Independent evolution of tetraloop in enterovirus oriL replicative element and its putative binding partners in virus protein 3C

Maria A. Prostova[1], Andrei A. Deviatkin[1], Irina O. Tcelykh[1,2], Alexander N. Lukashev[1,3] and Anatoly P. Gmyl[1,2,3]

[1] Chumakov Institute of Poliomyelitis and Viral Encephalitides, Moscow, Russia
[2] Lomonosov Moscow State University, Moscow, Russia
[3] Sechenov First Moscow State Medical University, Moscow, Russia

ABSTRACT

Background. Enteroviruses are small non-enveloped viruses with a (+) ssRNA genome with one open reading frame. Enterovirus protein 3C (or 3CD for some species) binds the replicative element oriL to initiate replication. The replication of enteroviruses features a low-fidelity process, which allows the virus to adapt to the changing environment on the one hand, and requires additional mechanisms to maintain the genome stability on the other. Structural disturbances in the apical region of oriL domain d can be compensated by amino acid substitutions in positions 154 or 156 of 3C (amino acid numeration corresponds to poliovirus 3C), thus suggesting the co-evolution of these interacting sequences in nature. The aim of this work was to understand co-evolution patterns of two interacting replication machinery elements in enteroviruses, the apical region of oriL domain d and its putative binding partners in the 3C protein.
Methods. To evaluate the variability of the domain d loop sequence we retrieved all available full enterovirus sequences (>6,400 nucleotides), which were present in the NCBI database on February 2017 and analysed the variety and abundance of sequences in domain d of the replicative element oriL and in the protein 3C.
Results. A total of 2,842 full genome sequences was analysed. The majority of domain d apical loops were tetraloops, which belonged to consensus YNHG (Y = U/C, N = any nucleotide, H = A/C/U). The putative RNA-binding tripeptide 154–156 (*Enterovirus C* 3C protein numeration) was less diverse than the apical domain d loop region and, in contrast to it, was species-specific.
Discussion. Despite the suggestion that the RNA-binding tripeptide interacts with the apical region of domain d, they evolve independently in nature. Together, our data indicate the plastic evolution of both interplayers of 3C-oriL recognition.

Subjects Evolutionary Studies, Genomics, Virology
Keywords *Enterovirus*, RNA-protein interaction, Tetraloop, Virus evolution

Submitted 8 August 2017
Accepted 16 September 2017
Published 6 October 2017

Corresponding author
Maria A. Prostova,
prostova_ma@chumakovs.su,
prostovna@gmail.com

Academic editor
Ana Grande-Pérez

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj.3896**

INTRODUCTION

Enteroviruses are small non-enveloped viruses with a plus strand genome about 7500 nt long which contains one open reading frame that encodes structural (capsid) and non-structural proteins, 5′ and 3′ NTRs (non translated regions), and polyA on the 3′ end

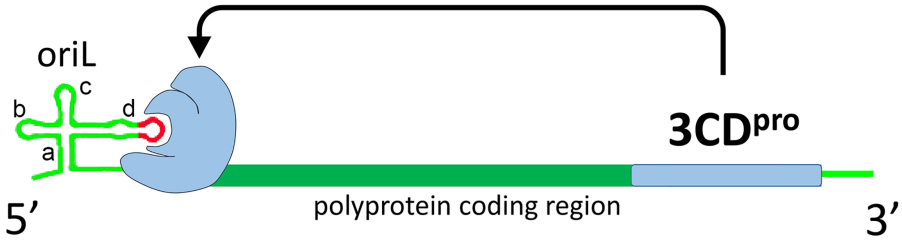

Figure 1 Schematic representation of interaction of poliovirus protein 3CD (colored with blue) with poliovirus genome replicative element oriL. Domains a, b, c and d of oriL are labeled. Apical region of domain d, corresponding to the tetraloop and its flanking base pair, is colored with red.

(*Palmenberg, Neubauer & Skern, 2010*) (Fig. 1). Most non-structural enterovirus proteins are polyfunctional. Protease 3CD is a precursor of polymerase 3D and plays a key role in the initiation of replication (*Harris et al., 1992*; *Gamarnik & Andino, 1998*; *Thompson & Peersen, 2004*). After translation by host cell ribosomal machinery, the genome is utilized for the synthesis of the (−) strand RNA, which, in turn, serves as a matrix for the synthesis of multiple daughter (+) strands. Non-translated regions of the genome and a coding sequence within the genomic region encoding the viral helicase 2C contain replicative elements, which interact with viral and host proteins. These RNA-protein complexes regulate initiation and further steps of replication. For poliovirus, the most clinically relevant member of the *Enterovirus* genus, there are at least three known RNA-protein complexes, which are formed with the replicative elements oriL, oriR and oriI during replication (Fig. 1).

Complex oriL with viral protein 3CD and the host protein PCBP2 is crucial for the transcription initiation (*Goodfellow et al., 2000*; *Vogt & Andino, 2010*; *Chase, Daijogo & Semler, 2014*). The element oriL has a cloverleaf-like secondary structure with four domains termed a (stem of the cloverleaf), b, c and d (leaves of the cloverleaf) (*Trono, Andino & Baltimore, 1988*; *Andino, Rieckhof & Baltimore, 1990*) (Fig. 1). Previously, it was demonstrated *in vitro* that 3CD (or 3C) of poliovirus, coxsackievirus B3 and bovine enterovirus 1 interacts with the apical loop and the flanking base pairs of hairpin d in the oriL element (*Andino, Rieckhof & Baltimore, 1990*; *Du et al., 2003*; *Ihle et al., 2005*) (Fig. 1).

The apical loops of domain d in genomes belonging to several viruses of *Enterovirus* genera were shown by NMR experiments to be tetraloops with a specific spatial structure, which belongs to the UNCG structural class of stable tetraloops (*Du et al., 2003*; *Du et al., 2004*; *Ihle et al., 2005*; *Melchers et al., 2006*). There are several known structural classes of tetraloops, three of which, named according to consensus sequences, contain tetraloops of extreme stability: UNCG (where N = any nucleotide), GNRA (where R = A/G) and gCUUGc (*Uhlenbeck, 1990*; *Cheong & Cheong, 2010*). Tetraloops of UNCG and GNRA classes are the most widely represented (*Woese et al., 1990*; *Cheong & Cheong, 2010*; *Bottaro & Lindorff-Larsen, 2017*). Previously, it was shown that only tetraloops of the UNCG structural class, but not tetraloops of GNRA or gCUUGc structural classes, can support effective replication of the poliovirus genome (*Prostova et al., 2015*). Moreover, the exact sequence of the apical region of poliovirus domain d was of less importance for effective

3CD-oriL recognition than its spatial structure (*Rieder et al., 2003*; *Prostova et al., 2015*). At the same time, structural disturbance in the apical region of oriL domain d of poliovirus could be compensated by amino acid substitutions in the tripeptide 154–156 of the 3C protein (here and hereafter amino acid numeration corresponds to poliovirus 3C protein) (*Andino et al., 1990*; *Prostova et al., 2015*). In addition to triplet 154–156, the conserved motif $_{82}$KFRDI$_{86}$ of the 3C protein also takes part in the oriL recognition (*Andino et al., 1990*; *Andino et al., 1993*; *Hämmerle, Molla & Wimmer, 1992*; *Shih, Chen & Wu, 2004*). To date, a comprehensive analysis of the diversity in domain d apical region and amino acid tripeptide sequences in the *Enterovirus* genus has not been conducted.

The replication of an enterovirus is a low-fidelity process, generating, on average, one mutation per genome (*Sanjuán et al., 2010*; *Acevedo, Brodsky & Andino, 2013*). The high probability of mutation allows the virus to adapt to a constantly changing environment on the one hand, but requires additional mechanisms to maintain genome stability on the other (*Wagner & Stadler, 1999*; *Lauring, Frydman & Andino, 2013*). The aim of the present study was to understand co-evolution patterns of two interacting replication machinery elements in enteroviruses, the apical domain d of oriL and the 3C protein.

