



Poorly Conserved P15 Proteins of Cileviruses Retain Elements of Common Ancestry and Putative Functionality: A Theoretical Assessment on the Evolution of Cilevirus Genomes

OPEN ACCESS

Edited by:

Ken Komatsu,
Tokyo University of Agriculture
and Technology, Japan

Reviewed by:

Ken-Taro Sekine,
University of the Ryukyus, Japan
Alejandro Olmedo-Velarde,
University of Hawaii, United States

*Correspondence:

Pedro L. Ramos-González
plrg1970@gmail.com
Juliana Freitas-Astua
juliana.astua@embrapa.br

†These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Plant Pathogen Interactions,
a section of the journal
Frontiers in Plant Science

Received: 07 September 2021

Accepted: 18 October 2021

Published: 05 November 2021

Citation:

Ramos-González PL, Pons T,
Chabi-Jesus C, Arena GD and
Freitas-Astua J (2021) Poorly
Conserved P15 Proteins of Cileviruses
Retain Elements of Common Ancestry
and Putative Functionality:
A Theoretical Assessment on
the Evolution of Cilevirus Genomes.
Front. Plant Sci. 12:771983.
doi: 10.3389/fpls.2021.771983

Pedro L. Ramos-González^{1†}, Tirso Pons^{2†}, Camila Chabi-Jesus^{1,3}, Gabriella Dias Arena¹ and Juliana Freitas-Astua^{1,4*}

¹ Laboratório de Biologia Molecular Aplicada, Instituto Biológico de São Paulo, São Paulo, Brazil, ² National Centre for Biotechnology (CNB-CSIC), Madrid, Spain, ³ Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo, Piracicaba, Brazil, ⁴ Embrapa Mandioca e Fruticultura, Cruz das Almas, Brazil

The genus *Cilevirus* groups enveloped single-stranded (+) RNA virus members of the family *Kitaviridae*, order *Martellivirales*. Proteins P15, scarcely conserved polypeptides encoded by cileviruses, have no apparent homologs in public databases. Accordingly, the open reading frames (ORFs) *p15*, located at the 5'-end of the viral RNA2 molecules, are considered orphan genes (ORFans). In this study, we have delved into ORFs *p15* and the relatively poorly understood biochemical properties of the proteins P15 to posit their importance for viruses across the genus and theorize on their origin. We detected that the ORFs *p15* are under purifying selection and that, in some viral strains, the use of synonymous codons is biased, which might be a sign of adaptation to their plant hosts. Despite the high amino acid sequence divergence, proteins P15 show the conserved motif [FY]-L-x(3)-[FL]-H-x-x-[LIV]-S-C-x-C-x(2)-C-x-G-x-C, which occurs exclusively in members of this protein family. Proteins P15 also show a common predicted 3D structure that resembles the helical scaffold of the protein ORF49 encoded by radinoviruses and the phosphoprotein C-terminal domain of mononegavirids. Based on the 3D structural similarities of P15, we suggest elements of common ancestry, conserved functionality, and relevant amino acid residues. We conclude by postulating a plausible evolutionary trajectory of ORFans *p15* and the 5'-end of the RNA2 of cileviruses considering both protein fold superpositions and comparative genomic analyses with the closest kitaviruses, negeviruses, nege/kita-like viruses, and unrelated viruses that share the ecological niches of cileviruses.

Keywords: *Kitaviridae*, orphan ORF, structure-based phylogenetic analysis, synonymous codon usage bias, horizontal gene transfer, *Dichorhavirus*, miniproteins, small ORF

INTRODUCTION

Orphan genes (ORFans) code for proteins with unrecognized homologs in any other species (Fischer and Eisenberg, 1999; Arendsee et al., 2014). They are present in a variety of organisms, including prokaryotes, eukaryotes, and viruses, reaching up to one-third of the genome of some microorganisms and 50% to two-thirds in mimiviruses (Fischer and Eisenberg, 1999; Kaur Saini and Fischer, 2007; Yin and Fischer, 2008; Tautz and Domazet-Lošo, 2011; Colson et al., 2017). The quick origin of ORFans may provide adaptive mechanisms conferring rapid fitness to changing environments (Entwistle et al., 2019). Most of the known proteins encoded by ORFans interact with conserved proteins like transcription factors or receptors, acting as toxins, or as modulators of metabolic or regulatory networks (Singh and Wurtele, 2020).

Kitaviridae, order *Martellivirales*, is a family of heterogeneous plant-infecting viruses displaying two, three, or four segments of single-stranded positive-sense RNA molecules as genomes, which have been assigned into the genera *Cilevirus*, *Higrevirus*, and *Blunervirus*, respectively (Quito-Avila et al., 2021). Taken together, kitavirids have been detected in a relatively narrow range of natural hosts, where they normally produce local infections mainly characterized by chlorotic and/or necrotic lesions that in some cases resemble the outcome of a hypersensitive-like response (Arena et al., 2016, 2020). However, regardless of their failure to consummate the systemic movement through the plants, kitavirids pose serious threats to major crops such as citrus (Ramos-González et al., 2018). Likely most kitaviruses are transmitted by mites, but an effective transmission by *Brevipalpus* mites has been only confirmed for cileviruses (Kitajima and Alberti, 2014; Quito-Avila et al., 2021).

