Title:	A hepatitis C virus *cis*-acting replication element forms a long-range RNA-RNA interaction with upstream RNA sequences in NS5B
Working title:	Long-range *cre* interaction in HCV
Authors:	Sinéad Diviney [1], Andrew Tuplin [1], Madeleine Struthers [1], Victoria Armstrong [2], Richard Elliott [2], Peter Simmonds [2] and David J. Evans [1]*
Affiliation:	1. Department of Biological Sciences, University of Warwick, Coventry, CV4 7AL 2. Department of Biomolecular Sciences, University of St. Andrews, KY16 9ST 3. Centre for Infectious Diseases, University of Edinburgh, Summerhall, Edinburgh, EH9 1QH
Keywords:	Structure, HCV, virus, bioinformatics, RNA secondary structure, replication
Correspondence:	David J. Evans, Department of Biological Sciences, University of Warwick, Coventry, CV4 7AL d.j.evans@warwick.ac.uk +44 (0)24765 74183
Abstract:	242 words
Main text:	7039 words

Abstract

The genome of hepatitis C virus (HCV) contains *cis*-acting replication elements (CREs) comprised of RNA stem-loop structures located in both the 5' and 3' non-coding regions (NCR), and in the NS5B coding sequence. Through the application of several algorithmically-independent bioinformatic methods to detect phylogenetically-conserved, thermodynamically-favoured RNA secondary structures, we demonstrate a long-range interaction between sequences in the previously described CRE (5BSL3.2, now SL9266) with a previously predicted unpaired sequence located 3' to SL9033, approximately 200 nucleotides upstream. Extensive reverse genetic analysis both supports this prediction and demonstrates a functional requirement in genome replication. By mutagenesis of the Con-1 replicon, we show that disruption of this alternative pairing inhibited replication, a phenotype that could be restored to wild-type levels through the introduction of compensating mutations in the upstream region. Substitution of the CRE with the analogous region of different genotypes of HCV produced replicons with phenotypes consistent with the hypothesis that both local and long-range interactions are critical for a fundamental aspect of genome replication. This report further extends the known interactions of the SL9266 CRE, which has also been shown to form a 'kissing loop' interaction with the 3' NCR (8), and suggests that cooperative long-range binding with both 5' and 3' sequences stabilises the CRE at the core of a complex pseudoknot. Alternatively, if the long-range interactions are mutually exclusive, the SL9266 CRE may function as a molecular switch controlling a critical aspect of HCV genome replication.

Introduction

Hepatitis C virus (HCV), a flavivirus in the genus *Hepacivirus*, possesses a positive (mRNA)-sense genome of approximately 9.6kb encoding a single polyprotein. This is cleaved co- and post-translationally to generate proteins that form the enveloped virus particle and those that replicate the genome. Polyprotein translation is initiated within a highly-structured internal ribosome entry site (IRES) occupying much of the 5' non-coding region (NCR). The 5' NCR also contains sequences required for genome replication (9, 19, 26) and, like functionally analogous regions in the 3' NCR, these form defined stem-loop structures that operate in *cis* and are known or suspected to recruit cellular or viral proteins (5, 10, 30). In addition to these *cis*-acting replication elements (CREs) in the non-coding extremes of the genome, there is evidence that additional RNA structures exist within the coding regions. The latter structure is of two types, phylogenetically conserved well-defined structures occupying the 5' and 3' regions of the sense strand of the coding region of HCV (23, 31, 32, 36), and a less well-characterised but much more extensive set of RNA secondary structures, collectively designated genome-scale ordered RNA structure (GORS), that spans the entire coding region of HCV (29).

The potential functional role(s) of phylogenetically-conserved RNA secondary structures in coding regions have been analysed extensively by reverse genetic analysis – predominantly using antibiotic-resistance or luciferase-encoding sub-genomic replicon system (18) – and more recently, in analysis of structures in the core-encoding region using the HCV virus replication system (17, 23, 25, 34). Several groups have reported that a short stem-loop structure in the NS5B-coding region – variously designated 5BSL3.2 or SL-V (16, 36) – has a clearly defined function in genome replication. This structure, henceforth designated SL9266 (see Materials and Methods for details of a unified numbering scheme), forms a stem-loop with two short base-paired helices,

separated by a 8nt. bulge loop on the 3' side, and capped with a 12 nt. terminal loop (16, 36). Extensive mutagenesis has demonstrated that the structural integrity of the element must be retained for replication. In addition, substitutions within the two unpaired loop or bulge regions can also be deleterious implying these also contribute important functions during replication. SL9266 therefore forms a *cis*-acting replication element, though its precise function during genome replication has yet to be determined. SL9266 is the penultimate of five phylogenetically conserved RNA structures in the region encoding NS5B. Limited mutagenesis of the upstream adjacent structure (SL9217) designated SL-VII (16) or 5BSL3.1 (36) have produced contradictory results and further studies are required to unequivocally demonstrate a role in genome replication.

Functional analysis of the SL9266 CRE and related RNA structures in the NS5B coding region necessitates the introduction of mutations that leave the underlying coding sequence intact. The restriction of mutagenesis to synonymous substitutions naturally places some limits on the substitutions that can be tested. However, Friebe and colleagues (8) have demonstrated that SL9266 can be functionally moved to the 3' NCR, albeit with a reduction in replication efficiency. This suggests that the function of this structure is at least partially position-dependent, but does allow more extensive mutagenic studies. The position-dependence could be due to a requirement for a spatially-dependent interaction with another region of the virus genome; indeed they have demonstrated a functionally required kissing-loop (tertiary RNA structure) interaction between the terminal unpaired loop of SL9266 and SL2 in the X tail of the 3' NCR (8).

We have developed novel bioinformatic strategies to detect phylogenetically-conserved long-range RNA-RNA interactions. These approaches are based upon well-established and accepted thermodynamic methodologies, but extend them to take advantage of the

wealth of sequence data available for HCV. Using this information we have investigated the structure and function of SL9266. We demonstrate that the relatively weak prediction of SL9266 using standard bioinformatic methods can be explained by the structure adopting an additional alternative and potentially metastable pairing with sequences situated approximately 200 nt. upstream. Mutagenesis of the two interacting sequences provides genetic support for the interaction, and also demonstrated some sequence specificity within SL9266. Duplex formation with the upstream sequences and the 3' Xtail involves distinct regions of SL9266, and the revised model presented here does not preclude the existence of a combined kissing loop interaction with SL2 in the 3'UTR, and a pseudoknot interaction of the CRE bulge sequence upstream to form a complex long-range pseudoknot.

Materials and Methods

Sequence alignments

Sequences datasets were initially compiled from all available epidemiologically unlinked variants of all 6 genotypes (those showing >2% sequence divergence from each other) that were >95% complete between nucleotide positions 9001 and 9377 and second set between 8204 and 9377. Nucleotides were numbered with reference to the H77 complete genome sequence, AF011753 (15). Representative subsets of sequences within each alignment were used for RNA structure determination. Accession numbers are provided in Supplementary Material.

Stem-loop nomenclature

Several methods have been used to describe stem-loops in NS5B and elsewhere in the HCV genome (16, 32, 36). Following the adoption of a standardised system for numbering HCV sequences (15), it had been proposed that stem-loops are numbered based on the position of the first 5' paired base in the structure (S. Lemon, Fields

Virology, 5th edition; in press). Accordingly, stem-loops previously referred to as 5BSL1, 5BSL2, 5BSL3.1 to 5BSL3.3 (36), SLIV to SLVII (16) or SL8828, SL8926, SL9011, SL9061 and SL9118 (16, 31, 32) are re-designated as SL9033, SL9132, SL9217, SL9266 and SL9324 respectively in the current study. Likewise, SL2 in the 3' Xtail is renumber SL9571.

