Published in final edited form as:
Virology. 2011 January 5; 409(1): 1–11. doi:10.1016/j.virol.2010.09.028.

Identification of tolerated insertion sites in poliovirus non-structural proteins

Natalya L. Teterina[a], **Chris Lauber**[b], **Kenneth S. Jensen**, **Eric A. Levenson**, **Alexander E. Gorbalenya**[b], and **Ellie Ehrenfeld**[a]

[a]Laboratory of Infectious Diseases, NIAID, NIH, Bethesda, Maryland 20892, USA, nteterina@niaid.nih.gov, eehrenfeld@niaid.nih.gov

[b]Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands, c.lauber@lumc.nl, a.e.gorbalenya@lumc.nl

Abstract

Insertion of nucleotide sequences encoding “tags” that can be expressed in specific viral proteins during an infection is a useful strategy for purifying viral proteins and their functional complexes from infected cells and/or for visualizing the dynamics of their subcellular location over time. To identify regions in the poliovirus polyprotein that could potentially accommodate insertion of tags, transposon-mediated insertion mutagenesis was applied to the entire nonstructural protein-coding region of the poliovirus genome, followed by selection of genomes capable of generating infectious, viable viruses. This procedure allowed us to identify at least one site in each viral nonstructural protein, except protein 2C, in which a minimum of five amino acids could be inserted. The distribution of these sites is analyzed from the perspective of their protein structural context and from the perspective of virus evolution.

Keywords

poliovirus; transposon-mediated insertion mutagenesis; viral protein tags; protein structure; enterovirus evolution

Introduction

The genome organization and the proteome of picornaviruses are defining characteristics of the family (Gorbalenya and Lauber, 2010). All viral proteins are encoded by a single piece of single-stranded RNA containing only one open reading frame in all but one species. The resulting polyprotein is proteolytically processed through a sequential cascade of primary, secondary and maturation cleavages to generate the full complement of structural and non-structural precursor and mature proteins required for virus proliferation. Extensive sequence data are available for many genera and species members, allowing comparative genomic analyses and accurate gene maps. There still remains, however, a dearth of information about the specific biochemical activities and roles contributed by the majority of viral non-structural proteins during the replication process.

To fill this gap will require detailed studies of purified viral proteins and viral complexes, their interactions with both cellular and other viral components, and the dynamics of their intracellular movements during the course of infection. Recent advances in protein purification as well as intracellular localization by live cell imaging often rely on our ability to "tag" the protein of interest. For the tagged proteins to function normally during infection, the tag should be placed in the full-length infectious genome, since picornaviral proteins are delivered to their replication sites as precursor forms and they often function only in cis, replicating the RNA from which the polyprotein was translated (Egger et al., 2000). However, identifying acceptable sites for insertion of sequences to generate tagged proteins from the viral precursor polyprotein during infection can technically be very challenging. Insertions may interfere with the folding of mature and precursor proteins or interprotein interactions, including presentation of cleavage sites, resulting in non-viable virus progeny. Rarely, sites in proteins that can accommodate small insertions compatible with viability of the viruses harboring these mutations have been identified by mutagenesis (Bernstein and Baltimore, 1988; Li and Baltimore, 1988). Finding these sites by trial and error, however, or even with the aid of protein structural information, is often tedious, with uncertainty of success. An alternative approach to this problem is to employ transposon-mediated insertion mutagenesis to generate a library of viral genomes containing small insertions at random sites throughout the coding sequence of interest. The library can be used subsequently to select viable genomes (Atasheva et al., 2007; Frolov et al., 2009; Liu et al., 2006; Moradpour et al., 2004). This approach has recently identified non-disruptable protein or RNA domains that perform essential functions in the hepatitis C virus (Arumugaswami et al., 2008) and potato X virus (Draghici et al., 2009) life cycles.

We employed this method to identify sites in poliovirus non-structural proteins that would tolerate small insertions and generate viable viruses. The ability to select for a plaque-forming phenotype among the large numbers of viral genomes harboring insertions is a highly powerful selection technique. Our objective in the present study was to create a set of viable viruses stably carrying insertions in each of the poliovirus non-structural proteins. This set could provide a guide for creating viruses with internally tagged proteins that could potentially be used to monitor sub-cellular localization and dynamics during infection, to facilitate studies of the structure and composition of viral replication complexes, or to identify interactions with binding partners of specific viral proteins in the infected cell. This report summarizes and rationalizes the set of sites that accommodated small 5-amino acid (aa) insertions that we identified within the poliovirus non-structural protein-coding region. Detailed studies of viruses constructed by placing different tags in the identified sites in two poliovirus proteins, 2A and 3A, have been published recently (Teterina et al., 2010) and in preparation, respectively.

Results and Discussion

Generation of a 15-nt insertion mutant virus library

We utilized transposon-mediated insertion mutagenesis to introduce 15-nt insertions randomly into three separate overlapping segments of the poliovirus genomic cDNA (see Fig. 1). The P1 region encoding the viral structural proteins was excluded from this analysis, since our focus was on non-structural proteins for which function information is lacking or at best incomplete. For each segment of the PV genome the pool of mutagenized cDNA fragments from each subclone was substituted for the corresponding segment of non-mutagenized cDNA in plasmids encoding full-length PV cDNA to create a library of mutagenized genomes. The majority of the inserted sequence was removed by restriction enzyme digestion and re-ligation, leaving only 15 nts, five of which were duplicated from the target DNA, coding for different amino acid sequences depending on the insertion reading frame (e.g., see Fig. 2A). Viral RNAs were transcribed from the entire library

represented by clones of PV cDNAs with insertions in one of three segments indicated in Fig. 1 and used to transfect HeLa cells, which were then overlaid with agar and incubated to allow plaques to develop from any viable viruses. The pools of transcripts displayed RNA infectivities of $10^3 – 10^4$ pfu/μg RNA for different transript pools, approximately 100 – 1000 times lower than wild-type RNA transcripts. This suggests that the majority of insertions produced a lethal phenotype. Individual plaques that developed after transfection were picked, viruses were propagated, and insertion sites in viable viruses that withstood passaging were identified by cDNA sequencing. Many of the insertions were identified in viruses isolated from more than one independently obtained virus plaque (see Fig. 2A), suggesting that the library we analyzed was representative. However, it is possible that not every potential location that could tolerate these insertions was identified in our screen.

Figure 1 shows the locations of the insertions in the non-structural region of the poliovirus polyprotein encoded by viable insertion mutants that we isolated. In all cases, the entire P2 and P3 regions of each isolate were sequenced to identify the insertion site. For most of these, we then introduced the identified insertions into a wild-type cDNA background to ensure that no mutations elsewhere in the viral genomes had occurred. All such reconstructed viruses manifested the same growth phenotypes as the initial isolates. For those insertions that were not reconstructed (those in 3C and 3D), there remains a possibility of an additional mutation in the 5′ NTR or P1 region. At least one site that tolerated a 5-aa insertion was identified for each viral protein except 2C. We found two sites in protein 2A; one site in protein 2B; none in 2C; there were multiple sites clustered in a small region near the N-terminus of 3A; two sites in 3B; multiple sites were clustered near the C-terminus of 3C, including a site at the junction of 3C and 3D; and two sites near either end of the linear protein sequence were identified in 3D.