With several strains and isolates belonging to three species already described at the biological and molecular level, *Cilevirus* is the best-studied genus of kitavirids (Freitas-Astúa et al., 2018; Quito-Avila et al., 2021). Virions of citrus leprosis virus C (CiLV-C), citrus leprosis virus C2 (CiLV-C2), and passion fruit green spot virus (PFGSV) are enveloped, short bacilliform particles that encapsidate two poly-adenylated RNA molecules (Locali-Fabris et al., 2006; Roy et al., 2013; Kitajima and Alberti, 2014; Ramos-González et al., 2020). The canonical genome of cileviruses contains six open reading frames (ORF) distributed in two RNA segments. ORFs *RdRp* (RNA dependent-RNA polymerase) and *p29* (putative coat protein) are in RNA 1 (≈ 9.0 kb); whilst the ORFs *p32* (movement protein), *p61*, *p24*, and *p15* are contained in the RNA2 (≈ 5.0 kb) (Figure 1). RNA2 in certain cileviruses also includes some accessory ORFs, for instance, *p7* in CiLV-C2, and *p11-13* in PFGSV. ORFs *p61* and *p24* encode proteins that are likely involved in the virion structure (Solovyev and Morozov, 2017). *P24* is a transmembrane (TM) protein that is also probably an integral component of the virus factory-like membranes, whereas *P61* is a putative glycoprotein that in the case of CiLV-C acts as a viral effector in plants (Kuchibhatla et al., 2014; Arena et al., 2016, 2020; Leastro et al., 2018; Vinokurov and Koloniuk, 2019). ORF *p61* is taxonomically restricted to cileviruses, whereas, besides in the kitavirids, *p24* is also present in an increasingly discovered number of

arthropod-infecting viruses including mosquito-specific viruses of the proposed group Negevirus (clades Nelorpivirus and Sandewavirus) (Vasilakis et al., 2013; Kuchibhatla et al., 2014; Kondo et al., 2020). *P24* displays the highly conserved motif SP24 (TM-SP24, structure protein of 24 kDa, Pfam code PF16504). The glycoproteins *P61* of cileviruses lack the motifs DiSA (Pfam code PF19226) and DiSB-ORF2_chro B (Pfam code PF16506) detected in the glycoproteins encoded by their taxonomically related arthropod-infecting viruses (Kuchibhatla et al., 2014). ORF *p15*, only present in cileviruses, encodes the smallest annotated polypeptide across the genomes of these viruses (Ramos-González et al., 2020).

In this study, based on an integrator process comprising genomics, population genetics, usage of synonymous codons, and tridimensional (3D) structures, we elaborated a hypothesis about the conserved putative functionality of the *P15* proteins across the genus. Following a chain of logical thinking, we theorize on the origin and evolution of the ORF *p15* considering the genomic context in the 5'-end of the RNA2 in cileviruses. A meaningful part of the study has been founded on the comparisons of the protein folds using the principles of structure-based phylogenetic analyses.

THE OPEN READING FRAMES *p15* OF CILEVIRUSES ARE UNDER PURIFYING SELECTION AND ENCODE POORLY CONSERVED PROTEINS

Citrus leprosis virus C is the best-characterized cilevirus at both molecular and epidemiological levels. Its population is subdivided into three clades (Ramos-González et al., 2016; Chabi-Jesus et al., 2021). Average identity values of genomic nucleotide sequences between CiLV-C isolates belonging to different lineages range from 85 to 89%. Contrastingly, a highly conserved segment of approximately 1.5 kb, showing almost 100% of identity, is present at the 5'-end of their RNA2 molecules and comprises the ORF *p15* (393 nts) (Figure 1).

Based on the analysis of 58 nucleotide sequences of *p15* (Supplementary Table 1) using MEGA v. 10.1.8 (Kumar et al., 2018) and DnaSP v. 6.12.03 (Rozas et al., 2017) software suites, we confirmed that CiLV-C *p15* variability is very low with an overall genetic distance (*D*) and nucleotide diversity (π) ≤ 0.01 (Figure 2). According to population analyses, this low variability may result from constant purges of the non-synonymous substitutions by purifying selection ($\omega < 1$), and/or intra-species recombination processes, as already described (Ramos-González et al., 2016; Chabi-Jesus et al., 2021). *P15* encoded by CiLV-C shows Cys residues that were earlier suggested as taking part in a putative Zn-finger structure (Ramos-González et al., 2016) (Figure 3). Transiently expressed *P15* of CiLV-C of the strain CRD enters into the cell nucleus likely by passive diffusion and the formation of homodimers and heterodimers with *P29* and the viral movement proteins can be detected in the cytosol of agroinfiltrated *Nicotiana benthamiana* plants (Leastro et al., 2018). When expressed from a viral vector, the infected