RNA structure prediction

RNA structures were predicted using MFOLD through the web interface at http://frontend.bioinfo.rpi.edu/applications/mfold/. Automated analysis of most energetically stable RNA structure using the program StructureDist v. 1.3 (available from http://www.picornavirus.org/). SFOLD analysis using the program Srna on the server at http://sfold.wadsworth.org/srna.pl. PFOLD analysis used the web interface at http://www.daimi.au.dk/~compbio/rnafold/. All programs were run with default settings.

Cell culture, plasmids and mutagenesis

Cell monolayers of the human hepatoma cell line Huh7 (kindly provided by Dr. R. Bartenschlager) were maintained in Dulbecco's modified minimal essential medium (DMEM, Invitrogen) supplemented with 10% foetal bovine serum, 1% non-essential amino acids, 100 U pencillin/100 µg streptomycin, and 2mM L-Glutamine (Invitrogen) (DMEM P/S). Cells were passaged after treatment with trypsin-EDTA and seeded at dilution of 1:3-1:5.

The parental, genotype 1b, neomycin-encoding replicon, designated pFK-I_{389}neo/NS3–3'/wt was generously provided by Dr. R. Bartenschlager and is fully described in Lohmann *et al*., (18). The cDNA was modified by the introduction of a previously described cell culture adaptive change of Serine for Isoleucine as residue 2204 of the polyprotein (1). A derivative replicon, designated pFKnt341-sp-PI-lucEI3420-9605/5.1, expressing a firefly luciferase reporter gene (kindly provided by GlaxoSmithKline, UK)

consisted (5' to 3') of the HCV 5' NCR, a 63nt. spacer, the poliovirus IRES and luciferase gene followed by an EMCV IRES and the NS3-NS5B-coding region and 3' NCR of HCV. Derivatives of both replicons carrying substitutions (GDD to GND) of the active site of the NS5B RNA dependent RNA polymerase were used as controls where appropriate.

All site-directed mutagenesis was conducted on a unique *Spe* I – *Xho* I fragment (nucleotides 5582 to 8005), sub-cloned in pBluescript II SK(+), using Stratagene QuikChange™ site–directed mutagenesis. All mutations were detected and confirmed by sequencing, rebuilt into the appropriate sub-genomic replicon and sequenced again.

Substitution of SL9266 with the analogous sequence of other HCV genotypes was achieved using a cassette system. Briefly, a 528 nt. *Kpn*I – *Spe*I fragment spanning SL9266 was sub-cloned into pBluescript II SK+ (Invitrogen) and used as a template for PCR with primers *Bsm*BI-1F (GCGTCTCTGTTCATGTGGTGCCTACTCC) and *Bsm*BI-2R (GCGTCTCTTAACCAGCAACGAACCAGCT). The blunt ends of the reaction product were ligated to create a plasmid in which SL9266 was precisely replaced with a stuffer fragment containing two *Bsm*B I restriction sites. This cassette vector was cleaved with *Bsm*B I and ligated with complementary oligonucleotides for the stem-loop sequences from other genotypes. The sequences are illustrated in Figure 6a. After sequencing the *Kpn*I-*Spe*I fragment was rebuilt into pFK-I_{389}neo/NS3–3'/wt.

In vitro RNA transcription and replicon analysis

1μg of *Sca* I-linearised replicon cDNA was used as template for the production of RNA *in vitro* using a T7 MEGAscript™ kit (Ambion) according to the manufacturing instructions. RNA was purified using RNAeasy (Qiagen), the integrity confirmed by agarose gel electrophoresis and quantified spectrophotometrically.

Huh7 cells were transfected by electroporation. Briefly, 400μl of trypsinised, washed, Huh7 cells at 1 x 10^7 cells/ml in phosphate-buffered saline were mixed with 5μg *in vitro* transcribed RNA in a pre-chilled 4mm cuvette, pulsed once (25 milliseconds, 250V, 950μF, square wave) using a Bio-Rad Gene Pulser Xcell™ unit, and transferred into 100mm dishes with 10ml of DMEM P/S media added. After 24 hours culture at 37ºC the media was replaced with media supplemented with 500μg/ml G418 (Geneticin, G418 sulphate, Invitrogen), and the media changed at 2-3 day intervals for the duration of the selection period. G418-resistant colonies were washed with PBS, fixed with 4% formaldehyde and visualised with Giemsa stain after about 3 weeks.

Luciferase-encoding replicon RNA (10μg) was transfected into Huh7 cells as described previously, transferred into 20ml of DMEM P/S and 4ml placed in five wells of a 6-well dish. At each timepoint (4, 24, 48 and 72 hours post transfection) cells in one well were washed with PBS, lysed with 0.5ml Glo-Lysis buffer (Promega) and stored frozen before analysis using the Bright-Glo™ Luciferase Assay system (Promega) and quantified on a Turner TL-20 luminometer.

Results

Synonymous substitutions within SL9266 define a *cis*-acting replication element

We investigated the role of base-pairing within the SL9266 stem-loop structure by introducing a limited number of nucleotide substitutions to the region. For each of the six mutants – designated SL9266*mut*1 to SL9266*mut*6 – modifications were at synonymous sites, and were generated in a neomycin-encoding sub-genomic replicon (Figure 1). MFOLD analysis (data not shown) indicated that the mutations introduced in *mut*1, 3, 4 and 6 probably disrupted the predicted structure of SL9266, but that introduction of *mut*2 and 5 had no structural consequences, being restricted to the unpaired terminal loop region. RNA generated *in vitro* was transfected into Huh7 cells

and analysed in a G418 transduction assay. Two of the six mutants analysed, *mut*3 and *mut*5, generated colony numbers consistent with replication levels at, or near, that of the positive control. The remaining four mutants (*mut*1, *mut*2, *mut*4 and *mut*6) failed to yield significant numbers of colonies following transfection and neomycin selection (Figure 1D). Of these, substitutions involving the 5' side of the lower duplex in SL9266 (*mut*1; Figure 1C) or the terminal loop (*mut*2) were lethal, presumably reflecting a requirement for stable base-pairing in the former, or the interaction with the 3' X tail in the latter, and were in agreement with several other published studies (8, 16, 36). Substitution of $A_{9281}U$ (*mut*5) alone did not impair replication, again consistent with other studies (see A68C in Figure 7 of 36), and appeared to complement the otherwise lethal substitution of $U_{9296}A$ (compare *mut*3, *mut*5 and *mut*6; Figure 1C&D). Although the non-viable phenotype of *mut*4 could probably be ascribed to the $U_{9296}A$ mutation disrupting the upper duplex of SL9266, two other substitutions in this mutant were located in the unpaired 3' bulge loop (Figure 1B&C). A potential functional role for this region of SL9266, also hinted at by the currently unexplained lack of viability of replicons bearing mutations of C_{9303} and/or A_{9305} (designated C90A and A92G in Figure 7 of reference 36), prompted us to investigate additional features of SL9266 and possible interactions of the unpaired regions of the CRE with flanking RNA sequences.

RNA secondary structure prediction.

Previous comparative analysis of minimum free energy structures of the NS5B region of HCV revealed a series of evolutionarily conserved stem-loops spanning the terminal 700 bases of the coding sequence (Figure 2A and references 16, 31, 32, 36). Using an automated method (StructureDist) to quantify frequencies of concordant and discordant pairings at individual sites on pairwise comparison of structure predictions for each sequence (31), substantial variability in the degree of conservation of the stem-loops was found between HCV genotypes (Figure 2B). Similar variability was observed within

sets of approximately equally energetically favoured structure predictions for individual sequences (data not shown). The most highly conserved predicted stem-loop was SL9324, while SL9266 was the least. Lack of conservation of the latter structure was unexpected and relevant to the investigation of its demonstrated role as a *cis*-acting replication element (8, 16, 36).

