*Virus Evolution*, 2021, 7(1): veab019

doi: 10.1093/ve/veab019
Research Article

# Ancient gene duplications in RNA viruses revealed by protein tertiary structure comparisons

Alejandro Miguel Cisneros-Martínez,[1] Arturo Becerra,[1] and Antonio Lazcano[1,2,*]

[1]Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, Mexico and [2]El Colegio Nacional, Donceles 104, Centro Histórico, Mexico City, Mexico

*Corresponding author: E-mail: alar@ciencias.unam.mx

## Abstract

To date only a handful of duplicated genes have been described in RNA viruses. This shortage can be attributed to different factors, including the RNA viruses with high mutation rate that would make a large genome more prone to acquire deleterious mutations. This may explain why sequence-based approaches have only found duplications in their most recent evolutionary history. To detect earlier duplications, we performed protein tertiary structure comparisons for every RNA virus family represented in the Protein Data Bank. We present a list of thirty pairs of possible paralogs with <30 per cent sequence identity. It is argued that these pairs are the outcome of six duplication events. These include the $\alpha$ and $\beta$ subunits of the fungal toxin KP6 present in the dsRNA *Ustilago maydis virus* (family *Totiviridae*), the SARS-CoV (*Coronaviridae*) nsp3 domains SUD-N, SUD-M and X-domain, the *Picornavirales* (families *Picornaviridae*, *Dicistroviridae, Iflaviridae and Secoviridae*) capsid proteins VP1, VP2 and VP3, and the *Enterovirus* (family *Picornaviridae*) 3C and 2A cysteine-proteases. Protein tertiary structure comparisons may reveal more duplication events as more three-dimensional protein structures are determined and suggests that, although still rare, gene duplications may be more frequent in RNA viruses than previously thought.

*Keywords*: gene duplications; RNA viruses.

## 1. Introduction

Many hypotheses on the evolutionary importance and the mechanisms of gene duplications were already established by cytologists and cytogeneticists since the first decades of the twentieth century (Taylor and Raes 2004). However, it was not until the publication of *Evolution by gene duplication* by Susumu Ohno (1970), when the idea of gene duplication as a major evolutionary force became widely acknowledged. During the following decades, with the advent of DNA sequencing techniques, a wealth of accumulated evidence contributed considerably to our understanding of gene duplications, allowing for the refinement of models that describe its mechanisms and underlying its evolutionary relevance (Taylor and Raes 2004). The rationale for understanding the evolutionary significance of gene duplications lies within the notion that evolution cannot proceed solely through point mutations because any mutation that alters the function of a coding gene would be deleterious. The solution to this conundrum is provided by Ohno (1970): ‘Only the cistron which became redundant was able to escape from the relentless pressure of natural selection, and by escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus’. The evolutionary significance of gene duplications is highlighted by the relatively high frequency at which they occur in the three major domains of life (17–44% in bacteria, ~30% in archaea and 30–65% in eukarya) (Zhang 2003). In fact, the identification of ancient gene duplications that appear to have happened before the divergence of the three domains of life (Becerra et al. 2007) suggests that it has been one of the most important mechanisms for increasing the size and complexity of genomes since the early stages of cell evolution (Lazcano 1995).

Previous studies have shown that gene duplications have been a relatively frequent event in dsDNA viral genome evolution (Shackelton and Holmes 2004). Many examples of gene duplications are known in *Adenoviridae* (Davison et al. 2003), *Herpesviridae* (McGeoch and Davison 1999) and *Poxviridae* (Hughes and Friedman 2005). A search on 201 dsDNA viruses found gene duplications in 42.3 per cent of its genomes. The 1,874 identified paralogs were distributed in 612 protein families with two to sixty-one members (Gao et al. 2017). Additionally, a positive correlation was found between paralog number and genome size, which can reach up to 2,473 kbp in *Pandoravirus salinus*. In sharp contrast, Simon-Loriere and Holmes (2013) detected gene duplications only in 19 out of 1,198 (1.6%) RNA viruses analysed. The twenty paralogs were distributed in eight protein families composed of two to three members (Table 1). These twenty pairs are likely to represent nine duplication events that include four cases in ssRNA(+) viruses: 1) the coat protein (CP) and the minor CP (CPm) in the family *Closteroviridae* (Boyko et al. 1992; Kreuze et al. 2002; Tzanetakis et al. 2005; Tzanetakis and Martin 2007; Simon-Loriere and Holmes 2013), as well as a tandem duplication of CPm in the *Grapevine leafroll-associated virus 1* (Fazeli and Rezaian 2000); 2) p25 and p26 proteins encoded by the third and fifth segments, respectively, in the family *Benyviridae* (Simon-Loriere and Holmes 2013); and 3) a tandem duplication of the genome-linked protein VPg in the *Foot-and-mouth disease virus* of the family *Picornaviridae* (Forss and Schaller 1982). In ssRNA(-), the two cases were found in the family *Rhabdoviridae*: 1) the G and Gns glycoproteins in some viruses of the genera *Ephemerovirus* (Walker et al. 1992; Blasdell et al. 2012) and *Hapavirus* (Gubala et al. 2010); and 2) U1 and U2 of unknown function (Simon-Loriere and Holmes 2013). Finally, in ssRNA(RT), three cases were found in the family *Retroviridae*: 1) orfA and orfB of *Walleye epidermal hyperplasia virus 2* (LaPierre et al. 1999); 2) orf1 and orf2 of *Xenopus laevis endogenous retrovirus* (Kambol et al. 2003); and 3) vpr and vpx in *Human immunodeficiency virus 2* and *Simian immunodeficiency virus—mnd 2* (Tristem et al. 1990). As of today, no gene duplication events have been reported in dsRNA viruses.

Detection of gene duplications in RNA viral genomes is complicated for a number of reasons. For instance, the number of paralogs is known to be positively correlated with the genome size (Gevers et al. 2004) and, with the exception of coronaviruses, RNA viruses tend to have smaller genomes (from ~2 to ~33 kbp) compared with dsDNA viruses (from ~5 to ~2,500 kbp) (Campillo-Balderas et al. 2015). The genome sizes of RNA viruses may be limited by the high error rate of RNA replicases (Reanney 1982; Holmes 2009). As described by Eigen (1971), nucleic acids need a minimum replication fidelity to preserve the genetic information, where a higher fidelity allows a higher information content. This implies that the amount of genetic information is limited by the precision of the copying process. If the genome grows beyond the error, threshold deleterious mutations would quickly appear (Eigen 1971; Maynard Smith and Szathmáry 1995). In fact, an inverse relationship between mutation rate and genome size has been observed from viroids to eukaryotes, in which RNA viruses appear as the second biological entities with the highest mutation rates and the shortest genomes (Gago et al. 2009; Holmes 2011). Paradoxically, to evolve an accurate replication machinery requires more coding capabilities and thus a larger genome. This so-called Eigen's paradox (Maynard Smith and Szathmáry 1995) might imply that most RNA virus genomes are irrevocably limited to remain small. As can be inferred from Sol Spiegelman's *in vitro* RNA replication and evolution experiment from 1970 (Maynard Smith and Szathmáry 1995), replication efficiency is another pressure that selects for smaller genomes, which could also affect RNA viruses that benefit from a faster replication in a context of competition with the host and other viruses for cellular resources.

Other factors that could underline the RNA viruses genome size restrictions include the shape and size of the capsid and the impossibility of unwinding large dsRNA structures formed during the replication in viruses lacking a helicase domain (Reanney 1982; Holmes, 2009). Finally, and as expected, in a directed evolution experiment that tested the stability and fitness effect of different duplicated genes in artificial constructs derived from the plant-infecting ssRNA(+) *Tobacco etch virus* (family *Potyviridae*), Willemsen et al. (2016) observed a fitness reduction and the deletion of the duplicated gene. As a further explanation to the deleterious effect of gene duplications, they suggested that the correct processing of the polyprotein and a greater cellular resource requirement to express a larger genome could be important factors contributing to the constraints on RNA virus genome size. Some of these issues may also apply to ssDNA viruses which, from an evolutionary perspective, have been thought to behave similarly to RNA viruses, showing small genome sizes and little gene duplication (Boyko et al. 1992; Holmes 2009).

The current evidence of gene duplications in RNA viruses comes mainly from protein primary structure which, given the previously mentioned high mutation rate, can only recognize the most recent duplications (Simon-Loriere and Holmes 2013). As argued here, protein tertiary structure comparisons can broaden the known universe of paralogous proteins in RNA viruses. Our results suggest that in fact gene duplications might be more stable in RNA genomes than previously thought.