MATERIALS AND METHODS

Formation and filtration of sets of full genomes

All available nucleotide sequences (as of February 2017) containing the *Enterovirus* genus with length 8000>n>6800 were extracted from the NCBI database. For every species, a multiple sequence alignment was conducted using MAFFT version 7 with default settings (*Katoh & Standley, 2013*). Sequences that contained more than 50 N characters in succession and sequences that were annotated as "Modified_Microbial_Nucleic_Acid", were removed from alignments. All sequences that differed from any other sequence in the dataset by less than 1% of the nucleotide sequence were omitted in order to reduce the bias caused by over-represented sequences.

Analysis of tetraloop and amino acid variety in the sets of genomes

For analysis of domain d sequence variety, the multiple sequence alignments were used. The relevant region of multiple sequence alignment and the respective names of sequences were analysed in Microsoft Excel. To analyse correlation of the domain d loop and tripeptide of 3C sequences the same alignments were translated in the protein 3C coding region. The resulting amino acid sequences that corresponded to tripeptides 154–156 (poliovirus 3C numerations) were analysed using Microsoft Excel. An amino acid frequency plot was created via the WebLogo server using the set of filtered genomes for every species (*Crooks et al., 2004*). To do this, the multiple sequence alignment of filtered genomes of every species was translated in the region that codes protein 3C, while positions 71–89 and 147–160 were saved in separate MAS files, which were then used to produce logos.

Domain d secondary structure

The domain d secondary structure was folded using the Vienna RNA Websuite server with subsequent manual editing (*Gruber et al., 2008*; *Lorenz et al., 2011*). Algorithm accounting for minimum free energy and partition function was used.

Table 1 Number of full genome sequences that contained oriL region and number of unique domain d sequences before and after filtration. For *Enterovirus E* and *F* number of unique tetraloops is shown separately for first and the second oriL.

Species	Number of full genome sequences	Number of full genome sequences after 1% nucleic identity filtration	Number of unique tetraloops before filtration	Number of unique tetraloops after filtration
Enterovirus A	1052	564	17	16
Enterovirus B	339	244	18	18
Enterovirus C	747	274	15	12
Enterovirus D	419	57	7	6
Enterovirus E	12	10	6 5	6 5
Enterovirus F	13	10	4 3	4 3
Enterovirus G	10	8	6	6
Enterovirus H	3	2	2	2
Enterovirus J	8	5	3	3
Rhinovirus A	159	118	8	8
Rhinovirus B	50	37	7	7
Rhinovirus C	38	37	6	6

RESULTS

Sample characteristics

To evaluate the variability of the domain d loop sequence we retrieved all available complete genome (8000>n>6800 nucleotides) enterovirus (EV) sequences that were present in the NCBI database on February 2017. Representatives of *Enterovirus A* (1173 sequences in total), *Enterovirus B* (414), *Enterovirus C* (773), *Enterovirus D* (462), *Enterovirus E* (12), *Enterovirus F* (13), *Enterovirus G* (15), *Enterovirus H* (3), *Enterovirus J* (7), *Rhinovirus A* (202), *Rhinovirus B* (76) and *Rhinovirus C* (51) species were analysed. As expected, genomes of epidemiologically significant viruses were the most represented in the database. For example, 66% of *Enterovirus A* species genomes belonged to the EV71 type, the causative agent of hand, foot and mouth disease (*Solomon et al., 2010*); most *Enterovirus C* species sequences (78%) belonged to poliovirus; and most *Enterovirus D* species sequences (98.7%) represented EV68, an aetiological agent of severe respiratory illness (*Oermann et al., 2015*). The number of genome sequences of each species that contained the oriL region is shown in Table 1.

Sequences of apical regions in oriL domain d and the amino acids involved in RNA recognition in 3C protein were analysed. In genomes of *Enterovirus E* and *Enterovirus F* species with two oriLs (*Pilipenko, Blinov & Agol, 1990*; *Zell et al., 1999*) sequences of both oriLs were analysed (Table 1). To reduce the bias towards particular loop sequences present in a large set of closely related genomes, which, for example, belonged to one outbreak, all sequences that differed from any other sequence in the dataset by less than 1% of the nucleotide sequence were omitted. After curation, the sizes of the largest data sets decreased dramatically, but the number of unique loop sequences in every set did not change significantly (Table 1). Unique tetraloop variants were lost for *Enterovirus A*

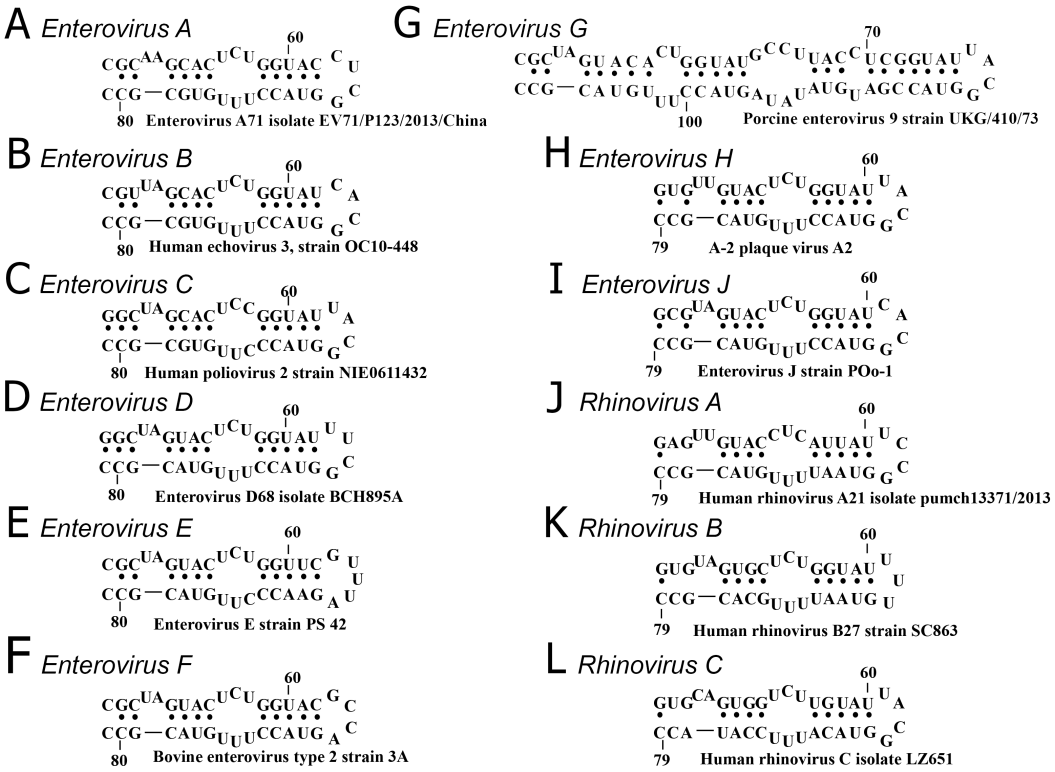

Figure 2 Secondary structure of oriL domain d of distinct enterovirus species. (A–I) secondary structure of oriL domain d in *Enterovirus A-J* genome. For *Enterovirus E* and *F* domain d of the first oriL is shown. Secondary structure of domain d of Porcine enterovirus 9 strain UKG/410/73 was folded with use as reference (*Krumbholz et al., 2002*); (J–L) secondary structure of oriL domain d in *Rhinovirus A-C* genome.

(tetraloop UGUG), *Enterovirus C* (tetraloops CCCG, CAUG and UGUG) and *Enterovirus D* (tetraloop UUGG). This indicates that, even among closely related genomes, the tetraloop sequence can vary. Indeed, in several outbreaks caused by EV71 or PV1, closely related genomes contained different apical domain d sequences (not shown). It should be noted that the filtration of the dataset using a 95% sequence identity threshold resulted in a dramatic loss of unique tetraloop variants (107 genomes out of 1,052 were left after filtration, while 13 unique tetraloop variants out of total 17 variants were detected in the filtered set).

Variability in the oriL domain d apical loop sequence

The secondary structure of domain d was conserved in all species, except *Enterovirus G*, in which an elongated domain d was observed (Fig. 2) (*Krumbholz et al., 2002*).