To investigate whether there were alternative RNA structures or pairings underlying this observed lack of conservation of SL9266, RNA structures were predicted for 26 NS5B nucleotide sequences each representing different (up to 4) subtypes within the six genotypes of HCV using the program SFOLD. This generates a statistical sample of secondary structures from the Boltzmann ensemble of RNA secondary structures using Turner free energy rules (7). The relatedness of structures to each was determined using the Diana method for cluster analysis followed by calculation of the Calinski and Harabasz index to determine the optimal number of centroids for which consensus structures can be calculated, as previously described (6). Each sequence submitted to SFOLD generated between 2 – 6 centroids, whose consensus structure predictions were compared to the previously described RNA structure for the NS5B region (Figure 3). Despite the variability in RNA pairings between centroids for individual sequence of the wide range of genotypes analysed, 4 of the 5 stem-loops were frequently (SL9033, SL9132, and SL9217) or invariably (SL9324) found among sampled structures, generally containing equivalent pairings to the predicted structures for HCV genotype 1a (black filled boxes) or pairings restricted to bases around the terminal loop (grey filled boxes). However, consistent with the more variable structure predictions in this region visualised by StructureDist (Figure 2B), only around a third (26) of the 71 consensus structures of the centroids contained pairings that matched those of SL9266 (Figure 3). Alternative structures for this region frequently retained the pairing of the terminal stem and loop (bases 9274-9297, but with a partially overlapping longer range

pairing of bases forming the bulge loop and part of the 3' lower stem of SL9266 (bases 9296 – 9306) with upstream predicted unpaired regions in NS5B. In approximately one half of these alternative structures (labelled "A" in Figure 3), bases between 9296 – 9306 formed a duplex with the predicted unpaired bases between structures SL9033 and SL9132 (bases 9106 – 9123). Analogous alternative conformations were found in predicted structures for all 6 genotypes of HCV, and frequently alternated with the standard structure in the Boltzmann ensemble for individual sequences. Similar frequencies of standard and alternative pairings were observed when longer sequences spanning position 8301 to the 3'NCR were analysed by SFOLD (data not shown).

StructureDist and SFOLD use free energy minimisation algorithms (*e.g*. MFOLD) to predict candidate RNA structures. Given the poor resolution of RNA structure in the HCV CRE, we used an independent, non-energy minimising algorithm that makes better use of the substantial comparative sequence information available for HCV (14, 24). The method, implemented as PFOLD, combines an explicit evolutionary model of RNA sequences with a probabilistic model for secondary structures. A stochastic context-free grammar is used to produce a prior probability distribution of RNA structures. For the analysis, a set of 40 NS5B sequences between positions 9001 – 9377 from genotype 1b and further sets of 20 sequences from genotypes 1, 2, 3, 4 and 6 containing as diverse a range of subtypes as possible, were analysed individually and in combination by PFOLD (Figures 4 and 5). For the set of genotype 1b sequences, pairing predictions corresponded to those of the standard structure, with robust prediction of SL9266 (upper left half in Figure 4), and the four other stem-loops predicted for NS5B. Similar results were obtained for pairing predictions of alignments of each genotype individually (Figure 5A). Intriguingly, analysing the combined dataset of all 5 genotypes produced a distinct pairing for the HCV CRE corresponding to the alternative pairing found by SFOLD (lower right, labelled 'Alt' in Figure 4). By analysing

alignments of each combination of 2, 3 and 4 genotypes, a relationship was found between sequence diversity and frequency of detection of standard and alternative RNA structures (Figure 5A). Representative comparisons of duplexes formed in the alternative pairing for a range of HCV genotypes are shown in Figure 5B. The region of maximum potential interaction (9121-9107 with 9291-9305, see bar graph Figure 5B) can be divided into two areas; a less well conserved region (on the left in Figure 5B) involving sequences already implicated in forming the 3' side of the upper duplex of SL9266, and a highly conserved block of five nucleotides centred around 9110 and 9302. To functionally test the relevance of the predicted alternative pairing we undertook further mutagenesis studies.

Substitution of SL9266 with the analogous region of alternate genotypes

Of the two previously defined interactions of SL9266, one is local, forming the interrupted base-pairing of the CRE (16, 36), whereas the second is long-range, involving an interaction with the X tail SL2 (8). Within SL9266, the nucleotides in the terminal loop that base-pair with the 3'NCR are very highly conserved (8). Similarly, sequences occupying the bulge loop of SL9266 are highly conserved, whereas those forming the upper and lower duplexes show more variability. This accounts for the different level of conservation of base-pairing between the left and right hand sides of the interaction depicted with the upstream sequences depicted in Figure 5B. Assuming SL9266 folds similarly in each genotype of HCV, we reasoned that replacement of SL9266 in the sub-genomic replicon (1b genotype) with the analogous structures from other HCV genotypes might allow us to determine whether just some or all of sequences between 9291 and 9305 were also involved in the alternative pairing we predict.

Using a *Bsm* BI-based cassette system (see Materials and Methods) we precisely replaced the regions between nucleotides 9266 and 9312 with complementary

oligonucleotides corresponding to the analogous sequences of other genotypes of HCV. Inevitably, due to the sequence variation inherent in HCV, this strategy resulted in changes to the encoded NS5B polypeptide sequence (see Figure 6A). All modifications were made in a neomycin-expressing replicon that, in parallel with appropriate controls, was independently transfected into Huh7 cells and selected with G418. Of the eight substitutions made, five were tolerated well, generating approximately equivalent colony numbers to the positive control after G418 selection. The remaining three substitutions – of genotypes 3b (Tr), 4a (ED43) and 6g (JK046) – produced markedly reduced colony numbers indicating the modifications introduced within SL9266 were incompatible with replication.

It seemed unlikely that the differences in the replication phenotypes of the chimaeric replicons were due to introduction of incompatible residues to the NS5B polypeptide, with the possible exception of the 3b (Tr) sequence. The latter contains two amino acid substitutions ($G_{558}N$ and $P_{569}S$; Figure 6A) not present in the other sequences analysed. In the remaining genotype swaps, amino acid substitutions were restricted to just three residues of NS5B, with both viable and non-viable chimaeric replicons containing the same changes, implying that they alone do not account for the phenotype. For example, the replication-deficient replicon containing 4a (ED43) sequences has substitutions at 556, 564 and 566; of these, $S_{556}G$ is in 2b (HCJ8), $L_{564}M$ is in 5a (EUH1480) and $R_{566}H$ is in 1a (HP-H), all of which are replication competent. Therefore, unless particular combinations of these changes are deleterious, it seemed probable that the poor replication of 6g (JK046) and 4a (ED43) must be mediated at the level of RNA, either by disruption of an RNA-RNA interaction, or alteration of a sequence motif bound by a cellular or viral protein(s).

Replication competence of the chimeric replicon did not correlate directly with either invariant or covariant (underlined in Figure 6A) base-pairing within the upper duplex

region of SL9266, or the covariation within the alternative interaction with the upstream sequence (in bold in Figure 6B). For example, 1a (HP-H) and 4a (ED43) were identical to the control 1b replicon in the upper duplex of SL9266, but only the former could replicate. Similarly, 6g (JK046) contains two compensating changes in the upper duplex but cannot replicate, whereas 6a (EUHK2) and 5a (EUH1480) had the same covariance in the upper duplex and were replication competent. Within the region forming the bulge loop of SL9266, none of the chimeras changed the highly conserved 5'-GCCCG motif. However, of the six that contained variation within this region of SL9266 (namely 1a GLA, 1a HP-H, 2b HCJ8, 4a ED43, 6a EUHK2 and 6g JK046) two of the non-viable chimeras 4a (ED43) and 6g (JK046) lacked any covariant changes within this region, whereas 1a (GLA), 1a (HP-H) and 6a (EUHK2) all contained at least one covariant substitution that could be involved in base-pairing to the upstream sequence (highlighted in bold in Figure 6A). All chimeras also introduced covariant changes at C_{9291} (to A or G), the 5' nucleotide within the SL9266 sequence that could pair with U_{9121} (Figure 5B & 6A), though there was not a correlation between viability of the replicon and the particular substitution at this position.