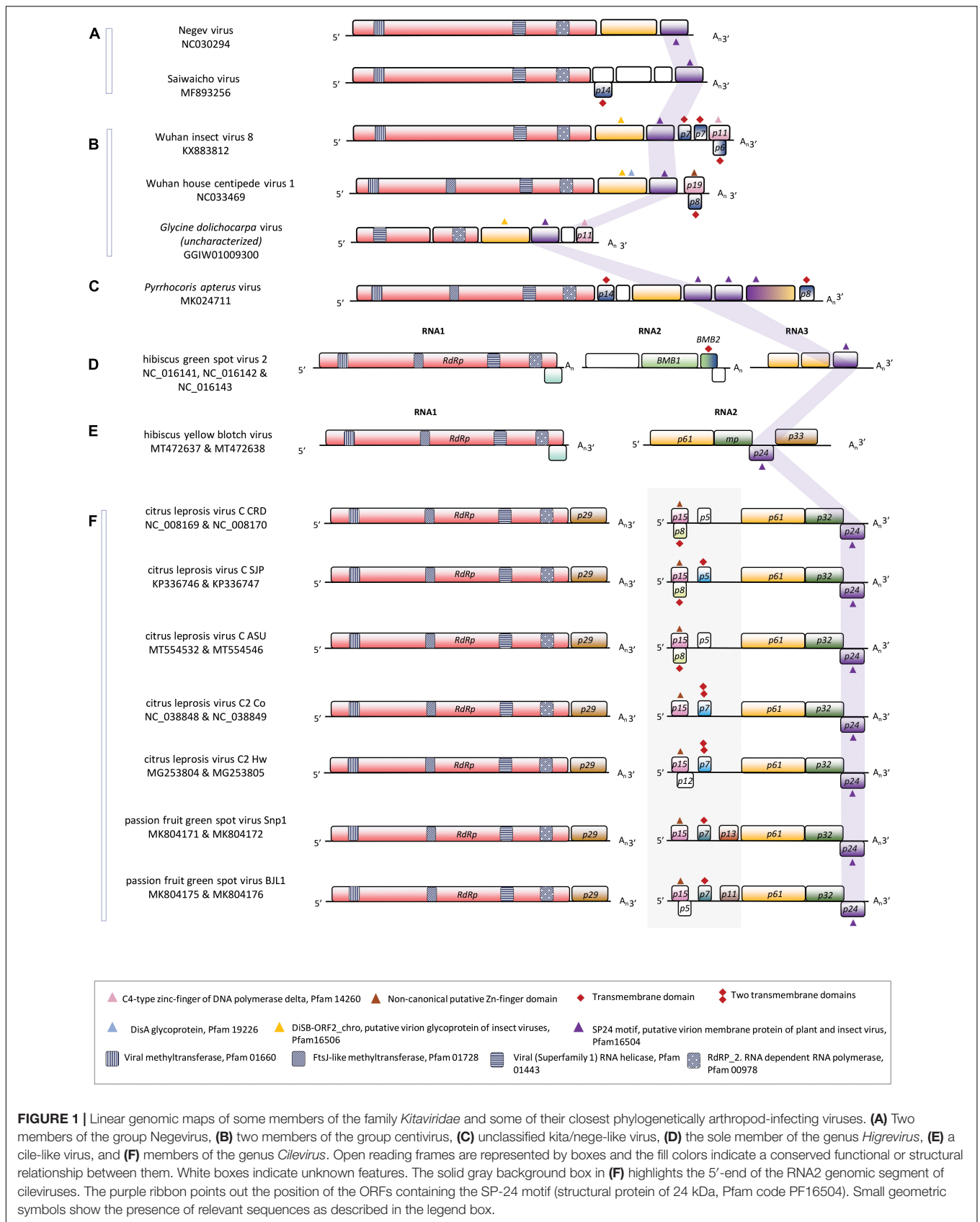


FIGURE 1 | Linear genomic maps of some members of the family *Kitaviridae* and some of their closest phylogenetically arthropod-infecting viruses. **(A)** Two members of the group Negevirus, **(B)** two members of the group centivirus, **(C)** unclassified kita/nege-like virus, **(D)** the sole member of the genus *Higrevirus*, **(E)** a cile-like virus, and **(F)** members of the genus *Cilevirus*. Open reading frames are represented by boxes and the fill colors indicate a conserved functional or structural relationship between them. White boxes indicate unknown features. The solid gray background box in **(F)** highlights the 5'-end of the RNA2 genomic segment of cileviruses. The purple ribbon points out the position of the ORFs containing the SP-24 motif (structural protein of 24 kDa, Pfam code PF16504). Small geometric symbols show the presence of relevant sequences as described in the legend box.

N. benthamiana plants show stunted growth and enhanced necrosis in their younger leaves suggesting pathophysiological disorders caused by its ectopic expression. A possible role as an RNA silencing suppressor was also suggested for P15 of CiLV-C strain CRD (Leastro et al., 2020).