## 2. Methods

### 2.1 Data selection

On 17 July 2020, we performed an advanced search on RCSB Protein Data Bank (PDB) (www.rcsb.org) (Berman et al. 2000) based on the following criteria:

**Table 1.** Gene duplications are much more frequent in dsDNA compared with RNA viruses.

| | dsDNA viruses (Gao et al. 2017) | RNA viruses (Simon-Loriere and Holmes 2013) |
|---|---|---|
| Viruses with duplicated genes | 85/201 (42.3%) | 19/1198 (1.6%) |
| Number of paralogous pairs | 1874 | 20 |
| Number of paralogous families | 612 | 8 |
| Family size | 2 to 61 members | 2 to 3 members |

The table summarizes the results described by Gao et al. (2017) and Simon-Loriere and Holmes (2013), respectively.

- Source Organism Taxonomy Name equals Riboviria
- Polymer Entity Distinct Taxonomy Count = 1
- Experimental Method equals X-RAY DIFFRACTION
- Resolution ≤ 3Å
- Polymer Entity Type equals Protein
- Polymer Entity Sequence Length ≥ 80
- Polymer Entity Mutation Count = 0

The search resulted in a table (Supplementary data S1) describing different features (such as Entity ID, Number of Entities (Protein), PDB ID, Source Organism, Taxonomy ID, Macromolecule Name, Resolution Å, R Work, Deposition Date, Structure Title, Chain Length, Number of Polymer Residues, Entity Polymer Type, Structure Keywords, PubMed Central ID, PubMed ID, DOI) of 4,049 protein entities. To select representative structures, the information was sorted on the basis of:

- PDB in alphabetical order
- Deposition date (most recent)
- Quality factor (highest)
  - $\frac{1}{\mathit{Resolution}\ \text{Å} - R\ \mathit{Work}}$
- Macromolecule name in alphabetical order
- Source organism in alphabetical order

The manual selection resulted in 1,112 representative entities corresponding to 961 PDB IDs (Supplementary data S2). The corresponding sequences were downloaded from RCSB PDB and further redundancy was filtered with a three-step iterative hierarchy clustering with CD-HIT (90% and 60% sequence identity) and PSI-CD-HIT (30% sequence identity) (Li and Godzik 2006). The clustering resulted in 305 protein entities corresponding to 297 PDB IDs (Supplementary data S3 and S4), which were downloaded from RCSB PDB and parsed with the Perl module ParsePDB.pm (Bulheller and Hirst 2009) to retrieve only the corresponding chains. Taxonomic annotation based on the NCBI taxonomy (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/ accessed July 2020) revealed that the most frequent entries in the dataset corresponded to ssRNA(+), followed by ssRNA(-), dsRNA and ssRNA(RT) viral families and two ssRNA(+) satellite viruses.

### 2.2 Tertiary structure alignments

Structural comparisons were carried out for each viral family. Families with only one representative structure (*Picobirnaviridae, Alphaflexiviridae, Alphatetraviridae, Mesoniviridae, Permutotetraviridae, Potyviridae, Tospoviridae*) and viruses with no family assigned (satellite viruses) were excluded from the analysis. A total of 2,406 tertiary structure alignments (Supplementary data S5) were automatically conducted via the FATCAT rigid algorithm (Ye and Godzik 2003). Structural similarity was evaluated with a modification of the structural alignment score (SAS) (Subbiah et al 1993) that takes into account the alignment coverage of each structure defined as:

$$\mathit{Lsas} = \frac{100\mathrm{RMSD}(L1 + L2)}{2\mathit{Naln}^2}$$

Where Lsas stands for length-weighted SAS, RMSD is the root mean square deviation between α-carbon atoms, L1 and L2 correspond to the lengths of the superposed structures and Naln is the number of aligned residues. 257 alignments with Lsas < 10 (Supplementary data S6) were manually analyzed to evaluate the homology type between pairs. Likely paralogs were defined with Lsas < 5. Structural alignments were visualized with UCSF Chimera (Pettersen et al. 2004).

### 2.3 Structure similarity trees

Structural models 2acf-A (X-domain), 1b35-C (VP3), 3q3y-A (3C) and both chains in 4gvb (KP6α and KP6β) were queried for a PDB search within the DALI server (http://ekhidna2.biocenter.helsinki.fi/dali/ accessed January 2021) (Holm 2020). Models selection for the structure similarity tree analyses was performed as follows: 1) for the KP6 subunits only the shared hits with Z ≥ 4 on the PDB25 report; 2) for the X-domain hits with Z ≥ 12 on the PDB90 report; 3) for VP3 hits with Z ≥ 7 on the PDB25 report (*Secoviridae* models 1a6c, 1bmv and 7chk were excluded to avoid long branch attraction artifacts); and 4) for the 3C protease nonviral hits with Z ≥ 11 on the PDB25 report plus viral hits with Z ≥ 11 on the PDB90 report (Supplementary data S8). In each case, protein and species redundancies were omitted. The models were edited with UCSF Chimera to retain only the corresponding chains. For the proteases, only the carboxy terminal domains were used. Multiple structure alignments were performed with the STAMP algorithm (Russell and Barton 1992) within the MultiSeq tool (Roberts et al. 2006) in VMD 1.9.3 (Humphrey et al. 1996) with default parameters for the KP6 subunits and the 3C protease and its respective relatives, and with scanscore = 0 for VP3 and its related structures. The X-domain-related structures were compared through pairwise against all alignments with the MatchMaker tool within Chimera using a BLOSUM30 matrix and the Smith-Waterman algorithm. Structure similarity was assessed with the Match → Align tool in Chimera from which RMSD and number of aligned residues were retrieved to compute SAS (100RMSD/Naln) as a geometric distance measure. The resulting distance matrices were introduced into the FITCH algorithm (Fitch and Margoliash 1967) within the PHYLIP 3.695 package (Felsenstein 1989) for tree construction with global branch-swapping rearrangements and the jumble option to randomize 100 times the input order. For the capsid proteins, the tree was rooted on *Solemoviridae* and *Tombusviridae* single jelly roll capsid proteins as outgroup. For the rest, the root was placed using the MAD method (Tria et al. 2017). Finally, tree visualization was made with FigTree 1.4.2 (Rambaut 2014).

## 3. Results

A total of 30 pairs of likely paralogous proteins was found (Supplementary data S7). Table 2 shows 12 representative pairs with the lowest Lsas. As argued below, it is possible that these cases represent six duplication events: 1) one that led to the α and β subunits of the *Ustilago maydis virus* (UmV) (family *Totiviridae*) KP6 toxin, which is potentially the first confirmed case of gene duplication in a dsRNA virus; 2) the SARS-Unique domain (SUD) N and M domains found in sarbecoviruses (family *Coronaviridae*) that may come from the more widely distributed coronavirus X-domain; 3) a duplication event that probably gave rise to the chymotrypsin-related 2A cysteine protease in the genus *Enterovirus* (family *Picornaviridae*) from the greater distributed 3C cysteine protease; and 4) VP1, VP2 and VP3, that probably originated after two duplication events during the dawn of the order *Picornavirales* (Liljas et al. 2002).

**Table 2.** Likely paralogs detected by protein tertiary structure comparisons.

| Genome type | Order | Family | Pair | RMSD | Naln | Lsas |
|---|---|---|---|---|---|---|
| dsRNA | *Ghabrivirales* | *Totiviride* | KP6α and KP6β | 1.52 | 62 | 2.9854 |
| ssRNA(+) | *Nidovirales* | *Coronaviridae* | SUD-M and X-domain | 2.82 | 123 | 3.9982 |
| | *Picornavirales* | *Dicistroviridae* | VP1 and VP2 | 3.14 | 173 | 2.6596 |
| | | | VP1 and VP3 | 3.11 | 208 | 2.0164 |
| | | | VP2 and VP3 | 3.13 | 208 | 2.0112 |
| | | *Iflaviridae* | VP1 and VP2 | 3.02 | 173 | 2.5831 |
| | | | VP1 and VP3 | 3.17 | 206 | 2.4987 |
| | | | VP2 and VP3 | 3.07 | 182 | 3.1373 |
| | | *Picornaviridae* | VP1 and VP2 | 3.19 | 165 | 2.8883 |
| | | | VP1 and VP3 | 3.15 | 179 | 2.3398 |
| | | | VP2 and VP3 | 3.08 | 167 | 2.8438 |
| | | | 3C and 2 A | 2.75 | 140 | 2.3151 |

The table shows 12 representative pairs with the lowest Lsas. Detailed information on the 30 protein pairs and PDB IDs is available in Supplementary data S7.