Results obtained with the chimeric replicons suggested that the RNA-RNA interactions within SL9266 and the proposed alternative upstream pairing were non-trivial. We therefore specifically examined the upstream interaction in a more focussed manner by further site-directed mutagenesis.

Critical interactions between SL9266 and the upstream sequence

Mutations were introduced singly, or in combination, to SL9266 or the upstream sequence located around nt. 9110. In each instance substitutions were selected to leave the encoded NS5B polypeptide unchanged, thereby excluding the possibility that the resulting phenotype was due to the introduction of an incompatible amino acid into the virus polymerase. The majority of the mutations introduced were within the SL9266

sub-terminal bulge loop, or the complementary sequence around 9110, though additional changes were also made in the sequences implicated in forming the 3' side of the upper duplex in SL9266. These, or the complementary changes 3' to 9110, were designed to test the extent of the alternative interaction proposed by our bioinformatic analysis.

In the upstream sequence (Figure 7A, left hand panel), substitutions at C_{9108} and G_{9110} were incompatible with replication, whereas substitution of $U_{9107}C$, $C_{9113}A$ or a combination of changes at $A_{9114}C$ and $A_{9116}U$, also in combination with $C_{9113}A$, were tolerated well. Within the sequences that contribute to the upper duplex or bulge loop of SL9266 (Figure 7A, right hand panel), substitutions of $U_{9296}A$, alone or in combination with $U_{9299}G$ and $C_{9303}A$, prevented replication. This phenotype is presumably attributable to the change at U_{9296} which disrupts the stability of the upper duplex. Of the other single substitutions constructed, only $U_{9299}G$ had no impact on replication, with changes of C_{9302} and C_{9303} all preventing colony formation in the G418 transduction assay.

Mutations in the upstream- and SL9266-region were also combined to test whether complementary substitutions could restore the replication phenotype to resemble that of the parental replicon (Figure 7B). In addition, combinations of substitutions were introduced to determine the influence of increasing the potential hydrogen bonding between the upstream region and SL9266 sequences. Of the combinations constructed, four that restored the predicted ability to base-pair G_{9110} and C_{9302} all generated significant numbers of G418-resistant colonies after transduction and selection. The demonstration that individual substitutions of G_{9110} or C_{9302} that disrupted the predicted base-pairing prevented replication, whereas all but one in which duplex formation could occur (summarised in Figure 7C) were replication-competent, provides strong support for the interaction of these regions. Double substitution of

nucleotides $C_{9110}U$ and $C_{9303}A$ did not restore replication capacity. Furthermore, all combinations of mutations that included $U_{9296}A$ were incapable of replicating (Figure 7B); this included substitutions at 9113, 9114 and 9116, the addition of which significantly increased the potential for hydrogen bonding between the upstream and SL9266 sequences. This result suggested that disruption of the upper duplex of SL9266 by $U_{9296}A$ could not be compensated by strengthening the predicted interaction with upstream sequences.

The majority of mutations constructed in the neomycin-encoding replicon were also rebuilt into a replicon carrying a luciferase reporter gene. Huh7 cells were transfected and a time-course of luciferase activity determined over 3 days (Figure 7D). Of those tested, the mutants could be divided into three broadly defined groups. With the exception of single mutations involving nucleotides G_{9110}, C_{9302} or C_{9303}, all the replicons harbouring mutations that prevented replication in the G418-colony forming assay (see Figure 7A) exhibited a phenotype similar to that of the negative control (which lacks an NS5B active site). This group included the mutation of $U_{9296}A$, the double mutations of $C_{9113}A$ + $U_{9296}A$ and all the triple mutants tested. In contrast, replicons that had generated colony numbers similar to the parental 1b replicon (+ive control) generated luciferase activities indistinguishable from the parental luciferase-encoding replicon. These included $C_{9113}A$, $U_{9299}G$, and the double mutant $A_{9114}C/A_{9116}U$. Significantly, this group also included the double mutant $G_{9110}U/C_{9302}A$ (Figure 7D). The final group were intermediate in phenotype, exhibiting a steady decline of luciferase activity over the second and third day of the time course, but at a lower rate to those that resembled the polymerase-defective negative control. Although we only tested a limited representative range of substitutions predicted to be involved in the highly conserved (Figures 4 & 5b) upstream interaction, it was notable that all those exhibiting an intermediate phenotype were from this group. This included $G_{9110}U$, $C_{9302}A$ and $C_{9303}A$ (Figure 7D). One

explanation for this could be an increase in RNA stability. However, since this phenotype was only observed in mutants in which the RNA structure was destabilised we suspect that the enhanced translation may be explained by something other than an increase in RNA stability.

Discussion

Many viral proteins are multi-functional, for example controlling aspects of the virus replication cycle and the intracellular milieu. Increasingly, studies are demonstrating that the virus genome also has multiple functions, particularly in the small RNA and DNA viruses where coding capacity is limited. In the case of the small positive-strand RNA viruses the genome must act both as a template for translation and replication. At least on the input genome, before a pool of progeny genomes have been generated, these are mutually exclusive processes. In certain examples, additional functions ascribed to the RNA genome include subversion of the innate immune response, temporal and spatial control of the replication process, and encapsidation (12, 28, 35, 37). Functional specificity is provided by the evolutionary conservation of binding determinants, often in a structural context. The accurate prediction of stem-loop and higher-order structures therefore provides primary information on key functional domains of the virus genome.

Well-established thermodynamic methods exist to predict two-dimensional RNA structure (e.g. MFOLD, see 20, 38) which we have extended, implemented in the program StructureDist, to extract the additional information present in large datasets of related sequences. Using this and an alternative thermodynamic approach, SFOLD (7), we investigated structures in the terminal 700 nucleotides of the HCV coding region – an area of the genome in which we had previously identified at least five well-conserved stem-loop elements (31). One of the five structures predicted, an interrupted stem-loop

starting at 9266 (SL9266) shown in previous studies to be a *cis*-acting replication element, was only poorly predicted. An alternative non-thermodynamic method (PFOLD, see 14, 24) robustly predicted SL9266 in genotype 1, but analysis of all 6 genotypes of HCV indicated a hitherto unsuspected interaction of sequences within SL9266 and a region located approximately 200 nt. in a 5' direction (see Figure 4).

The finding of poor RNA structure conservation of the HCV replication element among alternative structures showing similar folding free energies (StructureDist and SFOLD), may arise from either an incorrect structure prediction for the HCV CRE using thermodynamic methods or because there is more than one (metastable) RNA structure in this region. The evidence that the alternative folding better accommodates sequence variability between genotypes using PFOLD even though the standard structure was predicted for individual genotypes provides further evidence for possible alternation in RNA structure in this genome region. Unfortunately, none of the structure prediction methods are able to incorporate tertiary RNA structure interactions, such as pseudoknots or kissing-loop interactions, in predicted structure models. These interactions may have significant stabilising or de-stabilising influences on the two predicted structures for the HCV CRE. Variability in prediction outcomes in this study may therefore result from incomplete prediction of potential pairings in this region of the HCV genome.