Growth properties of viruses carrying 15 nt insertions

Several viruses carrying a 15-nt insertion in the genome manifested reduced growth and altered plaque morphology, although some appeared to grow as well as wild-type, as shown in Fig. 2B. The virus isolates are designated by the name of the protein containing the insertion followed by the number of the amino acid residue after which the five amino acids were inserted. The nucleotide and deduced amino acid sequences of the regions containing the insertions are shown in Fig. 2A. Insertions at different positions within the same protein may affect growth differently (e.g., compare 3B-10 with 3B-17). Note that the mechanism of transposon-mediated mutagenesis results in a duplication of five nucleotides derived from the original gene sequence. Thus, since the nucleotide insertion reaction may occur in different reading frames, insertions at the same amino acid position in the protein may generate different amino acid insertion sequences, albeit of the same 5 aa length (e.g., compare 3C-183a and 3C-183b in Fig. 2A). Such pairs of viruses containing different 5 aa insertions at the same position in the same protein also may display quite different plaque sizes (e.g., compare 3D-53a with 3D-53b and 3C-179a and 3C-179b in Fig. 2B). We consistently observed heterogeneity in plaque sizes, even among wild-type viruses (Fig. 2B). Purification and amplification of viruses from small vs. large-size plaques generated by a given mutant virus stock did not result in different insert sequences or stabilities. Occasionally, an extremely poorly growing virus stock generated a rare large plaque (e.g., see 2B-23), which may indicate partial or complete deletion or mutation of the insertion sequence, although this was not confirmed specifically for 2B-23.

Characteristics of tolerated insertion sites

In cases where protein structural information is available, the locations of insertions in viable viruses could be mapped to external surface loops or flexible, non-structured regions of the protein. Figure 3 shows the collection of PV proteins for which structures have been

resolved either by x-ray crystallography or NMR, with the positions of the amino acids preceding the 5-aa insertions present in viable viruses indicated in red. The structural contexts of the identified insertion sites are described below for each protein.

2A—In the enterovirus genus, to which PV belongs, the 2A protein is a small cysteine proteinase that catalyzes the cleavage of its own N –terminus *in cis*, thereby releasing the capsid protein precursor from the remainder of the nascent polyprotein as the first step in the processing cascade. 2A also cleaves a small number of cellular proteins whose activities affect virus growth in a variety of ways, and has been implicated directly in the process of viral RNA replication, independent of its protease functions (see (Teterin et al., 2010) and refs therein). Two insertion sites were identified after selection for viable viruses, between amino acid residues 50-51 or 144-145. The 5-aa insertions at either of these sites had little effect on virus growth (Fig. 2B). The insertion at residue 50 is located in a loop region that separates two domains of the protein (Fig. 3a); the 2A-144 insertion is located near the C-terminal end of the protein in a region with apparently very little stable structure (Fig. 3a). We have utilized these sites for placement of fluorescent proteins in protein 2A; a detailed study of those viruses has been presented previously (Teterina et al., 2010).

2B—The enteroviral protein 2B (~100 aa) and its relatively stable precursor 2BC, have been implicated in the remodeling of intracellular membranes to form the vesicle-like structures that support viral RNA replication (Aldabe and Carrasco, 1995; Cho et al., 1994; Taylor and Kirkegaard, 2007). The protein has been shown to bind membranes via an amphipathic helix located near the N-terminus (aa residues 35-52 for PV), and may cause impairment of Golgi trafficking and reduction in Ca++ levels by forming pores (Agirre et al., 2002; de Jong et al., 2008).

The insertion site identified in this study was located between aa residues 23 and 24, and therefore does not disrupt the essential amphipathic helix domain. No structural data are available for the 2B protein; thus no impact of an insertion near the N-terminus and upstream of the amphipathic helix can be predicted. Plaques formed by viruses harboring the 15 nt insertion are quite small, and the viruses were not further characterized.

3A—The 3A proteins from the different picornavirus genera display a high degree of variability in size, sequence and function. They all contain at least one hydrophobic domain, which apparently serves to anchor the protein and its precursors in the membrane structures where viral RNA replication occurs. Expression of PV or CVB3 3A proteins alone causes disruption of ER-to-Golgi secretory transport and disassembly of the Golgi complex (Cornell et al., 2006; Doedens and Kirkegaard, 1995; Wessels et al., 2005), and 3A may contribute to the membrane remodeling that generates the vesicle-like replication complex scaffold (Suhy et al., 2000). Modulation of membrane transport or structure involves regions near the N-terminus of 3A, which are group-specific among other picornaviral genera 3A proteins. The enteroviral 3AB (or larger) precursor protein(s) appear to participate directly in viral RNA synthesis reactions, perhaps as a donor of 3B (VPg) for RNA chain initiation (Paul et al., 1998) or as a cofactor for 3Dpolymerase and/or 3CD functions (Lama et al., 1994; Molla et al., 1994).

The structure of the soluble portion of the PV 3A protein, whose 24-aa hydrophobic C-terminal domain was deleted, was determined by NMR (Strauss et al., 2003). This truncated protein forms a symmetric dimer in solution with unstructured N- and C-termini (Fig. 3b). Within this unstructured N-terminus, we found a series of five closely neighboring sites, after aa residues 2, 6, 9, 10 or 11 (Fig. 3b), that tolerated small insertions to generate stable viruses. We used these sites to place several different tags in the 3A protein, and detailed descriptions of these viruses will be presented elsewhere (in preparation). A virus known as

3A-2 harboring a single amino acid insertion at residue 15 in the 3A sequence was isolated some years ago (Bernstein and Baltimore, 1988), and has been utilized subsequently by several groups to study 3A function. We consider that insertion to have been placed in the same generally unstructured region as the cluster identified in this study.

3B—The 3B polypeptide (also called VPg) is covalently attached to the 5′ ends of viral RNAs via a phosphodiester linkage between a conserved tyrosine residue (aa 3 of 3B) and the -phosphate of the terminal uridylylate residue of the RNA. Both 3B and its precursor 3AB are also found free in the cytoplasm of infected cells. A di-uridylylated form of 3B serves as primer to initiate synthesis of both positive and negative RNA strands (Takegami et al., 1983). The poliovirus protein is small (22 residues), basic (pI around 10), with some hydrophobic patches. Its structure has been determined by NMR (Schein et al., 2006): although the peptide is relatively unstructured in aqueous buffer, a single conformer can be obtained in the presence of trimethylamine N-oxide, which shows a large loop region from residues 1-14 with the reactive tyrosine projecting outward, and a C-terminal helix comprised of residues 18-21 that aligns residues of conserved amino acids on one face.

Two sites in 3B that tolerated 5-aa insertions were identified from the transposon screen -- between aa residues 10-11 and between 17-18. The latter formed larger plaques than the former (Fig. 2B). These sites are located within the large, flexible loop and just at the start of the terminal helical region, respectively (Fig. 3c).