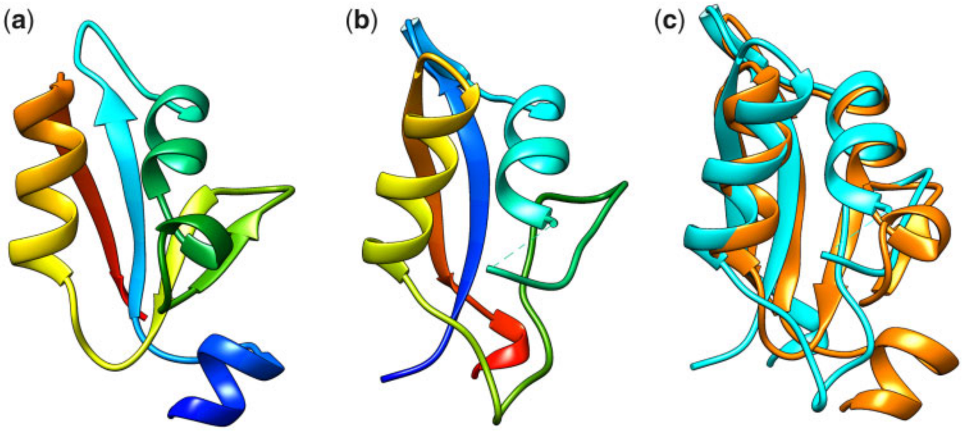

**Figure 1.** Tertiary structure of KP6α (a) and KP6β (b). Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (c) 3D alignment between KP6α (orange) and KP6β (cyan) as made by FATCAT rigid. PDB ID 4GVB.

### 3.1 *Totiviridae* KP6α-KP6β

As described elsewhere (Allen et al. 2013), KP6 is a viral toxin present in UmV. This virus is found only in fungi containing resistance genes that allow them to compete with other strains not resistant to the toxin. Thus, UmV acts as a symbiont which is only transmitted from one cell to another through mitosis or meiosis. KP6 is a heterodimer whose subunits are encoded on a single satellite dsRNA. Both subunits are translated as a single polypeptide that undergoes protease cleavage on a 31 amino acid linker between the former amino (KP6α) and carboxy (KP6β) terminal domains. KP6α and KP6β are 77 and 74 residues long, respectively, both of which fold into a α/β sandwich structure consisting of a four-stranded antiparallel β-sheet and a pair of antiparallel α-helices. The major differences between these structures are the presence of an extra N-terminal helix in KP6α, and longer α2-β2 and β3-α3 loops in KP6β (Fig. 1).

Although no clear statement regarding the paralogous relationship between the genes was made, a 3D alignment of the two KP6 subunits had been reported by Allen et al. (2013) with very similar results as those reported here with FATCAT rigid. Interestingly, upon structural database search, the only similar proteins to KP6α and KP6β are individual domains within larger cellular proteins, with KP6 being the only protein showing this kind of heterodimer (Allen et al. 2013). This suggests that the virus did not acquire an already duplicated protein, but that the gene encoding it underwent a duplication event in the virus after the acquisition of a single domain.

### 3.2 *Sarbecovirus* Sud and X-domain

SARS-CoV Nsp3 is translated as a polyprotein that contains an acidic domain, an X-domain, the SUD, a papain-like cysteine protease domain and other domains including a transmembrane region. The X-domain is a homodimeric phosphatase that removes the 1′ phosphate group of ADP-ribose-1′-phosphate (ADRP). Its structure is mainly defined by seven central β strands surrounded by six α helices (three on each side of the sheet), conforming a three-layered α/β/α topology as seen in proteins belonging to the Macro-H2A fold. The innermost five β strands are parallel while the outermost two strands are antiparallel (Saikatendu et al. 2005). SUD is a domain present only in SARS-CoV and closely related sarbecoviruses. It is known to be endowed with two macrodomains, N and M, similar to ADRP which is also present in other coronaviruses and even in viruses belonging to different families. However, SUD domains lack phosphatase activity, and have been shown to bind to oligonucleotides forming G-quadruplex secondary structures. The structure of SUD-N consists of six β strands and four α helices, while SUD-M is made of six β strands and five α helices. In both domains, the β sheet has five parallel strands and only one antiparallel β3 strand (Tan et al. 2009).

Although the homologous relations between SUD-N, SUD-M and the X-domain was not discussed by Tan et al. (2009), they had in fact described their structural similarity. They superimposed SUD-N and SUD-M with an RMSD of 3.3 Å, and found a conserved Leu-Glu-Glu-Ala motif at the N-terminal end of helix α4. We have confirmed the structural similarity between SUD-M and the X-domain (Fig. 2). Tan et al. observed better RMSD values between SUD-M and X-domain (2.3 Å) than between SUD-N and X-domain (2.7 Å), which is consistent with our results. Given the taxonomic distribution of these domains, the adjacent position of the three domains in the Nsp3 polyprotein and the clear structural similarity, we posit that SUD-N and SUD-M are likely paralogs that resulted from two duplication events that started with the duplication of the X-domain.

### 3.3 *Picornavirales* capsid proteins VP1, VP2 and VP3

The order *Picornavirales* comprises families such as *Dicistroviridae*, *Iflaviridae*, *Marnaviridae*, *Picornaviridae* and *Secoviridae*. In most viruses belonging to the family *Picornaviridae*, the capsid genes are translated into a single polyprotein which is proteolytically cleaved into VP0, VP3 and VP1. VP0 is then self-cleaved into VP4 and VP2, except in the genera *Kobuvirus* and *Parechovirus*, in which the equivalent of VP4 remains as a N-terminal extension of VP2 (Sabin et al. 2016; Kalynych et al. 2016a). In other families, such as *Dicistroviridae* and *Iflaviridae*, VP4 is cleaved from the N-terminal region of VP3 (Liljas et al. 2002; Kalynych et al. 2016b). In viruses of the family *Secoviridae* there is no equivalent to VP4, and in some cases the polyprotein is partially cleaved into a large and small subunit made of two and one domains, respectively (e.g. *Comovirus*), or not cleaved at all (e.g. *Nepovirus*). VP1, VP2 and VP3, or its equivalent domains (A, C and B) are the main building blocks of the *Picornavirales* capsids. These proteins are jelly-roll β barrels of approximately 250 residues long made of eight anti-parallel strands (B-I). Sixty copies of each domain assembly to form a T = p3 icosahedral capsid with a diameter of approximately 30 nm (Rossmann and Johnson 1989).

As discussed in qualitative terms by Chandrasekar and Johnson (1998) and Liljas et al. (2002), VP1, VP2 and VP3 tertiary structures are remarkably similar. As shown in Fig. 3, this similarity is particularly clear after visual inspection of the capsid proteins of the family *Dicistroviridae*, in which VP1, VP2 and VP3 do not have large insertions and the N-terminal arm conformation is conserved. Given that VP1, VP2 and VP3 of different families of the order *Picornavirales* appear to be related, it is likely that VP1, VP2 and VP3 arose after two duplication events prior to the divergences of the *Picornavirales* families (Liljas et al. 2002). This hypothesis is further supported by our quantitative analysis, which shows that the 3D structures of VP1 and VP2, VP1 and VP3, and VP2 with VP3 of the *Dicistroviridae* capsid proteins, align with 3.14, 3.11 and 3.13 RMSD along 173, 208 and 208 residues, respectively (Table 2). The selective advantage of these duplication events might be related to a rapid assembly of the capsid or to the interactions with the cell receptors and the host immune systems.

### 3.4 *Enterovirus* 3C and 2A cysteine-proteases

The 3C and 2A picornains are cysteine-proteases responsible for the viral polyprotein processing. According to the MEROPS peptidase database (https://www.ebi.ac.uk/merops/ accessed July 2020) (Rawlings et al. 2018), both picornains are part of the C3 family. Based on fold similarity and catalytic triad arrangement, this family is classified alongside other serine and cysteine-protease families into clan PA. Members of this clan show the same protein fold described in chymotrypsin (family S1), which consists of two homologous six-stranded antiparallel β barrels with a catalytic triad, His-Asp-Ser (His-Asp-Cys or His-Glu-Cys in some viral proteases), located in the barrel interface (Lesk and Fordham 1996). Each barrel is made of two structural motifs composed of three antiparallel β strands connected by two loops, with strands named from A1 to F1 (N-terminal domain) and from A2 to F2 (C-terminal domain), respectively. The

