Recurrent Viral Capture of a Host-Derived Immune Antagonist

Stephen A. Goldstein and Nels C. Elde

Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA
Howard Hughes Medical Institute, 4000 Jones Bridge Rd, Chevy Chase, MD 20815, USA

Co-corresponding authors – sgoldstein@genetics.utah.edu, nelde@genetics.utah.edu

Abstract

Phosphodiesterases (PDEs) encoded by nidoviruses and rotaviruses (RVs) promote viral replication and pathogenesis. Putatively acquired by horizontal transfer of cellular PDE ancestor genes, viral PDEs inhibit the OAS-RNase L antiviral pathway, a key effector component of the innate immune response to dsRNA. Although the functional significance of these proteins is well-characterized, the evolutionary history of these proteins is less clear. We used phylogenetic approaches to identify at least five independent PDE acquisition events in viral evolutionary history. In addition, we assess that these PDE-encoding genes were horizontally transferred between coronavirus genera. Three viral clades within *Nidovirales*, merbecoviruses (MERS-CoV), embecoviruses (OC43), and toroviruses encode independently acquired PDEs, and a clade of rodent alphacoronaviruses acquired the embecovirus PDE via a relatively recent horizontal transfer event. Among rotaviruses, the Rotavirus A PDE was acquired independently from Rotavirus B and G PDEs, which share a common ancestor. Conserved motif analysis links all viral PDEs to a common ancestor among the mammalian AKAP7 proteins despite overall low levels of sequence conservation across these proteins. Additionally, we used ancestral sequence reconstruction and structural modeling to reveal that sequence and structural divergence are not well-correlated among these proteins. Specifically, merbecovirus PDEs are as structurally divergent from the ancestral protein and the solved human AKAP7 PDE structure as they are from each other. In contrast, comparisons of Rotavirus B and G PDEs reveal virtually unchanged structures despite evidence for loss of function in one, suggesting functionally significant changes that lie outside conserved catalytic sites. These findings highlight the complex and volatile evolutionary history of viral PDEs and provide a new framework to facilitate future studies on structural and sequence determinants of PDE function.

Introduction

Horizontal gene transfer (HGT) is a major force in historical and ongoing evolution among diverse viruses and hosts. Within the virosphere, the capture and re-purposing of host genes has shaped evolutionary history spanning from the origins of viruses themselves (1) to more specialized interactions with host immune pathways. Viruses in distinct unrelated orders, Nidoviruses and Rotaviruses, have independently captured host mRNAs encoding

eukaryotic LigT-like 2H-phosphoesterases, specifically the mammalian AKAP7 central domain phosphodiesterase (PDE) (2). PDEs encoded by betacoronaviruses, toroviruses, and group A rotaviruses have been shown to function as potent antagonists of the OAS-RNase L pathway with important roles *in vitro* and *in vivo* promoting viral replication and pathogenesis (3–8). In contrast to the extensive research on PDE function, the origin and evolutionary history of viral PDEs remain obscure. Based on structural similarity the PDEs encoded by mouse hepatitis virus (9) and group A rotaviruses (10,11) have been linked to the PDE domain of mammalian AKAP7 proteins. Previously predicted structures for PDEs encoded by MERS-CoV (6), toroviruses (5) and group B and G rotaviruses (11) exhibit an AKAP7 PDE-like fold and retain catalytic histidine (HxS/T) motifs.

The interface between viral PDEs and the OAS-RNase L pathway is an ideal opportunity to study evolutionary pressures on the acquisition, fixation, and evolution of host-derived viral genes. The restriction of rotavirus (8) and coronavirus (3,7,12) replication in the absence of an active PDE along with the presence of these genes across diverse viruses provide strong evidence for the fitness advantage conferred by PDE acquisition. The capacity for viruses to accommodate increased genome size is a key determinant of their ability to evolve via HGT, though PDEs may act as functional protein sub-domains of as few as ~115 amino acids, so size may not always be a major limiting factor. Considered broadly, the genomic architecture of viral PDEs is of two types. PDEs encoded by rodent alphacoronaviruses (13), embecoviruses (5), and merbecoviruses (6) are ~240-300 amino acids long, translated off of a discrete viral mRNA, presumably exist in the cell as autonomous proteins, and exhibit high amino acid sequence divergence relative to conserved regions of the replicase gene . In contrast, the rotavirus (4) and torovirus (5) PDEs are C–terminal subdomains of larger proteins - rotavirus VP3 or the polyprotein1a of toroviruses, and exhibit similar divergence as their parental polyprotein.

In this study we aimed to elucidate the evolutionary history of viral PDEs. Despite similarity in function and predicted structure with each other and mammalian AKAP7, the possibility of a single horizontal transfer event giving rise to PDEs across nidovirus and rotavirus diversity can be excluded based on our analysis. We used a combination of phylogenetics, conserved motif analysis, and AlphaFold-based structural predictions to characterize the origin of viral PDEs, their diversification following horizontal transfer, and the relationship of divergent viral PDEs to each other and mammalian AKAP7. These findings are a profound advance in our understanding of the history of these critical immune antagonists and provide a detailed characterization of how host-derived viral genes originate and evolve in new, viral contexts.

Results

Viral PDEs derive from several AKAP-like horizontal transfer events

To begin elucidating the evolutionary relationships among viral PDEs we conducted phylogenetic analyses of their amino acid sequences. We aligned 173 viral PDE amino acid

sequences (18 torovirus, 10 alphacoronavirus, 55 *Embecovirus*, 28 *Merbecovirus*, 8 rotavirus G, 20 rotavirus B, 35 rotavirus A) using MAFFT which, due to high sequence divergence, involved intensive manual refinement. We then constructed a maximum-likelihood phylogenetic tree, which we chose not to root due to the lack of an single feasible outgroup or common ancestor (**Figure 1A**). Likely due to high sequence divergence, the exact position of only some branches could be inferred with high bootstrap support, but the topology is sufficiently robust to reveal critical elements of PDE evolutionary history. The PDE tree contains five major branches likely corresponding to five acquisition events; rotavirus A (RVA) VP3, the rotavirus B (RVB) and rotavirus G (RVG) VP3 single common ancestor, toroviruses, embecoviruses, and merbecoviruses. The latter three, despite all being with the order *Nidovirales*, were expected to result from independent events due to being non-syntenic, high divergence, and lack of a recent common viral ancestor (**Figure 1B**). The rodent alphacoronavirus PDE exhibits high sequence identity with the *Embecovirus* PDE and is syntenic. Additionally, *Embecovirus* PDEs exhibit higher within-clade diversity, suggesting an older origin and providing support for an earlier hypothesis that the alphacoronavirus PDE was acquired via recombination from embecoviruses (13), but in contrast to other proposals that they were acquired independently (**Figure 1A**) (14) .

Despite being syntenic C-terminal additions to the rotavirus VP3 protein, phylogenetic analysis produced extremely long branch lengths between PDEs encoded by rotavirus RVA and those encoded by RVB and RVG, which cluster together on the ML tree. This tree topology suggests at least two horizontal transfer events, but leaves some ambiguity as to whether RVB and RVG acquired a similar PDE independently or whether a single RVB/G common ancestor experienced a single acquisition. To clarify this we constructed a maximum-likelihood tree of 137 rotavirus VP3 genes representing all known rotavirus genotypes (RVA, RVB, RVC, RVG, RVH, RVI, RVJ) with the PDE domain trimmed from the RVA, RVB, and RVG VP3s (**Figure 2C**). The topology of this tree indicated that the RVB and RVG VP3s derive from a recent common ancestor, suggesting a parsimonious scenario in which that ancestor acquired a PDE which diversified at the same rate as the rest of the gene. The RVA VP3 is extremely divergent from the RVB/G VP3, mirroring their relationship in the PDE phylogeny. Despite their synteny, the similarly distant relationship between the trimmed VP3s and the PDEs supports two horizontal transfer events, not one.

Previous work had identified mammalian AKAP7 central domain (CD) PDEs as structural and functional homologs of viral PDEs (15,16), but identifying conserved sequence to confidently link the viral and cellular proteins has been elusive. We used MAFFT to align human AKAP7 (*Hs*AKAP7) and representative viral PDEs and manually scanned for conserved motifs to anchor the alignment. In addition to the conserved catalytic motifs, we identified five other motifs present in *Hs*AKAP7 and at least some viral PDEs, though with some degenerate positions (**Figure 1D**). We did not find any similar sequence conservation between viral PDEs and other eukaryotic 2H-PDEs such as USB1, CGI-18, or LENG9.

