

A peer-reviewed version of this preprint was published in PeerJ on 6 October 2017.

[View the peer-reviewed version](https://doi.org/10.7717/peerj.3896) (peerj.com/articles/3896), which is the preferred citable publication unless you specifically need to cite this preprint.

Prostova MA, Deviatkin AA, Tcelykh IO, Lukashev AN, Gmyl AP. 2017. Independent evolution of tetraloop in enterovirus oriL replicative element and its putative binding partners in virus protein 3C. PeerJ 5:e3896 <https://doi.org/10.7717/peerj.3896>

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

Maria A Prostova ^{Corresp., 1}, Andrei A Deviatkin ¹, Irina O Tcelykh ^{1,2}, Alexander N Lukashev ^{1,3}, 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

Corresponding Author: Maria A Prostova
Email address: prostova_ma@chumakovs.ru

Background. Enteroviruses are small non-enveloped viruses with (+) 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 low fidelity, 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 (>6400 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.

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

4 Maria A. Prostova¹, Andrei A. Deviatkin¹, Irina O. Tcelykh^{1,2}, Alexander N. Lukashev^{1,3},
5 Anatoly P. Gmyl^{1,2,3}

6 1 - Chumakov Institute of Poliomyelitis and Viral Encephalitides, Moscow, Russian
7 Federation

8 2 - Lomonosov Moscow State University, Moscow, Russian Federation

9 3 - Sechenov First Moscow State Medical University, Moscow, Russian Federation

10

11 Corresponding author:

12 Maria A. Prostova¹

13 Email address:

14 prostovna@gmail.com

15

16

Abstract

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

Background. Enteroviruses are small non-enveloped viruses with (+) 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 low

fidelity, 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 (>6400 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

35 apical loops were tetraloops, which belonged to consensus YNHG (Y=U/C, N=any nucleotide,
36 H=A/C/U). The putative RNA-binding tripeptide 154-156 (*Enterovirus C* 3C protein
37 numeration) was less diverse than the apical domain d loop region and, in contrast to it, was
38 species-specific.

39

40 **Discussion.** Despite the suggestion that the RNA-binding tripeptide interacts with the apical
41 region of domain d, they evolve independently in nature. Together, our data indicate the plastic
42 evolution of both interplayers of 3C-oriL recognition.

43

44

Introduction

45 Enteroviruses are small non-enveloped viruses with a plus strand genome about 7500 nt
46 long, which contains one open reading frame that encodes structural (capsid) and non-structural
47 proteins, 5' and 3' NTRs (non translated regions), and polyA on the 3' end (Palmenberg,
48 Neubauer and Skern, 2010). (Figure 1). Most non-structural enterovirus proteins are
49 polyfunctional. Protease 3CD is a precursor of polymerase 3D and plays a key role in the
50 initiation of replication (Harris *et al.*, 1992; Gamarnik and Andino, 1998; Thompson and
51 Peersen, 2004). After translation by host cell ribosomal machinery, the genome is utilized for the
52 synthesis of the (-) strand RNA, which, in turn, serves as a matrix for the synthesis of multiple

53 daughter (+) strands. Non-translated regions of the genome and a coding sequence within the
54 genomic region encoding the viral helicase 2C contain replicative elements, which interact with
55 viral and host proteins. These RNA-protein complexes regulate initiation and further steps of
56 replication. For poliovirus, the most clinically relevant member of the *Enterovirus* genus, there
57 are at least three known RNA-protein complexes, which are formed with the replicative elements
58 oriL, oriR and oriI during replication (Figure 1).

59 Complex oriL with viral protein 3CD and the host protein PCBP2 is crucial for the
60 transcription initiation (Goodfellow *et al.*, 2000; Vogt and Andino, 2010; Chase, Daijogo and
61 Semler, 2014). The element oriL has a cloverleaf-like secondary structure with four domains
62 termed a (stem of the cloverleaf), b, c and d (leaves of the cloverleaf) (Trono, Andino and
63 Baltimore, 1988; Andino, Rieckhof and Baltimore, 1990) (Figure 1). Previously, it was
64 demonstrated *in vitro* that 3CD (or 3C) of poliovirus, coxsackievirus B3 and bovine enterovirus
65 1 interacts with the apical loop and the flanking base pairs of hairpin d in the oriL element
66 (Andino, Rieckhof and Baltimore, 1990; Du *et al.*, 2003; Ihle *et al.*, 2005) (Figure 1).

67 The apical loops of domain d in genomes belonging to several viruses of *Enterovirus*
68 genera were shown by NMR experiments to be tetraloops with a specific spatial structure, which
69 belongs to the UNCG structural class of stable tetraloops (Du *et al.*, 2003, 2004; Ihle *et al.*, 2005;
70 Melchers *et al.*, 2006). There are several known structural classes of tetraloops, three of which,
71 named according to consensus sequences, contain tetraloops of extreme stability: UNCG (where

72 N=any nucleotide), GNRA (where R=A/G) and gCUUGc (Uhlenbeck, 1990; Cheong and
73 Cheong, 2010). Tetraloops of UNCG and GNRA classes are the most widely represented (Woese
74 *et al.*, 1990; Cheong and Cheong, 2010; Bottaro and Lindorff-Larsen, 2017). Previously, it was
75 shown that only tetraloops of the UNCG structural class, but not tetraloops of GNRA or
76 gCUUGc structural classes, can support effective replication of the poliovirus genome (Prostova
77 *et al.*, 2015). Moreover, the exact sequence of the apical region of poliovirus domain d was of
78 less importance for effective 3CD-oriL recognition than its spatial structure (Rieder *et al.*, 2003;
79 Prostova *et al.*, 2015). At the same time, structural disturbance in the apical region of oriL
80 domain d of poliovirus could be compensated by amino acid substitutions in the tripeptide 154-
81 156 of the 3C protein (here and hereafter amino acid numeration corresponds to poliovirus 3C
82 protein) (Andino *et al.*, 1990; Prostova *et al.*, 2015). In addition to triplet 154-156, the conserved
83 motif ₈₂KFRDI₈₆ of the 3C protein also takes part in the oriL recognition (Andino *et al.*, 1990,
84 1993; Hämmerle, Molla and Wimmer, 1992; Shih, Chen and Wu, 2004). To date a
85 comprehensive analysis of the diversity in domain d apical region and amino acid tripeptide
86 sequences in the *Enterovirus* genus has not been conducted.

87 The replication of an enterovirus is a low-fidelity process, generating, on average, one
88 mutation per genome (Sanjuán *et al.*, 2010; Acevedo, Brodsky and Andino, 2013). The high
89 probability of mutation allows the virus to adapt to a constantly changing environment on the
90 one hand, but requires additional mechanisms to maintain genome stability on the other (Wagner

91 and Stadler, 1999; Lauring, Frydman and Andino, 2013). The aim of the present study was to
92 understand co-evolution patterns of two interacting replication machinery elements in
93 enteroviruses, the apical domain d of oriL and the 3C protein.