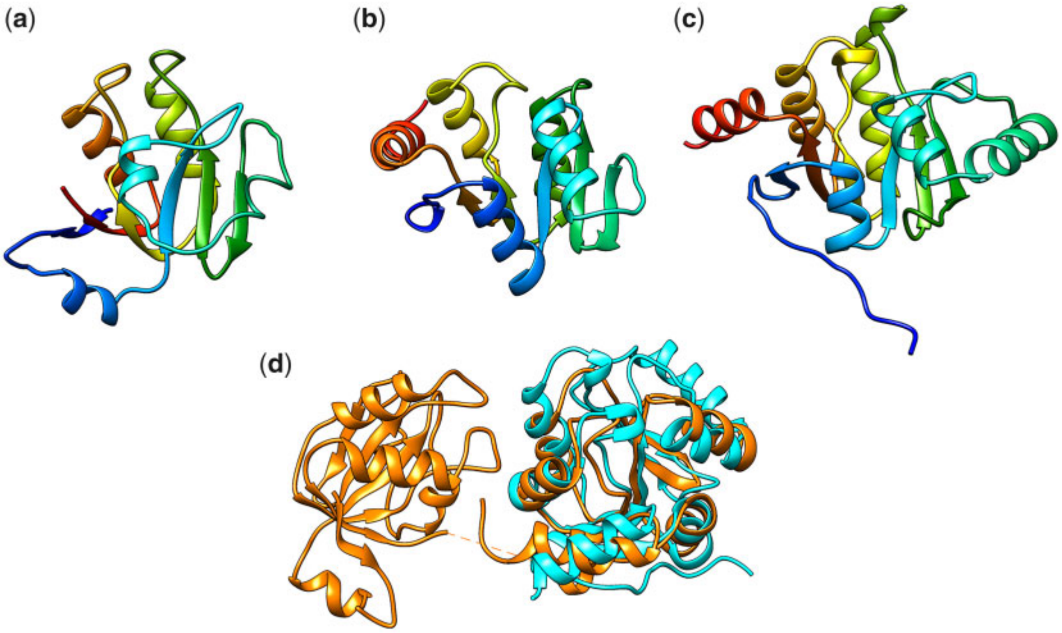

**Figure 2.** Tertiary structure of SUD-N (a), SUD-M (b) and X-domain (c) of SARS-CoV. Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (d) 3D alignment between SUD (orange) and X-domain (cyan) as made by FATCAT rigid. PDB IDs 2WCT and 2ACF.

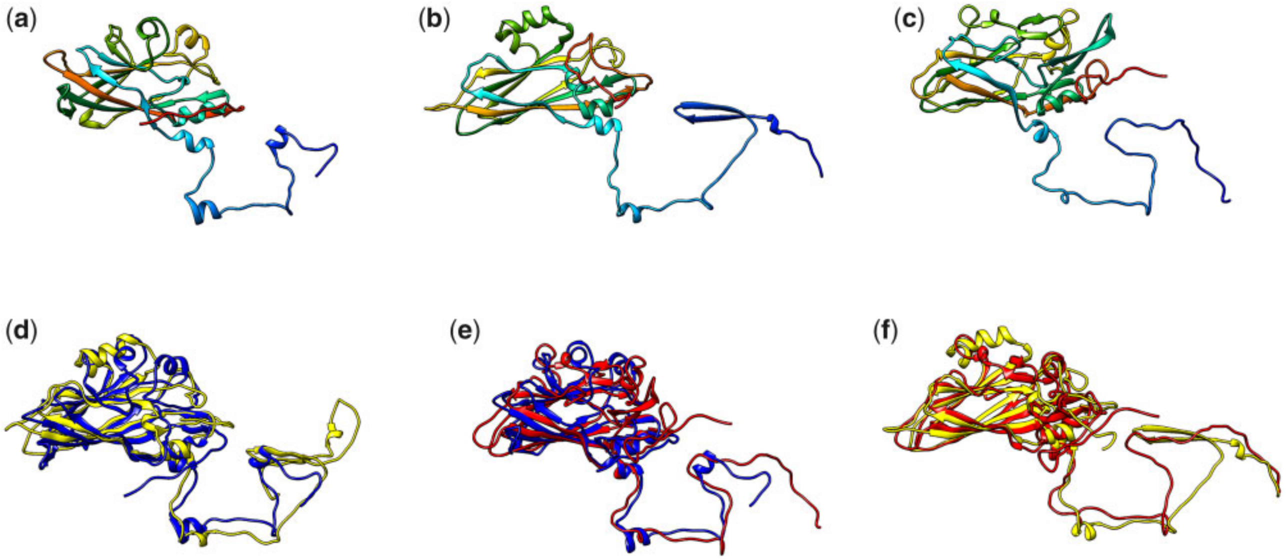

**Figure 3.** Tertiary structure of *Dicitroviridae* (a) VP1, (b) VP2 and (c) VP3. Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (d–f) Pairwise 3D alignments between VP1 (blue), VP2 (yellow) and VP3 (red) as made by FATCAT rigid. PDB IDs (a, b) 1B35, (c) 5CDD, (d) 1B35 and 5CDD, (e) 1B35 and 5CDD and (f) 1B35 and 3NAP.

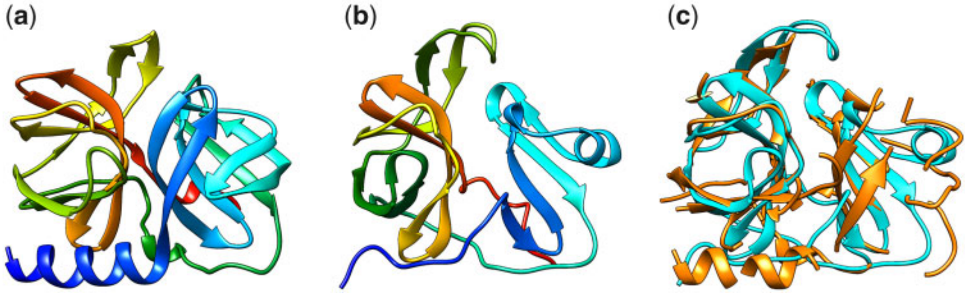

**Figure 4.** Tertiary structure of *Enterovirus* 3C (a) and 2A (b). Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (c) 3D alignment between 3C (orange) and 2A (cyan) as made by FATCAT rigid. PDB IDs 3Q3Y and 3W95.

catalytic residues are located in different loops named accordingly with the corresponding residue. The histidine loop is located between strands C1 and D1, the aspartate loop between strands E1 and F1 and the serine loop between strands C2 and D2 (James et al. 1978; Petersen et al. 1999).

Picornains differ from chymotrypsin in that both have a shorter D1-E1 (which does not bind to calcium) and A2-B2 loops (James et al. 1978; Matthews et al. 1994). Another major difference is in the B2-C2 loop (referred as methionine loop in S1 proteases) that presents an abrupt extended turn in picornains instead of the characteristic helix found in chymotrypsin (James et al. 1978; Allaire et al. 1994; Matthews et al. 1994). The major difference between 3C and 2A is the absence of A1 and D1 strands in 2A, whose domain 1 consists of only four β stands, which makes it ∼40 resides shorter than 3C (Fig. 4). Additionally, 2A has a pair of cysteines, close to the strand A2, that mediates zinc ion coordination together with another cysteine and a histidine located in D2-E2 loop. This coordination is very similar to the one found in hepacivirin (family S29) of *Hepatitis C virus* (family *Flaviviridae*) and might play a stabilizing role such as the disulfide bond found in chymotrypsin in a similar position (Petersen et al. 1999).

It has been suggested that 3C and 2A proteases are paralogs based on a pairwise sequence alignment in which, despite the low sequence identity (13%), the catalytic residues and predicted secondary structure elements appear to be conserved (Bazan and Fletterick 1988). Based on a qualitative structure comparison, the paralogous relationship of 3C and 2A has also been suggested (Petersen et al. 1999). Our analysis adds quantitative support to the duplication hypothesis.

Information about the function and taxonomic distribution of both proteases can provide insights into the evolutionary relevance of this duplication. The 3C protease is responsible for most of the polyprotein processing, whereas 2A can only catalyze its own cleavage from the structural proteins. 3C is found in every genera of the family *Picornaviridae*, while 3C-like proteases are found in all the families of the order *Picornavirales* and in some related families such as *Caliciviridae, Coronaviridae* and *Potyviridae* (King et al. 2011). On the other hand, there are up to five different types of 2A proteins in the family *Picornaviridae*: 1) the 2A protease (2A$^{pro}$) typical of the genus *Enterovirus*; 2) the 2A$^{npgp}$ protein typical of *Cardiovirus*, *Senecavirus*, *Aphthovirus*, *Teschovirus* and *Erbovirus*, which produces an analogous effect to the 2A$^{pro}$ cleavage through ribosomal skipping in a conserved sequence motif NPGP; 3) the 2A$^{H\text{-}box/NC}$ protein typical of *Parechovirus*, *Kobuvirus* and *Tremovirus*, which lacks proteolytic activity and is related to a family of proteins involved in the control of cell proliferation (Hughes and Stanway 2000); 4) the