Given the variable within-clade branch lengths, we wanted to measure the divergence of viral PDEs relative to other regions of the genome. Coronavirus PDEs were significantly more divergent between viruses than the RdRp encoding-region of the viral polyprotein, whereas PDEs encoded as sub-domains of larger proteins such as the torovirus and rotavirus PDEs seemingly evolve at similar rates to the RdRp or N-terminal part of VP3, respectively (**Figure S1**).

Finally, a PDE encoded by Orf4 of a shrew alphacoronavirus was recently reported (14). On a phylogenetic tree this protein is positioned alone on the end of a long branch (**Figure S2A**), but with very support due to the obvious undersampling. Its phylogenetic positioning remains uncertain. Several of the key motifs, including the second catalytic motif, are degraded, indicating that this protein is almost certainly non-functional (**Figure S2B**).

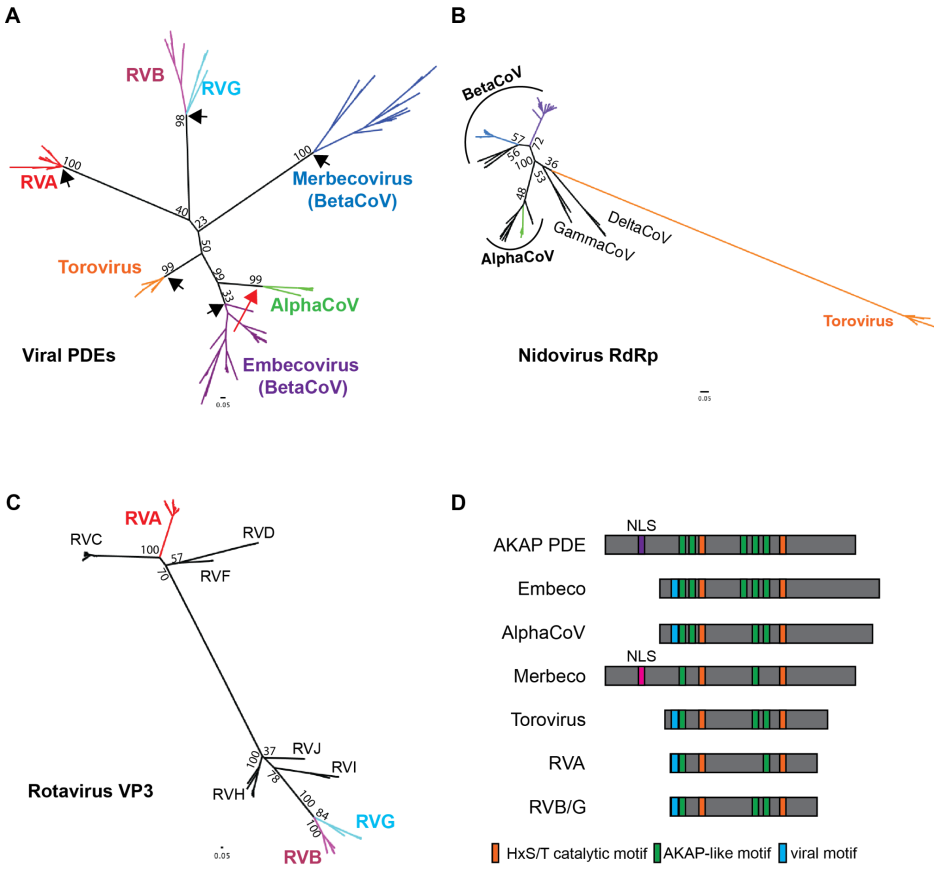

Figure 1. Phylogenetic and sequence analysis of viral PDEs reveals multiple independent acquisitions and conservation of ancestral motifs A) Unrooted maximum-likelihood (ML) Tree of viral phosphodiesterase amino acid sequences showing five PDE acquisitions from a cellular ancestor (black arrows) and one acquisition via recombination (red arrow) B) Unrooted ML tree of coronavirus and torovirus full length RdRp amino acid sequences C) Unrooted ML tree of rotavirus VP3 amino acid sequences, with the C-terminal domain PDEs trimmed from the RVA, RVB, and RVG sequences D) Schematic showing the conservation or absence of AKAP-like motifs in viral phosphodiesterases.

Ancestral Sequence Reconstruction of Nidovirus PDEs

To further elucidate the evolutionary history of Nidovirus PDEs we decided to conduct ancestral sequence reconstruction coupled with downstream analysis at major nodes. We primarily focused on *Embecovirus* NS2 and *Merbecovirus* NS4b proteins given the higher degree of within-clade sequence diversity among these proteins which likely indicates a combination of more thorough sampling and an older acquisition. We aligned 55 *Embecovirus* NS2 sequences with the *Hs*AKAP7 PDE domain. We then constructed a maximum-likelihood tree of *Embecovirus* NS2 protein sequence with an *Hs*AKAP7 outgroup rooting and used this tree as the basis for ancestral reconstruction in FastML (**Figure 2A**). We aligned the *Hs*AKAP7 and ancestral *Embecovirus* NS2 sequences from major nodes and calculated amino acid identity. The deepest ancestral node exhibited substantially higher amino acid identity to *Hs*AKAP7 than shallower nodes or observed extant sequences.

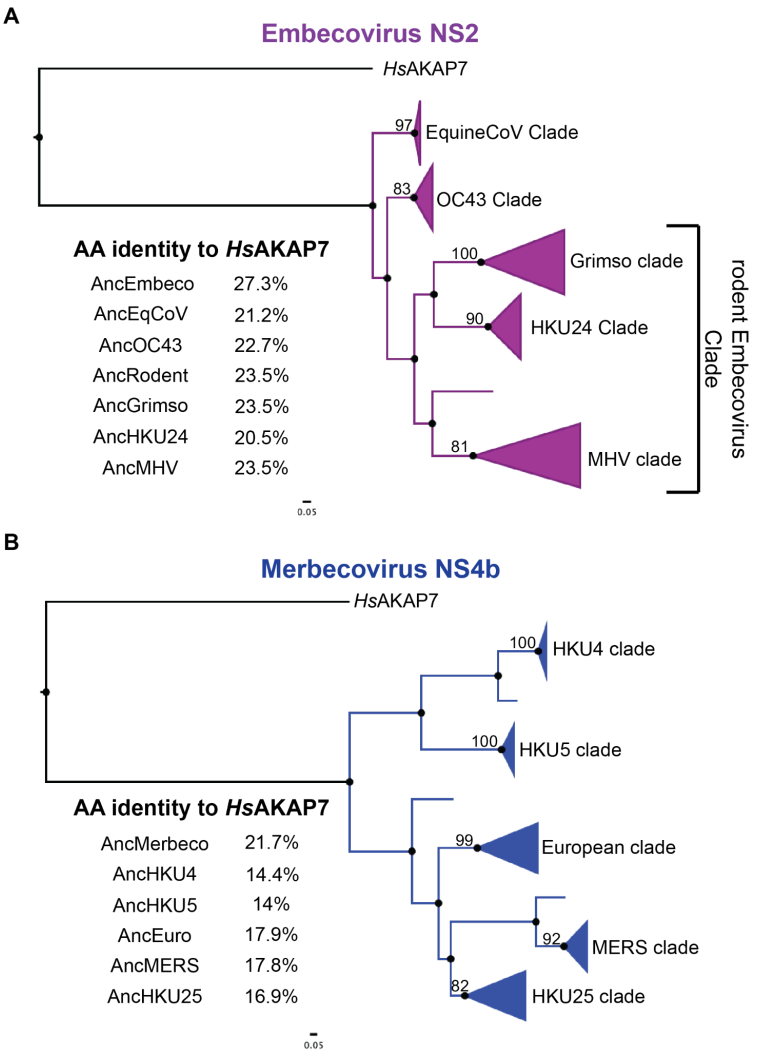

Figure 2. Ancestral sequence reconstruction (ASR) of Embecovirus and Merbecovirus PDEs. A) ML tree of 55 Embecovirus PDE sequences, collapsed into major clades, rooted with the *Hs*AKAP7 PDE domain sequence (AA 82-287) and sequence identity to *Hs*AKAP7 of reconstructed sequences at major nodes. B) ML tree of 27 Merbecovirus PDE sequences,

collapsed into major clades, rooted with the *Hs*AKAP7 PDE domain sequence (AA 82-287) and sequence identity to *Hs*AKAP7 of reconstructed sequences at major nodes.