Analyses conducted in this work indicated that despite the low number of isolates available for CiLV-C2 and PfGSV (Supplementary Table 1), intra-species variability of their *p15* is almost 10-fold higher than among the CiLV-C isolates. Yet, as well as for CiLV-C, ORFs *p15* from CiLV-C2 and PfGSV show signatures of purifying selection ($\omega < 1$). P15 proteins of these two viruses are phylogenetically closer to each other and more distant from that in CiLV-C (Figure 2).

Pairwise comparisons of the P15 amino acid (aa) sequences reveal very low identity values, approximately 14 or 20%, between the proteins from CiLV-C and PfGSV or CiLV-C2, respectively. Higher values, still moderate, approximately 55% aa sequence identity, are detected in the comparisons between P15 from CiLV-C2 and PfGSV (Ramos-González et al., 2016, 2020). Remarkably, the aa pairwise sequence identity values among P15 of CiLV-C and those from CiLV-C2 and PfGSV meaningfully deviate from those commonly observed within the cilevirus proteomes. The identity values between the P15 of PfGSV and CiLV-C are less than half of those shown by the next less conserved proteins among cileviruses, i.e., P29 and P61 (31–32%) (Ramos-González et al., 2020).

OPEN READING FRAME *p15* DISPLAYS SYNONYMOUS CODON USAGE BIAS THAT MIGHT REFLECT ADAPTION TO HOST-SPECIFIC CODON USAGE PATTERNS

Codon usage bias (CUB) refers to differences in the frequency of occurrence of synonymous codons in coding sequences (Behura and Severson, 2013). In viruses, CUB reflects changes in gene expression that occur as a consequence of, for instance, the interplay and co-evolution with their hosts, which may increase the viral fitness (Biswas et al., 2019; He et al., 2019; Khandia et al., 2019). Several indices, such as the Effective Number of codons (ENc), the Relative Synonymous Codon Usage (RSCU), COdon Usage Similarity Index (COUSIN), and Relative Codon Deoptimization Index (RCDI) are used to quantify the CUB (Sharp and Li, 1987; Wright, 1990; Bourret et al., 2019). These indices enable the comparison of genes within a genome and across different genotypes and species revealing elements about their evolution.

Effective Number of codons values range from 20 to 61, with a value of 20 indicating extreme bias, whereas, inversely, 61 indicates non-biased codon usage (Wright, 1990). Meanwhile, an RSCU value higher or lower than 1 denotes that the codon has a positive or negative CUB, respectively, and those with an RSCU value equal to the unit are randomly chosen. In this study, ENc and RSCU values were calculated using the

web servers COUSIN¹ and CAIcal², respectively. Analyses of both ENc and RSCU of the cilevirus ORFs revealed codon usage bias. Even though the means of the ENc values of ORFs both at virus and genus levels were generally higher than 50, some viral isolates and ORFs, e.g., PfGSV_Snp1 and ORF *p15*, had mean values lower than 50 which are considered an indicator of skewed codon usage. Within this group, the ORF *p15* of CiLV-C2_Hw displayed a distinctly low ENc value, ≈ 35 (Figure 4A and Supplementary Table 2), which is generally accepted as a mark of genes with significant codon usage bias (Comeron and Aguadé, 1998; Butt et al., 2016). The values of RSCU were computed for every codon of each ORF by the respective genomes (Supplementary Table 3). When the RSCU patterns were hierarchically clustered, most of the ORFs were grouped in a large branch of the generated dendrogram. ORFs *p15*, however, were generally distributed in two small branches shared also with the ORFs *p32* and *p24* of a few isolates (Figure 4B). Differently, the ORF *p15* of CiLV-C2_Hw was clustered together with the ORFs *RdRp* and *p61* of that isolate, suggesting that they could have undergone a comparable mutational selection.

To get further insight into the codon preferences of *p15*, we assessed and compared the CUB of each ORF in each cilevirus strain using COUSIN¹ and the ratio RCDI/eRCDI² (RCDI/expected RCDI) (Puigbò et al., 2010). It should be noted that while ENc and RSCU values exclusively rely on the coding sequence of the analyzed gene, COUSIN and the ratio RCDI/eRCDI assess the similarity of the codon usage patterns among the coding sequences of at least two genes. In this study, the COUSIN score (Bourret et al., 2019) estimates the codon usage preferences of a viral ORF compared with those of a reference, the viral host, normalized over a null hypothesis of equal usage of synonymous codons. Sequences with scores ≥ 1 display adaptation to the host cellular machinery and higher values might reflect an increasing level of expression. RCDI reflects the similarity of the codon usage between a given coding sequence and a reference genome (Mueller et al., 2006). A lower RCDI value shows the best rate of viral gene translation in the host as well as being indicative of the possible co-evolution of virus and host genomes. The eRCDI value is the RCDI corresponding to a random sequence generated with similar G + C content and amino acid composition to the analyzed sequence and provides a threshold that allows for statistical analysis of the calculated RCDI values. A lower ratio RCDI/eRCDI means a better viral ORF adaptation to its host.