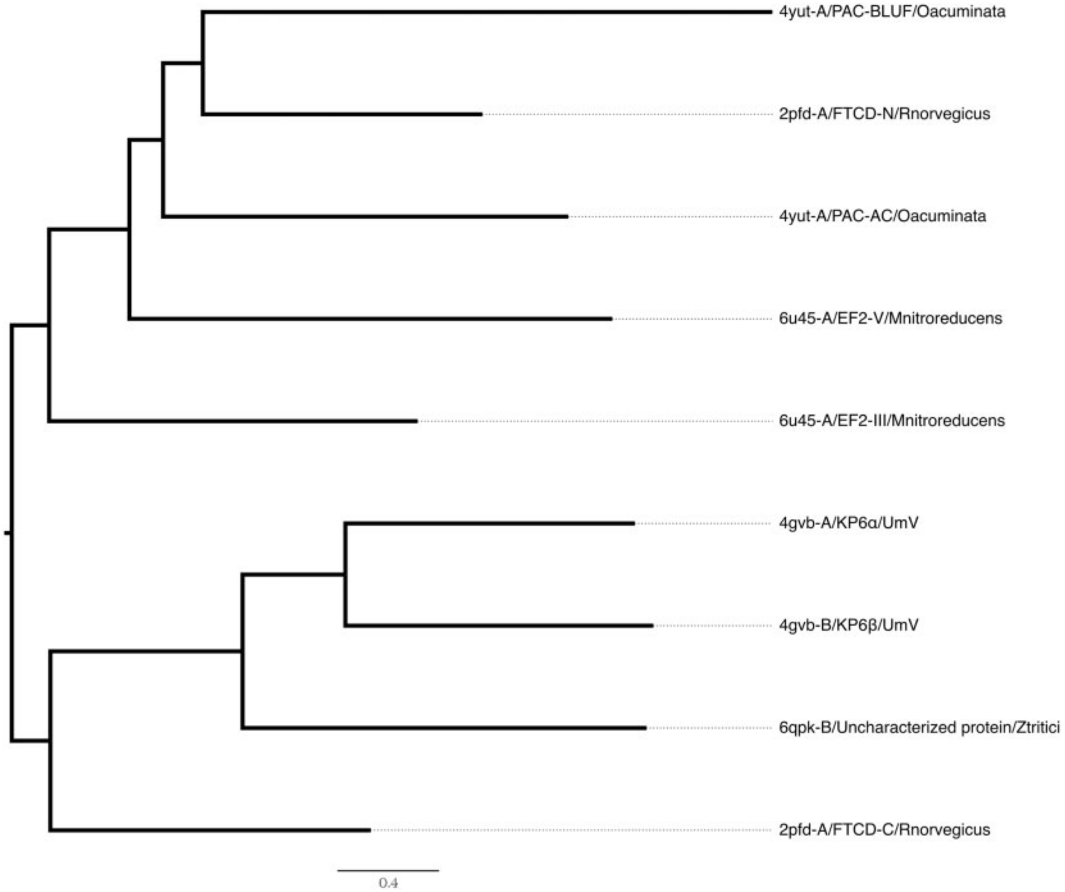
4yut-A/PAC-BLUF/Oacuminata
2pfd-A/FTCD-N/Rnorvegicus
4yut-A/PAC-AC/Oacuminata
6u45-A/EF2-V/Mnitroreducens
6u45-A/EF2-III/Mnitroreducens
4gvb-A/KP6α/UmV
4gvb-B/KP6β/UmV
6qpk-B/Uncharacterized protein/Ztritici
2pfd-A/FTCD-C/Rnorvegicus
0.4

**Figure 5.** Structure similarity tree supporting a close relationship between KP6α and KP6β. PDB IDs with its chain, protein names and organisms are indicated on each leaf. PAC-AC and PAC-BLUF stand for photoactivated adenylate cyclase-adenylate cyclase domain and blue light using flavin domain, respectively, FTCD equates to Formimidoyl Transferase Cyclo Deaminase amino (-N) and carboxy (-C) terminal domains and EF2 corresponds to Elongation Factor 2 domains III and V. Organisms are: UmV = *Ustilago maydis virus*, Ztritici = *Zymoseptoria tritici*, Rnorvegicus = *Rattus norvegicus*, Oacuminata = *Oscillatoria acuminata* and Mnitroreducens = *Methanoperedens nitroreducens*.

2A AIG1-like protein with possible NTPase function, located between a $2A^{npgp}$ and a $2A^{H\text{-}box/NC}$, only found in *Avihepatovirus* (Tseng et al. 2007); and 5) a 2A protein of unknown function unrelated to the previous ones only found in the genus *Hepatovirus* (King et al. 2011). This suggests that the 2A protease is a synapomorphy with a particular selective advantage on the genus *Enterovirus* (and possibly on the closely related genus *Sapelovirus*), and that the polyprotein processing activity of $2A^{pro}$, which can be replaced by the $2A^{npgp}$ or 3C, may be in fact dispensable for most picornaviruses.

Additional functions have been associated with both picornains. On the one hand, 3C has been associated to the viral replication initiation complex formation via 5′-untranslated region binding and to the host transcription inhibition through the degradation of the H3 histone, the TATA-binding protein or some transcription factors. On the other hand, $2A^{pro}$ has been associated with the host translation inhibition by means of eIF4G degradation (Bazan and Fletterick 1988; Porter 1993; Matthews et al. 1994; Petersen et al. 1999). Degradation of eIF4G allows the virus to impair the host protein translation while it takes advantage of the translation machinery through its internal ribosome entry site (IRES). It has been suggested that picornaviruses lacking a $2A^{pro}$ have a strong IRES for ribosome binding that can compete with an intact host initiation factor complex, whereas *Enterovirus* IRES binding is weak, so that these viruses depend on eIF4G inhibition to gain access to the host translation machinery (Petersen et al. 1999).

## 4. Discussion

Since gene homology within a genome can result from recombination and not from paralogous duplications, we performed searches against protein structure databases for each case reported here to construct structure similarity trees as a means to distinguish between the different possible scenarios. In all four cases, the trees display topologies consistent with paralogous relationships (see Figs 5 and 6 and Supplementary Figs S1 and S2). This is specially clear for the KP6 subunits KP6α and KP6β, which group together with its closest relative being an uncharacterized protein from an ascomycete fungus (Fig. 5). The capsid proteins VP1, VP2 and VP3, form three monophyletic groups, each showing a similar inner topology, which suggest two paralogous duplication events prior to the divergence of the order *Picornavirales* (Fig. 6) Although the possibility of independent gains of these proteins cannot be ruled out completely, the phylogeny depicted in Fig. 6 strongly supports their origin through gene duplication events.

It has been argued that an important difference between RNA and dsDNA viruses is the number of gene duplications (Holmes 2009). However, the newly detected cases of paralogous