We conducted the same analysis with 28 *Merbecovirus* NS4b sequences, which was of particular interest due to the wide range in size of these proteins (246-285 AA), largely due to a highly variable N-terminus. Ancestral sequence reconstruction produced an AncMerbecovirus (AncMerbeco) NS4b with 21.7% identity to *Hs*AKAP7, again substantially higher than for ancestral sequences at reconstructed nodes or among observed sequences (**Figure 2B**). The AncMerbeco NS4b is 246 AA long, the same as human MERS-CoV isolates and significantly shorter than the 285 AA NS4b encoded by most HKU4 isolates, a subject of further analysis and discussion below. We conducted similar analyses with the torovirus PDE and rodent alphacoronavirus NS2 PDE, with comparable results (**Figure S3**).

Nidovirus PDE structural models show similar structural divergence from *Hs*AKAP

We then conducted analysis of Nidovirus ancestral PDE structural models in comparison to the *Hs*AKAP7 PDE structure to determine if there were differences in their amino acid identity or root mean square deviation (RMSD), a measure of the similarity in their structures. AncEmbeco (**Figure 3A**) and AncAlphaCoV (**Figure 3B**) NS2 amino acid identity to *Hs*AKAP was the same and their RMSDs relative to the *Hs*AKAP structure were nearly equivalent, supporting the phylogenetic evidence for a single origin of these proteins. In contrast, AncMerbeco NS4b has the lowest amino acid identity to *Hs*AKAP and highest RMSD (**Figure 3C**), supporting its independent origin which, given the substantial within-clade diversity of these proteins (**Figure 1A**), may have been particularly ancient. In contrast, the torovirus PDE encoded on the 3' end of Orf1a has relatively high identity to *Hs*AKAP PDE (~39%) and the lowest RMSD value (**Figure 3D**). Coupled with the limited within-clade diversity among torovirus PDEs (**Figure 1A**, **Figure S1G**) this suggests the torovirus PDE was acquired independently and relatively recently.

Relative to the *Hs*AKAP7 PDE structure, which comprises just the central 206 AA of the 315 AA *Hs*AKAP7 protein, Nidovirus PDEs have extended N-termini (Merbeco) or C-termini (Embeco and AlphaCoV). To control for the possibility that RMSD values might be inflated by poor modeling of these domains we modeled the core PDE domains of Nidovirus PDEs. The AncMerbeco core PDE was not modeled by AlphaFold2, so we substituted the shallower AncMERS PDE. Overall, similarity to *Hs*AKAP7 PDE was generally unchanged regardless of whether the full-length or core viral PDE structural model was analyzed (**Figure S4**).

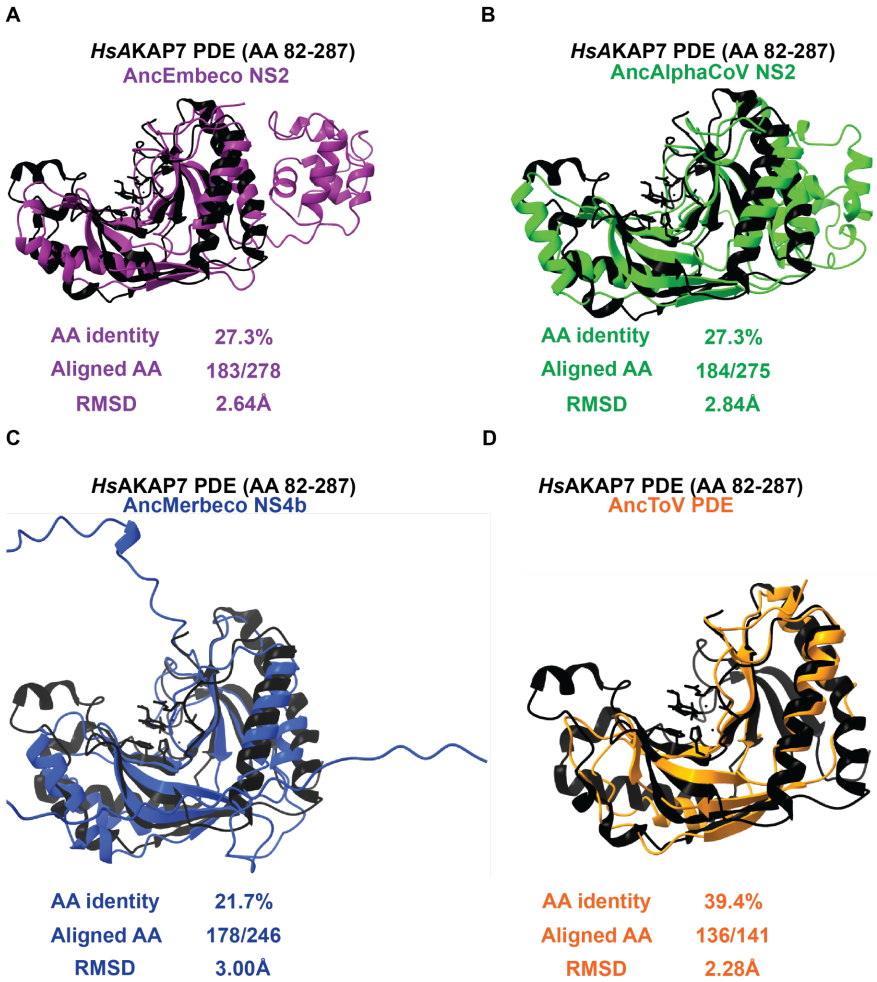

Figure 3. Structural similarity between *Hs*AKAP7 and Nidovirus PDEs A) Overlay of the AncEmbecovirus PDE structure predicted in AlphaFold2 with the solved structure of the *Hs*AKAP PDE. The number of aligned amino acids (from the query sequence) and the RMSD were calculated using FatCat. B) Overlay of the AncAlphacoronavirus PDE structure predicted in AlphaFold with the solved structure of the *Hs*AKAP PDE. The number of aligned amino acids (from the query sequence) and the RMSD were calculated using FatCat. C) Overlay of the AncMerbecovirus PDE structure predicted in AlphaFold with the solved structure of the *Hs*AKAP PDE. The number of aligned amino acids (from the query sequence) and the RMSD were calculated using FatCat. D) Overlay of the AncTorovirus PDE structure predicted in AlphaFold with the solved structure of the *Hs*AKAP PDE. The number of aligned amino acids (from the query sequence) and the RMSD were calculated using FatCat.

Motif and structural analysis supports embecovirus-to-alphacoronavirus horizontal transfer of NS2

We then sought to better elucidate the seemingly close relationship between embecovirus and alphacoronavirus NS2 proteins suggested by the phylogenetic analysis (**Figure 1A**). Given that these viruses fall into different coronavirus genera, a single introduction event is implausible, leaving the possibility of a recombination event or acquisition of a similar cellular ancestor. We first analyzed conservation of key motifs between AncEmbeco NS2

and the reconstructed ancestral PDE sequences from other viral clades. Compared to AncEmbeco NS2, AncAlphaCoV NS2 had the fewest amino acid differences, with only three outside the catalytic His motif in which the second residue frequently toggles among PDEs (**Figure 4A**), supporting the phylogenetic inference that the embecovirus and alphacoronavirus NS2 proteins are closely related. We then conducted a series of overlays of AlphaFold2-predicted structural models of reconstructed PDEs as well as the solved *Hs*AKAP7 PDE structure. Specifically, we used a "core" PDE domain structural model to ensure we captured similarities/differences in the PDE fold itself, and not have the results confounded by disordered regions or N-terminal/C-terminal domains that AlphaFold2 failed to model. This better matches the *Hs*AKAP7 PDE structure which was solved for just the 206 AA PDE central domain, not the entire protein.