The variety and occurrence of various loops in the apical region of domain d in all species of the *Enterovirus* genera were analysed in filtered sets of full genome sequences. Most domain d apical loops were tetraloops (i.e., they consisted of four nucleotides) (Table 2). However, triloops (3-nucleotide loop) could be found in genomes of *Enterovirus C* and *Rhinovirus A* and *B* species, whereas pentaloops (5-nucleotide loop) were detected in genomes of Enterovirus E species (Table 2).

Table 2 Occurrence of domain d apical sequences in filtered sets of full genomes of different enterovirus species and serotypes. Tetraloops CCCG, UGUG, CAUG and UUGG that were unique for species *Enterovirus A, B, C* and *D* and were lost upon filtration, were added to maintain the diversity of loop sequence and are shown in blue. The gradient coloring from red to green represents abundance heat map for the genomes with different domain d sequence.

Loop sequence	Enterovirus														Rhinovirus		
	A				B	C			D	E	F	G	H	J	A	B	C
	all	EV71	EV71 C4 genotype	non EV71		all	PV	non PV									
Triloops																	
CCG						1		1									
CAG						1		1									
UCU															1	5	
UUU																17	
UAU																8	
AUU																4	
UGU																1	
UUC																1	
GAU																1	
YNMG Tetraloops																	
UACG	85	28		57	51	106	64	42				3	1	2	38		15
UGCG	114	2		112	31	43	31	12				1	1		2		
UUCG	16	16	14		3				50						6		6
UCCG	2			2	11	1	1								53		10
CACG	48	28		20	98	101	54	47	1			1		2	5		
CGCG	3	2		1	3	13	6	7				1					
CUCG	132	127	126	5	5	2		2	2						1		3
CCCG	40	39	28	1	16	1		1	1						12		2
UAAG	10	10			2												
UGAG	22	22			1												
UUAG																	
UCAG																	
CAAG	1	1			4	1		1						1			
CGAG	1			1		2		2									
CUAG																	
CCAG																	
YACG					1												
YNUG Tetraloops																	
UAUG	54	1		53											1		
UGUG	1			1	1	1		1									
UUUG									1								
UCUG					1												
CAUG					9	1		1									
CGUG	1			1	3	2		2									
CUUG	34	34	35		3				2								
CCUG	1	1	1		1	1		1									
GYYA Tetraloops																	
GCUA										2	13						
GCCA											3						
GUUA										2	3	1					
Other tetraloops																	
UUGG									1								
CUUC																	1
AUUA											1						
Pentaloops																	
GCUUA										7							
GUUUA										2							
GCCUA										4							
GCGUA										1							
GCGUA										1							
GAUUA										1							
GUCUA										1							

1
2
11
31
51
132

The most common loop sequences belonged to consensuses YNMG (Y = C/U, N = any, M = A/C; tetraloops with UNCG class spatial structure belong to this consensus) and YNUG (tetraloops with UNCG class and gCUUGc class spatial structures belong to this consensus) (Table 2). Consensus YNMG and consensus YNUG together corresponded to 24 unique sequence variants. Interestingly, in our dataset of 2,842 full genomes, four tetraloops out of 24 possible variants were never found in the domain d apical region:

UUAG, UCAG, CUAG and CCAG (Table 2). Thus, dinucleotides UA and CA are likely to be avoided at positions 2 and 3 of the tetraloop in enterovirus genomes.

In *Enterovirus A* species, 17 out of 24 possible unique tetraloop sequences were identified (Table 2, Table S2). Twelve unique loops of *Enterovirus A* belonged to consensus YNMG, while the other five belonged to consensus YNUG. The most abundant tetraloop in Enterovirus A genomes, in contrast to other species, was CUCG (Table 2, Table S1). This is explained by the prevalence of this tetraloop in the EV71 C4 genotype (Table 2, Table S1). The frequency of other tetraloop sequences varied significantly (Table 2, Table S2). One tetraloop (UGUG) was lost upon filtration. Such sequences here and below were manually added to the final data set to maintain diversity of loop sequences within the species, as well as provide comprehensive information about sequences in apical domain d in viable viruses (Table 2). Interestingly, EV71 sequences contained 13 out of 17 tetraloop variants, which were detected in the *Enterovirus A* genus (Table 2). In other words, the diversity of tetraloops in one discrete lineage in general resembles its diversity in the unification of different discrete lineages.

In Enterovirus B genomes, 18 unique tetraloops out of 24 possible were found. Twelve of these tetraloops belonged to consensus YNMG and six to consensus YNUG (Table 2, Table S3). The most abundant tetraloops were CACG (98 genomes), UACG (51 genomes) and UGCG (31 genomes), which were also present among the most abundant tetraloops of Enterovirus A species.

In genomes of the *Enterovirus C* species, nine unique tetraloops belonged to the YNMG consensus and four to the YNUG consensus. Three unique tetraloops were lost upon filtration and added to the final data set (CCCG, UGUG, CAUG) (Table 2, Table S4). Two genomes annotated in the NCBI data base as Human coxsackievirus A21, strain Coe, (accession number D00538) and Human coxsackievirus A21, strain BAN00-10467, (accession number EF015031) contained triloops CAG and CCG, respectively. The most abundant tetraloops in EV-C species were UACG (106), CACG (101), UGCG (43) (Table 2, Tables S1 and S4), which corresponds to the Sabin vaccine strains of poliovirus serotypes 2, 3 and 1, respectively. To evaluate bias caused by the redundant number of vaccine strain sequences in the data set, we subtracted genomes of vaccine/vaccine derived poliovirus strains from the analysed set. As a result, tetraloops UACG, CACG and UGCG were still the most frequent variants (Table 2, Table S1).

Only 57 Enterovirus D genomes out of 419 were left after 1% identity filtration. Fifty genomes belonged to Human enterovirus 68, the aetiological agent of respiratory illness. All genomes of this type contained loop UUCG in the domain d apical region. Other tetraloops were UUUG (1), CUCG (2), CCCG (1), CUUG (2) and CACG (1) (Table 2, Table S5). One tetraloop (UUGG) was lost upon filtration and manually added to the final data set.

Species *Enterovirus E* and *F* have two oriLs in the 5′ NTR, generally with similar sequences in the apical region of domain d (*Pilipenko, Blinov & Agol, 1990*; *Zell et al., 1999*) (Table S6). As such, we united sequences from the first and the second oriL of these viruses in the heat map (Table 2). Domain d loops in 10 genomes of Enterovirus F were tetraloops, while, in 10 Enterovirus E genomes, there were both tetraloops (first oriL) and pentaloops

(first and second oriL) (Table 2, Table S6). There were four diverse tetraloop sequences in oriLs of *Enterovirus E* and *F* with no obvious preference between these species. These sequences were GCUA, GUUA, GCCA, AUUA (Table 2, Table S6). Tetraloop AUUA was found once in the first oriL domain d of EV-F (strain PS87/Belfast, accession number DQ092794) (Table 2, Table S6). There were six diverse pentaloop sequences in domain d of Enterovirus E genomes—GCUUA, GUUUA, GCCUA, GCGUA, GAUUA, GUCUA (Table 2, Table S6).

All domain d loops in genomes of *Enterovirus G, H* and *J* species were tetraloops; all except one tetraloop variant belonged to consensus YNMG (Table 2, Table S7). One Enterovirus G representative had a GUUA tetraloop sequence (strain LP 54, accession number AF363455), similar to loops of *Enterovirus E* and *F* species (Table 2). This genome had only one oriL with the same domain d length as that of Enterovirus G genomes (*Krumbholz et al., 2002*).

All except one (isolate V38_URT-6.3m, accession number JF285329) of the full genomes of Rhinovirus A species and all full genomes of Rhinovirus C species had tetraloops in the apical regions of domain d (Table 2). Tetraloops of these viruses in almost all cases belonged to consensus YNMG, with one exception found in Rhinovirus C (tetraloop CUUC, isolate JAL-1, accession number JX291115) (Table 2, Table S8). All loops in the apical region of Rhinovirus B domain d were triloops (Table 2, Table S8).