We investigated the relevance of the two predicted conformations of SL9266 to HCV replication by site directed mutagenesis of a sub-genomic replicon encoding either a neomycin resistance marker or luciferase reporter gene. The definition of SL9266 as a functional CRE was supported by limited site-directed mutagenesis (Figure 1CD, Figure 7AB). Disruption of the lower duplex (in *mut*1) or the sequences (*mut*2) implicated in the "kissing loop" interaction with SL2 (now SL9571 see 15) in the 3' Xtail prevented replication in agreement with results already published (8, 16, 36). Three of the

mutants (*mut*3, *mut*4 and *mut*6) had substitutions of $U_{9296}A$, a substitution that in our more extensive mutagenic analysis (Figure 7) was always incompatible with replication. However, our results suggest that the additional presence of $A_{9281}U$ (compare *mut*3, *mut*5 and *mut*6 in Figure 1CD) could somehow compensate for the otherwise lethal substitution of $U_{9296}A$. Our present understanding of SL9266, together with knowledge of interactions of SL9266 with the 3' UTR or the upstream sequences demonstrated here, does not explain how substitution of 9281 (unpaired in the terminal loop of SL9266) compensates for a mutation that destabilises the upper duplex of the CRE.

More extensive modification of SL9266 was achieved by substituting the entire structure with the analogous region of other genotypes of HCV. These modifications were intended to allow the distinction between the importance of interactions within the SL9266 structure, and those involving more distant sequences. Of the representative genotypes chosen, the sequence variation was unevenly distributed within the SL9266 structure, presumably reflecting evolutionary conservation of certain features. Significantly, all of the introduced sequences were invariant between nucleotides 9284 and 9290 (inclusive) in the terminal loop thereby excluding the possibility that the resulting phenotype of the chimeric replicons were due to disruption of the 'kissing loop' interaction with SL9571in the 3' NCR (8). Other regions of significant conservation existed within the 3' side of the bulge loop (9300-9304), and the central region of each of the two duplexes either side of the bulge loop. Unsurprisingly, considering the predicted structure of SL9266, there was good evidence for covariation within the region (underlined in Figure 6A), in particular at 9267/9312 and 9275/9296. All but one of the chosen sequences included an $A_{9281}U$ substitution and all also carried a change at 9291 that created the potential to interact with U_{9121} in the upstream region. The resulting phenotype of replicons in the G418-transduction assay (Figure 6B) indicated that there was a good correlation between the overall level of retained base-pairing – both within

SL9266 and between SL9266 and the upstream sequence around position 9110 – and viability of the chimeric replicon. Chimeras either generated good numbers of colonies, broadly equivalent in number to the unmodified replicon, or very limited numbers of G418-resistant colonies; the latter phenotype is consistent with the introduced mutation being grossly sub-optimal for replication, with the appearance of a limited number of colonies due to the acquisition of one or more compensatory mutations that restore replicative capacity. These are considered non-viable without the adaptive changes. The non-viable chimeras exhibited only 43% (6g JK046), 40% (4a ED43) or 30% (3b Tr) covariation, whereas all viable chimeras contained >50% covariation. For example, 70% of the 10 nucleotide changes between the genotype 1b parental replicon and the 6a (EUHK2) chimera were covariant – 5 within duplex regions of SL9266, at 9267, 9268, 9275, 9296 and 9311, and a further two, at 9291 and 9299, with regard to the upstream alternative interaction proposed here. Although based on a limited sample size these results suggest that both the SL9266 CRE and the interaction of SL9266 sequences with the upstream region were important for replication. These studies also demonstrated that there was no absolute requirement for a U at 9296; the viable chimeric replicons 2b (HCJ8), 5a (EUH1480) and 6a (EUHK2) all had a substitution at 9296, but also carried a covariant change at 9275 that retained the base-pairing in the upper stem of SL9266 (Figure 6A). However, base-pairing of 9275/9296, for example in 6g (JK046), was alone not sufficient for replication. In this chimera – encoding an identical NS5B polypeptide to the viable 1a HP-H construct (Figure 6A) – it is presumed that the overall reduced level of conserved base-pairing within SL9266 and between the bulge loop of SL9266 and the upstream sequences rendered the chimera non-viable.

Despite demonstrating that replicons chimeric for the SL9266 CRE exhibiting divergence of ~20% in this region were still replication competent, the distribution of substitutions within the replaced sequence meant that further site-directed mutagenesis

was required to determine the contribution of individual nucleotides to the predicted RNA-RNA interactions with the upstream region. Individual substitutions of U_{9107}, C_{9113} and U_{9299} were not detrimental to replicon activity, whether determined by luciferase activity or the generation of G418-resistance (Figure 7). Of these, C_{9113} and U_{9299} are juxtaposed in the predicted long-range interaction, but are not complementary in the majority of sequences. In contrast, a possible base-pair between 9107 and 9305 is highly conserved, but apparently not necessary for replication (see $U_{9107}C$ in Figure 7A and Figure 5B). Although substitution of 9107 had no apparent effect, modification of A_{9305} in isolation in a previous study (A92G/C/U in Figure 7 of 36) generated a wild-type phenotype when converted to a C, reduced colony numbers when a U and no colonies as a G. This suggests qualitative differences between the potential A-U or G-U pairing of 9107/9305 or, more likely, that 9305 is possibly involved in another RNA or protein interaction that has yet to be defined.

Although covariation of 9275/9296 (see Figure 6A) could be accommodated without destroying replication, all individual substitutions of A_{9296}, or combinations of mutations that included a change of A_{9296} were incapable of replicating (Figure 7A&B). This included the combination of $A_{9296}U$ with substitutions at 9113, 9114 and 9116. The latter were designed to increase potential hydrogen bonding between sequences within SL9266 and the upstream region. We interpret this to mean that additional bonding between these more distant regions cannot compensate for disruption of the upper duplex of SL9266.

The remaining substitutions involved the highly conserved five nucleotide 5'-GCCCG motif occupying the sub-terminal bulge loop of SL9266 and the perfect complementarity to a 5'-CGGGC sequence centred on nucleotide 9110. Individual synonymous substitutions in both regions, of $C_{9108}A$, G_{9110} to U, A or C, $C_{9303}A$ or C_{9302} to U, A or G all prevented colony formation in the G418 transduction assay. Of these, only $C_{9302}U$

was predicted to retain any capacity to base-pair with the upstream region. Interestingly, despite using standardised transfection conditions as with the chimeric SL9266 exchanges, point mutations in this region did not generate any colonies in our assays. Although not tested, this implies these mutants were incapable of generating revertant colonies under G418 selection. We went on to investigate the effect of substitutions in both parts of the predicted interacting sequence. In every case, dual mutations that restored the potential for base-pairing between position 9110 and 9302 resulted in a replication competent phenotype (Figure 7B&C). Individually, both 9110 and 9302 were substituted for each possible alternative nucleotide, indicating no sequence specificity at either position. It was perhaps surprising therefore that the single substitution of $C_{9302}U$, which left a potential interaction with G_{9110}, was incapable of replicating when a $G_{9110}A/C_{9302}U$ double mutant was viable. This strongly implies that a canonical Watson-Crick may be essential in this position to ensure the interaction of the two interacting regions. This conclusion is supported by an analysis of a large dataset of divergent HCV sequences, corresponding to available complete genome sequences of all six genotypes of HCV, in which none were identified with a G-U at this position (the distribution was A-U 12% and G-C 88%; data not shown). The requirement to retain synonymous substitutions prevented an individual mutation being introduced to restore complementarity between 9303 and 9109 (which respectively form the first and second nucleotides in arginine and serine codons).