We first overlaid AncEmbeco and AncAlphaCoV NS2 and calculated their amino acid identity along with an RMSD value to measure the similarity of the structural models (**Figure 4B**). The AncEmbeco/AncAlphaCoV RMSD was comparable to the RMSD of AncEmbeco and AncHKU24, a shallower node in the *Embecovirus* PDE phylogeny (**Figure 4B, 4C**) although the amino acid identity was substantially lower. Comparing the AncEmbeco and AncAlphaCoV NS2 models to the *Hs*AKAP PDE structure demonstrated that the viral PDEs are much closer to each other in amino acid identity and RMSD than either is to *Hs*AKAP7 PDE (**Figure 4D, 4E)**. Though it's possible this is due to convergence in their structural evolution, the combination of higher amino acid identity with structural model similarity strongly support a common origin, which could only reflect recombination. Finally, we compared the AncEmbeco NS2 model to the AncMERS NS4b model (AlphaFold2 failed to predict the AncMerbeco core PDE), showing that their divergence is comparable to that between the NS2 PDEs and *Hs*AKAP7 PDE (**Figure 4F**), supporting a single rather than multiple origins of the embecovirus and alphacoronavirus PDEs.

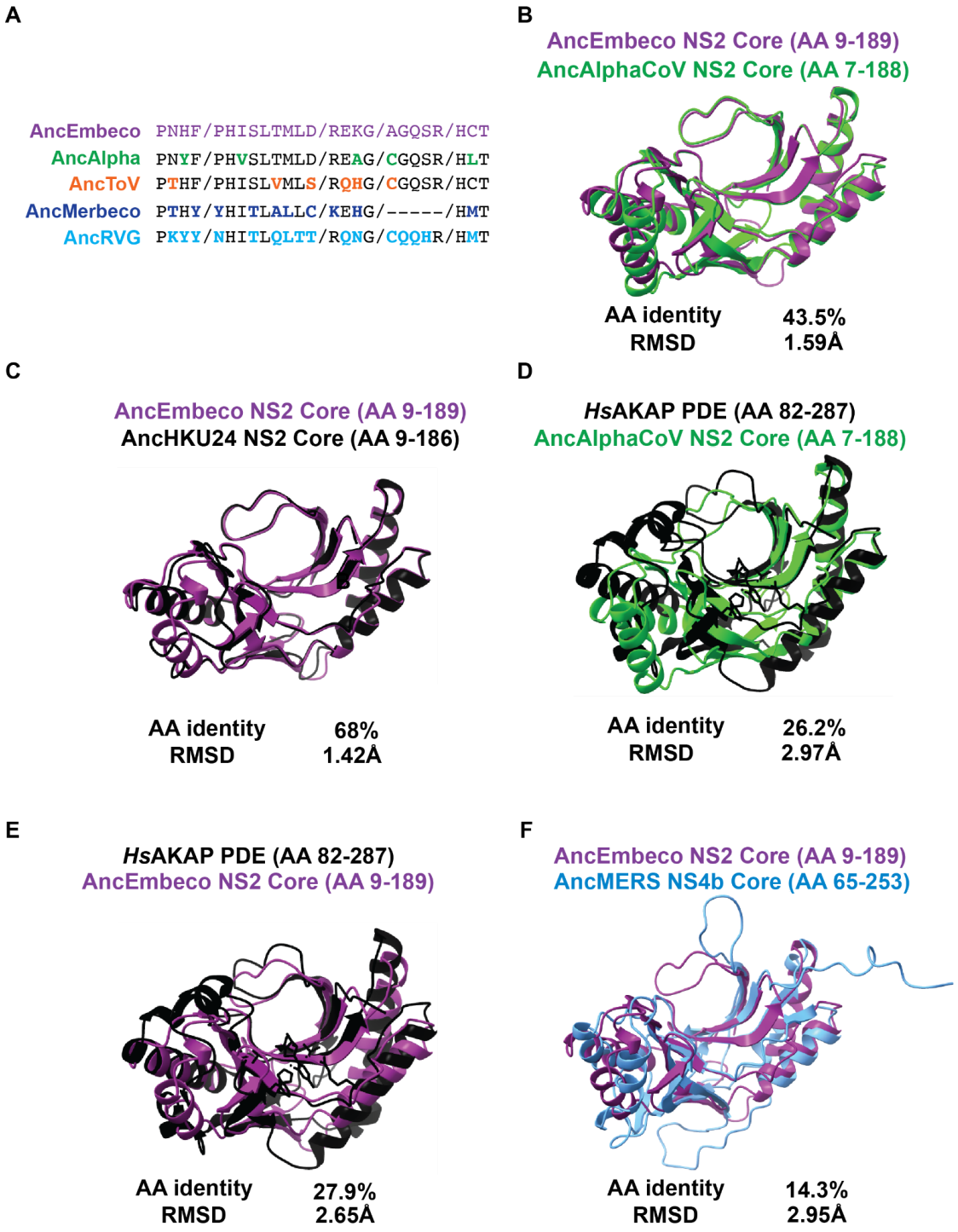

Figure 4. Comparative analysis of AncEmbeco and AncAlphaCoV PDEs. A) Motif analysis with AncEmbeco as the reference. Amino acid differences in key motifs are bolded and in the color corresponding to the viral PDE. B) Overlay and RMSD comparison of AncEmbeco NS2 and AncAlphaCoV NS2. C) Overlay and RMSD comparison of AncEmbeco NS2 and AncHKU24 NS2. D) Overlay and RMSD comparison of *Hs*AKAP PDE and AncAlphaCoV NS2. E) Overlay and RMSD comparison of *Hs*AKAP PDE and AncEmbeco NS2. F) Overlay and RMSD comparison of AncEmbeco NS2 and AncMERS NS4b.

Merbecovirus PDEs exhibit significant intra-clade diversification

We next sought to characterize the post-acquisition evolution and diversification of viral PDEs and focused on embecovirus and merbecovirus PDEs due to their breadth of sampling, 55 and 28 sequences respectively, but different evolutionary trajectories. We first looked at intra-clade variation of embecovirus PDEs. Relative to the AncEmbeco NS2, two reconstructed sequences at shallower nodes have relatively high sequence conservation (**Figure 5A**) and the modeled structures have a low RMSD value, suggesting minimal structural change throughout the course of embecovirus NS2 evolution (**Figure 5B**). We then compared how the radiation of embecovirus NS2 altered structure across the phylogeny by overlaying the modeled structures of the reconstructed AncMHV and AncGrimso NS2 proteins (**Figure 5C**). This confirmed that among embecoviruses sequence and structural divergence is relatively constrained.

In contrast, merbecovirus NS4b has undergone significantly greater sequence divergence and structural variation. This is despite similar sequence diversity in the RdRps of embecoviruses and merbecoviruses, suggesting their radiation has occurred over similar timescales (**Figure S1A-B, S1E-F**). Relative to the reconstructed AncMerbeco NS4b, evolution has occurred via point mutation and indels. There has been notable extension, truncation, insertion, and deletion in the N-terminus of the protein (**Figure 5D**), likely facilitated by the fact that NS4b is encoded on a bicistronic ORF4ab viral sgRNA, which enables it to tolerate mutations that move the start codon upstream or downstream of the ancestral site, all within the overlapping region with Orf4a. However, even accounting for sequence divergence, the structural variation among merbecovirus NS4b proteins is notable. AncMerbeco and AncHKU25 NS4b (**Figure 5E**) (reconstructed at a shallower node) exhibit 69.4% amino acid identity, comparable to AncEmbeco and AncGrimso (**Figure 5B**) yet their RMSD is 2.71 Å compared to 1.83 Å for the embecovirus NS2 proteins. Derived reconstructed sequences AncMERS and AncHKU4 NS4b exhibit just 26.6% amino acid identity and an RMSD >3 Å, further indicating enhanced diversification of the PDEs encoded by this group of viruses (**Figure 5F**).