3C—The 3C proteinases, either as mature proteins or within the context of larger precursors such as 3CD, catalyze the majority of the cleavage events occurring during processing of the viral polyprotein, and additionally target a specific set of cellular proteins that impact a variety of host cell metabolic processes. Structures of 3C proteinases from a number of picornavirus genera have been determined to atomic resolution (HAV (Allaire et al., 1994); rhino (Matthews et al., 1994); polio (Mosimann et al., 1997); FMDV (Birtley et al., 2005). They have chymotrypsin-like folds and catalytic triads, except for an active site cysteine in place of serine and, in a subset of picornaviruses including enteroviruses, active site glutamate instead of aspartate residues (Gorbalenya et al., 1989). The proteins additionally manifest specific RNA binding activities at a site distinct from the proteinase catalytic site (Seipelt et al., 1999).

The carboxyl-end of the 183-residue 3C protein is a short three-residue helix followed by three terminal residues that are flexible and disordered (Fig. 3d). Viable viruses with 15-nt insertions at multiple positions within those terminal six residues were isolated, after residues 179, 181, 182 or 183, the latter of which represents the C-terminus of the wt 3C coding sequence and therefore serves as the 3C/3D processing site. Since the precursor 3CD displays both proteinase and RNA binding activities, and since the domain structures of both 3C and 3D are relatively unchanged in the precursor (Marcotte et al., 2007), the C-terminal 3C residues serve as a linker to 3D and are tolerant to insertions (Fig. 3f). It was surprising to us that insertions at the very end of 3C (e.g., isolates 3C-183a and b) apparently did not interfere significantly with the cleavage site specificity or processing efficiency. The scissile bond for 3C/3D cleavage is always Q/G in the PV polyprotein. Assuming that it remains so for the 3C-183a or 3C-183b variants that we isolated, cleavage of 3CD in isolate 3C-183a would generate a 3C protein with a 5-aa extension at its C-terminus (see Figs. 2A and 3f). Cleavage of 3CD in isolate 3C-183b, however, could yield either a similar extended 3C terminus (albeit with different amino acids), or a 3D protein with a 5-aa extension at its N-terminus, since there are two potentially scissile Q/G sites generated by the 15-nt insertion in 3C-183b. If both Q/G sites were proteolytically cleaved, the 5-aa insert would be deleted, leaving both 3C and 3D cleavage products with authentic termini; however, the uncleaved 3CD protein would contain the insertion (Fig. 2A). We believe it unlikely that an N-

terminally extended 3D would function as an active polymerase, since the N-terminus of 3D becomes buried in a pocket at the base of the fingers domain post-cleavage, causing changes in molecular flexibility at the active site (Campagnola et al., 2008;Marcotte et al., 2007). The scissile bonds utilized by these mutant viruses were not determined; however, the differences in plaque size observed after placement of different insertions in the same position might be related to differences in cleavage site or cleavage efficiency.

3D—Protein 3D is an RNA-dependent RNA polymerase (RdRP), the key subunit in the replication complex that catalyzes the synthesis of plus- and minus-strand viral RNAs. Given the importance of this enzyme to the unique replication mechanism of RNA viruses, several RdRPs have been subjected to intensive biochemical and molecular biological investigation, and the three-dimensional structures of multiple enzymes from different viruses have been solved by X-ray crystallography (see (Campagnola et al., 2008) and refs therein).

Genomes carrying insertions at either of two different locations in the 3D protein were isolated from plaque-forming viruses. Insertions were found after residue 53 (Fig. 3e), within a disordered segment of the protein chain in a region with sequence conserved among picornaviral 3D proteins but which has no apparent counterpart in other types of polymerase enzymes. Within the palm domain wherein lies the catalytic site, 5-aa insertions were identified after positions 362 and 364, in a short linker between two beta strands (Fig. 3e). None of these insertions apparently caused significant disruption in the overall fold of the protein. We did not find an insertion located in the surface-exposed loop (residues 257-263) between two helices of the 3D structure. In this loop, the corresponding region in the Coxsackie B3 virus 3D protein contains one amino acid more, and the rhinovirus 3D protein contains one amino acid less, than the polio protein, from which it had been concluded that some sequence/structure variation can be tolerated (Campagnola et al., 2008). However, of 11 isolates that were found to contain 5-aa inserts in 3D, only the two regions after residue 50 or near position 362 were isolated.

Relationship of insertion site locations to virus evolution

Although the locations of tolerated insertion sites identified in this study correlated with regions of apparent structural flexibility in the proteins, not all external surface loops or unstructured regions were found to accept the transposon-mediated insertions. Such domains may be structurally constrained due to their engagement in protein-protein or other inter-molecular interactions that would preclude interruption by insertions. These considerations prompted us to examine whether the identified insertion sites reflected regions of variability that occurred during virus evolution. If this were the case, identification of regions of dissimilarity (non-conservation) (Gorbalenya and Lauber, 2010) among the sequences of proteins from closely related viruses may serve to predict where insertions might be tolerated. To this end, we assessed the conservation of the non-structural proteins by calculating the average similarity of each protein in a protein alignment and then plotted positional conservation along the polyprotein alignment using a sliding window of 5 aa with 3-aa overlap (Fig. 4). We analyzed conservation at three different levels of virus evolutionary diversity that are defined as species (Fig. 4A) and supraspecies (Fig. 4B) and genus (Fig. 4C). PV belongs to the *Human enterovirus C* species formed by 14 serotypes (HEV-C level). This species is one of 4 human enterovirus species that we treated as a supraspecies cluster (HEV). They, together with 7 other species, form the enterovirus genus (Entero). The average similarity of each non-structural protein was calculated for each of these three levels (indicated by the dashed lines in Fig. 4 for each protein) and the positional conservation was then plotted along the polyprotein alignment.

Amino acid conservation varies not only among the individual viral proteins, but also locally within protein sub-domains, resulting in alternating peaks and valleys in the conservation plot of the P2-P3 part of the polyprotein alignment (Fig. 4). Sequences whose conservation fall below the protein's average (dashed lines) appear as downward spikes or valleys (similarity scores lower than that indicated by the dashed line), whereas regions that are more highly conserved than the average for the protein appear as spikes above the dashed line. There is good correlation between peak distributions in the three plots, although their precise positions and numbers vary. This variation is due to site- and branch-dependent evolution in the members contributing to the three datasets of increasing sequence diversity (Gorbalenya and Lauber, 2010) and unpublished data). Practically, access to all three rather than a single plot allows for minimizing an effect of sequence sampling on the identification of sites prone to mutation during natural evolution.

The insertion sites identified in this study are indicated on the conservation plots by gray vertical bars in Fig. 4. The locations that accommodated insertions in viable viruses resided at or very near valleys within sequences whose conservation fell below the protein's average (dashed lines) in one or more plots. Thus, insertions are tolerated at places that are prone to mutation in poliovirus and closely related viruses in the course of their natural evolution from common ancestors. We did not, however, find viable viruses with insertions located in many regions that demonstrate high degrees of sequence variability. This could be due to some non-randomness in the transposon insertion reaction, our failure to screen enough viruses to find all possible acceptable sites, and/or structural constraints on the specific sequences and their size that may be accommodated in a given, albeit poorly-conserved, region. These constraints could be host-dependent: selection for protein interactions with host factors that vary among different virus hosts may impose specificity on precisely those viral protein sequences that have diverged during evolution to enable infection of different hosts.