Our results strongly support a long-range interaction between highly conserved sequences located in the sub-terminal bulge loop of SL9266 and a similarly conserved upstream region around nucleotide 9110 that is not implicated in any evolutionary conserved RNA structure. Additional supporting data for the importance of this interaction comes from the study by Friebe *et al*., who constructed a $G_{9300}A$ substitution (designated bulge-G>A, see reference 8) in a replicon with a duplication of SL9266

sequences and the flanking regions within the 3' NCR. This substitution rendered the replicon non-viable and because G_{9300} was now non-coding this could not be attributed to a defect in NS5B. In one construct, P1-ins3.2 (8), SL9266 alone was duplicated in the 3' NCR of a replicon bearing synonymous substitutions that disrupted the native SL9217, SL9266 and SL9324 structures in the NS5B coding region. Although this replicon exhibited 10- to 15- fold lower replication activity than wild type, it implies that the distance separating sequences around 9110 and the complementary functional SL9266 sequences are not absolutely critical for replication.

The data available from our analysis, and re-interpretation of previous studies of SL9266 (8, 16, 36), cannot unequivocally demonstrate whether formation of SL9266 and either or both of the upstream and downstream interactions are mutually exclusive events or could occur simultaneously. A number of scenarios are possible; the rather weak (as evidenced by the poor bioinformatic prediction) SL9266 structure could be stabilised by interaction with either or both sequences around 9110 and SL9571 to form a complex extended pseudoknot containing four duplexed regions. Alternatively, interaction of sequences normally unpaired within SL9266 with the 3' NCR and the 9110 region could destabilise, or prevent formation of, SL9266, thereby forming a molecular switch capable of adopting at least two conformations. Intermediates between these two examples – separately involving the 3' NCR or the upstream region – are also possible. Further mutagenic and functional studies will be needed to distinguish between these various possibilities. Considering the available data we currently favour a model in which SL9266 interacts, at least some of the time, with both the upstream and downstream sequences to form an extended pseudoknot structure, as illustrated in Figure 8. In our model we define the upstream interaction as involving complementarity between 5'-CGGGC and 5'-GCCCG sequences centred on nucleotides 9110 and 9302 respectively. Good evidence to support this interpretation

includes the primary involvement of single stranded regions of SL9266 in the long-range interactions. Furthermore, the phenotype exerted by the majority of substitutions introduced to SL9266 in this and previous studies can be interpreted as affecting either SL9266 *per se* or one or other of the long-range interactions. Sequences within the region 9108-9112/9300-9304 are highly conserved; of 192 divergent HCV sequences analysed all exhibited G_{9109} to C_{9303} and C_{9112} to G_{9300} pairings. There was a single – presumably unpaired – variant of C_{9108} to A_{9304}, the remainder being C_{9108} to G_{9304}, and another singleton of G_{9111} to U_{9301} with all others in the dataset being G to C pair at this position (data not shown). The variation of 9110 and 9302 is listed above. This conservation of Watson-Crick pairings presumably explaining the inhibition of replication mediated by the C_{9303}A substitution constructed by You and colleagues (their substitutions of C90, see reference 36). Overall there is less variation or co-variation in the unpaired regions of SL9266, compared with the lower and upper duplexes of the stem-loop (data not shown and reference 36). The lack of covariation in the pentanucleotide motif forming the upstream interaction described here is presumably a consequence of the juxtaposition of the third base 'wobble' position of the codons in these regions; almost all variation is restricted to substitution of a G_{9110}-C_{9302} pair by an A-U pair in genotype 6 sequences.

Many viruses are known to possess pseudoknots that contribute essential functions during the replication cycle. In most viruses, pseudoknots located within coding regions are primarily involved in translational control, in particular -1 frameshifting (2). However, there is no evidence for such a role in HCV, and the previously demonstrated positional independence of SL9266 would argue strongly against any such function. Instead it seems likely that the RNA structure forming SL9266, together with interactions of the unpaired loop sequences of SL9266 and both upstream and downstream regions, has one of more functions in genome replication. Precedents exist in bacteriophages,

several plant viruses and some animal RNA viruses. The first identified pseudoknot, the tRNA-like sequence (TLS) of turnip yellow mosaic virus (27), has multiple functions including recruitment of a nucleotidyl transferase for genome completion and genome circularisation (or at least juxtaposition of the 5' and 3' ends) probably via interaction with eIF1a, and consequent enhancement of translation. The TLS is also implicated in the switch from translation of the input genome to replication by competitive binding with newly synthesised viral polymerase and may also have a role in late replication functions such as encapsidation (4, 11, 21, 22). Genome circularisation by the TLS is probably protein-mediated, but long-range RNA-RNA interactions that form pseudoknots can critically influence the global folding of RNA. Such interactions form the core of the ribosome (reviewed in 2) and are also known to occur in virus genomes. In bacteriophage Qβ a pseudoknot spanning 1.2kb of the genome recruits the 3' end of the genome to the internally-bound viral replicase (13). Similarly, recruitment of the replicase to the 3' end of porcine reproductive and respiratory syndrome virus requires a long-range (~300 nt.) pseudoknot (33).

Considering the important role in replication of the complex pseudoknot proposed here it is perhaps unsurprising that the RNA structures in the 3' end of the HCV coding region (3) and SL9266, forming the core of the pseudoknot, interact with NS5B in *in vitro* assays (16). Although further investigation is required to define the function(s) of this complex RNA structure in the translation and replication of the HCV genome, our demonstration of important 5' interactions with the sub-terminal bulge loop of SL9266 provides a structural basis on which these studies can be based.

Acknowledgements

We thank Dr. R. Bartenschlager for the neomycin-encoding replicon, GlaxoSmithKline for the luciferase-encoding replicon, the Medical Research Council for financial support (DJE and PS) and MRC/GlaxoSmithKline for a CASE PhD. studentship (to RME) for VA.

Supplementary data

Table 1S

AF011751, HEC278830, HPCHCJ1, AF511948, HPCPLYPRE, AF511949, AF511950, AF207754, AB049091, HPCCGENOM, HPCJCG, AF207753, AB049088, AF207752, HPCUNKCDS, AY460204, HPVHCVN, AF207758, AF054250, AF207756, AF207767, AB049097, AB049100, AB080299, D85516, AF207763, AF165060, HCU01214, D11355, D89872, AF207762, AF207765, AF207770, AY045702, AF207759, AF165062, AF165058, AF165050, AY587844, AF165052, AF207772, AF207769, AB049099, HPCJRNA, AB049093, HPCHUMR, AF207774, AF139594, AB049098, HPCRNA, AF207766, AB049101, AF208024, HPCK1R1, HPCGENANT, AF176573, AF356827, HCVPOLYP, AF207761, AF313916, AJ238800, AB049090, AF207760, HPCK1R2, AF207764, HCVJK1G, AB191333, HPCPP, D89815, AF165064, AF165056, AB049096, AB016785, AF165048, AF207773, AB049095, AF207771, AF207757, AF165046, AB049094, AF165054, AB049087, AF207768, AB049092, AB049089, D50484, AJ132997, HCV132996, HCU45476, AF483269, HPCJ491, AF165059, HPCJTA, AF165061, AF165057, AF165049, AF165051, HCJ238799, AF165063, AF165055, AF165047, AF207755, AF165045, AF165053, HPCK1R3, AJ851228, AY651061, AY051292, HPCCGS, AF290978, AF271632, AB047639, AY746460, AB047642, AB047643, AF238483, AB047640, AF238481, AF169005, AF169002, AF169004, AB047644, AF238485, AF238484, AF169003, AB047645, AF238482, AB047641, AF177036, HPCPOLP, AB031663, DQ155561, D50409, AY232733, AY232747, AB030907, AY232749, AY232743, AY232735, AY232731, AF238486, AY232741, AY232739, AY232737, AY232745, HPCJ8G, DQ437509, HCVCENS1, AF046866, HPCEGS, HPCK3A, HPCJK049E, HPCFG, DQ418785, DQ418782, DQ418787, DQ516084, DQ418789, DQ418783, DQ418788, DQ418784, HCV4APOLY, DQ516083, DQ418786, AF064490, HCV1480, DQ278894, DQ278893, DQ278891, D84264, D84265, DQ314806, HPCJK046E, DQ278892, DQ314805, D84263, D84262, DQ480524, DQ480523, DQ480513, DQ480515, DQ480521, DQ480519, AY859526, HCV12083, DQ480518, DQ480520, DQ480516, DQ480522, DQ480512, DQ480517, DQ480514

Figure legends

Figure 1

SL9266 is a *cis*-acting replication element in hepatitis C virus

A) The generic organisation of the hepatitis C sub-genomic replicon expressing either a luciferase reporter gene, or neomycin selection marker, is shown, together with an indication of the location of SL9266 in the region encoding the C-terminus of NS5B.