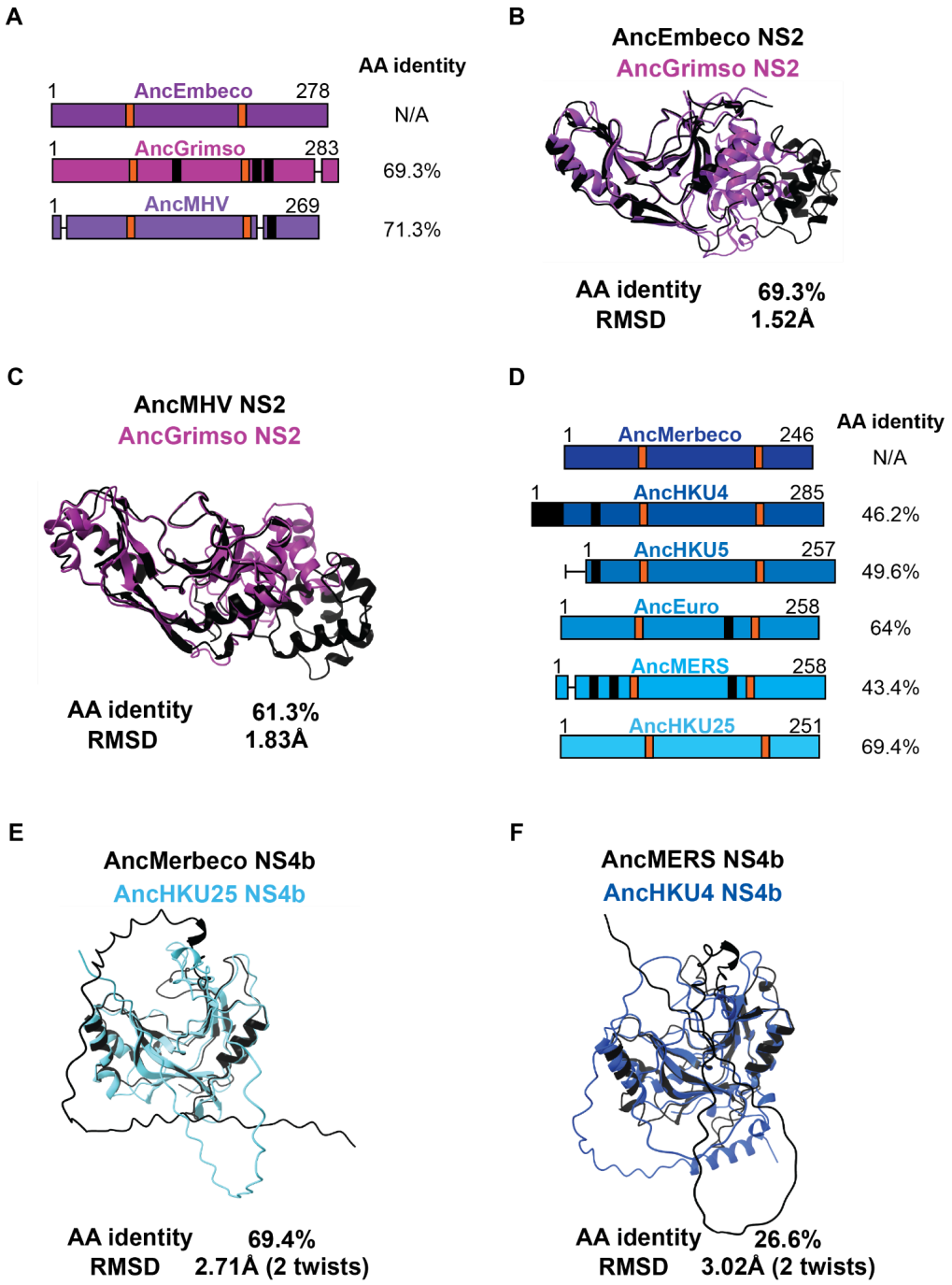

Figure 5. Variable sequence and structural PDE diversity within Nidovirus subgenera A) Schematic of Embecovirus NS2 at key ancestral nodes with catalytic motifs (orange), deletions (horizontal lines), and insertions (black) indicated along with protein length and amino acid identity to the AncEmbeco sequence. B) Overlay of AlphaFold predicted structures of the reconstructed AncEmbeco and derived AncGrimso sequences with amino acid identity and the RMSD as calculated in FatCat. C) Overlay of AlphaFold predicted structures of the AncMHV and AncGrimso sequences with amino acid identity and the RMSD as calculated in FatCat. D) Schematic of Merbecovirus NS4b at key ancestral nodes with catalytic motifs (orange), deletions (horizontal lines), and insertions (black) indicated along with protein length and amino acid identity to the AncMerbeco sequence. E) Overlay of AlphaFold predicted structures of the AncMerbeco sequence and a shallower node

(AncHKU25) with similar AA identity to each other as between the sequences in panel B, with RMSD from a FatCat comparison. F) Overlay of AlphaFold predicted structures of shallower ancestral nodes with different AA sequence identity and RMSDs from FatCat comparisons.

Sequence and structural analysis of rotavirus PDEs supports two horizontal transfer events

Rotavirus PDE and VP3 phylogenies suggest that a PDE was acquired twice, in the RVA VP3 ancestor and in an RVB/RVG VP3 common ancestor (**Figure 1A, 1C**). To confirm this and further elucidate the evolutionary history of rotavirus PDEs we conducted ancestral sequence reconstruction and structural modeling of ancestral RVA and both ancestral and observed RVB and RVG PDEs. As with the nidovirus PDEs, ancestral sequence reconstruction was successful in producing sequences with highest amino acid identity to *Hs*AKAP at the deepest node and higher than any observed sequences (**Figure 6A, 6C**). The AncRVA and AncRVB/G PDEs exhibited similar amino acid identity and RMSDs to *Hs*AKAP7 of 34-36% and ~2.4-2.8 Å (**Figure 6B, 6D**). However, their amino acid identity to each other was considerably lower than to *Hs*AKAP7 PDE and their modeled structures were also relatively dissimilar, with RMSD >3 Å (**Figure 6E**). This result strongly suggested that they are similarly diverged from *Hs*AKAP7 and each other, further supporting multiple origins of the rotavirus PDEs, despite their synteny. Comparison of modeled structures of reconstructed RVB and RVG sequences at shallow nodes provided a stark contrast (**Figure 6F**). These reconstructed PDE sequences had nearly 50% amino acid identity and an RMSD <1 Å, consistent with the phylogenetic inference that the RVB and RVG PDE have a recent common ancestor as does the N-terminal portion of their VP3 protein. This is most parsimoniously explained by a single horizontal transfer event giving rise to all PDEs encoded by these viruses.