The values of COUSIN and the ratio RCDI/eRCDI were calculated using the set of host plants whose codon usage data could be directly retrieved from, or calculated using available genomic information in, public databases. Codon usage frequencies of some natural and experimental hosts of cileviruses, i.e., citrus (*Citrus × sinensis*), passion fruit (*Passiflora* spp.), orchid (*Oncidium* spp.) (Kitajima et al., 2010), *Arabidopsis thaliana* (Arena et al., 2017), and *Phaseolus vulgaris*

¹<http://cousin.ird.fr/index.php>

²<https://ppuigbo.me/programs/CAIcal/>

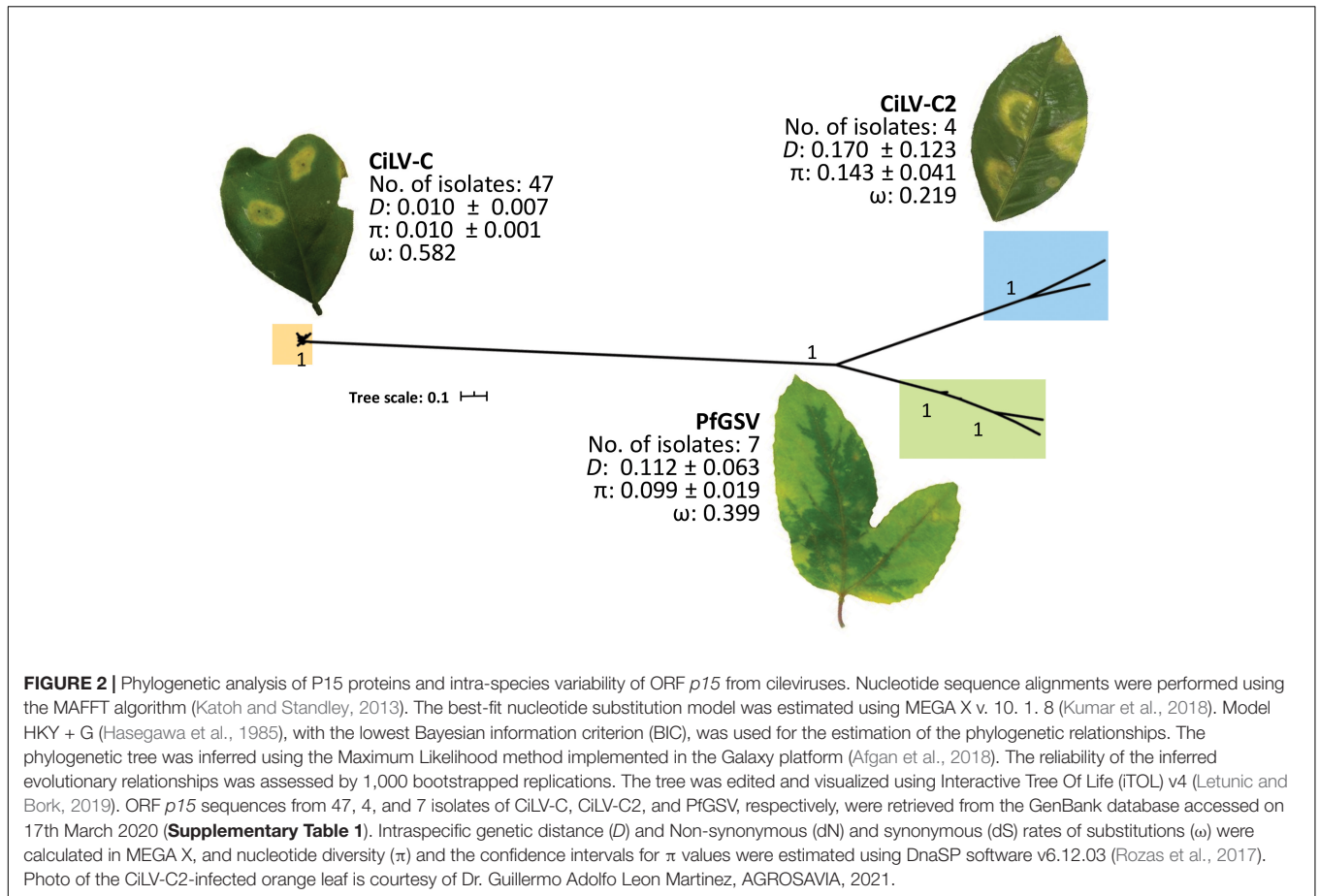


FIGURE 2 | Phylogenetic analysis of P15 proteins and intra-species variability of ORF *p15* from cileviruses. Nucleotide sequence alignments were performed using the MAFFT algorithm (Kato and Standley, 2013). The best-fit nucleotide substitution model was estimated using MEGA X v. 10. 1. 8 (Kumar et al., 2018). Model HKY + G (Hasegawa et al., 1985), with the lowest Bayesian information criterion (BIC), was used for the estimation of the phylogenetic relationships. The phylogenetic tree was inferred using the Maximum Likelihood method implemented in the Galaxy platform (Afgan et al., 2018). The reliability of the inferred evolutionary relationships was assessed by 1,000 bootstrapped replications. The tree was edited and visualized using Interactive Tree Of Life (ITOL) v4 (Letunic and Bork, 2019). ORF *p15* sequences from 47, 4, and 7 isolates of CiLV-C, CiLV-C2, and PFGSV, respectively, were retrieved from the GenBank database accessed on 17th March 2020 (Supplementary Table 1). Intraspecific genetic distance (D) and Non-synonymous (dN) and synonymous (dS) rates of substitutions (ω) were calculated in MEGA X, and nucleotide diversity (π) and the confidence intervals for π values were estimated using DnaSP software v6.12.03 (Rozas et al., 2017). Photo of the CiLV-C2-infected orange leaf is courtesy of Dr. Guillermo Adolfo Leon Martinez, AGROSAVIA, 2021.