Thus, the secondary structure of domain d was very similar among species of the *Enterovirus* genus, with the exception of *Enterovirus G* species (Fig. 2). The apical region of domain d has a high diversity of sequences; however, in species of *Enterovirus A, B, C, D, G, H* and *J* and *Rhinovirus A* and *C,* it mostly corresponds to the same consensus, that is, YNHG (Y = C/U, N = any, H = A/C/U).

Variety of RNA-recognition tripeptide of 3C

Two motifs of protein 3C are involved in RNA recognition and interact with oriL: the conservative motif KFRDI (positions 82–86 of poliovirus 3C) and the putative RNA-binding tripeptide (positions 154–156 in poliovirus 3C) (*Andino et al., 1990*; *Andino et al., 1993*; *Hämmerle, Hellen & Wimmer, 1991*; *Shih, Chen & Wu, 2004*). Substitutions in the putative RNA-binding tripeptide are known to compensate for disturbance in the apical region of domain d, such that the RNA-binding tripeptide is a putative candidate to co-evolve with the domain d loop (*Andino et al., 1990*; *Prostova et al., 2015*). There are other amino acids that have been found to affect oriL-3CD interaction, but tripeptide 154–156 (*Enterovirus C* 3C protein numbering here and below) is the only one that is proven to compensate structural disturbance in the domain d apical region (*Andino et al., 1990*; *Andino et al., 1993*). To evaluate the possible co-evolution between the domain d tetraloop and its putative interaction partners in protein 3C, relevant sequences in the filtered full genome data sets were analysed.

Motif $_{82}$KFRDI$_{86}$ was conserved in all species, as well as amino acids Glu 71 and Cys 147 of the protease catalytic triad (Fig. 3). In overwhelming majority of cases in second position of the putative RNA-binding tripeptide (position 155) was Gly.

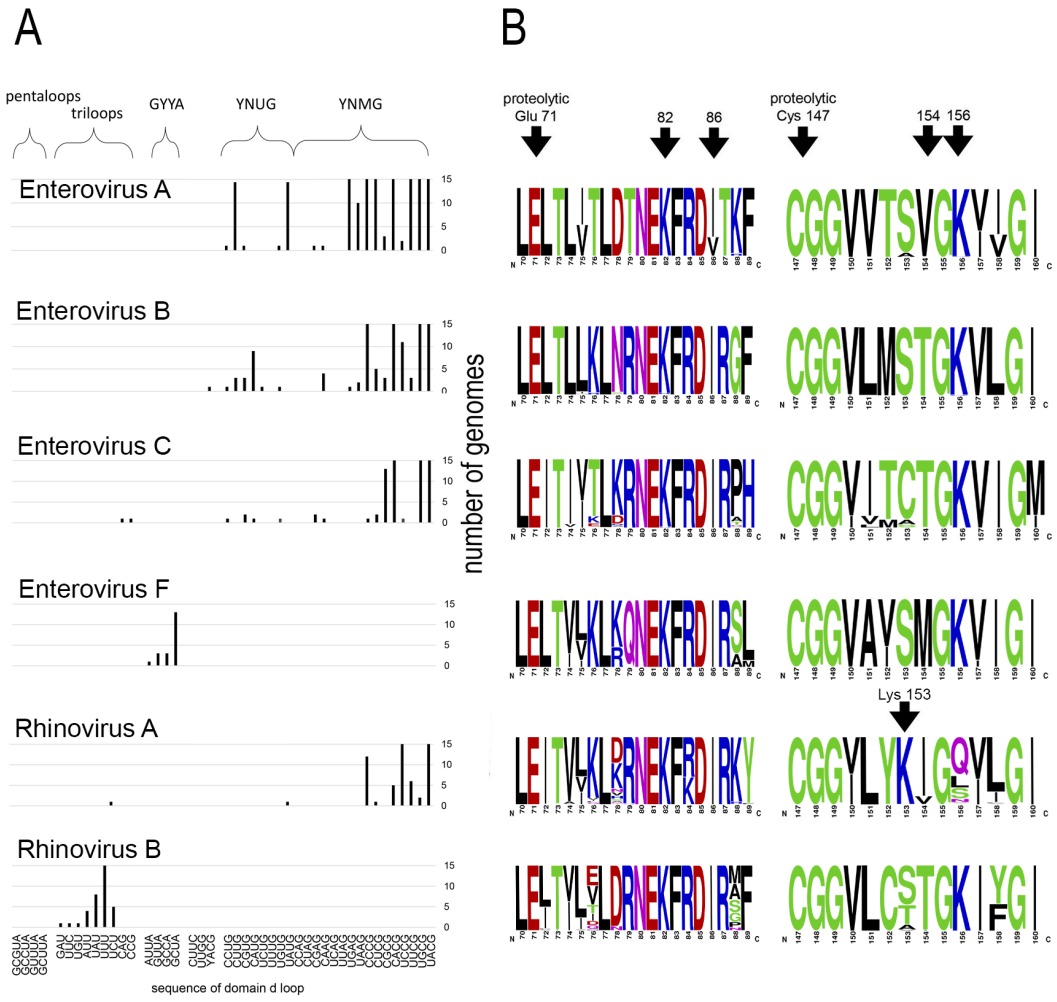

Figure 3 Distribituion of domain d loop sequence and amino acid motifs in the 3C protein. (A) Distribution of domain d loop sequences. The regions corresponding to tetraloop consensuses, triloops and pentaloops are shown. Number of genomes cut off at 15 for clear view of sequence distribution. (B) The frequency plot of amino acid sequence of 3C in species of genus *Enterovirus*. The amino acid sequence logo was done with the WebLogo server (*Crooks et al., 2004*). Arrows indicate amino acids of the proteolytic triade (Glu71 and Cys 147), the first and the last amino acids of motif $_{82}$KFRDI$_{86}$, the putative RNA-binding tripeptide 154–156 of 3C and Lys153 in the protein 3C of *Rhinovirus A*.

No mutual dependence between loop sequences and tripeptide sequences was found within enterovirus genomes of the same species. For example, *Enterovirus A* genomes contained 17 unique variants of the tetraloop sequence, whereas the predominant fraction of 3C sequences (548 out of 564) contained the conservative tripeptide VGK at positions 154–156 (Fig. 3, Table S9). It is noteworthy that, this tripeptide was not found exclusively only in genomes of the EV71 serotype, although genomes of this serotype prevailed in the data set. Other *Enterovirus A* genomes contained tripeptides VGR (seven out of 564), TGK (four out of 564), IGK (three out of 564), VGE (one out of 564) and SRK (one out of 564) (Fig. 3, Table S9). Genomes with tripeptides other than VGK contained no peculiarities of the domain d loop sequence (Table S9). This observation confirms that the specific loop

sequence is not likely to be the main subject for recognition by the RNA-binding tripeptide. Similarly, all or almost all genomes of *Enterovirus B* (242 out of 244), *Enterovirus C* (272 out of 274), *Enterovirus D* (all), *Enterovirus G* (seven out of 8), *Enterovirus H* (a total of two: genomes—one with TGK, one with TGR), *Enterovirus J* (all) and *Rhinovirus B* (36 out of 37) species contained tripeptide TGK at positions 154–156 of the 3C protein (Fig. 3, Tables S7, S9 and S10). Alternative tripeptides were TGR in two genomes of *Enterovirus B* and one genome of *Enterovirus H*; IGK in one genome of *Enterovirus C* species and in one genome of *Rhinovirus B* species; PGK in one genome of *Enterovirus C* species; and MGK in one genome of *Enterovirus G* species (Tables S7, S9 and S10).

Genomes of *Enterovirus E* and *F* species contained two oriLs with tetraloops in domain d mostly of consensus GYYA or pentaloops of consensus GHBUA, where H = A/C/U and B = U/C/G. All genomes contained tripeptide MGK at positions 154–156 of protein 3C (Fig. 3, Table S7). Interestingly, a similar loop-tripeptide pair was found in one genome of *Enterovirus G* species (strain LP 54, accession number AF363455). It contained tetraloop GUUA in domain d of its single oriL and tripeptide MGK in 3C. Unlike this unique genome, other genomes of *Enterovirus G* species contained tetraloops of YNMG consensus and tripeptide TGK in the protein 3C.

