were renamed to include a sequence family number followed by a letter if the sequence was found as one of several variants. Six of the eight sequence families identified in the second round had been previously identified in the first round (1, 2 and 4–7). Since only nine of the 18 first round plants were analyzed for sat-RNA C sequences, the remaining two sequences identified in the second round probably originated from the unanalyzed plants. Sequence families 1–4 were highly represented in nearly every second round plant, while sequence families 5–7 were only sporadically represented. Sequence family 8, containing only a single member, was unusual, as it was the majority species cloned from plant 60 at 14 d.p.i., but only one of 24 clones sequenced at 34 d.p.i. and not represented in any other plant.

The computer derived secondary structures of the 3′-ends of second round sequence families 1–7 are shown in Figure 4. The stability of the hairpins ranged from –3.3 to –12.0 kcal/mol. The most highly represented sequences (1a, 2a, 3 and 4a) had hairpin stabilities ranging from –5 to –12 kcal/mol. All sequences had the non-randomized 3′-terminal 6 nt present as single-stranded tails. Most base changes in family variants maintained or promoted additional base pairing in the stem. For example, clone 1a, which was identical to clone 5-3 from the first round, contained a C:A bulge near the base of the hairpin. Variants of 1a contained single base changes leading to either a U:A base pair (1b) or a C:G base pair (1c). Clone 2a, which was identical to clone 5-1 from the first round, contained a G:C base pair in the upper stem, while variant 2b contained two alterations that resulted in replacement of this base pair with a U:A base pair. Sequence family 4 contained five members, with clone 4a originally identified in the first round (39-1b). One of the family variants (4b) had a single base

Table 2. Second round *in vivo* selection

Name	Past name[a]	Sequence[b]	Plant[c]									
			1		2		3		4		5		6	
			14d	34d	14d	34d	14d	34d	14d	34d	14d	34d	14d	34d
WT		GCCUCCCUCCUCGGACGGGGGG												
1a	5-3	GCCCUCCACACUGCUGAGAAGG	9	4		3		8		9		6	3	13
1b		GCCuUCCACACUGCUGAGAAGG	1	2		1		3						
1c		GCCCUCCACACUGCUGAGAgGG								2				
2a	5-1	GCCGGGCAGCAUAUGCUCCUGG	5	7		16		1		9			3	1
2b		GCCGGGCAGuAUAUaCUCCUGG								2				
3		CAGGGCUACCUACUUGGUGGCC	2	3		2		6		4		1	4	1
4a	39-1b	GGACCAGCUGAAAUAAGCUGUC	3	2				2				5	3	4
4b		GGACCAGCUGAAAUAAuGCUGUC												1
4c		GGACuAGCUGAAAUAAGCUGUC		1										1
4d		GGACCAGuUGAAAUAAaCUGUC												1
4e		GGACCAGCUGAAAUAAaCUGUC								1				
5	31-2	CCCGGGGUGUUACACAAUACCU	1											
6a	4-1	AGCCUCCUAAUACCAAUGGAAG											1	
6b		AGCCUCCUAAUACCAAUGGAgG		1										
7a	31-3a	CAGCCUUCUAUCUUGGUGAAGG												1
7b		CAGCCUUCUAcCUUGGUGAAGG						1						
8		CUAGGUGACCCCUCGGGGAAAC											9	1
Total clones sequenced:			21	20		22		21		27		12	23	24

[a]From Table 1.
[b]Sequences families are separated by dividers. Lower case letters denote differences from an arbitrarily selected 'parental' clone.
[c]The number of clones with the sequence shown found in plants 1–6 at 14 or 34 d.p.i. is shown.