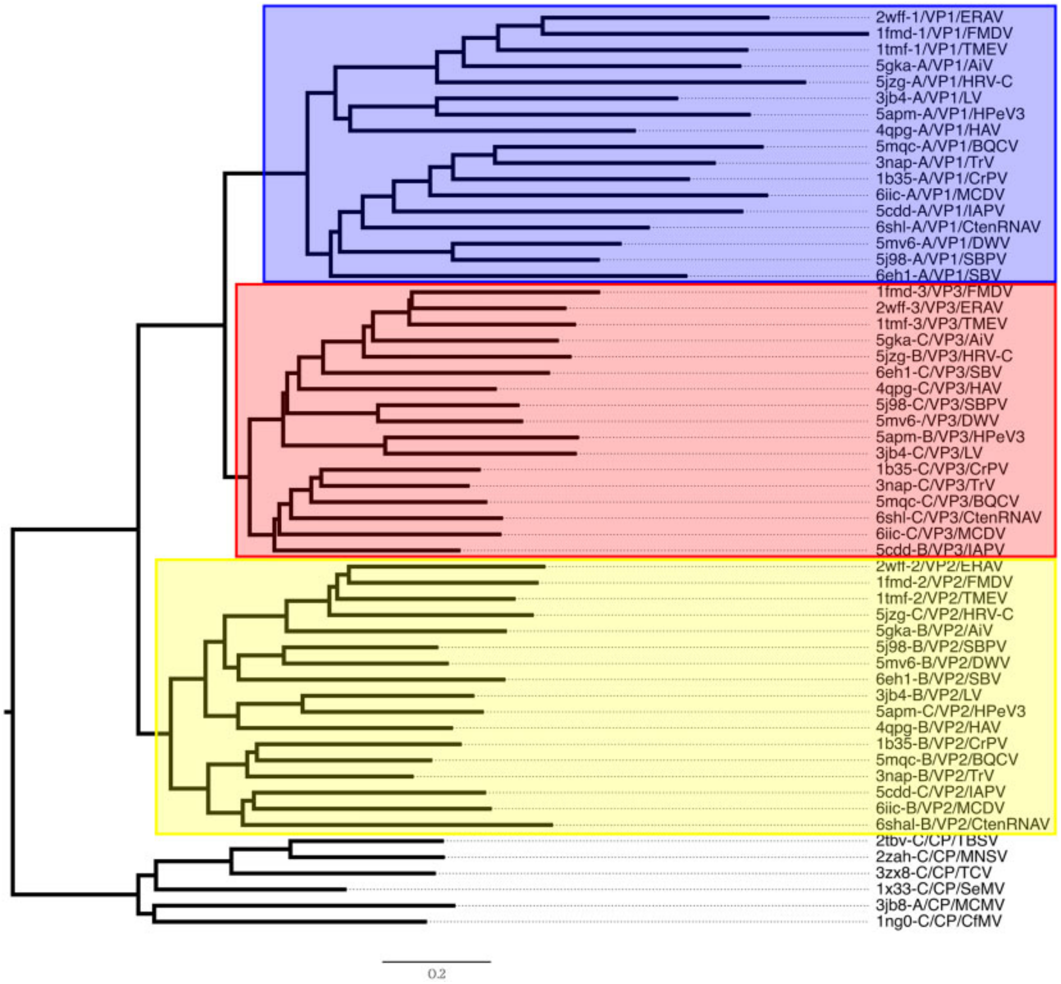

**Figure 6.** Structure similarity tree of proteins related to VP1, VP2 and VP3. PDB ids with its chain, protein names and organisms are indicated on each leaf. VP1, VP2 and VP3 proteins are highlighted by a blue, yellow and red box, respectively. Organisms are: ERAV = *Equine rhinitis A virus*; FMDV = *Foot and mouth disease virus*; TMEV = *Theiler's encephalomyelitis virus*; AiV = *Aichi virus*; HRV-C = *Human rhinovirus-C*; LV = *Ljungan virus*; HPeV3 = *Human parechovirus 3*; HAV = *Hepatitis A virus*; BQCV = *Black queen cell virus*; TrV = *Triatoma virus*; CrPV = *Cricket paralysis virus*; MCDV = *Mud crab dicistrovirus*; IAPV = *Israeli acute paralysis virus*; CtenRNAV = *Chaetoceros tenuissimus RNA virus*; DWV = *Deformed wing virus*; SBPV = *Slow bee paralysis virus*; SBV = *Sacbrood virus*; TBSV = *Tomato bushy stunt virus*; MNSV = *Melon necrotic spot virus*; TCV = *Turnip crinkle virus*; SeMV = *Sesbania mosaic virus*; MCMV = *Maize chlorotic mottle virus*; CfMV = *Cocksfoot mottle virus*.

proteins in RNA viruses reported here suggests that gene duplication may be a more frequent phenomenon on these viruses than previously thought. The number of detected paralogs using 3D protein comparison methodology discussed here is expected to increase as more viral protein structures are determined. Unfortunately, due to the lack of X-ray determined models, we were unable to apply our methodology to confirm the cases reported by Simon-Loriere and Holmes (2013). The addition of structural models determined with techniques other than X-ray crystallography (like NMR or Cryo-EM) may increase the size of the analyzed database.

Our inability to detect some previously reported duplications is an indication of the limits of our approach. Examples of suggested duplicated domains that are not discussed in our analysis include the shell (S) and protruding (P) domains of *Tombusviridae* capsid protein (Jones et al. 1989), as well as the P1 and P2 domains of *Hepeviridae* and *Caliciviridae* capsid proteins (Guu et al. 2009). It is worth pointing out that we also found structural similarity between the *Macrobrachium rosenbergii nodavirus* capsid P domain and the *Black beetle virus* capsid S domain (Lsas = 7.7631), both of which have a jelly-roll topology (Wery et al. 1994; Chen et al. 2019), which suggests that some nodaviruses may have undergone a duplication similar to the one suggested by Jones et al. (1989) for tombusviruses. Other similar structures that might indicate duplication events but will require further analysis are the *Porcine reproductive and respiratory syndrome virus* (*Arteriviridae*) nsp1α and nsp1β papain-like cysteine protease domains (Sun et al. 2009; Xue et al. 2010) (Lsas = 6.2778), the coronavirus 3C-like protease and nsp9 (Sutton et al. 2004) (Lsas = 13.912), the ssRNA(-) *Human respiratory syncytial virus* (*Pneumoviridae*) NS1 and matrix protein (Chatterjee et al. 2017) (Lsas = 7.4347), the retrovirus capsid N-terminal domain and C-terminal domain (Lsas = 7.0739–8.1531) and the retrovirus reverse transcriptase-ribonuclease H (RT-RNaseH) connection domain, RNaseH domain and integrase (INT) (Lsas INT-RNaseH = 5.7483–6.4967) (Malik and Eickbush 2001), although it has been suggested that the later have independent evolutionary histories (Koonin et al. 2015). Finally, It is important to note that, despite their low sequence similarity, coronavirus papain-like proteases PL1pro and PL2pro have been suggested to be paralogs (Lee et al. 1991; Herold et al. 1999; Ziebuhr et al. 2000; Ziebuhr et al. 2001). This case was not detected by our method because both sequences were clustered together by PSI-CD-HIT. Specifically, the *Swine acute diarrhea syndrome coronavirus* PL2pro (PDB: 6L5T) was selected as the representative protein for the cluster in which the *Porcine transmissible gastroenteritis coronavirus* PL1pro (PDB: 3MP2) and the *Porcine epidemic diarrhea virus* PL2pro (PDB: 6NOZ) were included with 27.404 per cent and 44.872 per cent sequence identity, respectively (Supplementary data S4).

Given the positive correlation between paralog number and genome size (Gevers et al. 2004), we would have expected to find more duplicated genes in viruses with larger genomes. For example, viruses of the family *Coronaviridae* with genomes that can reach more than 30kb (Campillo-Balderas et al. 2015) or viruses with segmented genomes which, on average, tend to be larger than non-segmented RNA genomes (Holmes 2009). However, most of the detected cases belong to monopartite viruses with genomes not larger than 20kb. This might suggest different growth mechanisms or even a sample bias. For the particular case of the coronaviruses, in which we have detected three likely paralogs, it has been suggested that their large genomes are possible because they encode a HEL domain helicase and an ExoN domain 3′-5′ exoribonuclease which are involved in RNA duplex unwinding, and proofreading and repair, respectively (Gorbalenya et al. 2006; Holmes 2009). The presence of a helicase domain could explain the number of duplicated genes detected so far in the order *Picornavirales*. Interestingly, it has been shown that large single and multi-domain protein families are less frequent in viruses compared to cellular organisms. Also, the percentage of multi-domain proteins belonging to viruses tends to be smaller than the percentage of single-domain proteins (Forslund et al 2019). Considering that the reported duplication events in RNA viruses only involve single domains of around 300 residues or less, it is possible that RNA viruses can preserve gene duplications only if the genetic redundancy is comprised of relatively small sequences. This appears to be the case presented in Willemsen et al. (2017), where the artificial insertion of the relatively small gene 2b, which codes for a redundant function, is preserved despite the predicted fitness cost of a growing genome. Although gene duplication and horizontal gene transfer imply a different homology origin, the fitness effects related to the genome size limitations are predicted to be the same. Therefore, functional redundancy may also be beneficial both after the recruitment of external sequences or following a duplication event, which is consistent with our results as well as with other gene duplication reports (Simon-Loriere and Holmes 2013) where at best we can only see slight indications of functional diversification.

Studies on RNA virus genome size increase due to recombination and/or paralogous duplications provide insights into how hypothetical cellular RNA genomes grew during early stages of cell evolution and increased their coding capacity from a small number of coding genes to the hundreds or thousands of genes that eventually led to a complex organism such as the last common ancestor (Becerra et al. 2007). Despite the obvious differences, some features of RNA virus genomes can be used as models to understand the hypothetical RNA/protein World cellular genomes. For instance, early proteinic polymerases were probably just as error prone as current viral RNA-dependent RNA polymerases (Reyes-Prieto et al. 2012; Jácome et al. 2015), whose palm subdomain is probably one of the oldest structural domains still recognizable in today’s viruses and cells, and may actually be a relic from the RNA/protein World (Jácome et al. 2015).

If we add the paralogous pairs reported here to the 20 pairs listed by Simon-Loriere and Holmes (2013) at least fifty pairs distributed in twelve paralogous families composed of two to three members derived from fifteen duplication events can be considered. Although this numbers might still indicate that gene duplications are infrequent in RNA viruses, it is remarkable that gene duplications still occur and are maintained, which indicates that although genome growth tends to reduce the virus fitness (Willemsen et al. 2016), paralogous genes can be maintained whenever the acquired benefits are greater than the cost of a longer genome. Gene duplications can be conspicuous in only some RNA viruses. For instance, the three capsid proteins and the tandem repeat of VPg represent a high proportion of genes originated by gene duplication in the genome of *Foot-and-mouth disease virus*.