These observations show that insertions are often tolerated at places that are prone to mutation in PV and closely related viruses in the course of their natural evolution from common ancestors. However due to other factors that grossly modulate the outcome of insertion probing, this tendency alone is not sufficient to predict where markers, tags or other practical insertions will be tolerated.

Failure to identify viruses with insertions in 2C or the 3′NTR

Our results of sequencing genomes from viruses that grew after transfection of cells with RNAs containing transposon-mediated insertions in either the P2 or P3 region (see Fig. 1 for mutagenized segments) yielded no insertion in the 2C coding regions or in the 3′ NTR. Previous work had resulted in the isolation of two mutant polioviruses containing either a 4-aa insertion following residue 255 or a 6-aa insertion after residue 263, both in a region of 2C in between a SFIII helicase-like domain and a Zn –finger domain (Li and Baltimore, 1988). Although no 3-dimensional structure has been resolved for 2C to help assess the structural context of these previously identified insertions, we did see multiple sites with negative similarity scores near that region (Fig. 4), suggesting that sequence variability has occurred during evolution. We did not attempt to reconstruct potential transposon-mediated 15-nt insertions in that region of 2C to determine their viabilities. The fact that insertion sites identified in this study may not always correlate with those few reported previously supports the notion that there are insertion- and site-specific constraints.

We also found no insertions in the 3′NTR, despite its sequence having been included in the P3 segment mutagenesis screen (Fig. 1). A viable virus containing an 8-nt insertion was described previously (Sarnow et al., 1986), approximately 50 nts upstream of the poly(A) sequence, suggesting that at least some small insertions could be tolerated in this region.

Since such insertions would not have been useful for potentially tagging viral proteins, we did not pursue this analysis.

Placing useful tag sequences in identified insertion-tolerant sites

Useful tag sequences such as fluorescent proteins (FPs) or different epitope tags can be substituted for the 5-aa transposon-mediated insertions used to identify the insertion-tolerant sites. However, as both size and specific amino acid sequences can affect protein function and virus growth, selection of tags that can replace the transposon insertion in each of the identified sites remains highly empirical. For example, the sequences of a tetracysteine motif that binds a fluorescent ligand or the significantly larger FPs, DsRed or GFP (25-30 kDa), were accepted at either of the two insertion sites identified in PV protein 2A (~17 kDa), although growth of PV2A50-FP was more impaired than PV2A144-FP (Teterina et al., 2010). In protein 3A, a 9-aa HA epitope tag or a 12-aa tetracysteine motif sequence was well tolerated in the N-terminal region; however, viruses encoding a similar size FLAG epitope (8 aa) or c-myc epitope (10 aa) at the same position were recovered with significantly lower efficiency (Teterina et al., in preparation). Viruses carrying a HA-epitope tag (8 aa) insertion after residue 17 of protein 3B (Fig. 3c) displayed normal virus growth properties, while placement of the c-myc epitope at the same site significantly reduced virus growth, and a tetracysteine motif sequence (6 aa) at the same position rapidly deleted one amino acid. The latter motif was also excluded from protein 2B, and only viruses with deletion of most of this motif were recovered.

The biases against specific sequences in some but not all sites that readily accommodated the 5-aa insertions left by the transposon mutagenesis reaction are not predictable. It is highly likely that some sites that did not accept the transposon insertion might have accepted an insertion of a different sequence or other properties. As a result, use of a different transposon or introduction of a different sequence by another means might reveal additional insertion-tolerant sites in a given genome region. For this reason, it is not possible to consider a genome region “saturated” for insertion-tolerant sites, even if the region appears saturated in a particular transposon screen, unless it has been screened with multiple insertion sequence probes.

Conclusions

Using random transposon-mediated insertion mutagenesis, we have screened the non-structural protein coding region of the poliovirus genome and selected for viable viruses that contained 5-aa insertions. The locations of sites that tolerated these insertions were not distributed evenly throughout the viral non-capsid polyprotein sequence, but rather tended to cluster in about 9 regions. One or two regions that tolerated small insertions were identified in each viral non-capsid protein except for 2C, for which we found no insertions in plaque-forming viruses. The tolerated insertions were always located in external surface loops or unstructured regions of the three-dimensional protein folds; however, not all such loops or flexible peptide domains accepted insertions. Similarly, insertions were often tolerated at places that are prone to mutation in PV and closely related viruses in the course of their natural evolution from common ancestors. However, many more places of less-than-average conservation did not accept insertions, indicating that this tendency alone is not sufficient to predict where markers, tags or other practical insertions will be accepted for purposes of imaging or biochemical analyses. Thus, structural considerations and analysis of evolutionary conservation may be useful to decrease the genetic space to be explored by experimental mutagenesis to identify potential sites for protein insertion. This combination of knowledge-based and random probing approaches is likely to provide the most efficient strategy for tagging picornaviral proteins in future efforts.