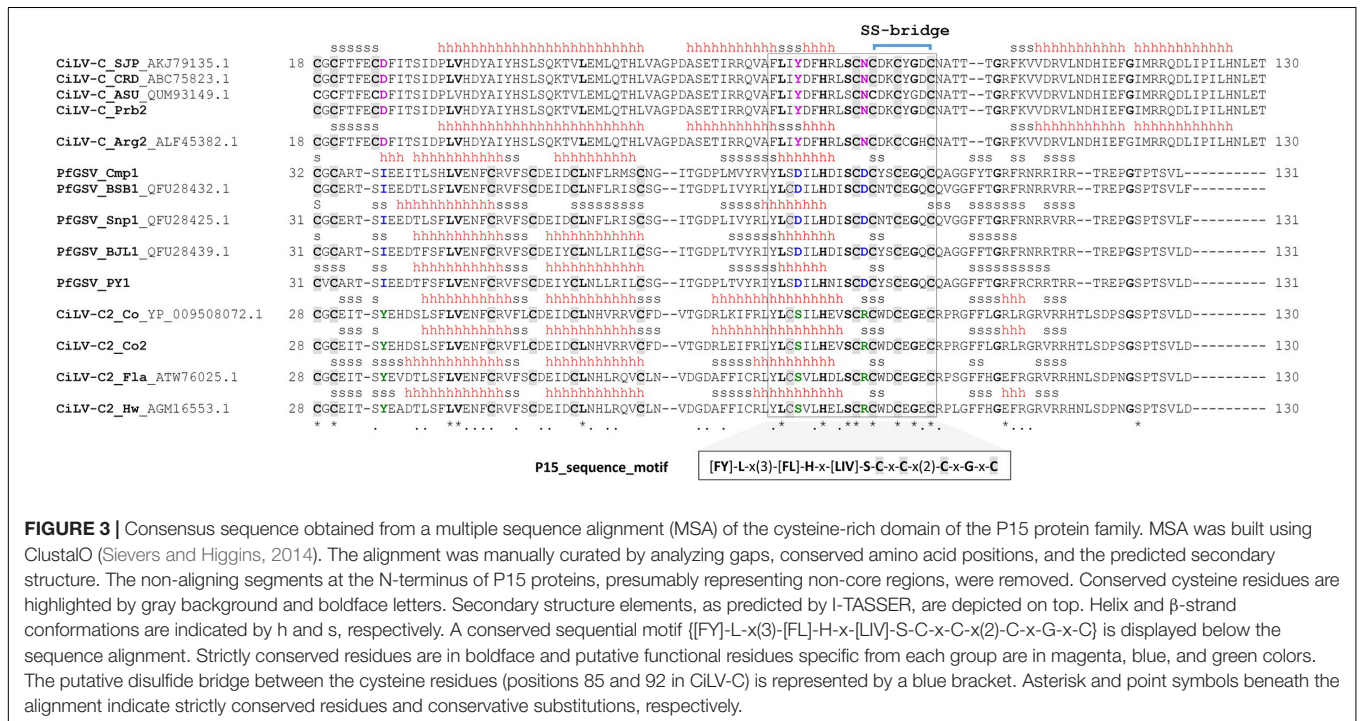


FIGURE 3 | Consensus sequence obtained from a multiple sequence alignment (MSA) of the cysteine-rich domain of the P15 protein family. MSA was built using ClustalO (Sievers and Higgins, 2014). The alignment was manually curated by analyzing gaps, conserved amino acid positions, and the predicted secondary structure. The non-aligning segments at the N-terminus of P15 proteins, presumably representing non-core regions, were removed. Conserved cysteine residues are highlighted by gray background and boldface letters. Secondary structure elements, as predicted by I-TASSER, are depicted on top. Helix and β -strand conformations are indicated by h and s, respectively. A conserved sequential motif $\{[FY]-L-x(3)-[FL]-H-x-[LIV]-S-C-x-C-x(2)-C-x-G-x-C\}$ is displayed below the sequence alignment. Strictly conserved residues are in boldface and putative functional residues specific from each group are in magenta, blue, and green colors. The putative disulfide bridge between the cysteine residues (positions 85 and 92 in CiLV-C) is represented by a blue bracket. Asterisk and point symbols beneath the alignment indicate strictly conserved residues and conservative substitutions, respectively.