94

95 **Materials and methods**

96 **Formation and filtration of sets of full genomes**

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

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

106 For analysis of domain d sequence variety, the multiple sequence alignments were used.

107 The relevant region of multiple sequence alignment and the respective names of sequences were

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

116 **Domain d secondary structure**

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

120 **Results**

121 **Sample characteristics**

122 To evaluate the variability of the domain d loop sequence we retrieved all available
123 complete genome (8000>n>6800 nucleotides) enterovirus (EV) sequences that were present in
124 the NCBI database on February 2017. Representatives of *Enterovirus A* (1173 sequences in

125 total), *Enterovirus B* (414), *Enterovirus C* (773), *Enterovirus D* (462), *Enterovirus E* (12),
126 *Enterovirus F* (13), *Enterovirus G* (15), *Enterovirus H* (3), *Enterovirus J* (7), *Rhinovirus A* (202),
127 *Rhinovirus B* (76) and *Rhinovirus C* (51) species were analysed. As expected, genomes of
128 epidemiologically significant viruses were the most represented in the database. For example,
129 66% of *Enterovirus A* species genomes belonged to the EV71 type, the causative agent of hand,
130 foot and mouth disease (Solomon *et al.*, 2010); most *Enterovirus C* species sequences (78%)
131 belonged to poliovirus; and most *Enterovirus D* species sequences (98.7%) represented EV68, an
132 aetiological agent of severe respiratory illness (Oermann *et al.*, 2015). The number of genome
133 sequences of each species that contained the oriL region is shown in Table 1.

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

163 possible variants were never found in the domain d apical region: UUAG, UCAG, CUAG and
164 CCAG (Table 2). Thus, dinucleotides UA and CA are likely to be avoided at positions 2 and 3 of
165 the tetraloop in enterovirus genomes.

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

178 In *Enterovirus B* genomes, 18 unique tetraloops out of 24 possible were found. Twelve
179 of these tetraloops belonged to consensus YNMG and six to consensus YNUG (Table 2, Table
180 S3). The most abundant tetraloops were CACG (98 genomes), UACG (51 genomes) and UGCG
181 (31 genomes), which were also present among the most abundant tetraloops of *Enterovirus A*

182 species.

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

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

199 Species *Enterovirus E* and *F* have two oriLs in the 5'NTR, generally with similar
200 sequences in the apical region of domain d (Pilipenko, Blinov and Agol, 1990; Zell *et al.*, 1999)

201 (Table S6). As such, we united sequences from the first and the second oriL of these viruses in
202 the heat map (Table 2). Domain d loops in 10 genomes of *Enterovirus F* were tetraloops, while,
203 in 10 *Enterovirus E* genomes, there were both tetraloops (first oriL) and pentaloops (first and
204 second oriL) (Table 2, Table S6). There were four diverse tetraloop sequences in oriLs of
205 Enterovirus E and F with no obvious preference between these species. These sequences were
206 GCUA, GUUA, GCCA, AUUA (Table 2, Table S6). Tetraloop AUUA was found once in the
207 first oriL domain d of EV-F (strain PS87/Belfast, accession number DQ092794) (Table 2, Table
208 S6). There were six diverse pentaloop sequences in domain d of Enterovirus E genomes –
209 GCUUA, GUUUA, GCCUA, GCGUA, GAUUA, GUCUA (Table 2, Table S6).

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

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

220 domain d were triloops (Table 2, Table S8).

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

226 **Variety of RNA-recognition tripeptide of 3C**

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

239 Motif $_{82}\text{KFRDI}_{86}$ was conserved in all species, as well as amino acids Glu 71 and Cys
240 147 of the protease catalytic triad (Figure 3). Always in second position of the putative RNA-
241 binding tripeptide (position 155) was Gly.

242 No mutual dependence between loop sequences and tripeptide sequences was found
243 within enterovirus genomes of the same species. For example, *Enterovirus A* genomes contained
244 17 unique variants of the tetraloop sequence, whereas the predominant fraction of 3C sequences
245 (548 out of 564) contained the conservative tripeptide VGK at positions 154-156 (Figure 3,
246 Table S9). It is noteworthy that, this tripeptide was not found exclusively only in genomes of the
247 EV71 serotype, although genomes of this serotype prevailed in the data set. Other *Enterovirus A*
248 genomes contained tripeptides VGR (seven out of 564), TGK (four out of 564), IGK (three out
249 of 564), VGE (one out of 564) and SRK (one out of 564) (Figure 3, Table S9). Genomes with
250 tripeptides other than VGK contained no peculiarities of the domain d loop sequence (Table S9).
251 This observation confirms that the specific loop sequence is not likely to be the main subject for
252 recognition by the RNA-binding tripeptide. Similarly, all or almost all genomes of *Enterovirus B*
253 (242 out of 244), *Enterovirus C* (272 out of 274), *Enterovirus D* (all), *Enterovirus G* (seven out
254 of 8), *Enterovirus H* (a total of two: genomes – one with TGK, one with TGR), *Enterovirus J*
255 (all) and *Rhinovirus B* (36 out of 37) species contained tripeptide TGK at positions 154-156 of
256 the 3C protein (Figure 3, Table S7, S9, S10). Alternative tripeptides were TGR in two genomes
257 of *Enterovirus B* and one genome of *Enterovirus H*; IGK in one genome of *Enterovirus C*

258 species and in one genome of *Rhinovirus B* species; PGK in one genome of *Enterovirus C*
259 species; and MGK in one genome of *Enterovirus G* species (Table S7, S9, S10).

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

268 Rhinovirus genomes contained tetraloops, mostly of consensus YNMG (*Rhinovirus A*
269 and *C* species) or triloops (*Rhinovirus A* and *B*) (Table 2). *Rhinovirus A* genomes with tetraloops
270 in domain d contained tripeptides in 3C with positively charged amino acid before the tripeptide,
271 but not in its final position, as in case of genomes of *Enterovirus A-C* species (Figure 3, Table
272 S11, S12). The sequence of tripeptides, which did not depend on the tetraloop sequence, was, in
273 descending order, IGQ (the most abundant, 65 genomes out of 119), IGL (20 genomes out of
274 119), IGS, VGS, IGN, VGQ, IGV and VGH (Table S11). In the case of the *Rhinovirus A*
275 genome with triloop UCU in domain d (isolate V38_URT-6.3m, accession number JF285329),
276 protein 3C contained tripeptide TGK without positively-charged amino acid before it (Table

277 S11). All genomes of *Rhinovirus B* species contained triloops in domain d, with all but one (with
278 IGK) containing tripeptide TKG in 3C. Genomes of *Rhinovirus C* contained tetraloops mostly of
279 consensus YHCG (H=all but G) and tripeptides in 3C without a positively charged amino acid at
280 the last position (TGN, VGN, TGH) or outside of the tripeptide (Table S12). One genome
281 contained tetraloop CUUC paired with most abundant tripeptide, that is, TGN (23 out of 37
282 genomes)(Table S12).