Materials and Methods

Generation of 15-nt insertion PV plasmid libraries

Plasmid pPVM containing infectious full-length cDNA of PV1 with several unique restriction sites introduced in the P2 region was a generous gift from A. Paul and E. Wimmer, Stony Brook School of Medicine, NY. Plasmid pXpA (Herold and Andino, 2000) containing full-length cDNA of PV1 was a gift from R. Andino, UCSF. Plasmid pXpA-X was created from pXpA by introducing a single nucleotide change to create a unique XhoI site at position 4242 in pXpA. The RNA genome with this point mutation showed infectivity of ~10^6 pfu/μg, similar to that of wild-type PV RNA. In order to concentrate transposon insertions in specific subregions of the viral polyprotein, we prepared several subclones on which we performed mutagenesis reactions separately. A subclone pPVsub-P2 was engineered by cloning the NheI – Xho I fragment from pPVM (nts 2470 - 4433 in PVM cDNA) into pSP-72 vector (Promega) digested with XbaI and XhoI. A subclone pPVsub-P2/3 was generated by substituting the NheI – HindIII fragment (nts 2470 - 6056 in PV cDNA) from pXpA-X for the region between the XbaI and HindIII sites of pGEM-3Zf(+) (Promega). A subclone pPVsub-P3 was generated by insertion of the BglII – EcoRI fragment (nt 5601 through the poly(A) sequence) from pXpA between the BglII and EcoRI sites of the pSP-72 vector. Plasmids pPVsub-P2 and pPVsub-P2/3 were subjected to Tn7 transposon-mediated mutagenesis *in vitro* (GPS-LS Linker Scanning System, New England Biolabs). This system allows essentially random and single hit insertion of a Tn7-based transprimer into a plasmid of interest (Biery et al., 2000). A pGPS5 transprimer donor plasmid and kanamycin selection were used for mutagenesis of pPVsub-P2, whereas pGPS4 and chloramphenicol selection were used for pSub-P2/P3. Standard DNA transposition reactions (20 μl) contained 100 ng of plasmid DNA and were performed according to the manufacturer's protocol. A total of 4.8×10^4 or 5.1×10^5 bacterial colonies were obtained from pPVsub-P2 or pPVsub-P2/P3 reactions, respectively. Mutant plasmids were isolated from the pooled bacterial libraries. To remove the transposon DNA fragment, 7 μg of the pooled mutant plasmids were subjected to PmeI digestion to remove the selective antibiotic resistance gene, recircularized by ligation and used for bacterial transformation. This resulted in pPVsub-P2-μ and pPVSsub-P2/P3-μ libraries of mutants with randomly inserted 15-nt sequence 5′TGTTTAAAGANNNNN-3′, where N represents 5 nts duplicated from the adjacent upstream target DNA. As insertion of this sequence in one reading frame generates a termination codon, only insertions in two frames might be found in viable viruses. These secondary libraries were also represented by ~8×10^4 and 5×10^5 clones, respectively. Mutant plasmid DNAs were isolated from the pooled secondary libraries. Plasmid DNA from the pPVsub-P2-μ library was cut with SnaBI and XhoI and the 1479 bp fragment with random insertions was ligated to pPVM cut with the same enzymes. Similarly, plasmid DNA from pPVsub-P2/P3 was cut with XhoI and BglII and the 1359 bp fragment with random insertions was ligated to pXpA-X vector cut with the same enzymes. Ligated DNA was used for transformation to produce final libraries pPVM-μP2 or pXpA-μP2/P3 respectively. The final libraries generated after transfer of mutagenized sub-clones into plasmids containing full-length PV cDNA were of the same complexity as the primary sub-clone mutagenized libraries (5×10^4 and 5.2×10^5 clones, respectively).

Plasmid pPVsub-P3 was subjected to Mu transposon-mediated mutagenesis *in vitro* (MGS kit, Finnzymes). A total of 5.8×10^5 clones were obtained and the mutant plasmids were isolated from the pooled bacterial library as described above. The transposon DNA fragment was removed by digestion of 5 μg of the pooled mutant plasmids with NotI restriction enzyme. Plasmid DNA was recircularized and used to transform bacterial cells to produce a library with random insertions of the 15-nt sequence 5′-TGCGGCCGCANNNNN-3′ in pPVsub-P3 plasmid. This sequence does not encode termination codons in any frame. This secondary library was represented by 5.6×10^5 clones. Plasmid DNA isolated from the

pooled library was digested with BglII and EcoRI and the 1921 bp fragment with random insertions was used to substitute for the corresponding fragment in pXpA. Transformation of bacterial cells with the ligated product yielded the pXpA-μP3 library, represented by more than 5.5×10^5 clones..

All libraries were created from independent transposon reactions at least two times.

Selection of viable viruses encoding 15-nt insertions

A total of 5 μg of pPVM-μP2, pXpA-μP2/P3 or pXpA-μP3 library DNAs were linearized by digestion with EcoRI downstream of the PV poly(A) sequence and used for *in vitro* transcription with T7 RNA polymerase. HeLa cell monolayers in six-well plates were transfected in triplicate with serial dilutions of 3 μg RNA transcripts using TransIt-mRNA transfection kit (Mirus Bio) as described previously (Teterina et al., 2010). Two to 3 hours after transfection the medium was replaced with DMEM containing 0.4% agarose and plates were incubated at 37°C for 2, 3 or 4 days in order to allow recovery of viruses with different growth properties. Individual viral plaques were selected and virus from each plaque was used to infect fresh HeLa monolayers in 35 mm plates to produce virus stocks. For each library, ~50 independent virus stocks were analyzed.

Determination of insertion sites

Viral RNA isolation and reverse transcription were performed as described previously (Teterina et al., 2010). The cDNAs were used as templates for PCR amplification using Phusion High-Fidelity PCR kit (Finnzymes) and PV-specific primers to produce fragment encoding the entire P2 and P3 regions. Purified 4060 bp PCR product was used for sequence analysis using 16 PV-specific primers to resolve all non-structural protein coding sequences and the 3′ untranslated region. Some viruses lacking any insertions were also recovered, presumably due to contamination during the library sub-cloning steps.

Datasets for analysis of variation during virus evolution

Three datasets representing different evolutionary scales were analyzed. The first dataset includes 14 virus sequences all belonging to the picornavirus species *Human enterovirus C* - one representative for each of the eleven Coxsackie A virus plus three poliovirus serotypes were taken. The second dataset includes one representative for each of the four Human enterovirus species. The third dataset includes one representative for each of the eleven species of the genus *Enterovirus*. For each dataset a multiple amino acid alignment by Muscle (Edgar, 2004) covering the non-structural proteins (P2+P3 region) was extracted from a polyprotein alignment of all available enterovirus sequences that was prepared using Viralis software platform ((Gorbalenya et al.); Gorbalenya, unpublished). Sequences with the following Genbank/RefSeq accession numbers were used: HEV-C: AF499635-43, AF546702, D90457, M12197, K01392, NC_002058 (PVM); HEV: AF119796, AY896766, NC_001430, NC_002058; Entero: NC_009996, DQ473512, NC_001490, NC_003988, NC_001859, NC_004441, AF119796, NC_010415, AY896766, NC_001430, NC_002058.

Protein conservation

A measure of protein similarity was used to estimate conservation along the non-structural part of the polyprotein. It is based on the Blosum62 amino acid substitution matrix (Henikoff and Henikoff, 1994) and was calculated utilizing the R package Bio3D (Grant et al., 2006). The similarity measure was compiled for the three datasets representing different evolutionary scales and plotted along the alignments.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, NIAID, and by the Netherlands Bioinformatics Center (NBIC Biorange SP3.2.2). We thank Igor Sidorov, Alexander Kravchenko and Dmitry Samborskiy for administering the Viralis software platform.

References

Agirre A, Barco A, Carrasco L, Nieva JL. Viroporin-mediated membrane permeabilization. Pore formation by nonstructural poliovirus 2B protein. J Biol Chem. 2002; 277(43):40434–41. [PubMed: 12183456]

Aldabe R, Carrasco L. Induction of membrane proliferation by poliovirus proteins 2C and 2BC. Biochem. Biophys. Res. Commun. 1995; 206(1):64–76. [PubMed: 7818552]

Allaire M, Chernaia MM, Malcolm BA, James MN. Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteinases. Nature. 1994; 369(6475):72–6. [PubMed: 8164744]

Arumugaswami V, Remenyi R, Kanagavel V, Sue EY, Ngoc Ho T, Liu C, Fontanes V, Dasgupta A, Sun R. High-resolution functional profiling of hepatitis C virus genome. PLoS Pathog. 2008; 4(10):e1000182. [PubMed: 18927624]

Atasheva S, Gorchakov R, English R, Frolov I, Frolova E. Development of Sindbis viruses encoding nsP2/GFP chimeric proteins and their application for studying nsP2 functioning. J Virol. 2007; 81(10):5046–57. [PubMed: 17329335]