B) The thermodynamically-predicted structure of SL9266.

C) Genetic analysis of synonymous mutations introduced to sub-genomic replicons. The sequence of SL9266 is shown with the third 'wobble' position of each triplet underlined. Underneath the sequence the location of individual mutations (*mut*1 to *mut*6) are shown, together with their phenotype (+ indicating growth) after G418 selection. The shaded boxes joined by horizontal brackets and lines indicate the duplex regions (lower [pale] and upper [dark]) of SL9266.

D) The phenotype of SL9266 neomycin-encoding replicon mutants *mut*1 to *mut*6 in a G418-selection assay.

Figure 2

Stem-loop structures in the NS5B-encoding region of HCV.

A) Predicted RNA secondary structures in the terminal 350 bases of the HCV coding sequence (in NS5B). Structures numbered according their position in the H77 reference sequence, using standard nomenclature for stem-loops (see Methods)

B) Frequencies of concordant pairing (left-hand y-axis) predictions and predicted unpaired bases (right hand y-axis) at each nucleotide position (x-axis) on pairwise comparison of most energetically favoured RNA structures predicted by MFOLD (38) for a set of 150 sequences representative of HCV genotypes 1-6. Frequencies were compiled using StructureDist v.1.3 (31). The location of each of the five predicted stem-loop structures is indicated above the graph. The location of the alternative upstream paired region is indicated as a black bar labelled 'Alt'.

Figure 3

SFOLD analysis of HCV NS5B sequences.

Numbers of consensus structures in 72 centroids generated by SFOLD from a total of 26 HCV NS5B sequences (positions 9001 – 9377) corresponding to standard stem-loop structures (see Figure 2A; filled black) or containing partial structure (filled grey). Frequencies of alternative pairings of the 3' side of SL9266 to upstream sequences are labelled as Alternative or Other boxes.

Figure 4

PFOLD analysis of HCV NS5B sequences.

Coordinates (dotplot) of pairing predictions for consensus structures predicted for alignments of HCV genotype 1b sequences (upper left) or HCV genotypes 1-6 (lower right) using PFOLD. The size of dots corresponds to reliability of pairing predictions. The positions of standard predicted structures and base-pairing forming the alternative RNA structure (Alt) are shown as grey filled ellipses.

Figure 5

Alternative interactions of SL9266 sequences in a range of HCV genotypes.

A) Frequencies of RNA structure prediction by PFOLD corresponding to the standard model (light grey filled boxes) or containing the alternative pairing (black filled boxes). The x-axis records the number of different genotypes in each alignment; numbers above bars records the number of different genotype combinations tested by PFOLD. For example, there are ten possible combinations of the five genotypes tested, all of which were analysed and the results presented in column two of the graph.

B) Comparison of duplexes formed in the alternative pairing for representative sequences of HCV genotypes 1-6. Genomic numbering for upstream and downstream bases shown at the top and bottom of the figure respectively. The location of known interactions of genotype 1b SL9266 are indicated at the top of the figure; 'KL' indicates the location of sequences forming a kissing loop interaction with the 3' Xtail (8), 'SL9266 Upper' and 'SL9266 Lower' indicate the 3' side of the upper and lower duplexes of SL9266. The grey block highlights the area of maximal conserved base-pairing (nucleotides 0291 – 9305 and 9121 –

9107; indicated in a simple bar chart at the bottom of the figure, each bar representing a single nucleotide in the aligned sequences) forming the predicted alternative interaction of sequences within SL9266 and the upstream region.

Figure 6

Exchange of SL9266 with the analogous region of other genotypes of HCV.

A) The SL9266 nucleotide sequence is shown (left) together with the nucleotide differences introduced by exchange with the sequences from a range of genotypes indicated. At the top, and emphasised with a dark shaded box, is the kissing loop interaction between the terminal loop of SL9266 and the 3'NCR (8). At the bottom, and highlighted by a pale shaded box, is the predicted interaction between SL9266 and upstream sequences centred around 9110. Underlined nucleotides in the SL9266 or upstream sequences indicate the third base 'wobble' position of codons. The upper and lower duplexes that form SL9266 are indicated by horizontal joined brackets (see also Figure 1C). Nucleotides underlined in the alternative genotype sequences retain the ability to form these duplexes. Nucleotides in bold within the dark shaded box retain (or acquire) the potential to base-pair with the upstream sequence. The NS5B amino acid sequences altered by exchange of SL9266 with the analogous region from other genotypes is indicated on the right hand side of the panel.

B) G418-selection assay of SL9266 substitutions for the sequences from the genotypes indicated.

Figure 7

Mutational analysis of the alternative interaction of sequences within SL9266.

A) The phenotype of neomycin-encoding replicons containing mutations within the upstream region (nt. 9107 – 9121; left panel) or within the sequences that form part of SL9266 (nt. 9291 – 9305; right panel). For each named mutant a photograph of a stained dish after G418 selection is shown next to the sequence indicating the impact on the alternative interaction predicted bioinformatically. For consistency with other figures, the upstream sequence is the lower sequence depicted. Substitutions are indicated in bold, as are additional or changed hydrogen bonding interactions. The total number of hydrogen bonds that could form between the sequences shown are indicated in the column headed 'H'. The regions of the SL9266 sequence that form the 3' side of the upper duplex of SL9266 is underlined. The positive control replicon is shown at the top of the figure. 'GND' indicates a control replicon containing active-site mutations within the NS5B polymerase (see Materials and Methods).

B) The phenotype of neomycin-encoding replicons containing substitutions in both upstream and SL9266 sequences.

C) Summary of changes made at nts. 9110 and 9302, with a (+) indicating a replication phenotype similar to that of a positive control, and a (-) indicating no apparent replication. 'nd' indicates not done.

D) Replication phenotype of luciferase-encoding sub-genomic replicons bearing mutations at nucleotides 9110, 9113, 9114, 9296, 9299, 9302, 9303 and combinations thereof. The average of two or three independent repeats at each time point are plotted.

Figure 8

Proposed structure of a complex pseudoknot in hepatitis C virus

A) Connected horizontal lines above and below a linear representation of the HCV genome (dotted line) indicate the interactions involved in formation of SL9266 (above) and the long-range interactions (below) with sequences located 5’ and 3’ to SL9266. The position of evolutionarily conserved stem loop structures in the NS5B coding region and the Xtail in the 3’ NCR are also indicated.

B) Schematic of a complex pseudoknot involving SL9266 and long-range interactions between the sub-terminal bulge loop and sequences centred on nucleotide 9110 and the SL9266 terminal loop and complementary sequences in SL9571.