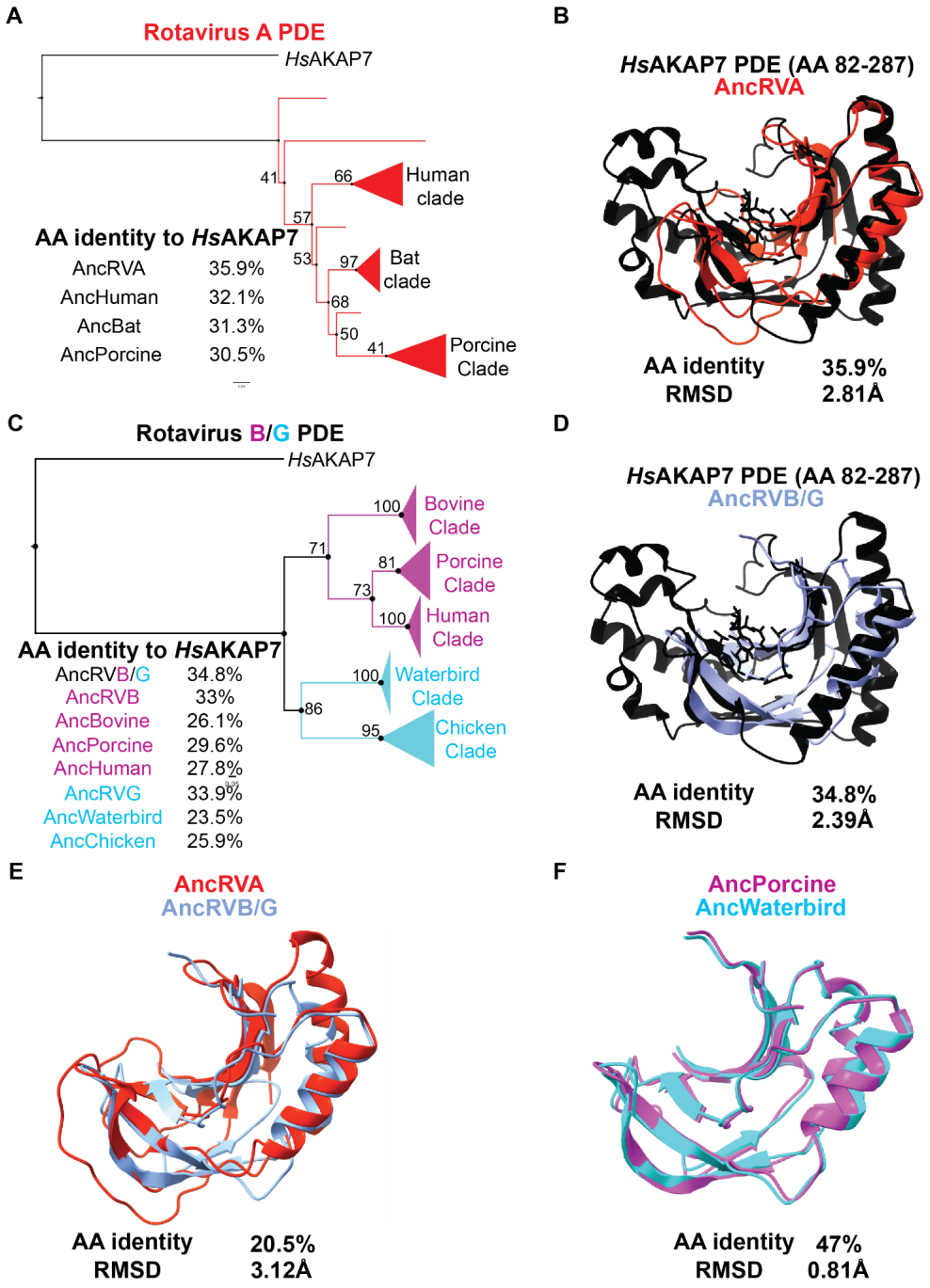

Figure 6. Ancestral sequence reconstruction and structural analysis of Rotavirus PDEs. A) Ancestral sequence reconstruction (ASR) of RVA PDEs at major nodes with AA identity to *Hs*AKAP B) Overlay of the AncRVA PDE structure predicted in AlphaFold with the solved structure of the *Hs*AKAP PDE. C) ASR of RVB/G PDEs at major nodes with AA identities to the hAKAP7 PDE sequence. D) Overlay of the AncRVB/G PDE structure predicted in AlphaFold with the solved structure of the *Hs*AKAP PDE. E) Overlay of the AncRVA PDE structure modeled in AlphaFold with the structural model of the AncRVB/G PDE. E) Overlay of the AncRVB Porcine clade PDE model structure and AncRVG Waterbird clade PDE structural model.

RVG PDE loss of function is not correlated with major structural changes

A previous study compared enzymatic activity of observed RVA, RVB, and RVG PDEs, finding that the RVG PDE tested was inactive (11). This RVG contained an inactivating T90N amino acid change in the second catalytic motif, but enzymatic activity was not restored by repair of this mutation. To determine whether sequence or structural variation might underlie this loss of activity we modeled the structure of RVB^{active} and RVG^{inactive} along with the reconstructed ancestral sequences. The AncRVG PDE and the AncChicken RVG sequences contain identical key motifs and threonine at position 90 (T90) (**Figure 7A**). In contrast, RVG^{inactive} and other viruses in its subclade have not only an amino acid change at position 90, but also unique substitutions in key motifs (**Figure 7A**), though whether these have functional significance is unclear. We then compared structural models to determine whether structural divergence might underlie the RVG loss-of-function. Modeled structures of RVB^{active} and RVG^{inactive} are similarly diverged from the AncRVB/G model in both sequence identity and RMSD (**Figure 7B-C)** and are comparably similar to each other (**Figure 7D**), all with an RMSD <1 Å , indicating a very high degree of similarity. Thus, while subtle structural changes might underlie the loss of function in at least some RVG PDEs, none are obvious. It is possible that some of the amino acid changes in key motifs are involved as well.

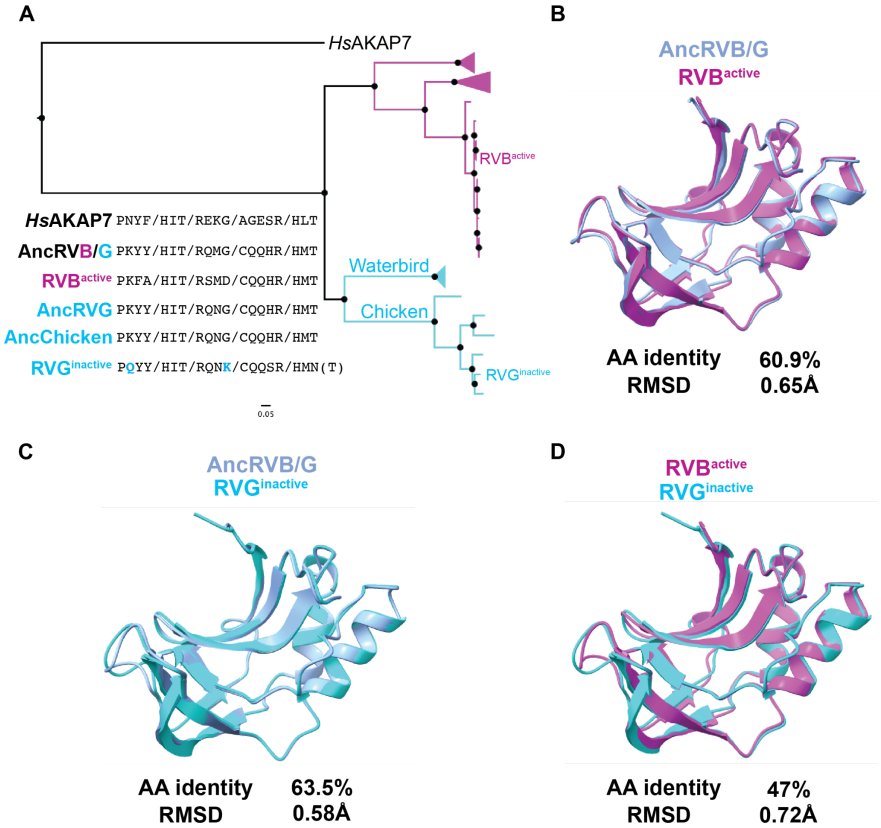

Figure 7. Sequence motif and structural analysis of RVB and RVG functional divergence. A) ML tree of RVB and RVG PDEs, with previously assayed PDEs (Ogden et. al. 2015) indicated B) Overlay and similarity metrics for the predicted structures of the AncRVB/G PDE and the active RVB PDE tested by Ogden et. al. C)

Overlay and similarity metrics for the predicted structures of the AncRVB/G PDE and the inactive RVG PDE tested by Ogden et. al. D) Overlay and similarity metrics for the predicted structures of the active RVB PDE and the inactive RVG PDE tested by Ogden et. al. RMSDs for all overlays were calculated in FatCat.

Discussion

In this study we have comprehensively traced and characterized the acquisition and evolution of viral phosphodiesterases. These proteins are among the clearest examples of acquisition of a host gene by RNA viruses, in contrast to more abundant examples among DNA viruses. However, their amino acid sequence can tolerate significant substitutions without loss of function, so primary sequence identity within and especially across clades of viruses is very low, and often even lower compared to putative AKAP7 cellular ancestors. Although previous work has linked these proteins to an AKAP7-like ancestor (2,16) via comparisons among solved *Hs*AKAP7, MHV NS2, and RVA PDE structures (9–11,17) this provides only a limited view of PDE evolution given challenges of sequence comparison and analysis of only observed, extant PDEs.

We employed several approaches to overcome these challenges. Our phylogenetic analysis is a major expansion on prior efforts (16), incorporating 173 PDE sequences from embecoviruses, merbecoviruses, toroviruses, alphacoronaviruses, RVAs, RVBs, and RVGs. Although outgroup rooting is infeasible and some branches are inferred with only low-to-moderate support, the topology of the tree overall is sufficiently robust to support confident conclusions about the acquisition history of and relationships between viral PDEs (**Figure 1A**). The PDE tree topology features several very long branches, suggesting multiple independent horizontal transfer events and one virus-to-virus transfer via recombination. In contrast to embecovirus and alphacoronavirus PDEs which are syntenic and we assess to have a single origin, RVA and RVB/G PDEs are syntenic but have distinct histories. Comparison of the rotavirus PDE and VP3 phylogenies supports two introductions, one into an ancestral RVA VP3 and one into an RVB/G VP3 common ancestor (**Figure 1C**). The question of why two independent PDE acquisitions landed in exactly the same genomic position remains open, but their phylogenetic distance and structural dissimilarity minimizes the possibility of one acquisition followed by recombination, in contrast to embecoviruses and alphacoronaviruses. It is possible that among rotavirus genome segments, the VP3 segment is particularly capable of accommodating new genome content and that acquisitions on other segments are selected against. This could be experimentally tested using rotavirus infectious clone systems (8,18).