Baxter NJ, Roetzer A, Liebig HD, Sedelnikova SE, Hounslow AM, Skern T, Waltho JP. Structure and dynamics of coxsackievirus B4 2A proteinase, an enyzme involved in the etiology of heart disease. J Virol. 2006; 80(3):1451–62. [PubMed: 16415022]

Bernstein HD, Baltimore D. Poliovirus mutant that contains a cold-sensitive defect in viral RNA synthesis. J. Virol. 1988; 62:2922–2928. [PubMed: 2839711]

Biery MC, Stewart FJ, Stellwagen AE, Raleigh EA, Craig NL. A simple in vitro Tn7-based transposition system with low target site selectivity for genome and gene analysis. Nucleic Acids Res. 2000; 28(5):1067–77. [PubMed: 10666445]

Birtley JR, Knox SR, Jaulent AM, Brick P, Leatherbarrow RJ, Curry S. Crystal structure of foot-and-mouth disease virus 3C protease. New insights into catalytic mechanism and cleavage specificity. J Biol Chem. 2005; 280(12):11520–7. [PubMed: 15654079]

Campagnola G, Weygandt M, Scoggin K, Peersen O. Crystal structure of coxsackievirus B3 3Dpol highlights the functional importance of residue 5 in picornavirus polymerases. J Virol. 2008; 82(19):9458–64. [PubMed: 18632862]

Cho MW, Teterina N, Egger D, Bienz K, Ehrenfeld E. Membrane rearrangement and vesicle induction by recombinant poliovirus 2C and 2BC in human cells. Virology. 1994; 202:129–145. [PubMed: 8009827]

Cornell CT, Kiosses WB, Harkins S, Whitton JL. Inhibition of protein trafficking by coxsackievirus b3: multiple viral proteins target a single organelle. J Virol. 2006; 80(13):6637–47. [PubMed: 16775351]

de Jong AS, de Mattia F, Van Dommelen MM, Lanke K, Melchers WJ, Willems PH, van Kuppeveld FJ. Functional analysis of picornavirus 2B proteins: effects on calcium homeostasis and intracellular protein trafficking. J Virol. 2008; 82(7):3782–90. [PubMed: 18216106]

Doedens JR, Kirkegaard K. Inhibition of cellular protein secretion by poliovirus proteins 2B and 3A. EMBO J. 1995; 14(5):894–907. [PubMed: 7889939]

Draghici HK, Pilot R, Thiel H, Varrelmann M. Functional mapping of PVX RNA-dependent RNA-replicase using pentapeptide scanning mutagenesis-Identification of regions essential for replication and subgenomic RNA amplification. Virus Res. 2009; 143(1):114–24. [PubMed: 19463728]

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–7. [PubMed: 15034147]

Egger D, Teterina N, Ehrenfeld E, Bienz K. Formation of the poliovirus replication complex requires coupled viral translation, vesicle production, and viral RNA synthesis. J Virol. 2000; 74(14):6570–6580. [PubMed: 10864671]

Frolov I, Garmashova N, Atasheva S, Frolova EI. Random insertion mutagenesis of sindbis virus nonstructural protein 2 and selection of variants incapable of downregulating cellular transcription. J Virol. 2009; 83(18):9031–44. [PubMed: 19570872]

Gorbalenya AE, Donchenko AP, Blinov VM, Koonin EV. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. FEBS Lett. 1989; 243(2):103–14. [PubMed: 2645167]

Gorbalenya, AE.; Lauber, C. Origin and evolution of the Picornaviridae proteome. In: Ehrenfeld, E.; Domingo, E.; Roos, R., editors. The Picornaviruses. ASM Press; Washington, DC: 2010.

Gorbalenya AE, Lieutaud P, Harris MR, Coutard B, Canard B, Kleywegt GJ, Kravchenko AA, Samborskiy DV, Sidorov IA, Leontovich AM, Jones TA. Practical application of bioinformatics by the multidisciplinary VIZIER consortium. Antiviral Res. 87(2):95–110. [PubMed: 20153379]

Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 2006; 22(21):2695–6. [PubMed: 16940322]

Henikoff S, Henikoff JG. Position-based sequence weights. J Mol Biol. 1994; 243(4):574–8. [PubMed: 7966282]

Herold J, Andino R. Poliovirus requires a precise 5′ end for efficient positive-strand RNA synthesis. J Virol. 2000; 74(14):6394–6400. [PubMed: 10864650]

Lama J, Paul AV, Harris KS, Wimmer E. Properties of purified recombinant poliovirus protein 3AB as substrate for viral proteinases and as co-factor for RNA polymerase 3Dpol. J Biol Chem. 1994; 269(1):66–70. [PubMed: 8276867]

Li J-P, Baltimore D. Isolation of poliovirus 2C mutants defective in viral RNA synthesis. J. Virol. 1988; 62(11):4016–4021. [PubMed: 2845120]

Liu S, Ansari IH, Das SC, Pattnaik AK. Insertion and deletion analyses identify regions of non-structural protein 5A of Hepatitis C virus that are dispensable for viral genome replication. J Gen Virol. 2006; 87(Pt 2):323–7. [PubMed: 16432018]

Marcotte LL, Wass AB, Gohara DW, Pathak HB, Arnold JJ, Filman DJ, Cameron CE, Hogle JM. Crystal structure of poliovirus 3CD protein: virally encoded protease and precursor to the RNA-dependent RNA polymerase. J Virol. 2007; 81(7):3583–96. [PubMed: 17251299]

Matthews DA, Smith WW, Ferre RA, Condon B, Budahazi G, Sisson W, Villafranca JE, Janson CA, McElroy HE, Gribskov CL, et al. Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site, and means for cleaving precursor polyprotein. Cell. 1994; 77(5):761–71. [PubMed: 7515772]

Molla A, Harris KS, Paul AV, Shin SH, Mugavero J, Wimmer E. Stimulation of poliovirus proteinase 3Cpro-related proteolysis by the genome-linked protein VPg and its precursor 3AB. J Biol Chem. 1994; 269(43):27015–20. [PubMed: 7929442]

Moradpour D, Evans MJ, Gosert R, Yuan Z, Blum HE, Goff SP, Lindenbach BD, Rice CM. Insertion of green fluorescent protein into nonstructural protein 5A allows direct visualization of functional hepatitis C virus replication complexes. J Virol. 2004; 78(14):7400–9. [PubMed: 15220413]

Mosimann SC, Cherney MM, Sia S, Plotch S, James MN. Refined X-ray crystallographic structure of the poliovirus 3C gene product. J Mol Biol. 1997; 273(5):1032–47. [PubMed: 9367789]

Paul AV, van Boom JH, Fillipov D, Wimmer E. Protein-primed RNA synthesis by purified poliovirus RNA polymerase. Nature. 1998; 393:280–284. [PubMed: 9607767]