## 5. Conclusions

New cases of paralogous proteins that probably diverged early on RNA virus evolution are reported here. These cases include both subunits of the cytotoxin KP6, SARS-CoV SUD and X-domain, *Enterovirus* cysteine proteases 3C and 2A, and *Picornavirales* VP1, VP2 and VP3 viral capsids. Due to the low sequence conservation, these cases could only be confirmed by quantitative tertiary structure comparisons. The number of known paralogous proteins in RNA viruses is likely to grow as more viral protein structures are determined. Overall, our results suggest that gene duplication is a more relevant mechanism for increasing the coding capacities of RNA genomes than previously thought.

## Acknowledgments

We thank the theoretical and methodological support provided by José Alberto Campillo Balderas, Ricardo Hernández Morales and Alejandro Rodrigo Jácome Ramírez. Alejandro Miguel Cisneros-Martínez is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 814975 from the Consejo Nacional de Ciencia y Tecnología (CONACYT).

## Funding

This study was funded by PAPIIT-UNAM (IN214421).

## Data availability

Availability of data and material: Figs S1 and S2 are available as supplementary material. Supplementary datasets (1–8) are available in a public repository: https://github.com/abb-GB/gene-duplication.git.

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Conflict of interest

None declared.

## References

Allaire, M. et al. (1994) 'Picornaviral 3C Cysteine Proteinases Have a Fold Similar to Chymotrypsin-like Serine Proteinases', *Nature*, 369: 72–6.

Allen, A., Chatt, E., and Smith, T. J. (2013) 'The Atomic Structure of the Virally Encoded Antifungal Protein, KP6', *Journal of Molecular Biology*, 425: 609–21.

Bazan, J. F., and Fletterick, R. J. (1988) 'Viral Cysteine Proteases Are Homologous to the Trypsin-like Family of Serine Proteases: Structural and Functional Implications', *Proceedings of the National Academy of Sciences of the United States of America*, 85: 7872–6.

Becerra, A. et al. (2007) 'The Very Early Stages of Biological Evolution and the Nature of the Last Common Ancestor of the Three Major Cell Domains', *Annual Review of Ecology, Evolution, and Systematics* , 38: 361–79.

Berman, H. M. et al. (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28: 235–42.

Blasdell, K. R. et al. (2012) 'Kotonkan and Obodhiang Viruses: African Ephemeroviruses with Large and Complex Genomes', *Virology*, 425: 143–53.

Boyko, V. P. et al. (1992) 'Coat Protein Gene Duplication in a Filamentous RNA Virus of Plants', *Proceedings of the National Academy of Sciences of the United States of America*, 89: 9156–60.

Bulheller, B. M., and Hirst, J. D. (2009) 'DichroCalc – Circular and Linear Dichroism Online', *Bioinformatics (Oxford, England)*, 25: 539–40.

Campillo-Balderas, J. A., Lazcano, A., and Becerra, A. (2015) 'Viral Genome Size Distribution Does Not Correlate with the Antiquity of the Host Lineages', *Frontiers in Ecology and Evolution*, 3: 143.

Chandrasekar, V., and Johnson, J. E. (1998) 'The Structure of Tobacco Ringspot Virus: A Link in the Evolution of Icosahedral Capsids in the Picornavirus Superfamily', *Structure (London, England : 1993)*, 6: 157–71.

Chatterjee, S. et al. (2017) 'Structural Basis for Human Respiratory Syncytial Virus NS1-Mediated Modulation of Host Responses', *Nature Microbiology*, 2:

Chen, N. C. et al. (2019) 'The Atomic Structures of Shrimp Nodaviruses Reveal New Dimeric Spike Structures and Particle Polymorphism', *Communications Biology*, 2: 72.

Davison, A. J., Benkő, M., and Harrach, B. (2003) 'Genetic Content and Evolution of Adenoviruses', *Journal of General Virology*, 84: 2895– 2908.

Eigen, M. (1971) 'Self-Organization of Matter and the Evolution of Biological Macromolecules', *Die Naturwissenschaften* , 58: 465–523.

Fazeli, C. F., and Rezaian, M. A. (2000) 'Nucleotide Sequence and Organization of Ten Open Reading Frames in the Genome of Grapevine Leafroll-Associated Virus 1 and Identification of Three Subgenomic RNAs', *The Journal of General Virology* , 81: 605–615.

Felsenstein, J. (1989) 'PHYLIP—Phylogeny Inference Package (Version 3.2)', *Cladistics*, 5: 164–166.

Fitch, W. M., and Margoliash, E. (1967) 'Construction of Phylogenetic Trees', *Science (New York, N.Y.)*, 155: 279–284.

Forslund, S. K., Kaduk, M., and Sonnhammer, E. L. L. (2019) 'Evolution of Protein Domain Architectures', in M., Anisimova (ed.) *Evolutionary Genomics. Statistical and Computational Methods*, 2nd edn, pp 469–504. New York: Humana.

Forss, S., and Schaller, H. (1982) 'A Tandem Repeat Gene in a Picornavirus', *Nucleic Acids Research*, 10: 6441–6450.

Gago, S. et al. (2009) 'Extremely High Mutation Rate of a Hammerhead Viroid', *Science (New York, N.Y.)*, 323: 1308.

Gao, Y. et al. (2017) 'Extent and Evolution of Gene Duplication in DNA Viruses', *Virus Research*, 240: 161–165.

Gevers, D. et al. (2004) 'Gene Duplication and Biased Functional Retention of Paralogs in Bacterial Genomes', *Trends in Microbiology*, 12: 148–154.

Gorbalenya, A. E. et al. (2006) '*Nidovirales*: Evolving the Largest RNA Virus Genome', *Virus Research*, 117: 17–37.

Gubala, A. et al. (2010) 'Ngaingan Virus, a Macropod-Associated Rhabdovirus, Contains a Secondglycoprotein Gene and Seven Novel Open Reading Frames', *Virology*, 399: 98–108.

Guu, T. S. Y. et al. (2009) 'Structure of the Hepatitis E Virus-like Particle Suggests Mechanisms for Virus Assembly and Receptor Binding', *Proceedings of the National Academy of Sciences of the United States of America*, 106: 12992–12997.

Herold, J., Siddell, S. G., and Gorbalenya, A. E. (1999) 'A Human RNA Viral Cysteine Proteinase That Depends upon a Unique $Zn^{2+}$-Binding Finger Connecting the Two Domains of a Papain-like Fold', *The Journal of Biological Chemistry* , 274: 14918–14925.

Holm, L. (2020) 'DALI and the Persistence of Protein Shape', *Protein Science : a Publication of the Protein Society*, 29: 128–140.

Holmes, E. C. (2009) *The Evolution and Emergence of RNA Viruses*. Oxford: Oxford University Press.

—— (2011) 'What Does Virus Evolution Tell Us about Virus Origins? ', *Journal of Virology*, 85: 5247–5251.

Hughes, A. L., and Friedman, R. (2005) 'Poxvirus Genome Evolution by Gene Gain and Loss', *Molecular Phylogenetics and Evolution*, 35: 186–195.

Hughes, P. J., and Stanway, G. (2000) 'The 2A Proteins of Three Diverse Picornaviruses Are Related to Each Other and to the H-rev107 Family of Proteins Involved in the Control of Cell Proliferation', *The Journal of General Virology* , 81: 201–207.

Humphrey, W., Dalke, A., and Schulten, K. (1996) 'VMD: Visual Molecular Dynamics', *Journal of Molecular Graphics*, 14: 33–38.

Jácome, R. et al. (2015) 'Structural Analysis of Monomeric RNA-Dependent Polymerases: Evolutionary and Therapeutic Implications', *PLoS ONE*, 10: e0139001.

James, M. N. G., Delbaere, L. T. J., and Brayer, G. D. (1978) 'Amino Acid Sequence Alignment of Bacterial and Mammalian Pancreatic Serine Proteases Based on Topological Equivalences', *Canadian Journal of Biochemistry*, 56: 396–402.

Jones, E. Y., Stuart, D. I., and Walker, N. P. C. (1989) 'Structure of Tumor Necrosis Factor', *Nature*, 338: 225–228.

Kalynych, S., Pálková, L., and Plevka, P. (2016a) 'The Structure of Human Parechovirus 1 Reveals an Association of the RNA Genome with the Capsid', *Journal of Virology*, 90: 1377–1386.

—— et al. (2016b) 'Virion Structure of Iflavirus Slow Bee Paralysis Virus at 2.6-Angstrom Resolution', *Journal of Virology*, 90: 7444–7455.

Kambol, R., Kabat, P., and Tristem, M. (2003) 'Complete Nucleotide Sequence of an Endogenous Retrovirus from the Amphibian, Xenopus laevis', *Virology*, 311: 1–6.

King, A. M. Q. et al. (2011) Virus Taxonomy. *Ninth Report of the International Committee on Taxonomy of Viruses*. London: Elsevier Academic Press

Koonin, E. V., Dolja, V. V., and Krupovic, M. (2015) 'Origins and Evolution of Viruses of Eukaryotes: The Ultimate Modularity', *Virology*, 479-480: 2–25.