Rhinovirus genomes contained tetraloops, mostly of consensus YNMG (*Rhinovirus A* and *C* species) or triloops (*Rhinovirus A* and *B*) (Table 2). *Rhinovirus A* genomes with tetraloops in domain d contained tripeptides in 3C with positively charged amino acid before the tripeptide, but not in its final position, as in case of genomes of *Enterovirus A–C* species (Fig. 3, Tables S11 and S12). The sequence of tripeptides, which did not depend on the tetraloop sequence, was, in descending order, IGQ (the most abundant, 65 genomes out of 119), IGL (20 genomes out of 119), IGS, VGS, IGN, VGQ, IGV and VGH (Table S11). In the case of the *Rhinovirus A* genome with triloop UCU in domain d (isolate V38_URT-6.3m, accession number JF285329), protein 3C contained tripeptide TGK without positively-charged amino acid before it (Table S11). All genomes of *Rhinovirus B* species contained triloops in domain d, with all but one (with IGK) containing tripeptide TGK in 3C. Genomes of Rhinovirus C contained tetraloops mostly of consensus YHCG (H = all but G) and tripeptides in 3C without a positively charged amino acid at the last position (TGN, VGN, TGH) or outside of the tripeptide (Table S12). One genome contained tetraloop CUUC paired with most abundant tripeptide, that is, TGN (23 out of 37 genomes) (Table S12).

Thus, dependence between apical domain d sequences and tripeptides in protein 3C within a species was not detected (Fig. 3). We can state that the tripeptide and motif KFRDI are almost non-variable within a species compared to the domain d loop sequence, but there is a specifically preferred tripeptide sequence for each species. Hence, tripeptide sequences are species-specific, while the domain d loop sequences are almost universal among *Enterovirus A, B, C* and *D* and *Rhinovirus A* and *C* species.

DISCUSSION

Most of the domain d apical loops in enterovirus genomes were represented by tetraloops. The most common variants of tetraloop sequences corresponded to consensuses YNHG

(Y = C/U, N = any, H = A/C/U) (Table 2). Similar results were obtained in our previous experimental work, where eight apical nucleotides of domain d of the poliovirus genome were randomized and viable variants were selected *in vitro,* with the majority of selected tetraloops belonging to consensus YNHG (*Prostova et al., 2015*). Some tetraloops of consensus YNHG were found in genomes in the NCBI database, but not among the variants selected *in vitro*, namely tetraloops CACG, CUCG, UAAG, UGAG, CAAG, CGAG, UGUG and CUUG (*Prostova et al., 2015*). Tetraloops UGAG, UGUG and CUUG were reconstructed with a U****G flanking base pair in the context of the poliovirus genome strain Mahoney, which supported effective virus replication (*Prostova et al., 2015*).

Conversely, tetraloops UUAG, UCAG and CCAG, found in domain d of selected *in vitro* viable poliovirus variants, were able to support virus reproduction; however, they were not found in naturally circulating viruses (*Prostova et al., 2015*). One tetraloop of the YNHG consensus (CUAG) was neither found in genomes from the NCBI database ($n = 2,842$), nor in the randomized poliovirus genomes selected *in vitro* ($n = 62$) (Table 2). Thermodynamic stability is unlikely to be the reason why this and other tetraloops were unrepresented as the melting temperature of stem loops with avoided tetraloops is within range of the melting temperature of YNHG tetraloops, which supported replication (*Proctor et al., 2002*). Moreover, tetraloops UUAG and UCAG are common in rRNA (*Woese et al., 1990*). Sample insufficiency cannot be excluded for both database and *in vitro* selected sets of genomes, but it is safe to conclude that these tetraloop variants are at least extremely rare. In any case, the fact that the incidence of these tetraloops is much less than for other tetraloops indicates that such variants are possibly less fit.

The most abundant tetraloops in the domain apical region of genomes from the NCBI database and variants selected *in vitro* could be compiled into consensuses UNCG and CNCG (Table 2, Table S1). At the same time, these tetraloops are most abundant in rRNA, and, with certain closing base pairs, among the most thermodynamically stable tetraloops (*Woese et al., 1990*; *Proctor et al., 2002*). Tetraloops of these consensuses and some other found tetraloops of the YNHG consensus form a specific spatial structure of the UNCG structural class of stable tetraloops (*Cheong, Varani & Tinoco, 1990*; *Varani, Cheong & Tinoco, 1991*; *Du et al., 2003*; *Du et al., 2004*).

Another set of tetraloops, which correspond to GNYA consensus, was found both in genomes of *Enterovirus E* and *F* and in genomes of viable polioviruses selected *in vitro* (*Prostova et al., 2015*). Tetraloop GCUA was able to support the effective replication of poliovirus and, together with tetraloop GUUA, is known to assume an UNCG fold (*Ihle et al., 2005*; *Melchers et al., 2006*; *Prostova et al., 2015*). In sum, these data suggest that the spatial structure, rather than the exact sequence, is the main subject for recognition by virus protein 3C. Structure-based recognition of tetraloops occurs in several known RNA-protein complexes. For example, tetraloops with a GNRA class structure in the context of bacteriophages P22 and λ genome transcription antitermination element boxB are specifically recognized by the bacteriophage N-protein arginine-rich motif (ARM) (*Cai et al., 1998*; *Legault et al., 1998*; *Schärpf et al., 2000*). Arginines and lysines of the ARM recognize the shape of the negatively charged phosphodiester backbone of the stem-loop and positions N-peptide for hydrophobic or stacking interaction with a non-conserved

nucleotide of the loop (*Cai et al., 1998*; *Legault et al., 1998*; *Schärpf et al., 2000*; *Thapar, Denmon & Nikonowicz, 2013*). Another example of structure-specific recognition is the complex of the double-stranded RNA-binding domain (dsRBD) of RNase Rnt1p and AGNN class tetraloop (*Chanfreau, Buckle & Jacquier, 2000*; *Lebars et al., 2001*; *Wu et al., 2001*; *Wu et al., 2004*; *Wang et al., 2011*; *Thapar, Denmon & Nikonowicz, 2013*). Motif dsRBD recognizes the phosphodiester backbone at the 3′ side of the tetraloop and its non-conserved third and fourth nucleotides (*Wu et al., 2004*; *Wang et al., 2011*; *Thapar, Denmon & Nikonowicz, 2013*).

The sequence to structure degeneracy (different RNA sequences are able to form similar spatial structure) is the known phenomenon (*Petrov, Zirbel & Leontis, 2013*; *Bottaro & Lindorff-Larsen, 2017*). Moreover, it is suggested to refrain from associating sequences with a particular fold (*D'Ascenzo et al., 2016*; *D'Ascenzo et al., 2017*). Together with the literature data, our result let us assume that sequence-structure degeneracy is a universal way in which RNA tetraloops are used in nature (*Lebars et al., 2001*; *Wu et al., 2004*; *Ihle et al., 2005*; *Petrov, Zirbel & Leontis, 2013*; *D'Ascenzo et al., 2016*; *D'Ascenzo et al., 2017*; *Bottaro & Lindorff-Larsen, 2017*).

It can be speculated that pentaloops in domain d of the *Enterovirus E* genome and triloops of domain d of rhinoviruses have the potential to comprise the same UNCG fold as some YNHG and GNYA tetraloops. For HRV14 domain d, it was shown that its triloop resembles the structure of the first and last two nucleotides of UNCG structural class tetraloops (*Headey et al., 2007*). There are pentaloops with four nucleotides that belong to consensus UNCG, GNRA or gCUUGc, which are able to form spatial structures of corresponding structural classes with the fifth bulged nucleotide (*Cai et al., 1998*; *Schärpf et al., 2000*; *Theimer, Finger & Feigon, 2003*; *Oberstrass et al., 2006*; *Liu et al., 2009*). It is possible that four nucleotides of the pentaloops in domain d of Enterovirus E species have a UNCG fold with one bulged nucleotide.

Tetraloops that did not belong to the YNHG or GNYA consensus were found in both sets of natural and *in vitro* selected genomes. However, in an experiment such variants were found to evolve towards the YNHG or GNYA consensus (*Prostova et al., 2015*). Apparently, tetraloops that do not belong to the YNHG or GNYA consensus are less fit in most settings and under experimental conditions. However, as these variants may still be found in a few naturally circulating viruses (consequently, they have emerged and been fixed), we speculate that they may be beneficial under specific replication conditions.