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

289 Discussion

290 Most of the domain d apical loops in enterovirus genomes were represented by tetraloops.
291 The most common variants of tetraloop sequences corresponded to consensus YNHG (Y=C/U,
292 N=any, H=A/C/U) (Table 2). Similar results were obtained in our previous experimental work,
293 where eight apical nucleotides of domain d of the poliovirus genome were randomized and
294 viable variants were selected *in vitro*, with the majority of selected tetraloops belonging to

295 consensus YNHG (Prostova *et al.*, 2015). Some tetraloops of consensus YNHG were found in
296 genomes in the NCBI database, but not among the variants selected *in vitro*, namely tetraloops
297 CACG, CUCG, UAAG, UGAG, CAAG, CGAG, UGUG and CUUG (Prostova *et al.*, 2015).
298 Tetraloops UGAG, UGUG and CUUG were reconstructed with a U****G flanking base pair in
299 the context of the poliovirus genome strain Mahoney, which supported effective virus replication
300 (Prostova *et al.*, 2015).

301 Conversely, tetraloops UUAG, UCAG and CCAG, found in domain d of selected *in vitro*
302 viable poliovirus variants, were able to support virus reproduction; however, they were not found
303 in naturally circulating viruses (Prostova *et al.*, 2015). One tetraloop of the YNHG consensus
304 (CUAG) was neither found in genomes from the NCBI database (n=2842), nor in the
305 randomized poliovirus genomes selected *in vitro* (n=62) (Table 2). Thermodynamic stability is
306 unlikely to be the reason why this and other tetraloops were unrepresented as the melting
307 temperature of stem loops with avoided tetraloops is within range of the melting temperature of
308 YNHG tetraloops, which supported replication (Proctor *et al.*, 2002). Moreover, tetraloops
309 UUAG and UCAG are common in rRNA (Woese *et al.*, 1990). Sample insufficiency cannot be
310 excluded for both database and *in vitro* selected sets of genomes, but it is safe to conclude that
311 these tetraloop variants are at least extremely rare. In any case, the fact that the incidence of
312 these tetraloops is much less than for other tetraloops indicates that such variants are possibly
313 less fit.

314 The most abundant tetraloops in the domain apical region of genomes from the NCBI
315 database and variants selected *in vitro* could be compiled into consensus UNCG and CNCG
316 (Table 2, Table S1). At the same time, these tetraloops are most abundant in rRNA, and, with
317 certain closing base pairs, among the most thermodynamically stable tetraloops (Woese *et al.*,
318 1990; Proctor *et al.*, 2002). Tetraloops of these consensus and some other found tetraloops of
319 the YNHG consensus form a specific spatial structure of the UNCG structural class of stable
320 tetraloops (Cheong, Varani and Tinoco, 1990; Varani, Cheong and Tinoco, 1991; Du *et al.*,
321 2003, 2004).

322 Another set of tetraloops, which correspond to GNYA consensus, was found both in
323 genomes of *Enterovirus E* and *F* and in genomes of viable polioviruses selected *in vitro*
324 (Prostova *et al.*, 2015). Tetraloop GCUA was able to support the effective replication of
325 poliovirus and, together with tetraloop GUUA, is known to assume an UNCG fold (Ihle *et al.*,
326 2005; Melchers *et al.*, 2006; Prostova *et al.*, 2015). In sum, these data suggest that the spatial
327 structure, rather than the exact sequence, is the main subject for recognition by virus protein 3C.
328 Structure-based recognition of tetraloops occurs in several known RNA-protein complexes. For
329 example, tetraloops with a GNRA class structure in the context of bacteriophages P22 and λ
330 genome transcription antitermination element boxB are specifically recognized by the
331 bacteriophage N-protein arginine-rich motif (ARM) (Cai *et al.*, 1998; Legault *et al.*, 1998;
332 Schärpf *et al.*, 2000). Arginines and lysines of the ARM recognize the shape of the negatively

333 charged phosphodiester backbone of the stem-loop and positions N-peptide for hydrophobic or
334 stacking interaction with a non-conserved nucleotide of the loop (Cai *et al.*, 1998; Legault *et al.*,
335 1998; Schärpf *et al.*, 2000; Thapar, Denmon and Nikonowicz, 2013). Another example of
336 structure-specific recognition is the complex of the double-stranded RNA-binding domain
337 (dsRBD) of RNase Rnt1p and AGNN class tetraloop (Chanfreau, Buckle and Jacquier, 2000;
338 Lebars *et al.*, 2001; Wu *et al.*, 2001, 2004; Wang *et al.*, 2011; Thapar, Denmon and Nikonowicz,
339 2013). Motif dsRBD recognizes the phosphodiester backbone at the 3' side of the tetraloop and
340 its non-conserved third and fourth nucleotides (Wu *et al.*, 2004; Wang *et al.*, 2011; Thapar,
341 Denmon and Nikonowicz, 2013).

342 The sequence to structure degeneracy (different RNA sequences are able to form similar
343 spatial structure) is the known phenomenon (Petrov, Zirbel and Leontis, 2013; Bottaro and
344 Lindorff-Larsen, 2017). Moreover, it is suggested to refrain from associating sequences with a
345 particular fold (D'Ascenzo *et al.*, 2016, 2017). Together with the literature data our result let us
346 assume that sequence-structure degeneracy is an universal way in which RNA tetraloops are
347 used in nature (Lebars *et al.*, 2001; Wu *et al.*, 2004; Ihle *et al.*, 2005; Petrov, Zirbel and Leontis,
348 2013; D'Ascenzo *et al.*, 2016, 2017; Bottaro and Lindorff-Larsen, 2017).

349 It can be speculated that pentaloops in domain d of the *Enterovirus E* genome and
350 triloops of domain d of rhinoviruses have the potential to comprise the same UNCG fold as some
351 YNHG and GNYA tetraloops. For HRV14 domain d, it was shown that its triloop resembles the

352 structure of the first and last two nucleotides of UNCG structural class tetraloops (Headey *et al.*,
353 2007). There are pentaloops with four nucleotides that belong to consensus UNCG, GNRA or
354 gCUUGc, which are able to form spatial structures of corresponding structural classes with the
355 fifth bulged nucleotide (Cai *et al.*, 1998; Schärpf *et al.*, 2000; Theimer, Finger and Feigon, 2003;
356 Oberstrass *et al.*, 2006; Liu *et al.*, 2009). It is possible that four nucleotides of the pentaloops in
357 domain d of Enterovirus E species have a UNCG fold with one bulged nucleotide.

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

365 A similar structure of domain d and its apical region suggests the free exchange of this
366 region between genomes of the same and different species of Enterovirus genera. Indeed, viable
367 intra and inter species recombinants for this region could be obtained *in vitro* (Muslin *et al.*,
368 2015; Bessaud *et al.*, 2016). To evaluate the relative impact of the high mutation rate and
369 recombination on domain d apical loop variability, sequences of EV71 C4 genotype viruses were
370 analysed. The natural recombination in EV71 genotype C4 is much less frequent than other

371 *Enterovirus A* types (Lukashev *et al.*, 2014); meanwhile only one recombinant genome
372 (accession number HQ423143) was detected in our data set. Therefore, the variability of its
373 domain d loop sequence reflects changes that were only accumulated via mutations. The
374 diversity of the domain d loop sequence of EV-71 C4 viruses was far less prominent than among
375 *Enterovirus A* genomes and represented only by five tetraloop sequence variants (Table 2). As
376 the most recent common ancestor of EV71 genotype C4 dates back about 20 years (McWilliam
377 Leitch *et al.*, 2012), this diversity, although limited, has only emerged very recently. On the other
378 hand, the high sequence variability of the domain d apical region in all enterovirus genomes was
379 possibly assisted by inter- and intra-species recombination events.