Recent computational advances have enabled previously impossible studies of deep evolutionary events, such as ancient horizontal gene transfers. Because sequence identities are so low among viral PDEs and between viral PDEs and AKAPs, we identified structural analyses of reconstructed ancestral sequences as having the potential to bolster relationships inferred phylogenetically as well as elucidate similarities or differences possibly inapparent in sequence space. We generated maximum-likelihood trees of orthologous viral PDEs as the basis of ancestral sequence reconstruction, which importantly for this study, has been shown as robust to low-to-moderate branch support

(19). We then used AlphaFold2 to generate modeled structures for ancestral sequences at key nodes, overlaid them across viral families and with *Hs*AKAP, and calculated structural similarities for these comparisons (20). These analyses were concordant with the phylogenetic inferences in supporting multiple horizontal transfer events and the single origins of the embecovirus/alphavirus PDEs and the RVB/G PDEs, respectively.

Our characterization of several horizontal transfer events of orthologous ancestral AKAPs and their common repurposing towards antagonism of the OAS-RNase L pathway is both illuminating and thought-provoking. The multiple acquisitions of PDEs that inhibit OAS-RNase L emphatically demonstrates the evolutionary pressure imposed by this antiviral pathway. Yet, these horizontal transfer events have been recurrent but phylogenetically restricted, occurring multiple times in rotaviruses and nidoviruses and seemingly nowhere else, though the OAS-RNase L pathways targets many more viral families (21). Among RNA viruses constraints on genome size due to error rate and packaging capacity may be a limiting factor (22), and the multifunctional influenza A virus NS1 protein does have RNase L-inhibitory activity, obviating an absolute need for a viral PDE (21). Coronaviruses have redundant OAS-RNase L inhibitors (7,23,24), but encoding a PDE apparently provides a strong additive selective advantage and is tolerated in the viral genome.

Less clear is why no PDEs have been found in DNA viruses, even viruses that have captured numerous host genes such as herpesviruses (25) and poxviruses (26–29). Vaccinia virus and related viruses do encode an OAS-RNase L antagonist (30–32), but redundancy in targeting antiviral pathways is well-documented in the case of vaccinia and PKR (33–35) where genome expansion is relatively unconstrained (36,37). Therefore, the presence of one antagonist is not expected to preclude the acquisition of another, as with coronavirus EndoU proteins and PDEs, or K3L and E3L of vaccinia.

Overall, this study demonstrates the power of combining phylogenetic analyses with structural comparisons to elucidate deep evolutionary history. Ancient evolutionary events can be challenging or infeasible to resolve with high confidence due to low sequence identity in extant homologs and orthologs. A clear example of this has been the relationships between eukaryotic and viral phosphodiesterases that antagonize the OAS-RNase L pathway. In significantly expanding the scope of phylogenetic analysis and combining this approach with analyses made possible by recent, major computational advances we are able to describe the evolutionary history of these proteins in comprehensive detail. This provides novel insight into the evolution of virus-host interactions in clinically significant viral families.

Methods

Phylogenetic Analysis

Viral PDEs

All alignments (**Supplementary Materials**) were generated using the MAFFT plug-in (38) with default parameters in Geneious Prime v2022.2.1. Alignments were manually refined as necessary. Maximum-likelihood trees with 100 replicates were generated using IQ-Tree 2 (39) on the Los Alamos National Laboratory server. Different substitution models (with 20 rate-categories) were necessary for different alignments to maximize tree stability and branch support. The all-viral PDE tree is unrooted, as no clear outgroup exists to all viral PDEs. ML trees for individual clades of viral PDEs (**Figure 2, S2, 6A, 6C**) are rooted with *Hs*AKAP7. Alignments (.fasta*) and IQ-Tree logs (.log*) containing detailed parameters and associated with every PDE tree can be found in the Supplementary Files. The ViralPDEs.fasta file contains accessions for all viral PDE sequences analyzed.

Nidovirus RdRps

To generate the Nidovirus RdRp tree (**Figure 1B**) we aligned 145 Nidovirus RdRp sequences representing toroviruses, alphacoronaviruses, betacoronaviruses, gammacoronaviruses, and deltacoronaviruses using MAFFT (default parameters) and generated a ML tree in IQ-Tree 2 using a GTR20 substitution model, 20 rate categories, and 100 bootstrap replicates. Raw files associated with the analysis are in the Supplementary Files.

Rotavirus VP3

We aligned 137 rotavirus VP3 amino acid sequences using MAFFT and trimmed the PDE domains from the RVA, RVB, and RVG sequences in the alignment. We then used IQ-Tree 2 to generate a maximum-likelihood tree using a JTT substitution model, 20 rate categories, and 100 bootstrap replicates.

Maximum-likelihood tree visualization

All .treefile outputs from IQ-Tree 2 were imported into FigTree v1.4.4 for visualization and coloring. Branch support values and labels were added manually in Adobe Illustrator 2023.

Sequence motif analysis

Some of the conserved motifs (**Figure 1D**) had been previously described (5,16) and we identified others by manual scanning of the viral PDE/AKAP alignments.

Ancestral Sequence Reconstruction and Analysis

We generated alignments of each viral PDE group with the *Hs*AKAP7 PDE domain (described above) and fed these alignments into FastML (40,41) along with the maximum-likelihood tree for each alignment generated with IQ-Tree 2. Ancestral sequences were then reconstructed using a JTT substitution model and other default

parameters (optimize branch length, Gamma distribution, marginal reconstruction) and exported for analysis. The raw files (outFile_seq_marginal.txt*, outTreefileAncestor.txt*, outTreeFileNewick.txt*) are available on FigShare and in the Supplementary Materials. After downloading the inferred ancestral sequences, we identified major nodes and aligned them to *Hs*AKAP7 to confirm that deeper nodes have higher amino acid percent identity to the eukaryotic reference PDE. We also analyzed evolutionary trajectory in conserved AKAP-like motifs.

Structural Modeling

We exported major ancestral nodes from each viral PDE family as .fasta files and analyzed these using AlphaFold2 (v2.1.2) 9 as implemented at the University of Utah Center for High Performance Computing (CHPC) with the following parameters:

```
# explicit run of run_alphafold pointing to the appropriate
databases on the RAM disk, for file passed in as $1
singularity exec --nv
/uufs/chpc.utah.edu/sys/installdir/alphafold/2.1.2/alphafold_
2.1.2.sif /app/run_alphafold.sh
--data_dir=/scratch/general/lustre/app-repo/alphafold
--uniref90_database_path=/scratch/general/lustre/app-repo/alp
hafold/uniref90/uniref90.fasta
--mgnify_database_path=/scratch/general/lustre/app-repo/alpha
fold/mgnify/mgy_clusters_2018_12.fa
--uniclust30_database_path=/dev/shm/af/uniclust30_2018_08
--bfd_database_path=/dev/shm/af/bfd_metaclust_clu_complete_id
30_c90_final_seq.sorted_opt
--pdb70_database_path=/dev/shm/af/pdb70
--template_mmcif_dir=/scratch/general/nfs1/app-repo/alphafold
/pdb_mmcif/mmcif_files
--obsolete_pdbs_path=/scratch/general/nfs1/app-repo/alphafold
/pdb_mmcif/obsolete.dat --fasta_paths=$1
--max_template_date=2020-05-14 --output_dir=temp_module
--use_gpu_relax
```

We then extracted the ranked_0.pdb model for each sequence for further analysis. Models were imported into and analyzed with UCSF ChimeraXv1.5 – developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases (42,43). Our primary analysis in ChimeraX involved overlays of predicted and solved structures to generate images that were then exported to Adobe Illustrator 2023.