A similar structure of domain d and its apical region suggests the free exchange of this region between genomes of the same and different species of *Enterovirus* genera. Indeed, viable intra and inter species recombinants for this region could be obtained *in vitro* (*Muslin et al., 2015*; *Bessaud et al., 2016*). To evaluate the relative impact of the high mutation rate and recombination on domain d apical loop variability, sequences of EV71 C4 genotype viruses were analysed. The natural recombination in EV71 genotype C4 is much less frequent than other *Enterovirus A* types (*Lukashev et al., 2014*); meanwhile, only one recombinant genome (accession number HQ423143) was detected in our data set. Therefore, the variability of its domain d loop sequence reflects changes that were only accumulated via mutations. The diversity of the domain d loop sequence of EV-71 C4

viruses was far less prominent than among *Enterovirus A* genomes and represented only by five tetraloop sequence variants (Table 2). As the most recent common ancestor of EV71 genotype C4 dates back about 20 years (*McWilliam Leitch et al., 2012*), this diversity, although limited, has only emerged very recently. On the other hand, the high sequence variability of the domain d apical region in all enterovirus genomes was possibly assisted by inter- and intra-species recombination events.

Interestingly, in contrast to the similar structure of domain d and the very similar distribution of its apical sequences in genomes of different enterovirus species, its putative RNA-recognition tripeptide of 3C is diverse (Fig. 3). Most *Enterovirus A* genomes contain tripeptide VGK in 3C, while there is a prevalence of the TGK tripeptide among genomes of *Enterovirus B, C* and *D* species (Fig. 3). Genomes of *Rhinovirus A* and *C* also contain common enterovirus tetraloops in the domain d apical region, but, in 3C, unlike other species, they contain tripeptides without positively charged amino acids (Fig. 3, Tables S11 and S12). Positively charged amino acids are often involved in the interaction with RNA, in particular, with phosphates of the RNA backbone. As such, they are of importance to RNA-protein recognition (*Jones et al., 2001*; *Bahadur, Zacharias & Janin, 2008*). In *Rhinovirus A* genomes, positively charged amino acid "jumped" from the last position of the tripeptide (position 156) to the position that precedes the tripeptide (position 153) (Fig. 3, shown by an arrow). The residue at position 153 starts and the residue at position 156 ends the reverse turn between beta strands dII and eII of protein 3C (*Mosimann et al., 1997*; *Matthews et al., 1999*; *Cui et al., 2011*). In a crystal structure of the Rhinovirus A2 protein 3C, the side chain of Lys153 (preceding the tripeptide) is positioned in a region similar to that of the side chain of Lys156 (in last position of the tripeptide) in the crystal structure of Enterovirus 71 and Poliovirus 1 proteins 3C (*Mosimann et al., 1997*; *Matthews et al., 1999*; *Cui et al., 2011*). Thus, Lys at position 153 of 3C has almost the same potential to interact with the RNA-ligand as Lys at position 156 (*Mosimann et al., 1997*; *Matthews et al., 1999*; *Cui et al., 2011*). Genomes of *Rhinovirus C* species do not contain a positively charged amino acid, either inside the tripeptide of the 3C protein, or in the neighbouring positions, possibly indicating that tripeptide 154–156 in the protein 3C of *Rhinovirus C* genome does not interact directly with RNA. Thus, 3C is able to recognize domain d of the oriL with tripeptides of a different sequence. In contrast to the domain d structure and its apical sequence, the tripeptide is species-specific. The diversity of the tripeptide, which is expected to recognize domain d, has several compatible explanations. Residue 154 of the tripeptide possibly does not interact with domain d directly. The tripeptide may be involved into a species-specific cooperative amino acid network (amino acid "epistasis"). Moreover, different tripeptides could reflect slightly different molecular mechanisms for domain d recognition.

The complexity of the tripeptide's role in domain d recognition can be shown in several examples. The 3C protein of different species with the same RNA-binding tripeptide is not guaranteed to bind the same structured domain d. Genomes of the *Rhinovirus B* contain triloops in the apical region of domain d, which are paired with tripeptide TGK in 3C, common for genomes with tetraloops. In contrast, protein 3C of the Coxsackie virus B3 (*Enterovirus B* species, containing tripeptide TGK) cannot recognize oriL sufficiently

well when domain d is capped with a triloop (*Zell et al., 2002*). This indicates that the sequence of the RNA-binding tripeptide is probably not the exclusive participant in oriL-3C recognition. In other words, different molecular mechanisms of oriL-3C recognition have evolved in every enterovirus species independently. For example, it was shown for Rhinovirus 14 (*Rhinovirus B* species) that protein 3C recognizes the stem region of domain d, rather than its apical loop (*Leong et al., 1993*). Another oriL-3C recognition mechanism is seemingly employed by *Enterovirus E* and *F* species, two oriLs of which play the same role as the single oriL in genomes of other enteroviruses (*Pilipenko, Blinov & Agol, 1990*; *Zell et al., 1999*). The apical loop of their domain d is a tetra- or pentaloop with a sequence that differs from the loop consensuses of other enteroviruses. The RNA-binding tripeptide in 3C is species-specific as well, and is always MGK (Table S6). Interestingly, one genome of *Enterovirus G* species had the same pair domain d loop: tripeptide of 3C, i.e., GUUA MGK. Domain d of *Enterovirus G* species is prolonged in comparison to the length of domain d in genomes of other species (*Krumbholz et al., 2002*) (Fig. 2). Tripeptide MGK in the 3C of *Enterovirus E, F* and *G* possibly indicates another molecular mechanism of oriL-3C recognition (*Krumbholz et al., 2002*). Therefore, we assume that, though putative RNA-binding, the tripeptide, in most cases, possibly interacts with the domain d apical region (since amino acid substitutions in it are known to compensate for structural disturbance in domain d); however, this interaction is not the only one that determines the evolution of oriL-3C interaction. Altogether, the data suggest that the independent evolution of the putative RNA-binding tripeptide of 3C and domain d of oriL occurs.

CONCLUSIONS

We analysed the variety and occurrence of the replication element oriL's functional loop and its protein ligand virus protease 3C. RNA-binding motifs of 3C are species-specific, in contrast to domain d loop sequences: the sequence variety of domain d loop is almost the same for *Enterovirus A, B, C* and *D* and *Rhinovirus A* and *C* species, whereas tripeptide sequence variety differs. The conservation of the tripeptide sequence within species, together with the almost universal diversity of tetraloop sequences among species, indicates the occurrence of the independent evolution of these two elements. Our results suggest the structure-based, rather than sequence-based, recognition of domain d by virus protein 3CD. These, together with the data reported in the literature, let us assume that the sequence-structure degeneracy is a universal way in which RNA tetraloops are used in nature.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the Russian Science Foundation (No. 15-15-00147). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Russian Science Foundation: 15-15-00147.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Maria A. Prostova conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Andrei A. Deviatkin conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Irina O. Tcelykh performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Alexander N. Lukashev conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Anatoly P. Gmyl conceived and designed the experiments, wrote the paper, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

Multiple alignment files (.fas), containing Enterovirus full genomes before and after filtration, are provided as Supplementary Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3896#supplemental-information.

REFERENCES

Acevedo A, Brodsky L, Andino R. 2013. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505(7485)**:686–690 DOI 10.1038/nature12861.

Andino R, Rieckhof GE, Achacoso PL, Baltimore D. 1993. Poliovirus RNA synthesis utilizes an RNP complex formed around the 5′-end of viral RNA. *EMBO Journal* **12(9)**:3587–3598.

Andino R, Rieckhof GE, Baltimore D. 1990. A functional ribonucleoprotein around the 5′ end of poliovirus. *Cell* **63**:369–380 DOI 10.1016/0092-8674(90)90170-J.

Andino R, Rieckhof GE, Trono D, Baltimore D. 1990. Substitutions in the protease (3Cpro) gene of poliovirus can suppress a mutation in the 5′ noncoding region. *Journal of Virology* **64(2)**:607–612.