380 Interestingly, in contrast to the similar structure of domain d and the very similar
381 distribution of its apical sequences in genomes of different enterovirus species, its putative RNA-
382 recognition tripeptide of 3C is diverse (Figure 3). Most *Enterovirus A* genomes contain tripeptide
383 VGK in 3C, while there is a prevalence of the TGK tripeptide among genomes of *Enterovirus B*,
384 *C* and *D* species (Figure 3). Genomes of *Rhinovirus A* and *C* also contain common enterovirus
385 tetraloops in the domain d apical region, but, in 3C, unlike other species, they contain tripeptides
386 without positively charged amino acids (Figure 3, Table S11, Table S12). Positively charged
387 amino acids are often involved in the interaction with RNA, in particular, with phosphates of the
388 RNA backbone. As such, they are of importance to RNA-protein recognition (Jones *et al.*, 2001;
389 Bahadur, Zacharias and Janin, 2008). In *Rhinovirus A* genomes, positively charged amino acid

390 “jumped” from the last position of the tripeptide (position 156) to the position that precedes the
391 tripeptide (position 153) (Figure 3, shown by an arrow). The residue at position 153 starts and
392 the residue at position 156 ends the reverse turn between beta strands dII and eII of protein 3C
393 (Mosimann *et al.*, 1997; Matthews *et al.*, 1999; Cui *et al.*, 2011). In a crystal structure of the
394 Rhinovirus A2 protein 3C, the side chain of Lys153 (preceding the tripeptide) is positioned in a
395 region similar to that of the side chain of Lys156 (in last position of the tripeptide) in the crystal
396 structure of Enterovirus 71 and Poliovirus 1 proteins 3C (Mosimann *et al.*, 1997; Matthews *et*
397 *al.*, 1999; Cui *et al.*, 2011). Thus, Lys at position 153 of 3C has almost the same potential to
398 interact with the RNA-ligand as Lys at position 156 (Mosimann *et al.*, 1997; Matthews *et al.*,
399 1999; Cui *et al.*, 2011). Genomes of *Rhinovirus C* species do not contain a positively charged
400 amino acid, either inside the tripeptide of the 3C protein, or in the neighbouring positions,
401 possibly indicating that tripeptide 154-156 in the protein 3C of *Rhinovirus C* genome does not
402 interact directly with RNA. Thus, 3C is able to recognize domain d of the oriL with tripeptides of
403 a different sequence. In contrast to the domain d structure and its apical sequence, the tripeptide
404 is species-specific. The diversity of the tripeptide, which is expected to recognize domain d, has
405 several compatible explanations. Residue 154 of the tripeptide possibly does not interact with
406 domain d directly. The tripeptide may be involved into a species-specific cooperative amino acid
407 network (amino acid “epistasis”). Moreover, different tripeptides could reflect slightly different
408 molecular mechanisms for domain d recognition.

409 The complexity of the tripeptide's role in domain d recognition can be shown in several
410 examples. The 3C protein of different species with the same RNA-binding tripeptide is not
411 guaranteed to bind the same structured domain d. Genomes of the *Rhinovirus B* contain triloops
412 in the apical region of domain d, which are paired with tripeptide TGK in 3C, common for
413 genomes with tetraloops. In contrast, protein 3C of the Coxsackie virus B3 (*Enterovirus B*
414 species, containing tripeptide TGK) cannot recognize oriL sufficiently well when domain d is
415 capped with a triloop (Zell *et al.*, 2002). This indicates that the sequence of the RNA-binding
416 tripeptide is probably not the exclusive participant in oriL-3C recognition. In other words,
417 different molecular mechanisms of oriL-3C recognition have evolved in every enterovirus
418 species independently. For example, it was shown for Rhinovirus 14 (*Rhinovirus B* species) that
419 protein 3C recognizes the stem region of domain d, rather than its apical loop (Leong *et al.*,
420 1993). Another oriL-3C recognition mechanism is seemingly employed by *Enterovirus E* and *F*
421 species, two oriLs of which play the same role as the single oriL in genomes of other
422 enteroviruses (Pilipenko, Blinov and Agol, 1990; Zell *et al.*, 1999). The apical loop of their
423 domain d is a tetra- or pentaloop with a sequence that differs from the loop consensuses of other
424 enteroviruses. The RNA-binding tripeptide in 3C is species-specific as well, and is always MGK
425 (Table S6). Interestingly, one genome of *Enterovirus G* species had the same pair domain d loop:
426 tripeptide of 3C, i.e., GUUA MGK. Domain d of *Enterovirus G* species is prolonged in
427 comparison to the length of domain d in genomes of other species (Krumbholz *et al.*, 2002)

428 (Figure 2). Tripeptide MGK in the 3C of *Enterovirus E, F* and *G* possibly indicates another
429 molecular mechanism of oriL-3C recognition (Krumbholz *et al.*, 2002). Therefore, we assume
430 that, though putative RNA-binding, the tripeptide, in most cases, possibly interacts with the
431 domain d apical region (since amino acid substitutions in it are known to compensate for
432 structural disturbance in domain d); however, this interaction is not the only one that determines
433 the evolution of oriL-3C interaction. Altogether, the data suggest that the independent evolution
434 of the putative RNA-binding tripeptide of 3C and domain d of oriL occurs.

435

436

Conclusions

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

446 RNA in which tetraloops are used in nature.

447

448

References

449 Acevedo, A., Brodsky, L. and Andino, R. (2013) 'Mutational and fitness landscapes of an

450 RNA virus revealed through population sequencing.', *Nature*. Nature Publishing Group,

451 505(7485), pp. 686–690. doi: 10.1038/nature12861.

452 Andino, R., Rieckhof, G. E., Achacoso, P. L. and Baltimore, D. (1993) 'Poliovirus RNA

453 synthesis utilizes an RNP complex formed around the 5' -end of viral RNA', *EMBO Journal*,

454 12(9), pp. 3587–3598.

455 Andino, R., Rieckhof, G. E. and Baltimore, D. (1990) 'A Functional Ribonucleoprotein

456 around the 5' End of Poliovirus', *Cell*, 63, pp. 369–380.

457 Andino, R., Rieckhof, G. E., Trono, D. and Baltimore, D. (1990) 'Substitutions in the

458 protease (3Cpro) gene of poliovirus can suppress a mutation in the 5' noncoding region.', *J*

459 *Virol*, 64(2), pp. 607–612.

460 Bahadur, R. P., Zacharias, M. and Janin, J. (2008) 'Dissecting protein-RNA recognition

461 sites.', *Nucleic Acids Res.*, 36(8), pp. 2705–2716. doi: 10.1093/nar/gkn102.