To quantify the similarities (or differences) between different model structures we conducted pairwise alignments in FATCAT 2.0 (44) to calculate RMSD values.

Figure Design

All figures were produced using Adobe Illustrator 2023.

Acknowledgements

We thank Zoe A. Hilbert for conceptual guidance and contributions. We thank Ian Boys for assistance implementing AlphaFold2 and FATCAT.

This work was supported by National Institutes of Health grants F32AI152341 (S.A.G.) and R35GM134936 (N.C.E.).

1. Dolja, V. V. & Koonin, E. V. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* **244,** 36–52 (2018).
2. Mazumder, R., Iyer, L. M., Vasudevan, S. & Aravind, L. Detection of novel members, structure-function analysis and evolutionary classification of the 2H phosphoesterase superfamily. *Nucleic Acids Res.* **30,** 5229–5243 (2002).
3. Zhao, L. *et al.* Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology. *Cell Host Microbe* **11,** 607–616 (2012).
4. Zhang, R. *et al.* Homologous 2',5'-phosphodiesterases from disparate RNA viruses antagonize antiviral innate immunity. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 13114–13119 (2013).
5. Goldstein, S. A. *et al.* Lineage A Betacoronavirus NS2 Proteins and the Homologous Torovirus Berne pp1a Carboxy-Terminal Domain Are Phosphodiesterases That Antagonize Activation of RNase L. *J. Virol.* **91,** (2017).

6. Thornbrough, J. M. *et al.* Middle East Respiratory Syndrome Coronavirus NS4b Protein Inhibits Host RNase L Activation. *MBio* **7,** e00258 (2016).
7. Comar, C. E. *et al.* Antagonism of dsRNA-Induced Innate Immune Pathways by NS4a and NS4b Accessory Proteins during MERS Coronavirus Infection. *MBio* **10,** (2019).
8. Song, Y. *et al.* Reverse Genetics Reveals a Role of Rotavirus VP3 Phosphodiesterase Activity in Inhibiting RNase L Signaling and Contributing to Intestinal Viral Replication In Vivo. *J. Virol.* **94,** (2020).
9. Sui, B. *et al.* Crystal structure of the mouse hepatitis virus ns2 phosphodiesterase domain that antagonizes RNase L activation. *J. Gen. Virol.* **97,** 880–886 (2016).
10. Brandmann, T. & Jinek, M. Crystal structure of the C-terminal 2’,5'-phosphodiesterase domain of group A rotavirus protein VP3. *Proteins* **83,** 997–1002 (2015).
11. Ogden, K. M. *et al.* Structural basis for 2’-5'-oligoadenylate binding and enzyme activity of a viral RNase L antagonist. *J. Virol.* **89,** 6633–6645 (2015).
12. Li, Y. & Weiss, S. R. Antagonism of RNase L Is Required for Murine Coronavirus Replication in Kupffer Cells and Liver Sinusoidal Endothelial Cells but Not in Hepatocytes. *J. Virol.* **90,** 9826–9832 (2016).
13. Wang, W. *et al.* Extensive genetic diversity and host range of rodent-borne coronaviruses. *Virus Evol* **6,** veaa078 (2020).
14. Forni, D. *et al.* Homology-based classification of accessory proteins in coronavirus genomes uncovers extremely dynamic evolution of gene content. *Mol. Ecol.* **31,** 3672–3692 (2022).
15. Gusho, E. *et al.* Murine AKAP7 has a 2’,5'-phosphodiesterase domain that can complement an inactive murine coronavirus ns2 gene. *MBio* **5,** e01312–e1314 (2014).
16. Asthana, A., Gaughan, C., Dong, B., Weiss, S. R. & Silverman, R. H. Specificity and Mechanism of Coronavirus, Rotavirus, and Mammalian Two-Histidine Phosphoesterases That Antagonize Antiviral Innate Immunity. *MBio* **12,** e0178121

(2021).

17. Gold, M. G., Smith, F. D., Scott, J. D. & Barford, D. AKAP18 contains a phosphoesterase domain that binds AMP. *J. Mol. Biol.* **375,** 1329–1343 (2008).
18. Komoto, S. *et al.* Generation of Infectious Recombinant Human Rotaviruses from Just 11 Cloned cDNAs Encoding the Rotavirus Genome. *J. Virol.* **93,** (2019).
19. Hanson-Smith, V., Kolaczkowski, B. & Thornton, J. W. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol. Biol. Evol.* **27,** 1988–1999 (2010).
20. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).
21. Li, Y. *et al.* Activation of RNase L is dependent on OAS3 expression during infection with diverse human viruses. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 2241–2246 (2016).
22. Holmes, E. C. Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* **11,** 543–546 (2003).
23. Comar, C. E. *et al.* MERS-CoV endoribonuclease and accessory proteins jointly evade host innate immunity during infection of lung and nasal epithelial cells. *Proc. Natl. Acad. Sci. U. S. A.* **119,** e2123208119 (2022).
24. Kindler, E. *et al.* Early endonuclease-mediated evasion of RNA sensing ensures efficient coronavirus replication. *PLoS Pathog.* **13,** e1006195 (2017).
25. Fu, M., Deng, R., Wang, J. & Wang, X. Detection and analysis of horizontal gene transfer in herpesvirus. *Virus Res.* **131,** 65–76 (2008).
26. Hughes, A. L. & Friedman, R. Poxvirus genome evolution by gene gain and loss. *Mol. Phylogenet. Evol.* **35,** 186–195 (2005).
27. Bratke, K. A. & McLysaght, A. Identification of multiple independent horizontal gene transfers into poxviruses using a comparative genomics approach. *BMC Evol. Biol.* **8,** 67 (2008).

28. Rahman, M. J. *et al.* LINE-1 retrotransposons facilitate horizontal gene transfer into poxviruses. *Elife* **11,** (2022).

29. Fixsen, S. M. *et al.* Poxviruses capture host genes by LINE-1 retrotransposition. *Elife* **11,** (2022).

30. Xiang, Y. *et al.* Blockade of interferon induction and action by the E3L double-stranded RNA binding proteins of vaccinia virus. *J. Virol.* **76,** 5251–5259 (2002).

31. Beattie, E. *et al.* Reversal of the interferon-sensitive phenotype of a vaccinia virus lacking E3L by expression of the reovirus S4 gene. *J. Virol.* **69,** 499–505 (1995).

32. Liu, S.-W., Katsafanas, G. C., Liu, R., Wyatt, L. S. & Moss, B. Poxvirus decapping enzymes enhance virulence by preventing the accumulation of dsRNA and the induction of innate antiviral responses. *Cell Host Microbe* **17,** 320–331 (2015).

33. Chang, H. W., Watson, J. C. & Jacobs, B. L. The E3L gene of vaccinia virus encodes an inhibitor of the interferon-induced, double-stranded RNA-dependent protein kinase. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 4825–4829 (1992).

34. Carroll, K., Elroy-Stein, O., Moss, B. & Jagus, R. Recombinant vaccinia virus K3L gene product prevents activation of double-stranded RNA-dependent, initiation factor 2 alpha-specific protein kinase. *J. Biol. Chem.* **268,** 12837–12842 (1993).

35. Davies, M. V., Chang, H. W., Jacobs, B. L. & Kaufman, R. J. The E3L and K3L vaccinia virus gene products stimulate translation through inhibition of the double-stranded RNA-dependent protein kinase by different mechanisms. *J. Virol.* **67,** 1688–1692 (1993).

36. Elde, N. C. *et al.* Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. *Cell* **150,** 831–841 (2012).

37. Cone, K. R., Kronenberg, Z. N., Yandell, M. & Elde, N. C. Emergence of a Viral RNA Polymerase Variant during Gene Copy Number Amplification Promotes Rapid Evolution of Vaccinia Virus. *J. Virol.* **91,** (2017).

38. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).

39. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37,** 1530–1534 (2020).

40. Ashkenazy, H. *et al.* FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* **40,** W580–4 (2012).

41. Moshe, A. & Pupko, T. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. *Bioinformatics* **35,** 2562–2568 (2019).

42. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27,** 14–25 (2018).

43. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30,** 70–82 (2021).

44. Li, Z., Jaroszewski, L., Iyer, M., Sedova, M. & Godzik, A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Res.* **48,** W60–W64 (2020).