Bahadur RP, Zacharias M, Janin J. 2008. Dissecting protein-RNA recognition sites. *Nucleic Acids Research* **36(8)**:2705–2716 DOI 10.1093/nar/gkn102.

Bessaud M, Joffret M-L, Blondel B, Delpeyroux F. 2016. Exchanges of genomic domains between poliovirus and other cocirculating species C enteroviruses reveal a high degree of plasticity. *Scientific Reports* **6(1)**:38831 DOI 10.1038/srep38831.

Bottaro S, Lindorff-Larsen K. 2017. Mapping the universe of RNA tetraloop folds. *Biophysical Journal* **113(2)**:257–267 DOI 10.1016/j.bpj.2017.06.011.

Cai Z, Gorin A, Frederick R, Ye X, Hu W, Majumdar A, Kettani A, Patel DJ. 1998. Solution structure of P22 transcriptional antitermination N peptide-box B RNA complex. *Nature Structural Biology* **5(3)**:203–212 DOI 10.1038/nsb0398-203.

Chanfreau G, Buckle M, Jacquier A. 2000. Recognition of a conserved class of RNA tetraloops by Saccharomyces cerevisiae RNase III. *Proceedings of the National Academy of Sciences of the United States of America* **97(7)**:3142–3147 DOI 10.1073/pnas.070043997.

Chase AJ, Daijogo S, Semler BL. 2014. Inhibition of poliovirus-induced cleavage of cellular protein PCBP2 reduces the levels of viral RNA replication. *Journal of Virology* **88(6)**:3192–3201 DOI 10.1128/JVI.02503-13.

Cheong C, Cheong H. 2010. RNA structure: tetraloops. In: *Encyclopedia of life sciences*. Chichester: John Wiley & Sons, Ltd.

Cheong C, Varani G, Tinoco IJ. 1990. Solution structure of an unusually stable RNA hairpin, 5′GGAC(UUCG)GUCC. *Nature* **346(6285)**:680–682 DOI 10.1038/346680a0.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Research* **14(6)**:1188–1190 DOI 10.1101/gr.849004.

Cui S, Wang J, Fan T, Qin B, Guo L, Lei X, Wang J, Wang M, Jin Q. 2011. Crystal structure of human enterovirus 71 3C protease. *Journal of Molecular Biology* **408(3)**:449–461 DOI 10.1016/j.jmb.2011.03.007.

D'Ascenzo L, Leonarski F, Vicens Q, Auffinger P. 2016. "Z-DNA like" fragments in RNA: a recurring structural motif with implications for folding, RNA/protein recognition and immune response. *Nucleic Acids Research* **44(12)**:5944–5956 DOI 10.1093/nar/gkw388.

D'Ascenzo L, Leonarski F, Vicens Q, Auffinger P. 2017. Revisiting GNRA and UNCG folds: U-turns versus Z-turns in RNA hairpin loops. *RNA* **23(3)**:259–269 DOI 10.1261/rna.059097.116.

Du Z, Yu J, Andino R, James TL. 2003. Extending the family of UNCG-like tetraloop motifs: NMR structure of a CACG tetraloop from coxsackievirus B3. *Biochemistry* **42(15)**:4373–4383 DOI 10.1021/bi027314e.

Du Z, Yu J, Ulyanov NB, Andino R, James TL. 2004. Solution structure of a consensus stem-loop D RNA domain that plays important roles in regulating translation and replication in enteroviruses and rhinoviruses. *Biochemistry* **43(38)**:11959–11972 DOI 10.1021/bi048973p.

Gamarnik AV, Andino R. 1998. Switch from translation to RNA replication in a positive-stranded RNA virus. *Genes & Development* **12**:2293–2304 DOI 10.1101/gad.12.15.2293.

Goodfellow I, Chaudhry Y, Richardson A, Meredith J, Almond JW, Barclay W, Evans DJ. 2000. Identification of a *cis*-acting replication element within the poliovirus coding region. *Journal of Virology* **74(10)**:4590–4600 DOI 10.1128/JVI.74.10.4590-4600.2000.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Research* **36(Web Server issue)**:W70–W74 DOI 10.1093/nar/gkn188.

Hämmerle T, Hellen CU, Wimmer E. 1991. Site-directed mutagenesis of the putative catalytic triad of poliovirus 3C proteinase. *The Journal of Biological Chemistry* **266(9)**:5412–5416.

Hämmerle T, Molla A, Wimmer E. 1992. Mutational analysis of the proposed FG loop of poliovirus proteinase 3C identifies amino acids that are necessary for 3CD cleavage and might be determinants of a function distinct from proteolytic activity. *Journal of Virology* **66(10)**:6028–6034.

Harris KS, Reddigari SR, Nicklin MJ, Hämmerle T, Wimmer E. 1992. Purification and characterization of poliovirus polypeptide 3CD, a proteinase and a precursor for RNA polymerase. *Journal of Virology* **66(12)**:7481–7489.

Headey SJ, Huang H, Claridge JK, Soares GA, Dutta K, Schwalbe M, Yang D, Pascal SM. 2007. NMR structure of stem-loop D from human rhinovirus-14. *RNA* **13(3)**:351–360 DOI 10.1261/rna.313707.

Ihle Y, Ohlenschläger O, Häfner S, Duchardt E, Zacharias M, Seitz S, Zell R, Ramachandran R, Görlach M. 2005. A novel cGUUAg tetraloop structure with a conserved yYNMGg-type backbone conformation from cloverleaf 1 of bovine enterovirus 1 RNA. *Nucleic Acids Research* **33(6)**:2003–2011 DOI 10.1093/nar/gki501.

Jones S, Daley DTA, Luscombe NM, Berman HM, Thornton JM. 2001. Protein—RNA interactions: a structural analysis. *Biochemistry* **29(4)**:943–954.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4)**:772–780 DOI 10.1093/molbev/mst010.

Krumbholz A, Dauber M, Henke A, Birch-Hirschfeld E, Knowles NJ, Stelzner A, Zell R. 2002. Sequencing of porcine enterovirus groups II and III reveals unique features of both virus groups. *Journal of Virology* **76(11)**:5813–5821 DOI 10.1128/JVI.76.11.5813-5821.2002.

Lauring AS, Frydman J, Andino R. 2013. The role of mutational robustness in RNA virus evolution. *Nature Reviews. Microbiology* **11(5)**:327–336 DOI 10.1038/nrmicro3003.

Lebars I, Lamontagne B, Yoshizawa S, Aboul-Elela S, Fourmy D. 2001. Solution structure of conserved AGNN tetraloops: insights into Rnt1p RNA processing. *The EMBO Journal* **20(24)**:7250–7258 DOI 10.1093/emboj/20.24.7250.

Legault P, Li J, Mogridge J, Kay LE, Greenblatt J. 1998. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell* **93(2)**:289–299 DOI 10.1016/S0092-8674(00)81579-2.

Leong LEC, Walker PA, Porter AG, Protease HR, Leon LEC, Walker PA, Porter AG, Leong LEC, Walker PA, Porter AG. 1993. Human rhinovirus-14 protease 3C

(3Cpro) binds specifically to the 5′-noncoding region of the viral RNA. *The Journal of Biological Chemistry* **268(34)**:25735–25739.

Liu P, Li L, Keane SC, Yang D, Leibowitz JL, Giedroc DP. 2009. Mouse hepatitis virus stem-loop 2 adopts a uYNMG(U)a-like tetraloop structure that is highly functionally tolerant of base substitutions. *Journal of Virology* **83(23)**:12084–12093 DOI 10.1128/JVI.00915-09.

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**:26 DOI 10.1186/1748-7188-6-26.

Lukashev AN, Shumilina EY, Belalov IS, Ivanova OE, Eremeeva TP, Reznik VI, Trotsenko OE, Drexler JF, Drosten C. 2014. Recombination strategies and evolutionary dynamics of the Human enterovirus A global gene pool. *The Journal of General Virology* **95(Pt 4)**:868–873 DOI 10.1099/vir.0.060004-0.