Sarnow P, Bernstein HD, Baltimore D. A poliovirus temperature-sensitive RNA synthesis mutant located in a noncoding region of the genome. Proc. Natl. Acad. Sci. USA. 1986; 83:571–575. [PubMed: 3003739]

Schein CH, Oezguen N, Volk DE, Garimella R, Paul A, Braun W. NMR structure of the viral peptide linked to the genome (VPg) of poliovirus. Peptides. 2006; 27(7):1676–84. [PubMed: 16540201]

Seipelt J, Guarne A, Bergmann E, James M, Sommergruber W, Fita I, Skern T. The structures of picornaviral proteinases. Virus Res. 1999; 62(2):159–68. [PubMed: 10507325]

Strauss DM, Glustrom LW, Wuttke DS. Towards an understanding of the poliovirus replication complex: the solution structure of the soluble domain of the poliovirus 3A protein. J Mol Biol. 2003; 330(2):225–34. [PubMed: 12823963]

Suhy DA, Giddings TH Jr. Kirkegaard K. Remodeling the endoplasmic reticulum by poliovirus infection and by individual viral proteins: an autophagy-like origin for virus-induced vesicles. J Virol. 2000; 74(19):8953–65. [PubMed: 10982339]

Takegami T, Kuhn RJ, Anderson CW, Wimmer E. Membrane-dependent uridylylation of the genome-linked protein VPg of poliovirus. Proc. Natl. Acad. Sci. USA. 1983; 80:7447–7451. [PubMed: 6324172]

Taylor MP, Kirkegaard K. Modification of cellular autophagy protein LC3 by poliovirus. J Virol. 2007; 81(22):12543–53. [PubMed: 17804493]

Teterina NL, Levenson EA, Ehrenfeld E. Viable polioviruses that encode 2A proteins with fluorescent protein tags. J Virol. 2010; 84(3):1477–88. [PubMed: 19939919]

Thompson AA, Peersen OB. Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. Embo J. 2004; 23(17):3462–71. [PubMed: 15306852]

Wessels E, Duijsings D, Notebaart RA, Melchers WJ, van Kuppeveld FJ. A proline-rich region in the coxsackievirus 3A protein is required for the protein to inhibit endoplasmic reticulum-to-golgi transport. J Virol. 2005; 79(8):5163–73. [PubMed: 15795300]

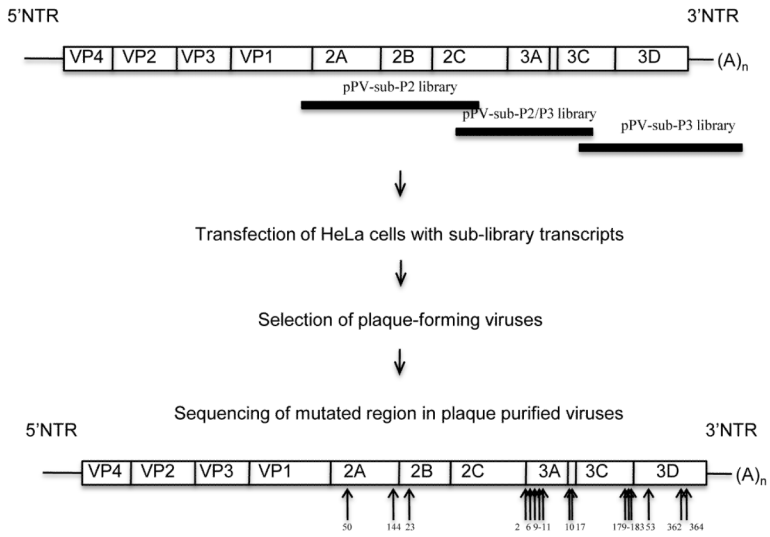

Figure 1.

Schematic presentation of the steps involved in isolation of PVs with 15-nt insertions. Insertion mutant libraries were generated in three PV sublcones shown with the positions of the corresponding fragments within the PV genome. Each mutant full-length plasmid library was transcribed *in vitro* and the pooled transcripts were used to transfect HeLa cells for genetic selection. Viable viruses were isolated from individual plaques and their RNA genomes were sequenced to determine the locations of the transposon-mediated insertions. The locations of the 15-nt insertions identified from independent virus isolates are indicated by the arrows at the bottom. The numbers indicate the amino acid residue of the corresponding PV protein after which insertion occurred.

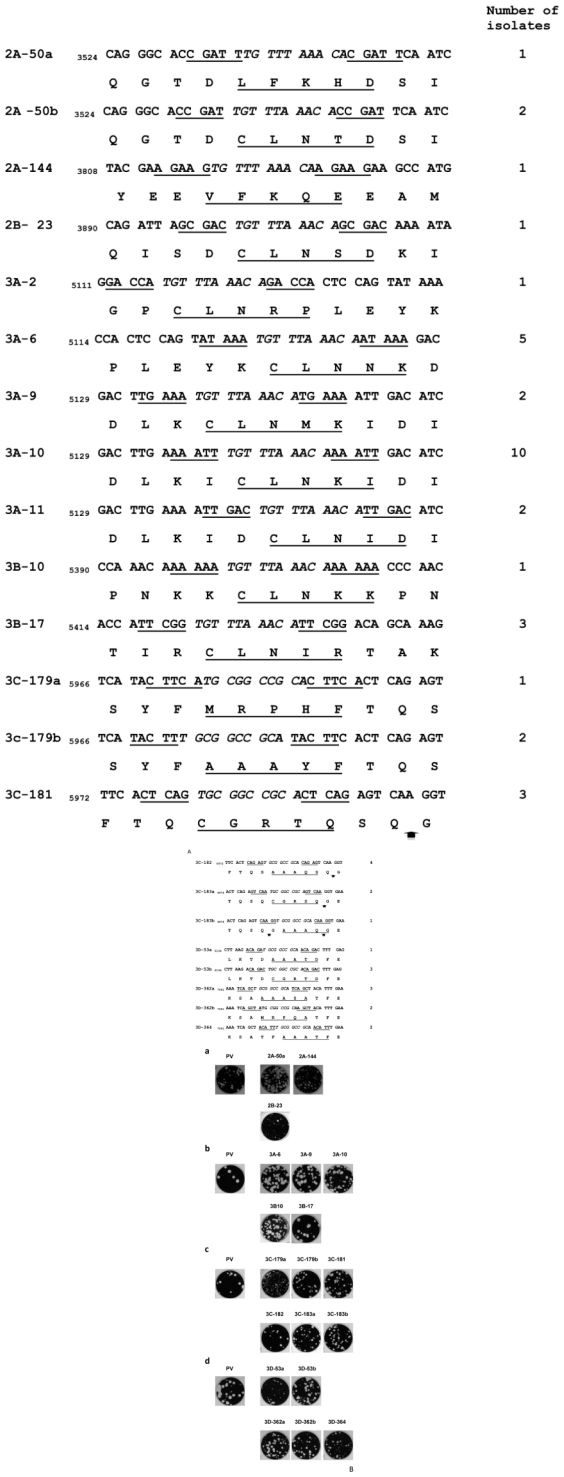

Figure 2.