462 Bessaud, M., Joffret, M.-L., Blondel, B. and Delpeyroux, F. (2016) 'Exchanges of

463 genomic domains between poliovirus and other cocirculating species C enteroviruses reveal a

464 high degree of plasticity', *Scientific Reports*, 6(1), p. 38831. doi: 10.1038/srep38831.

- 465 Bottaro, S. and Lindorff-Larsen, K. (2017) ‘Mapping the Universe of RNA Tetraloop
466 Folds’, *Biophysical Journal*, 113(2), pp. 257–267. doi: 10.1016/j.bpj.2017.06.011.
- 467 Cai, Z., Gorin, A., Frederick, R., Ye, X., Hu, W., Majumdar, A., Kettani, A. and Patel, D.
468 J. (1998) ‘Solution structure of P22 transcriptional antitermination N peptide-box B RNA
469 complex.’, *Nature structural biology*, 5(3), pp. 203–212.
- 470 Chase, A. J., Daijogo, S. and Semler, B. L. (2014) ‘Inhibition of poliovirus-induced
471 cleavage of cellular protein PCBP2 reduces the levels of viral RNA replication.’, *Journal of*
472 *virology*, 88(6), pp. 3192–3201. doi: 10.1128/JVI.02503-13.
- 473 Chanfreau, G., Buckle, M. and Jacquier, A. (2000) ‘Recognition of a conserved class of
474 RNA tetraloops by *Saccharomyces cerevisiae* RNase III.’, *Proceedings of the National Academy*
475 *of Sciences of the United States of America*, 97(7), pp. 3142–3147. doi:
476 10.1073/pnas.070043997.
- 477 Cheong, C. and Cheong, H. (2010) ‘RNA Structure: Tetraloops’, in *Encyclopedia of life*
478 *sciences*. Chichester: John Wiley & Sons, Ltd. doi: 10.1002/9780470015902.a0003135.pub2.
- 479 Cheong, C., Varani, G. and Tinoco, I. J. (1990) ‘Solution structure of an unusually stable
480 RNA hairpin, 5’GGAC(UUCG)GUCC.’, *Nature.*, 346(6285), pp. 680–682.
- 481 Crooks, G. E., Hon, G., Chandonia, J.-M. and Brenner, S. E. (2004) ‘WebLogo: A
482 Sequence Logo Generator’, *Genome Research*, 14(6), pp. 1188–1190. doi: 10.1101/gr.849004.
- 483 Cui, S., Wang, J., Fan, T., Qin, B., Guo, L., Lei, X., Wang, J., Wang, M. and Jin, Q.

484 (2011) ‘Crystal structure of human enterovirus 71 3C protease.’, *Journal of molecular biology*,
485 408(3), pp. 449–461. doi: 10.1016/j.jmb.2011.03.007.

486 D’Ascenzo, L., Leonarski, F., Vicens, Q. and Auffinger, P. (2016) “‘Z-DNA like”
487 fragments in RNA: a recurring structural motif with implications for folding, RNA/protein
488 recognition and immune response.’, *Nucleic acids research*, 44(12), pp. 5944–5956. doi:
489 10.1093/nar/gkw388.

490 D’Ascenzo, L., Leonarski, F., Vicens, Q. and Auffinger, P. (2017) ‘Revisiting GNRA and
491 UNCG folds: U-turns versus Z-turns in RNA hairpin loops’, *RNA*, 23(3), pp. 259–269. doi:
492 10.1261/rna.059097.116.

493 Du, Z., Yu, J., Andino, R. and James, T. L. (2003) ‘Extending the family of UNCG-like
494 tetraloop motifs: NMR structure of a CACG tetraloop from coxsackievirus B3.’, *Biochemistry*,
495 42(15), pp. 4373–4383. doi: 10.1021/bi027314e.

496 Du, Z., Yu, J., Ulyanov, N. B., Andino, R. and James, T. L. (2004) ‘Solution structure of
497 a consensus stem-loop D RNA domain that plays important roles in regulating translation and
498 replication in enteroviruses and rhinoviruses.’, *Biochemistry*, 43(38), pp. 11959–11972. doi:
499 10.1021/bi048973p.

500 Gamarnik, A. V and Andino, R. (1998) ‘Switch from translation to RNA replication in a
501 positive-stranded RNA virus’, *Genes & Dev.*, 12, pp. 2293–2304. doi: 10.1101/gad.12.15.2293.

502 Goodfellow, I., Chaudhry, Y., Richardson, A., Meredith, J., Almond, J. W., Barclay, W.

503 and Evans, D. J. (2000) 'Identification of a cis-acting replication element within the poliovirus
504 coding region.', *Journal of virology*, 74(10), pp. 4590–4600. doi: 10.1128/JVI.74.10.4590-
505 4600.2000.

506 Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. and Hofacker, I. L. (2008) 'The
507 Vienna RNA websuite.', *Nucleic acids research*, 36(Web Server issue), pp. W70-4. doi:
508 10.1093/nar/gkn188.

509 Hämmerle, T., Hellen, C. U. and Wimmer, E. (1991) 'Site-directed mutagenesis of the
510 putative catalytic triad of poliovirus 3C proteinase.', *The Journal of biological chemistry*, 266(9),
511 pp. 5412–5416.

512 Hämmerle, T., Molla, A. and Wimmer, E. (1992) 'Mutational analysis of the proposed
513 FG loop of poliovirus proteinase 3C identifies amino acids that are necessary for 3CD cleavage
514 and might be determinants of a function distinct from proteolytic activity.', *J Virol.*, 66(10), pp.
515 6028–6034.

516 Harris, K. S., Reddigari, S. R., Nicklin, M. J., Hämmerle, T. and Wimmer, E. (1992)
517 'Purification and characterization of poliovirus polypeptide 3CD, a proteinase and a precursor
518 for RNA polymerase.', *J Virol.*, 66(12), pp. 7481–7489.

519 Headey, S. J., Huang, H., Claridge, J. K., Soares, G. A., Dutta, K., Schwalbe, M., Yang,
520 D. and Pascal, S. M. (2007) 'NMR structure of stem-loop D from human rhinovirus-14.', *RNA*,
521 13(3), pp. 351–360. doi: 10.1261/rna.313707.

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

526 Jones, S., Daley, D. T. A., Luscombe, N. M., Berman, H. M. and Thornton, J. M. (2001)
527 ‘Protein – RNA interactions : a structural analysis’, *Biochemistry*, 29(4), pp. 943–954.

528 Katoh, K. and Standley, D. M. (2013) ‘MAFFT multiple sequence alignment software
529 version 7: improvements in performance and usability.’, *Molecular biology and evolution*, 30(4),
530 pp. 772–80. doi: 10.1093/molbev/mst010.

531 Krumbholz, A., Dauber, M., Henke, A., Birch-Hirschfeld, E., Knowles, N. J., Stelzner, A.
532 and Zell, R. (2002) ‘Sequencing of porcine enterovirus groups II and III reveals unique features
533 of both virus groups.’, *Journal of virology*, 76(11), pp. 5813–5821.