References

1. **Blight, K. J., A. A. Kolykhalov, and C. M. Rice.** 2000. Efficient initiation of HCV RNA replication in cell culture. Science **290:**1972-4.
2. **Brierley, I., S. Pennell, and R. J. Gilbert.** 2007. Viral RNA pseudoknots: versatile motifs in gene expression and replication. Nat Rev Microbiol **5:**598-610.
3. **Cheng, J. C., M. F. Chang, and S. C. Chang.** 1999. Specific interaction between the hepatitis C virus NS5B RNA polymerase and the 3' end of the viral RNA. J Virol **73:**7044-9.
4. **Choi, Y. G., and A. L. N. Rao.** 2003. Packaging of Brome Mosaic Virus RNA3 Is Mediated through a Bipartite Signal. Journal of Virology **77:**9750-9757.
5. **Clerte, C., and K. B. Hall.** 2006. Characterization of multimeric complexes formed by the human PTB1 protein on RNA. RNA **12:**457-475.
6. **Ding, Y., C. Y. Chan, and C. E. Lawrence.** 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. Rna **11:**1157-66.
7. **Ding, Y., and C. E. Lawrence.** 2003. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res **31:**7280-301.

8. **Friebe, P., J. Boudet, J. P. Simorre, and R. Bartenschlager.** 2005. Kissing-loop interaction in the 3' end of the hepatitis C virus genome essential for RNA replication. J Virol **79:**380-92.
9. **Friebe, P., V. Lohmann, N. Krieger, and R. Bartenschlager.** 2001. Sequences in the 5' nontranslated region of hepatitis C virus required for RNA replication. J Virol **75:**12047-57.
10. **Fukushi, S., M. Okada, T. Kageyama, F. B. Hoshino, K. Nagai, and K. Katayama.** 2001. Interaction of poly(rC)-binding protein 2 with the 5'-terminal stem loop of the hepatitis C-virus genome. Virus Res **73:**67-79.
11. **Giege, R.** 1996. Interplay of tRNA-like structures from plant viral RNAs with partners of the translation and replication machineries. Proc Natl Acad Sci U S A **93:**12078-81.
12. **Han, J.-Q., H. L. Townsend, B. K. Jha, J. M. Paranjape, R. H. Silverman, and D. J. Barton.** 2007. A Phylogenetically Conserved RNA Structure in the Poliovirus Open Reading Frame Inhibits the Antiviral Endoribonuclease RNase L. J. Virol. **81:**5561-5572.
13. **Klovins, J., V. Berzins, and J. van Duin.** 1998. A long-range interaction in Qbeta RNA that bridges the thousand nucleotides between the M-site and the 3' end is required for replication. Rna **4:**948-57.
14. **Knudsen, B., and J. Hein.** 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics **15:**446-54.
15. **Kuiken, C., C. Combet, J. Bukh, I. T. Shin, G. Deleage, M. Mizokami, R. Richardson, E. Sablon, K. Yusim, J. M. Pawlotsky, and P. Simmonds.** 2006. A comprehensive system for consistent numbering of HCV sequences, proteins and epitopes. Hepatology **44:**1355-61.
16. **Lee, H., H. Shin, E. Wimmer, and A. V. Paul.** 2004. cis-acting RNA signals in the NS5B C-terminal coding sequence of the hepatitis C virus genome. J Virol **78:**10865-77.
17. **Lindenbach, B. D., M. J. Evans, A. J. Syder, B. Wolk, T. L. Tellinghuisen, C. C. Liu, T. Maruyama, R. O. Hynes, D. R. Burton, J. A. McKeating, and C. M. Rice.** 2005. Complete replication of hepatitis C virus in cell culture. Science **309:**623-6.
18. **Lohmann, V., F. Körner, J. Koch, U. Herian, L. Theilmann, and R. Bartenschlager.** 1999. Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. Science **285:**110-3.
19. **Luo, G., S. Xin, and Z. Cai.** 2003. Role of the 5'-proximal stem-loop structure of the 5' untranslated region in replication and translation of hepatitis C virus RNA. J Virol **77:**3312-8.
20. **Markham, N. R., and M. Zuker.** 2005. DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res **33:**W577-81.
21. **Matsuda, D., and T. W. Dreher.** 2004. The tRNA-like structure of Turnip yellow mosaic virus RNA is a 3'-translational enhancer. Virology **321:**36-46.
22. **Matsuda, D., S. Yoshinari, and T. W. Dreher.** 2004. eEF1A binding to aminoacylated viral RNA represses minus strand synthesis by TYMV RNA-dependent RNA polymerase. Virology **321:**47-56.
23. **McMullan, L. K., A. Grakoui, M. J. Evans, K. Mihalik, M. Puig, A. D. Branch, S. M. Feinstone, and C. M. Rice.** 2007. Evidence for a functional RNA element in the hepatitis C virus core gene. PNAS **104:**2879-2884.
24. **Pedersen, J. S., I. M. Meyer, R. Forsberg, P. Simmonds, and J. Hein.** 2004. A comparative method for predicting and folding RNA secondary structures within protein coding regions. Submitted.

25. **Pietschmann, T., A. Kaul, G. Koutsoudakis, A. Shavinskaya, S. Kallis, E. Steinmann, K. Abid, F. Negro, M. Dreux, F. L. Cosset, and R. Bartenschlager.** 2006. Construction and characterization of infectious intragenotypic and intergenotypic hepatitis C virus chimeras. Proc Natl Acad Sci U S A **103:**7408-13.
26. **Reusken, C. B., T. J. Dalebout, P. Eerligh, P. J. Bredenbeek, and W. J. Spaan.** 2003. Analysis of hepatitis C virus/classical swine fever virus chimeric 5'NTRs: sequences within the hepatitis C virus IRES are required for viral RNA replication. J Gen Virol **84:**1761-9.
27. **Rietveld, K., R. van Poelgeest, C. W. A. Pleij, J. H. Van Boom, and L. Bosch.** 1982. The tRNA-like structure at the 3' terminus of Turnip Yellow Mosaic Virus RNA. Differences and similarities with canonical tRNA. Nucleic Acids Research **10:**1929-1946.
28. **Sasaki, J., and K. Taniguchi.** 2003. The 5'-End Sequence of the Genome of Aichi Virus, a Picornavirus, Contains an Element Critical for Viral RNA Encapsidation. J. Virol. **77:**3542-3548.
29. **Simmonds, P., A. Tuplin, and D. J. Evans.** 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA **10:**1337-1351.
30. **Spangberg, K., and S. Schwartz.** 1999. Poly(C)-binding protein interacts with the hepatitis C virus 5' untranslated region. J Gen Virol **80:**1371-1376.
31. **Tuplin, A., D. J. Evans, and P. Simmonds.** 2004. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. J Gen Virol **85:**3037-3047.
32. **Tuplin, A., J. Wood, D. J. Evans, A. H. Patel, and P. Simmonds.** 2002. Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. RNA **8:**824-841.
33. **Verheije, M. H., R. C. L. Olsthoorn, M. V. Kroese, P. J. M. Rottier, and J. J. M. Meulenberg.** 2002. Kissing Interaction between 3' Noncoding and Coding Sequences Is Essential for Porcine Arterivirus RNA Replication. Journal of Virology **76:**1521-1526.
34. **Wakita, T., T. Pietschmann, T. Kato, T. Date, M. Miyamoto, Z. Zhao, K. Murthy, A. Habermann, H. G. Krausslich, M. Mizokami, R. Bartenschlager, and T. J. Liang.** 2005. Production of infectious hepatitis C virus in tissue culture from a cloned viral genome. Nat Med **11:**791-6.
35. **Wang, S., and K. A. White.** 2007. Riboswitching on RNA virus replication. PNAS **104:**10406-10411.
36. **You, S., D. D. Stump, A. D. Branch, and C. M. Rice.** 2004. A cis-acting replication element in the sequence encoding the NS5B RNA-dependent RNA polymerase is required for hepatitis C virus RNA replication. J Virol **78:**1352-66.
37. **Yu, L., and L. Markoff.** 2005. The Topology of Bulges in the Long Stem of the Flavivirus 3' Stem-Loop Is a Major Determinant of RNA Replication Competence. J. Virol. **79:**2309-2324.
38. **Zuker, M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31:**3406-3415.