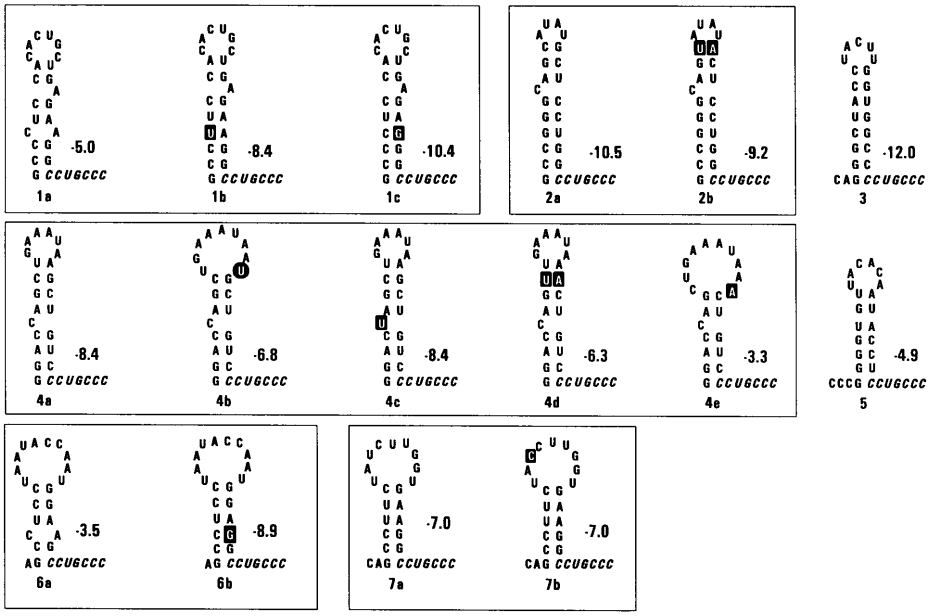

Figure 4. Computer derived secondary structures for second round selected sat-RNA C. The numbers below the hairpins denote the clone numbers from Table 2. In the hairpin for clone 4b, the base circled is an inserted nucleotide. See legend to Figure 3 for details.

Table 3. Third round *in vivo* selection

Name	Past name[a]	Sequence[b]	Plant[c]									
			7		8		9		10		11	
			14d	34d	14d	34d	14d	34d	14d	34d	14d	34d
WT		GCCUCCCUCCUCGGACGGGGGG										
1a	5-3	GCCCUCCACACUGCUGAGAAGG		1	1		1					8
1b		GCCuUCCACACUGCUGAGAAGG		5	7	6	8	13		4		1
1c		GCCCUCCACACUGCUGAGAgGG		19	2	3		2				5
1d		GCCuUCCACACU GAGAAGG										1
2a	5-1	GCCGGGCAGCAUAUGCUCCUGG		1	7	6	9	4		10		2
3		CAGGGCUACCUACUUGGUGGCC				1		2		3		
7b	31-3a	CAGCCUUCUAcCUUGGUGAAGG		1	1	2	1			1		
7c		CAGCCUUCcAUCUUGGUGAAGG										5
7d		CAGCCUUCUAUCUUaGUGAAGG								1		
Total clones sequenced:				27	18	18	19	21		19		22

[a]From Table 1.
[b]Sequence families are separated by dividers. Lower case letters denote differences from an arbitrarily selected 'parental' clone.
[c]The number of clones with the sequence shown found in plants 7–11 at 14 or 34 d.p.i. is shown.

insertion; 4c had a transition from a bulged C to a bulged U; 4e had a G→A transition; 4c had both the 4e alteration and a second alteration that resulted in the replacement of a G:C base pair with a U:A base pair in the upper stem. The single member of sequence family 8 could not be folded into a stable secondary structure. A portion of the sequence, however, had features (multiple C residues upstream of a purine-rich sequence) similar to the 11 and 14 base promoter sequences in the (–) strand of sat-RNA C, which also do not form discernible secondary structures (10).