534 Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A.,
535 McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T.
536 J. and Higgins, D. G. (2007) ‘Clustal W and Clustal X version 2.0.’, *Bioinformatics*, 23(21), pp.
537 2947–2948. doi: 10.1093/bioinformatics/btm404.

538 Lauring, A. S., Frydman, J. and Andino, R. (2013) ‘The role of mutational robustness in
539 RNA virus evolution.’, *Nature reviews. Microbiology*, 11(5), pp. 327–336. doi:
540 10.1038/nrmicro3003.

- 541 Lebars, I., Lamontagne, B., Yoshizawa, S., Aboul-Elela, S. and Fourmy, D. (2001)
542 ‘Solution structure of conserved AGNN tetraloops: insights into Rnt1p RNA processing.’, *The*
543 *EMBO journal*, 20(24), pp. 7250–7258. doi: 10.1093/emboj/20.24.7250.
- 544 Legault, P., Li, J., Mogridge, J., Kay, L. E. and Greenblatt, J. (1998) ‘NMR structure of
545 the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an
546 arginine-rich motif.’, *Cell*, 93(2), pp. 289–299.
- 547 Leong, L. E. C., Walker, P. A., Porter, A. G., Protease, H. R., Leon, L. E. C., Walker, P.
548 A., Porter, A. G., Leong, L. E. C., Walker, P. A. and Porter, A. G. (1993) ‘Human Rhinovirus-14
549 Protease 3C (3Cpro) Binds Specifically to the 5’-Noncoding Region of the Viral RNA’, *The*
550 *Journal of biological chemistry*, 268(34), pp. 25735–25739.
- 551 Liu, P., Li, L., Keane, S. C., Yang, D., Leibowitz, J. L. and Giedroc, D. P. (2009) ‘Mouse
552 hepatitis virus stem-loop 2 adopts a uYNMG(U)a-like tetraloop structure that is highly
553 functionally tolerant of base substitutions.’, *Journal of virology*, 83(23), pp. 12084–12093. doi:
554 10.1128/JVI.00915-09.
- 555 Lorenz, R., Bernhart, S. H., Höner Zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P.
556 F. and Hofacker, I. L. (2011) ‘ViennaRNA Package 2.0.’, *Algorithms for molecular biology*,
557 6(26). doi: 10.1186/1748-7188-6-26.
- 558 Lukashev, A. N., Shumilina, E. Y., Belalov, I. S., Ivanova, O. E., Eremeeva, T. P.,
559 Reznik, V. I., Trotsenko, O. E., Drexler, J. F. and Drosten, C. (2014) ‘Recombination strategies

560 and evolutionary dynamics of the Human enterovirus A global gene pool.’, *The Journal of*
561 *general virology*, 95(Pt 4), pp. 868–873. doi: 10.1099/vir.0.060004-0.

562 Matthews, D. a, Dragovich, P. S., Webber, S. E., Fuhrman, S. a, Patick, a K., Zalman, L.
563 S., Hendrickson, T. F., Love, R. a, Prins, T. J., Marakovits, J. T., Zhou, R., Tikhe, J., Ford, C. E.,
564 Meador, J. W., Ferre, R. a, Brown, E. L., Binford, S. L., Brothers, M. a, DeLisle, D. M. and
565 Worland, S. T. (1999) ‘Structure-assisted design of mechanism-based irreversible inhibitors of
566 human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus
567 serotypes.’, *Proceedings of the National Academy of Sciences of the United States of America*,
568 96(20), pp. 11000–11007.

569 McWilliam Leitch, E. C., Cabrerizo, M., Cardoso, J., Harvala, H., Ivanova, O. E., Koike,
570 S., Kroes, A. C. M., Lukashev, A., Perera, D., Roivainen, M., Susi, P., Trallero, G., Evans, D. J.
571 and Simmonds, P. (2012) ‘The association of recombination events in the founding and
572 emergence of subgenogroup evolutionary lineages of human enterovirus 71.’, *Journal of*
573 *virology*, 86(5), pp. 2676–2685. doi: 10.1128/JVI.06065-11.

574 Melchers, W. J. G., Zoll, J., Tessari, M., Bakhmutov, D. V, Gmyl, A. P., Agol, V. I. and
575 Heus, H. a (2006) ‘A GCUA tetranucleotide loop found in the poliovirus oriL by in vivo SELEX
576 (un)expectedly forms a YNMG-like structure: Extending the YNMG family with GYYA.’, *RNA*,
577 12(9), pp. 1671–1682. doi: 10.1261/rna.113106.

578 Mosimann, S. C., Cherney, M. M., Sia, S., Plotch, S. and James, M. N. (1997) ‘Refined

579 X-ray crystallographic structure of the poliovirus 3C gene product.’, *J Mol Biol*, 273(5), pp.
580 1032–1047. doi: 10.1006/jmbi.1997.1306.

581 Muslin, C., Joffret, M.-L., Pelletier, I., Blondel, B. and Delpyroux, F. (2015) ‘Evolution
582 and Emergence of Enteroviruses through Intra- and Inter-species Recombination: Plasticity and
583 Phenotypic Impact of Modular Genetic Exchanges in the 5’ Untranslated Region’, *PLOS*
584 *Pathogens*, 11(11), p. e1005266. doi: 10.1371/journal.ppat.1005266.

585 Oberstrass, F. C., Lee, A., Stefl, R., Janis, M., Chanfreau, G. and Allain, F. H.-T. (2006)
586 ‘Shape-specific recognition in the structure of the Vts1p SAM domain with RNA.’, *Nature*
587 *structural & molecular biology*, 13(2), pp. 160–167. doi: 10.1038/nsmb1038.

588 Oermann, C. M., Schuster, J. E., Conners, G. P., Newland, J. G., Selvarangan, R. and
589 Jackson, M. A. (2015) ‘Enterovirus D68. A focused review and clinical highlights from the 2014
590 U.S. Outbreak.’, *Annals of the American Thoracic Society*, 12(5), pp. 775–781. doi:
591 10.1513/AnnalsATS.201412-592FR.

592 Palmenberg, A., Neubauer, D. and Skern, T. (2010) ‘Genome organization and encoded
593 proteins.’, in Ehrenfeld, E., Domingo, E., and Roos, R. P. (eds) *The Picornaviruses*. ASM Press,
594 pp. 3–17.

595 Petrov, A. I., Zirbel, C. L. and Leontis, N. B. (2013) ‘Automated classification of RNA
596 3D motifs and the RNA 3D Motif Atlas’, *RNA*, 19(10), pp. 1327–1340. doi:
597 10.1261/rna.039438.113.