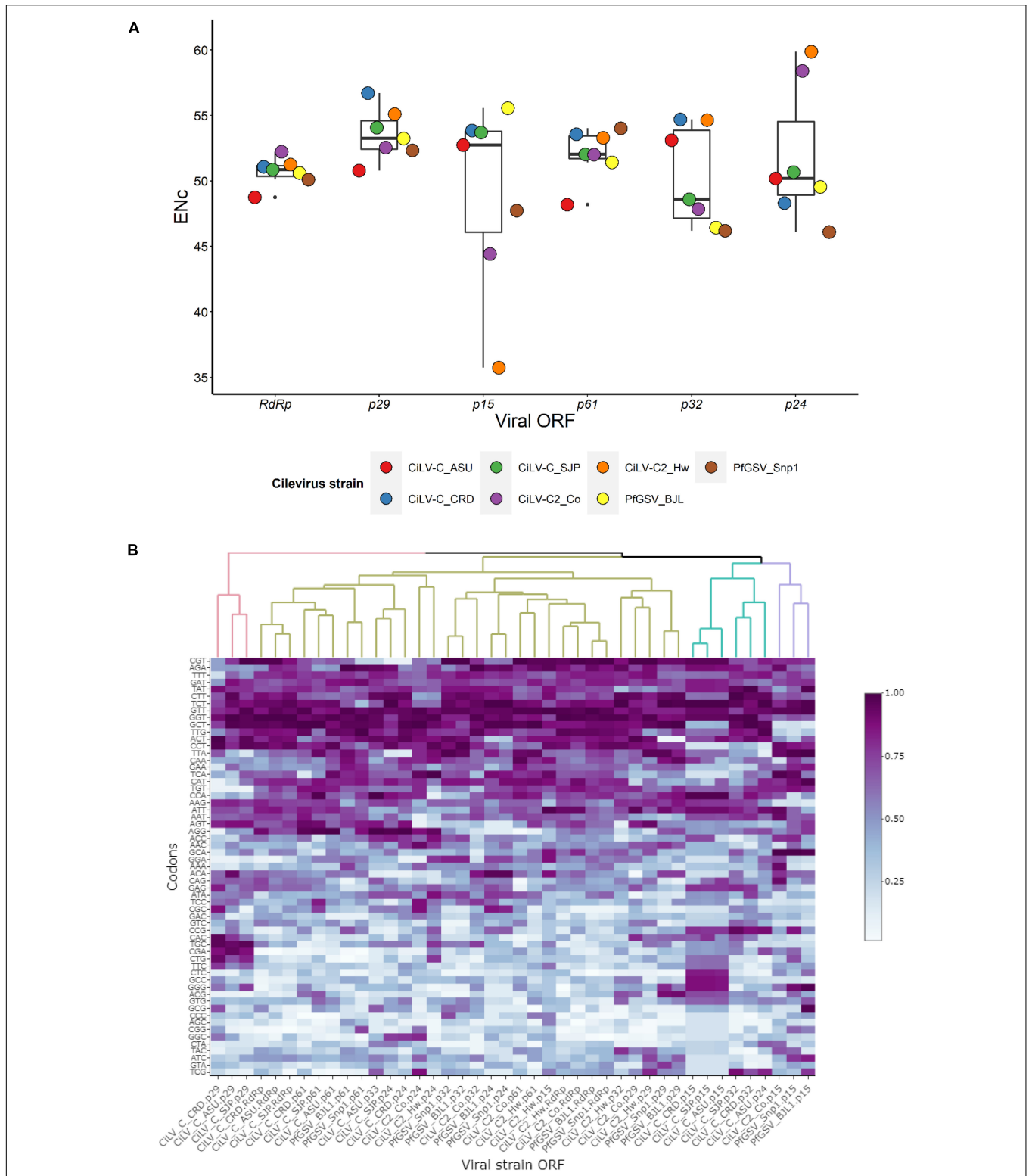


FIGURE 4 | Codon usage bias of cilevirus ORFs. **(A)** Boxplot of Effective Number of codons (ENC). ENC values were calculated using COUSIN webserver (<http://cousin.ird.fr/index.php>) and plotted in Rstudio (Rstudio_team, 2021). **(B)** RSCU profiles across ORF of seven different isolates of three species of cileviruses. RSCU values were obtained using the CAIcal webserver (<https://ppuigbo.me/programs/CAIcal/>). The heatmap was drawn in Rstudio using the “heatmapply” function (Galili et al., 2018). RSCU values were transformed using the “percentize” method and dist_method = “Euclidean” and hclust_method = “complete” were selected. Darker purple represents higher RSCU values.

(Garita et al., 2013) were retrieved from Codon Usage Database³ or calculated using countcodon v4⁴ (Supplementary Table 4). Coding sequence information of *Passiflora edulis*, approximately 300 ORFs, was retrieved from the CoGe platform⁵. Orchids were included as cilevirus natural hosts since the infection by PfGSV of plants from two orchids species collected in Brazil was molecularly confirmed (unpublished results). For the sake of the comparisons, the analyses were performed considering both the sum of natural and experimental host plants and two subsets of plants, each comprising natural or experimental hosts. Altogether, the analysis included plants belonging to five families, including perennial woody (citrus) and perennial herbaceous (orchids) plants, short-lived evergreen perennial vine (passion fruit), and herbaceous annual (common beans and Arabidopsis) plants. Values of COUSIN and the ratio RCDI/eRCDI were visualized and compared using hierarchical clustering. The accuracy of the assessment could be potentially increased as new annotated genomes of cilevirus host plants are available.