Equal amounts of RNA isolated from the six second round plants at 34 d.p.i. were combined and used to inoculate six turnip seedlings to initiate the third and final round of selection. Levels of sat-RNA at 14 d.p.i. ranged from ~25 to 100% of wild-type levels. Although the levels of sat-RNA were low in plants 9 and 12 at 14 d.p.i., by 34 d.p.i. all plants contained approximately equal levels of sat-RNA. Sat-RNA C was cloned from two plants at 14 and 34 d.p.i. and sat-RNA C in three plants was cloned at 34 d.p.i. (the remaining plant had no discernible genomic TCV RNA at 14 d.p.i.). The sequences of the 3′-end regions of the resultant clones are shown in Table 3. Only four sequence family 'winners' were present among the plants and all had been identified in the second round. Three of the sequence families (1–3) were highly represented in clones from second round plants. Clone 1a, which was the major sequence family 1 member cloned from second round plants, was a minor species in all but one third round plant, while variants 1b and 1c comprised a substantial portion of the clones from the third round plants. Newly identified variant 1d was similar to the 1b sequence but contained a 3 base deletion.

The computer derived secondary structures of the third round winners are shown in Figure 5A. The average stability of the hairpins increased from the second round and ranged from –7.0 to –12.0 kcal/mol. Three of the four sequence family winners (1, 2 and 7) had the same 3 bp at the base of the stem as wild-type sat-RNA C. Clone 1d and sequence family 7 share identical sequences for the entire stem region and nearly identical loop sizes (8 and 9 bases respectively), although the hairpin in sequence family 7 begins 2 bases into the 22 base randomized region. Clones 1d and 7c also contained the same 2 nt extending into the 5′-side of the loop, resulting in eight consecutive bases of identical sequence in one portion of the original randomized region between the two clones (Fig. 5B). To reach this convergence in sequence and structure the 1d sequence required a 1 base alteration and a 3 base deletion from the original 5-3 clone found in the first round, while the 7c sequence differed by a single base from the 31-3a clone found in the first round. The sequence and structural similarities between clones 1d and 7c and the sequence similarity at the base of the stem among nearly all

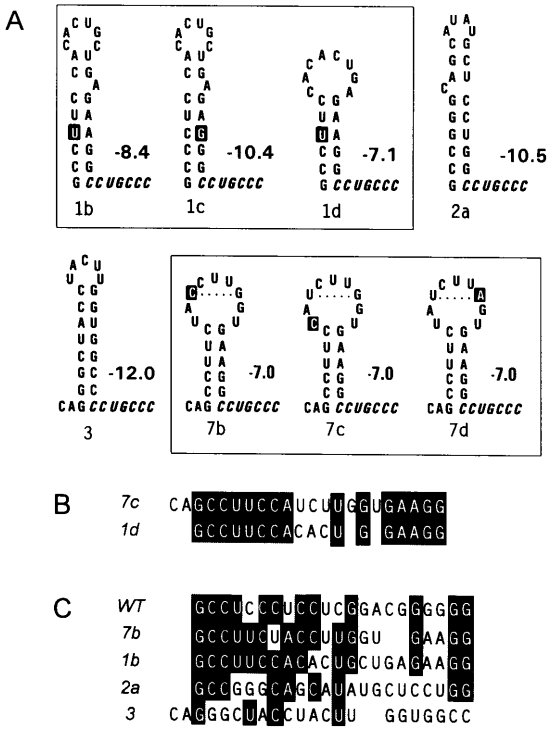

Figure 5. Third round selection winners. (**A**) Computer derived secondary structures for third round selected sat-RNA C. See legend to Figure 3 for details. A putative base pair in the hairpin loop of the sequence 7 family is indicated by a dotted line. (**B**) Sequence similarity between third round winners 7c and 1d. Similar sequences are boxed. (**C**) Sequence similarity between third round sequence families and wild-type sat-RNA C. Similar sequences are boxed.

third round winners and wild-type sat-RNA C (Fig. 5A and C) strongly suggest that both sequence and structure contribute to increased fitness of the sat-RNA.