Matthews DA, Dragovich PS, Webber SE, Fuhrman SA, Patick AK, Zalman LS, Hendrickson TF, Love RA, Prins TJ, Marakovits JT, Zhou R, Tikhe J, Ford CE, Meador JW, Ferre RA, Brown EL, Binford SL, Brothers MA, DeLisle DM, Worland ST. 1999. Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. *Proceedings of the National Academy of Sciences of the United States of America* **96(20)**:11000–11007 DOI 10.1073/pnas.96.20.11000.

McWilliam Leitch EC, Cabrerizo M, Cardosa J, Harvala H, Ivanova OE, Koike S, Kroes ACM, Lukashev A, Perera D, Roivainen M, Susi P, Trallero G, Evans DJ, Simmonds P. 2012. The association of recombination events in the founding and emergence of subgenogroup evolutionary lineages of human enterovirus 71. *Journal of Virology* **86(5)**:2676–2685 DOI 10.1128/JVI.06065-11.

Melchers WJG, Zoll J, Tessari M, Bakhmutov DV, Gmyl AP, Agol VI, Heus HA. 2006. A GCUA tetranucleotide loop found in the poliovirus oriL by *in vivo* SELEX (un)expectedly forms a YNMG-like structure: extending the YNMG family with GYYA. *RNA* **12(9)**:1671–1682 DOI 10.1261/rna.113106.

Mosimann SC, Cherney MM, Sia S, Plotch S, James MN. 1997. Refined X-ray crystallographic structure of the poliovirus 3C gene product. *Journal of Molecular Biology* **273(5)**:1032–1047 DOI 10.1006/jmbi.1997.1306.

Muslin C, Joffret M-L, Pelletier I, Blondel B, Delpeyroux F. 2015. Evolution and emergence of enteroviruses through intra- and inter-species recombination: plasticity and phenotypic impact of modular genetic exchanges in the 5′untranslated region. *PLOS Pathogens* **11(11)**:e1005266 DOI 10.1371/journal.ppat.1005266.

Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FH-T. 2006. Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nature Structural & Molecular Biology* **13(2)**:160–167 DOI 10.1038/nsmb1038.

Oermann CM, Schuster JE, Conners GP, Newland JG, Selvarangan R, Jackson MA. 2015. Enterovirus D68. A focused review and clinical highlights from the 2014 U.S. Outbreak. *Annals of the American Thoracic Society* **12(5)**:775–781 DOI 10.1513/AnnalsATS.201412-592FR.

Palmenberg A, Neubauer D, Skern T. 2010. Genome organization and encoded proteins. In: Ehrenfeld E, Domingo E, Roos RP, eds. *The picornaviruses.* Washington, D.C.: ASM Press, 3–17.

Petrov AI, Zirbel CL, Leontis NB. 2013. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA* **19(10)**:1327–1340 DOI 10.1261/rna.039438.113.

Pilipenko EV, Blinov VM, Agol VI. 1990. Gross rearrangements within the 5′-untranslated region of the picornaviral genomes. *Nucleic Acids Research* **18(11)**:3371–3375 DOI 10.1093/nar/18.11.3371.

Proctor DJ, Schaak JE, Bevilacqua JM, Falzone CJ, Bevilacqua PC. 2002. Isolation and characterization of a family of stable RNA tetraloops with the motif YNMG that participate in tertiary interactions. *Biochemistry* **41(40)**:12062–12075 DOI 10.1021/bi026201s.

Prostova MA, Gmyl AP, Bakhmutov DV, Shishova AA, Khitrina EV, Kolesnikova MS, Serebryakova MV, Isaeva OV, Agol VI. 2015. Mutational robustness and resilience of a replicative *cis*-element of RNA virus: promiscuity, limitations, relevance. *RNA Biology* **12(12)**:1338–1354 DOI 10.1080/15476286.2015.1100794.

Rieder E, Xiang W, Paul A, Wimmer E. 2003. Analysis of the cloverleaf element in a human rhinovirus type 14/poliovirus chimera: correlation of subdomain D structure, ternary protein complex formation and virus replication. *Journal of General Virology* **84(8)**:2203–2216 DOI 10.1099/vir.0.19013-0.

Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *Journal of Virology* **84(19)**:9733–9748 DOI 10.1128/JVI.00694-10.

Schärpf M, Sticht H, Schweimer K, Boehm M, Hoffmann S, Rösch P. 2000. Antitermination in bacteriophage lambda. The structure of the N36 peptide-boxB RNA complex. *European Journal of Biochemistry* **267(267)**:2397–2408 DOI 10.1046/j.1432-1327.2000.01251.x.

Shih S, Chen T, Wu C. 2004. Mutations at KFRDI and VGK domains of enterovirus 71 3C protease affect its RNA binding and proteolytic activities. *Journal of Biomedical Science* **11(2)**:239–248 DOI 10.1159/000076036.

Solomon T, Lewthwaite P, Perera D, Cardosa MJ, McMinn P, Ooi MH. 2010. Virology, epidemiology, pathogenesis, and control of enterovirus 71. *The Lancet. Infectious Diseases* **10(11)**:778–790 DOI 10.1016/S1473-3099(10)70194-8.

Thapar R, Denmon AP, Nikonowicz EP. 2013. Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins. *Wiley Interdisciplinary Reviews. RNA* **5(1)**:49–67 DOI 10.1002/wrna.1196.

Theimer CA, Finger LD, Feigon J. 2003. YNMG tetraloop formation by a dyskeratosis congenita mutation in human telomerase RNA. *RNA* **9**:1446–1455 DOI 10.1261/rna.5152303.activity.

Thompson AA, Peersen OB. 2004. Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. *The EMBO Journal* **23(17)**:3462–3471 DOI 10.1038/sj.emboj.7600357.

Trono D, Andino R, Baltimore D. 1988. An RNA sequence of hundreds of nucleotides at the 5′ end of poliovirus RNA is involved in allowing viral protein synthesis. *Journal of Virology* **62(7)**:2291–2299.

Uhlenbeck OC. 1990. Tetraloops and RNA folding. *Nature* **346(6285)**:613–614 DOI 10.1038/346613a0.

Varani G, Cheong C, Tinoco I. 1991. Structure of an unusually stable RNA hairpin. *Biochemistry* **30(13)**:3280–3289 DOI 10.1021/bi00227a016.

Vogt DA, Andino R. 2010. An RNA element at the 5′-end of the poliovirus genome functions as a general promoter for RNA synthesis. *PLOS Pathogens* **6(6)**:e1000936 DOI 10.1371/journal.ppat.1000936.

Wagner A, Stadler PF. 1999. Viral RNA and evolved mutational robustness. *The Journal of Experimental Zoology* **285(2)**:119–127 DOI 10.1002/(SICI)1097-010X(19990815)285:2<119::AID-JEZ4>3.0.CO;2-D.

Wang Z, Hartman E, Roy K, Chanfreau G, Feigon J. 2011. Structure of a yeast RNase III dsRBD complex with a noncanonical RNA substrate provides new insights into binding specificity of dsRBDs. *Structure* **19(7)**:999–1010 DOI 10.1016/j.str.2011.03.022.

Woese CR, Winker S, Gutell RR, Winkers S, Gutell RR. 1990. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proceedings of the National Academy of Sciences of the United States of America* **87(November)**:8467–8471.

Wu H, Henras A, Chanfreau G, Feigon J. 2004. Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Amino Acids* **101(22)**:8307–8312.

Wu H, Yang PK, Butcher SE, Kang S, Chanfreau G, Feigon J. 2001. A novel family of RNA tetraloop structure forms the recognition site for Saccharomyces cerevisiae RNase III. *The EMBO Journal* **20(24)**:7240–7249 DOI 10.1093/emboj/20.24.7240.

Zell R, Sidigi K, Bucci E, Stelzner A, Görlach M. 2002. Determinants of the recognition of enteroviral cloverleaf RNA by coxsackievirus B3 proteinase 3C. *RNA* **8(2)**:188–201 DOI 10.1017/S1355838202012785.

Zell R, Sidigi K, Henke A, Schmidt-Brauns J, Hoey E, Martin S, Stelzner A. 1999. Functional features of the bovine enterovirus 5′-non-translated region. *Journal of General Virology* **80**:2299–2309 DOI 10.1099/0022-1317-80-9-2299.