- 598 Pilipenko, E. V, Blinov, V. M. and Agol, V. I. (1990) ‘Gross rearrangements within the
599 5’-untranslated region of the picornaviral genomes.’, *Nucleic acids research*, 18(11), pp. 3371–
600 3375.
- 601 Proctor, D. J., Schaak, J. E., Bevilacqua, J. M., Falzone, C. J. and Bevilacqua, P. C.
602 (2002) ‘Isolation and characterization of a family of stable RNA tetraloops with the motif
603 YNMG that participate in tertiary interactions.’, *Biochemistry*, 41(40), pp. 12062–12075.
- 604 Prostova, M. A., Gmyl, A. P., Bakmutov, D. V, Shishova, A. A., Khitrina, E. V,
605 Kolesnikova, M. S., Serebryakova, M. V, Isaeva, O. V and Agol, V. I. (2015) ‘Mutational
606 robustness and resilience of a replicative cis-element of RNA virus: promiscuity, limitations,
607 relevance.’, *RNA biology*, 12(12), pp. 1338–1354. doi: 10.1080/15476286.2015.1100794.
- 608 Rieder, E., Xiang, W., Paul, A. and Wimmer, E. (2003) ‘Analysis of the cloverleaf
609 element in a human rhinovirus type 14/poliovirus chimera: correlation of subdomain D structure,
610 ternary protein complex formation and virus replication’, *J Gen Virol*, 84(8), pp. 2203–2216. doi:
611 10.1099/vir.0.19013-0.
- 612 Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. and Belshaw, R. (2010) ‘Viral
613 mutation rates.’, *Journal of virology*, 84(19), pp. 9733–9748. doi: 10.1128/JVI.00694-10.
- 614 Schärpf, M., Sticht, H., Schweimer, K., Boehm, M., Hoffmann, S. and Rösch, P. (2000)
615 ‘Antitermination in bacteriophage lambda. The structure of the N36 peptide-boxB RNA
616 complex.’, *Eur. J. Biochem.*, 267(267), pp. 2397–2408.

- 617 Shih, S., Chen, T. and Wu, C. (2004) ‘Mutations at KFRDI and VGK Domains of
618 Enterovirus 71 3C Protease Affect Its RNA Binding and Proteolytic Activities’, *Journal of*
619 *Biomedical Science*, pp. 239–248. doi: 10.1159/000076036.
- 620 Solomon, T., Lewthwaite, P., Perera, D., Cardosa, M. J., McMinn, P. and Ooi, M. H.
621 (2010) ‘Virology, epidemiology, pathogenesis, and control of enterovirus 71.’, *The Lancet.*
622 *Infectious diseases*, 10(11), pp. 778–790. doi: 10.1016/S1473-3099(10)70194-8.
- 623 Thapar, R., Denmon, A. P. and Nikonowicz, E. P. (2013) ‘Recognition modes of RNA
624 tetraloops and tetraloop-like motifs by RNA-binding proteins.’, *Wiley interdisciplinary reviews.*
625 *RNA*, 5(1), pp. 49–67. doi: 10.1002/wrna.1196.
- 626 Theimer, C. A., Finger, L. D. and Feigon, J. (2003) ‘YNMG tetraloop formation by a
627 dyskeratosis congenita mutation in human telomerase RNA’, *RNA*, 9, pp. 1446–1455. doi:
628 10.1261/rna.5152303.activity.
- 629 Thompson, A. A. and Peersen, O. B. (2004) ‘Structural basis for proteolysis-dependent
630 activation of the poliovirus RNA-dependent RNA polymerase.’, *The EMBO journal*, 23(17), pp.
631 3462–3471. doi: 10.1038/sj.emboj.7600357.
- 632 Trono, D., Andino, R. and Baltimore, D. (1988) ‘An RNA sequence of hundreds of
633 nucleotides at the 5’ end of poliovirus RNA is involved in allowing viral protein synthesis.’,
634 *Journal of virology*, 62(7), pp. 2291–2299.
- 635 Uhlenbeck, O. C. (1990) ‘Tetraloops and RNA folding’, *Nature*. Nature Publishing

636 Group, 346(6285), pp. 613–614. doi: 10.1038/346613a0.

637 Varani, G., Cheong, C. and Tinoco, I. (1991) ‘Structure of an unusually stable RNA
638 hairpin.’, *Biochemistry*, 30(13), pp. 3280–3289.

639 Vogt, D. A. and Andino, R. (2010) ‘An RNA element at the 5’-end of the poliovirus
640 genome functions as a general promoter for RNA synthesis’, *PLoS pathogens*, 6(6), p. e1000936.
641 doi: 10.1371/journal.ppat.1000936.

642 Wagner, A. and Stadler, P. F. (1999) ‘Viral RNA and evolved mutational robustness.’,
643 *The Journal of experimental zoology*, 285(2), pp. 119–127.

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

647 Woese, C. R., Winker, S., Gutell, R. R., Winkers, S. and Gutell, R. R. (1990)
648 ‘Architecture of ribosomal RNA : Constraints on the sequence of “tetra-loops”’, *Proceedings of*
649 *the National Academy of Sciences of the United States of America*, 87(November), pp. 8467–
650 8471.

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

654 Wu, H., Yang, P. K., Butcher, S. E., Kang, S., Chanfreau, G. and Feigon, J. (2001) ‘A

655 novel family of RNA tetraloop structure forms the recognition site for *Saccharomyces cerevisiae*

656 RNase III.', *The EMBO journal*, 20(24), pp. 7240–9. doi: 10.1093/emboj/20.24.7240.

657 Zell, R., Sidigi, K., Bucci, E., Stelzner, A. and Görlach, M. (2002) 'Determinants of the

658 recognition of enteroviral cloverleaf RNA by coxsackievirus B3 proteinase 3C.', *RNA (New*

659 *York, N.Y.*), 8(2), pp. 188–201. Available at:

660 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1370242&tool=pmcentrez&renderty>

661 [pe=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1370242&tool=pmcentrez&renderty) (Accessed: 19 December 2014).

662 Zell, R., Sidigi, K., Henke, A., Schmidt-Brauns, J., Hoey, E., Martin, S. and Stelzner, A.

663 (1999) 'Functional features of the bovine enterovirus 5'-non-translated region.', *J Gen Virol*, 80,

664 pp. 2299–2309.

665

666

667

Figure 1

Schematic representation of poliovirus genome and detailed representation of secondary structure of poliovirus replicative element oriL.

Protein 3C and its RNA-binding motifs $_{82}\text{KFRDI}_{86}$ and $_{154}\text{TGK}_{156}$ are shown (here and below amino acid numeration corresponds to poliovirus protein 3C) (Prostova et al., 2015 with modifications).

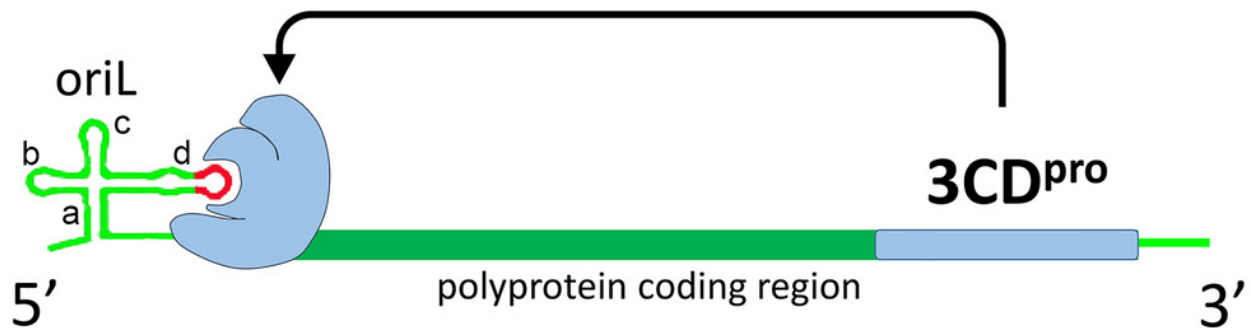


Figure 2

Secondary structure of oriL domain d of distinct enterovirus species.