Wild-type sat-RNA C contains a weak U:G base pair in the fourth position of the stem (counting from the base of the stem). Most of the third round winners (1b, 1d, 2a and 7b–d) also contained a weak U:G or U:A base pair in this position. To determine if a U:A base pair in the fourth position is preferred over a G:C base pair, direct competition was performed between clones 1b and 1c, which differ only in the identity of the base pair in this position (U:A or C:G respectively). Although equal amounts of 1b and 1c transcripts were used in the inoculation, all of the 24 clones sequenced from two plants at 19 days post-inoculation were clone 1b, which contains a U:A in the fourth position. This result suggests a preference for a weaker base pair at the fourth position of the stem, also given the surrounding sequences of these two clones.

Wild-type sat-RNA C, sequence family 7 and clone 1d have large loop sequences. NMR studies of hairpin loops have frequently indicated that loop bases form stable compact structures with stacked nucleotides and non-Watson–Crick base pairs (reviewed in 33). While it is not possible to predict if the large loops present in wild-type sat-RNA C and some third round winners form stable compact structures, base alterations in different members of sequence family 7 winners could potentially affect a single nucleotide base pairing in the loop (Fig. 5A, dotted lines). Wild-type sat-RNA C can also form a putative base pair (C:G) in a similar position in the loop. Since other third round winners (with the exception of variant 1d) had more compact loops and could not form a putative base pair in this position, it is not possible to determine at this time the importance of such a putative base pair in the larger loops of sat-RNA C and the family 7 members.

Our previous results (3) based on analysis of the hairpin using site-specific mutations indicated that biologically active promoters could have hairpins less stable than the wild-type, with loops of variable length and sequence and without an absolute need for the 6 base single-stranded tail. In addition, since compensatory mutations in the lower and upper stem did not abolish promoter activity, the conclusion was reached that the positioning of specific bases in the stem was not required to produce an active promoter. These previous conclusions are very similar to the conclusions from the current first round SELEX results. However, the addition of side-by-side competition introduced by further rounds of selection clearly demonstrates that increased fitness of the promoter is achieved by more stable hairpins with 6 base single-stranded tails, a preference for CG base pairs at the base of the stem and a weaker base pair in the fourth position of the stem.

Wild-type sat-RNA C is a better adapted template *in vivo* than third round winners

To determine how the third round winners compared with wild-type sat-RNA C in ability to accumulate *in vivo*, plants were inoculated with equal amounts of wild-type sat-RNA C and clones 1a, 1b, 2a or 7b. At 19 d.p.i. only wild-type sat-RNA C was cloned from two plants (19/19). These results indicate that the wild-type sequence is at a selective advantage compared with clones 1a, 1b, 2a or 7b. The lack of recovery of wild-type sequences in round three suggests that either this sequence was not present in the original population of randomized molecules or that the wild-type sat-RNA sequence and TCV were never present together in the initially inoculated cells, a condition required for amplification of any sat-RNA C molecules. Preliminary results analyzing a 12 base linear promoter on sat-RNA C involved in (+) strand synthesis indicates that the wild-type sequence can be recovered using this *in vivo* SELEX approach (Carpenter and Simon, unpublished).

In conclusion, we have established that *in vivo* genetic selection can be applied to analysis of *cis* sequences involved in accumulation of subviral RNAs and may be applicable to the study of such sequences in viral genomic RNAs. This technique has advantages over site-specific mutagenesis in that it allows a combination of side-by-side selection of numerous sequence possibilities and sequence evolution. Our finding that two clones in the third round of selection (1d and 7c) shared 15 of 22 nt in the randomized sequence region and had identical stem sequences indicates that sufficient sequence complexity was initially available to reach such sequence convergence. However, sat-RNA containing the wild-type sequence were not recovered, even though the wild-type sequence is at a selective advantage *in vivo* compared with selected third round winners. Since all plants in the first round contained sat-RNA with different randomized sequences, additional sequence complexity could be achieved by initial inoculation of substantially more plants.

ACKNOWLEDGEMENTS

We thank Dr Peter D.Nagy for critical reading of the manuscript. This work was supported by National Science Foundation grants MCB-9419303 and MCB-9630191 to A.E.S.