Kreuze, J. F., Savenkov, E. I., and Valkonen, J. P. T. (2002) 'Complete Genome Sequence and Analyses of the Subgenomic RNAs of Sweet Potato Chlorotic Stunt Virus Reveal Several

New Features for the Genus Crinivirus', *Journal of Virology*, 76: 9260–9270.

LaPierre, L. A. et al. (1999) 'Sequence and Transcriptional Analyses of the Fish Retroviruses Walleye Epidermal Hyperplasia Virus Types 1 and 2: Evidence for a Gene Duplication', *Journal of Virology*, 73: 9393–9403.

Lazcano, A. (1995) 'Cellular Evolution during the Early Archean: What Happened between the Progenote and the Cenancestor? ', *Microbiología SEM*, 11: 185–198.

Lee, H. J. et al. (1991) 'The Complete Sequence (22 Kilobases) of Murine Coronavirus Gene 1 Encoding the Putative Proteases and RNA Polymerase', *Virology*, 180: 567–582.

Lesk, A. M., and Fordham, W. D. (1996) 'Conservation and Variability in the Structures of Serine Proteinases of the Chymotrypsin Family', *Journal of Molecular Biology*, 258: 501–537.

Li, W., and Godzik, A. (2006) 'Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences', *Bioinformatics (Oxford, England)*, 22: 1658–9.

Liljas, L. et al. (2002) 'Evolutionary and Taxonomic Implications of Conserved Structural Motifs between Picornaviruses', *Archives of Virology*, 147: 59–84.

Malik, H. S., and Eickbush, T. H. (2001) 'Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses', *Genome Research*, 11: 1187–1197.

Matthews, D. A. et al. (1994) 'Structure of Human Rhinovirus 3C Protease Reveals a Trypsin-like Polypeptide Fold, RNA-Binding Site, and Means for Cleaving Precursor Polyprotein', *Cell*, 77: 761–771.

Maynard Smith, J., and Szathmáry, E. (1995) *The Major Transitions in Evolution*. Oxford: Oxford University Press.

McGeoch, D., and Davison, J. (1999) 'Molecular Evolutionary History of the Herpesviruses', in E., Domingo, R.G., Webster, H.F., Holland (eds.) *Origin and Evolution of Viruses*, pp. 441–465. London: Academic Press.

Ohno, S. (1970) *Evolution by Gene Duplication*. New York: Springer-Verlag.

Petersen, J. F. W. et al. (1999) 'The Structure of the 2A Proteinase from a Common Cold Virus: A Proteinase Responsible for the Shut-off of Host-Cell Protein Synthesis', *The EMBO Journal*, 18: 5463–5475.

Pettersen, E. F. et al. (2004) 'UCSF Chimera–a Visualization System for Exploratory Research and Analysis', *Journal of Computational Chemistry*, 25: 1605–12.

Porter, A. (1993) 'Replication and Inhibition of Host Cell Functions', *Journal of Virology*, 67: 6917–6921.

Rawlings, N. D. et al. (2018) 'The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison with Peptidases in the PANTHER Database', *Nucleic Acids Research*, 46: D624–D632.

Rambaut, A. (2014) *FigTree v1.4.2, a Graphical Viewer of Phylogenetic Trees*. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. <http://tree.bio.ed.ac.uk/software/figtree/> (accessed February 2021).

Reanney, D. C. (1982) 'The Evolution of RNA Viruses', *Annual Review of Microbiology*, 36: 47–73.

Reyes-Prieto, F. et al. (2012) 'Coenzymes, Viruses and the RNA World', *Biochimie*, 94: 1467–1473.

Roberts, E. et al. (2006) 'MultiSeq: Unifying Sequence and Structure Data for Evolutionary Analysis', *BMC Bioinformatics*, 7: 382.

Rossmann, M. G., and Johnson, J. E. (1989) 'Icosahedral RNA Virus Structure', *Annual Review of Biochemistry*, 58: 533–73.

Russell, R. B., and Barton, G. J. (1992) 'Multiple Protein Sequence Alignment from Tertiary Structure Comparisons: Assignment of Global and Residue Confidence Levels', *Proteins* , 14: 309–323.

Sabin, C. et al. (2016) 'Structure of Aichi Virus 1 and Its Empty Particle: Clues to Kobuvirus Genome Release Mechanism', *Journal of Virology*, 90: 10800–10810.

Saikatendu, K. S. et al. (2005) 'Structural Basis of Severe Acute Respiratory Syndrome Coronavirus ADP-Ribose-1"-Phosphate Dephosphorylation by a Conserved Domain of nsP3', *Structure (London, England : 1993)*, 13: 1665–1675.

Shackelton, L. A., and Holmes, E. C. (2004) 'The Evolution of Large DNA Viruses: Combining Genomic Information of Viruses and Their Hosts', *TRENDS in Microbiology*, 12: 458–465.

Simon-Loriere, E., and —— (2013) 'Gene Duplication is Infrequent in the Recent Evolutionary History of RNA Viruses', *Molecular Biology and Evolution*, 30: 1263–1269.

Subbiah, S., Laurents, D. V., and Levitt, M. (1993) 'Structural Similarity of DNA-Binding Domains of Bacteriophage Repressors and the Globin Core', *Current Biology : Cb*, 3: 141–148.

Sun, Y. et al. (2009) 'Crystal Structure of Porcine Reproductive and Respiratory Syndrome Virus Leader Protease Nsp$\alpha$', *Journal of Virology*, 83: 10931–10940.

Sutton, G. et al. (2004) 'The nsp9 Replicase Protein of SARS-Coronavirus, Structure and Funcional Insights', *Structure (London, England : 1993)*, 12: 341–353.

Tan, J. et al. (2009) 'The SARS-Unique Domain (Sud) of SARS Coronavirus Contains Two Macrodomains That Bind G-Quadruplexes', *PLoS Pathogens*, 5: e1000428.

Taylor, J., and Raes, J. (2004) 'Duplication and Divergence: The Evolution of New Genes and Old Ideas', *Annual Review of Genetics*, 38: 615–643.

Tseng, C. H., Knowles, N. J., and Tsai, H. J. (2007) 'Molecular Analysis of Duck Hepatitis Virus Type 1 Indicates That It Should Be Assigned to a New Genus', *Virus Research*, 123: 190–203.

Tria, F. D. K., Landan, G., and Dagan, T. (2017) 'Phylogenetic Rooting Using Minimal Ancestor Deviation', *Nature Ecology & Evolution*, 1: 193.

Tristem, M. et al. (1990) 'Origin of Vpx in Lentiviruses', *Nature*, 347: 341–342.

Tzanetakis, I. E., and Martin, R. R. (2007) 'Strawberry Chlorotic Fleck: Identification and Characterization of a Novel Closterovirus Associated with the Disease', *Virus Research*, 124: 88–94.

——, Postman, J. D., and Martin, R. R. (2005) 'Characterization of a Novel Member of the Family Closteroviridae from Mentha Spp', *Phytopathology*, 95: 1043–1048.

Walker, P. J. et al. (1992) 'The Genome of Bovine Ephemeral Fever Rhabdovirus Contains Two Related Glycoprotein Genes', *Virology*, 191: 49–61.

Wery, J. P. et al. (1994) 'The Refined Three-Dimensional Structure of an Insect Virus at 2.8 a Resolution', *Journal of Molecular Biology*, 235: 565–586.

Willemsen, A. et al. (2016) 'Predicting the Stability of Homologous Gene Duplications in a Plant RNA Virus', *Genome Biology and Evolution*, 8: 3065–3082.

—— et al. (2017) '2b or Not 2b: Experimental Evolution of Functional Exogenous Sequences in a Plant RNA Virus', *Genome Biol. Evol*, 9: 297–310.

Xue, F. et al. (2010) 'The Crystal Structure of Porcine Reproductive and Respiratory Syndrome Virus Nonstructural Protein Nsp1$\beta$ Reveals a Novel Metal-Dependent Nuclease', *Journal of Virology*, 84: 6461–6471.

Ye, Y., and Godzik, A. (2003) ‘Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists’, *Bioinformatics*, 19: ii246–ii255.

Zhang, J. (2003) ‘Evolution by Gene Duplication: An Update’, *Trends in Ecology & Evolution* , 18: 292–298.

Ziebuhr, J., Snijder, E. J., and Gorbalenya, A. E. (2000) ‘Virus-Encoded Proteinases and Proteolytic Processing in the *Nidovirales*’, *The Journal of General Virology* , 81: 853–879.

——, Thiel, V., and —— (2001) ‘The Autocatalytic Release of a Putative RNA Virus Transcription Factor from Its Polyprotein Precursor Involves Two Paralogous Papain-like Proteases That Cleave the Same Peptide Bond’, *The Journal of Biological Chemistry* , 276: 33220–33232.