Plaque-forming viruses containing 15-nt insertions. The insertions and corresponding viruses are designated by the name of the protein containing the insertion and the number of the amino acid residue after which 5 additional amino acids were inserted. Insertions after the same amino acid that were encoded by15 nts inserted in different reading frames of the

same or adjacent codons are named as “a” and “b” respectively. (A) Nucleotide and amino acid sequences in viruses with 15-nt insertions. For each insertion, the nucleotide (upper line) and the predicted amino acid (lower line) sequences are shown. The nucleotide numbers at the start of each sequence refer to the nucleotide numbers in the Mahoney strain of PV type 1 genome. The nucleotide sequences introduced by transposon insertion are in bold italics. In each case, five nts that were duplicated from the viral genome during transposon insertion are underlined, as are the five amino acid sequences introduced by the insertion. For the C-terminal insertions in protein 3C, the (potential) 3C-3D cleavage site(s) are indicated by arrowheads. The far right column indicates the number of independent isolates of each mutant that was recovered. (B) Plaque phenotypes of viruses with 15-nt insertions. Plaques were stained 48 h after infection with the indicated viruses. Virus plaque phenotypes were analyzed in four separate experiments (a – d), and each included a parental wild-type virus control for comparison. For panel (a), the parental wild-type virus was from pPVM, which forms slightly smaller plaques than that from pXpA, used as the parent for panels (b – d).

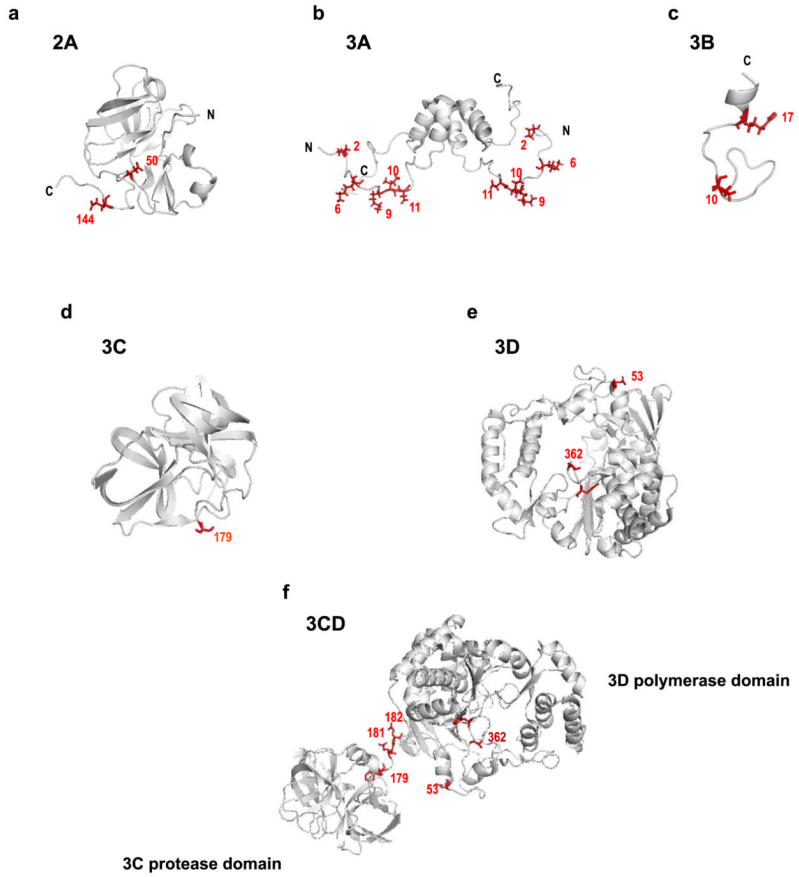

Figure 3.

The structures of PV proteins displaying sites of the insertions of 5 amino acids. Amino acid residues preceding the 5-aa insertions are indicated in red. Ribbon diagrams are shown of (a) a putative model of PV 2A protein modeled using the I-TASSER server was used to predict the PV 2A structure from the NMR-derived magnetic resonance structure of Coxsackievirus B4 2A protein (PDB ID: 1Z8R) (Baxter et al., 2006); (b) the predicted homodimer formed by the N-terminal soluble portion (60 amino acids) of PV protein 3A (PDB ID: 1NG7) (Strauss et al., 2003); (c) peptide 3B (PDB ID: 2BBL) (Schein et al., 2006); (d) protein 3C (PDB ID: 1L1N) lacking amino acids 180-182 at the C-terminus) (Mosimann et al., 1997); (e) protein 3D (PDB ID:1RA6 (Thompson and Peersen, 2004); and (f) protein 3CD (PDB ID: 2IJD) (Marcotte et al., 2007). The structure analysis and graphics generation were done using PyMOL Viewer .

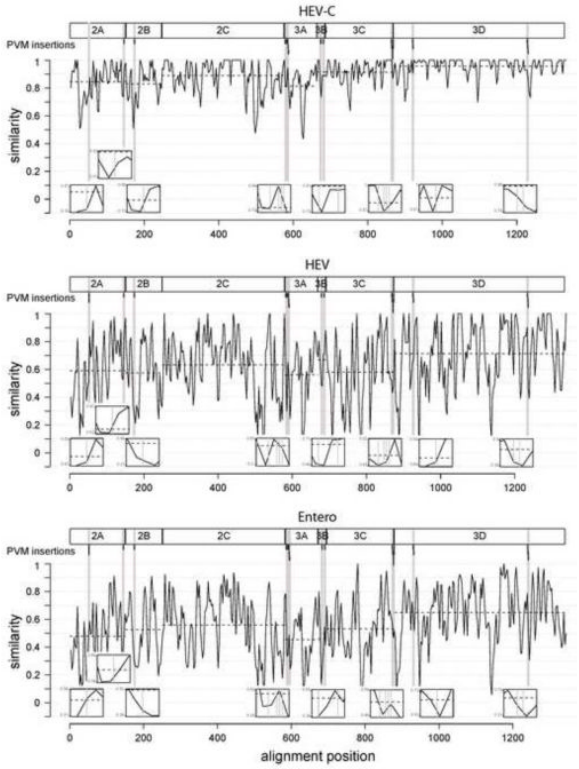

Figure 4.

Conservation of the non-structural polyprotein part and location of PVM insertions. A plot of the conservation along the non-structural part of the polyprotein alignment is shown for three evolutionary scales which include eleven Coxsackie A virus plus three poliovirus serotypes forming the species *Human enterovirus C* (A), one representative for each of the four Human enterovirus species (B) and one representative for each of the eleven species of the genus *Enterovirus* (C). The normalized similarity measure was compiled under the Blosum62 substitution matrix and smoothed using a sliding window with a size of 5 aa and an overlap of 3 aa positions. The mean similarity of single protein alignments is indicated by dashed horizontal lines and their positions are shown using rectangles and names on top. Below, the positions of PVM insertions resulting in viable virus are indicated by vertical bars. Neighboring insertions are grouped together and highlighted with grey background. Each intersection of an insertion with a similarity distribution curve is magnified in a respective inlet that zooms in on the intersection.