REFERENCES

1 Domingo,E. and Holland,J.J. (1994) In Morse,S.S. (ed.), *The Evolutionary Biology of Viruses*. Raven, New York, NY, pp. 161–184.
2 Pilipenko,E.V., Poperechny,K.V., Maslova,S.V., Melchers,W.J.G., Slot,H.J.B. and Agol,V.I. (1996) *EMBO J.*, **15**, 5428–5436.
3 Stupina,V. and Simon,A.E. (1997) *Virology*, **238**, 470–477.
4 Tsai,C.-H. and Dreher,T.W. (1992) *J. Virol.*, **66**, 5190–5199.
5 Duggal,R., Lahser,F.C. and Hall,T.C. (1994) *Annu. Rev. Phytopathol.*, **32**, 287–309.
6 Bujarski,J.L., Dreher,T.W. and Hall,T.C. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 5636–5640.
7 Deiman,B.A.L.M., Kortlever,R.M. and Pleij,C.W.A. (1997) *J. Virol.*, **71**, 5990–5996.
8 Dreher,T.W. and Hall,T.C. (1988) *J. Mol. Biol.*, **201**, 31–40.
9 Siegel,R.W., Adkins,S. and Kao,C.C. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 11238–11243.
10 Guan,H., Song,C. and Simon,A.E. (1997) *RNA*, **3**, 1401–1412.
11 Song,C. and Simon,A.E. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 8792–8796.
12 Song,C. and Simon,A.E. (1995) *J. Mol. Biol.*, **254**, 6–14.
13 Wang,J. and Simon,A.E. (1997) *Virology*, **232**, 174–186.
14 Carrington,J.C., Heaton,L.A., Zuidema,D., Hillman,B.I. and Morris,T.J. (1989) *Virology*, **170**, 219–226.
15 Oh,J.-W., Kong,Q., Song,C., Carpenter,C.D. and Simon,A.E. (1995) *Mol. Plant–Microbe Interact.*, **8**, 979–987.
16 Simon,A.E. And Howell,S.H. (1986) *EMBO J.*, **5**, 3423–3438.
17 Ellington,A.D. and Szostak,J.W. (1990) *Nature*, **346**, 818–822.
18 Tuerk,C. and Gold,L. (1990) *Science*, **249**, 505–510.
19 Gold,L., Polisky,B., Uhlenbeck,O. and Yarus,M. (1995) *Annu. Rev. Biochem.*, **64**, 763–797.
20 Houser-Scott,F., Ansel-McKinney,P., Cai,J.-M. and Gehrke,L. (1997) *J. Virol.*, **71**, 2310–2319.
21 Brown,D. and Gold,L. (1995) *Biochemistry*, **34**, 14765–14774.
22 Brown,D. and Gold,L. (1995) *Biochemistry*, **34**, 14775–14782.
23 Blumenthal,T. and Carmichael,G.G. (1979) *Annu. Rev. Biochem.*, **48**, 525–548.
24 Madhani,H.D. and Guthrie,C. (1994) *Genes Dev.*, **8**, 1071–1086.
25 Libri,D., Stutz,F., McCarthy,T. and Rosbash,M. (1995) *RNA*, **1**, 425–436.
26 Grimm,C., Lund,E. and Dahlberg,J.E. (1997) *EMBO J.*, **16**, 793–806.
27 Coulter,L.R., Landree,M.A. and Cooper,T.A. (1997) *Mol. Cell. Biol.*, **17**, 2143–2150.
28 Li,X.H., Heaton,L.A., Morris,T.J. and Simon,A.E. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9173–9177.
29 Carpenter,C.D. and Simon,A.E. (1996) *J. Virol.*, **70**, 478–486.
30 Kong,Q., Oh,J.-W., Carpenter,C.D. and Simon,A.E. (1997) *Virology*, **238**, 478–485.
31 Nagy,P.D., Carpenter,C.D. and Simon,A.E. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 1113–1118.
32 Qu,F. and Morris,T.J. (1997) *J. Virol.*, **71**, 1428–1435.
33 Shen,L.X., Cai,Z. and Tinoco,I.,Jr (1995) *FASEB J.*, **9**, 1023–1033.