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 Krumbholtz et al., 2002.

Enterovirus

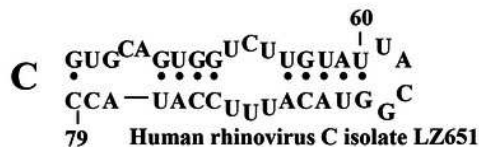
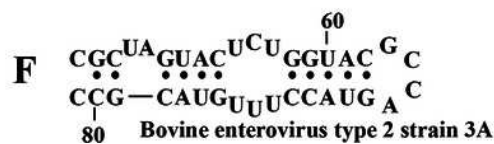
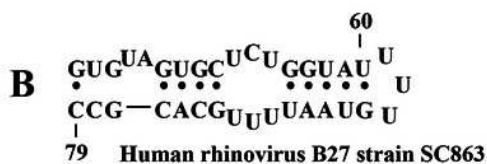
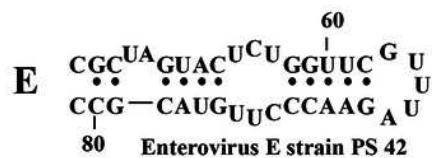
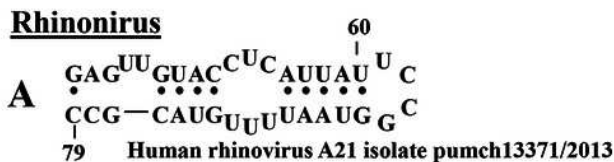
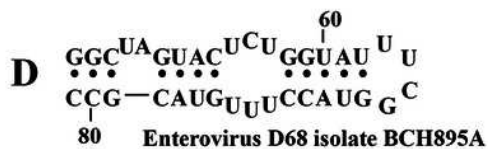
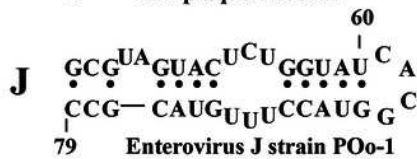
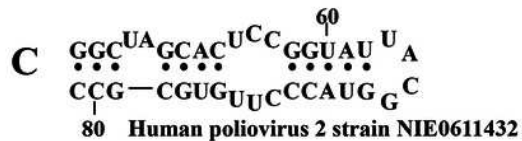
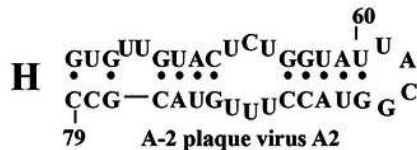
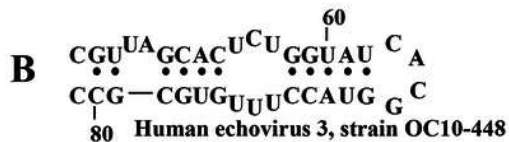
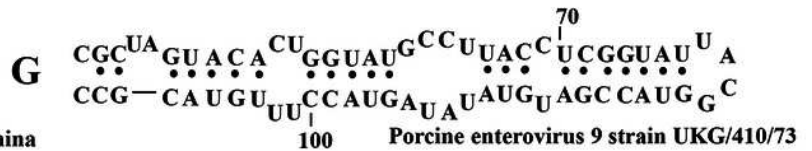
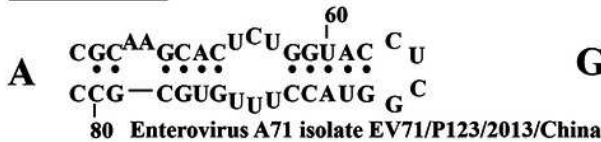
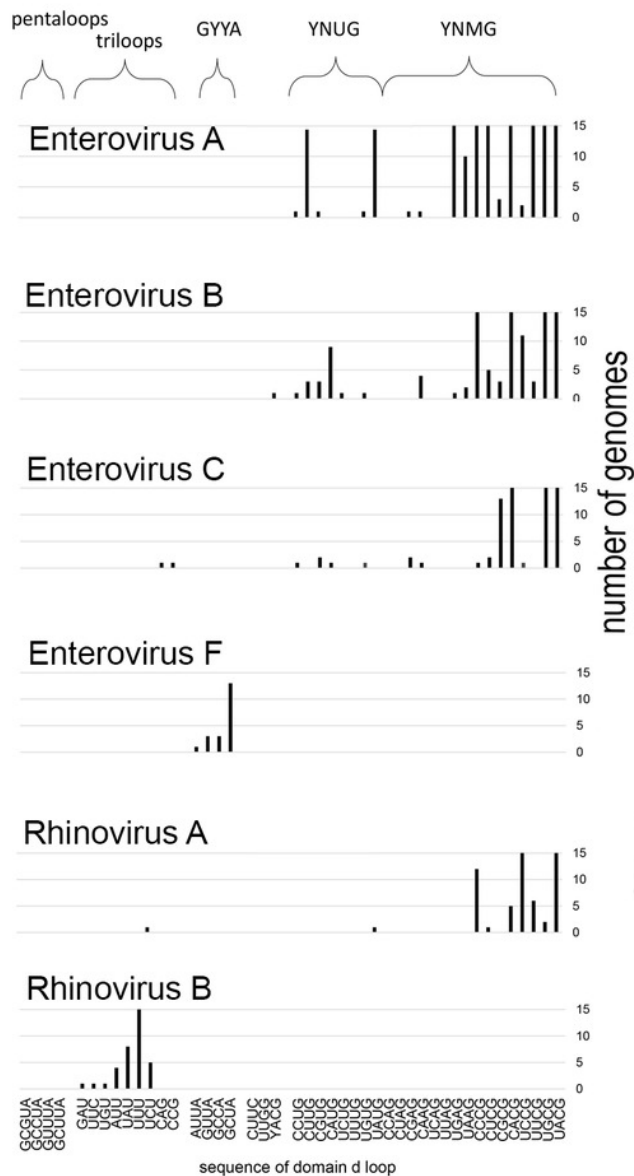


Figure 3

Distribution 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 consensus, 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 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 $_{82}\text{KFRDI}_{86}$, the putative RNA-binding tripeptide 154-156 of 3C and Lys153 in the protein 3C of *Rhinovirus A*.

A



B



Table 1 (on next page)

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	151	118	8		8	
Rhinovirus B	50	37	7		7	
Rhinovirus C	38	37	6		6	

1

Table 2 (on next page)

Occurrence of domain d apical sequences in filtered sets of full genomes of different enterovirus species.

Tetraloops CCCG, UGUG, CAUG and UUGG that were unique for species *Enterovirus A, C* and *D* and were lost upon filtration, were added and are shown in blue. The gradient coloring from red to green represents abundance heat map for the genomes with different domain d sequence.

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

6

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																	