Regardless of the viral species, the analysis by ORFs using the complete set of host plants revealed a >1 bias in the mean COUSIN values of *p61* (1.28 ± 0.44), *RdRp* (1.19 ± 0.36), *p32* (1.15 ± 0.37), and *p24* (1.09 ± 0.38), whereas mean values below or near to 1 were observed for the ORFs *p15* (0.89 ± 0.53) and *p29* (0.93 ± 0.35) (Supplementary Table 5A). COUSIN values of *p15*, however, showed the highest variability with a coefficient of variation surpassing 50%. Some CiLV-C2 ORFs reached COUSIN values higher than 1.5, whereas the minimum values, e.g., -0.15, corresponded to PfGSV isolates. The same trend of COUSIN values was obtained when calculations were performed with the two subgroups of plants conveniently separated into natural and experimental hosts (Supplementary Tables 5B,C). In a global analysis across the genus, the COUSIN profile of *p15* resembled that showed by *p32*, which encodes the viral movement protein (Figure 5A).

The study of the ratio RCDI/eRCDI of the cilevirus ORFs indicated that *RdRp* and *p61* bear the lowest mean values, followed in ascending order by *p24*, *p29*, and *p32*, and finally *p15* (Supplementary Table 6A). In the particular comparisons of *p15*, values from CiLV-C2 were collectively the lowest ones reached. In the evaluations using *P. vulgaris*, the ratios RCDI/eRCDI of *p15* genes from CiLV-C2 isolates were below the threshold of 0.9, similar to those held by other viral ORFs, e.g., *p29*, *p24*, and *p32*. The analyses suggest that the CUB of all those ORFs might be the outcome of equivalent processes of strain/host adaptation.

Regardless of the subgroup of studied plants, i.e., the subsets of natural, experimental, or the array containing all host plants, the mean values of the ratio RCDI/eRCDI of *p15* were higher than those shown by the remaining viral ORFs (Supplementary Tables 6B,C). At the genus level, when analyzed through hierarchical clustering, the RCDI/eRCDI profile of

p15 is unique among those shown by any other ORF of cileviruses (Figure 5B).

P15 PROTEINS DISPLAY A UNIQUE CONSERVED SEQUENCE MOTIF

Even though a low global sequence identity of less than 25% was observed for P15 proteins, the consensus sequence obtained from a multiple sequence alignment of the cysteine-rich domain reveals a stretch of highly conserved residues organized in the sequence motif {[FY]-L-x(3)-[FL]-H-x-[LIV]-S-C-x-C-x(2)-C-x-G-x-C} (Figure 3 and Supplementary Table 7). Besides four full-conserved residues of cysteines and one of histidine, the consensus sequence motif also contains three other invariable amino acids: leucine, serine, and glycine. A ScanProsite⁶ (de Castro et al., 2006) search using the conserved sequence motif in the UniProtKB/Swiss-Prot⁷ (release 07-Apr-2021: 564,638 entries) and UniProtKB/TrEMBL⁸ (release 07-Apr-2021: 214,406,399 entries) databases retrieved 30 hits spanned in 30 sequences, all of them exclusively corresponding to the P15 proteins from CiLV-C, CiLV-C2, and PfGSV. Additionally, no hits were retrieved from a similar search but against a randomized UniProtKB/Swiss-Prot database, i.e., reversed option. According to ScanProsite, the conserved sequence motif has an approximate number of expected random matches of 6.97×10^{-08} in 100,000 sequences (50,000,000 residues). Altogether, these results suggested a low probability of P15 matches by chance and, thus, highlighted the singular occurrence of the identified cysteine-rich domain in the P15 protein family.

THREE-DIMENSIONAL STRUCTURE PREDICTION SUGGESTS A HELICAL BUNDLE-LIKE SCAFFOLD FOR THE P15 PROTEINS

The existence of a conserved sequence motif among the P15 proteins dropped a hint that all of them might exhibit a similar fold. Interestingly, this motif overlaps a predicted α -turn- α motif in the secondary structure of P15 in CiLV-C and CiLV-C2, whereas in PfGSV the second α -helix is missing (Figure 3). The predictions by evaluation of sequence-structure fitness (Ouzounis et al., 1993) with proteins of known 3D structures of a representative group of ten P15 proteins from the three species of cileviruses were carried out by the iterative threading assembly refinement (I-TASSER) server⁹ (Yang and Zhang, 2015). For the computational structural analyses, experimentally determined 3D structures were downloaded from the RCSB PDB database¹⁰ and the protein 3D structure alignments were done using the

⁶<https://prosite.expasy.org/scanprosite/>

⁷<https://www.uniprot.org/statistics/Swiss-Prot>

⁸<https://www.uniprot.org/statistics/TrEMBL>

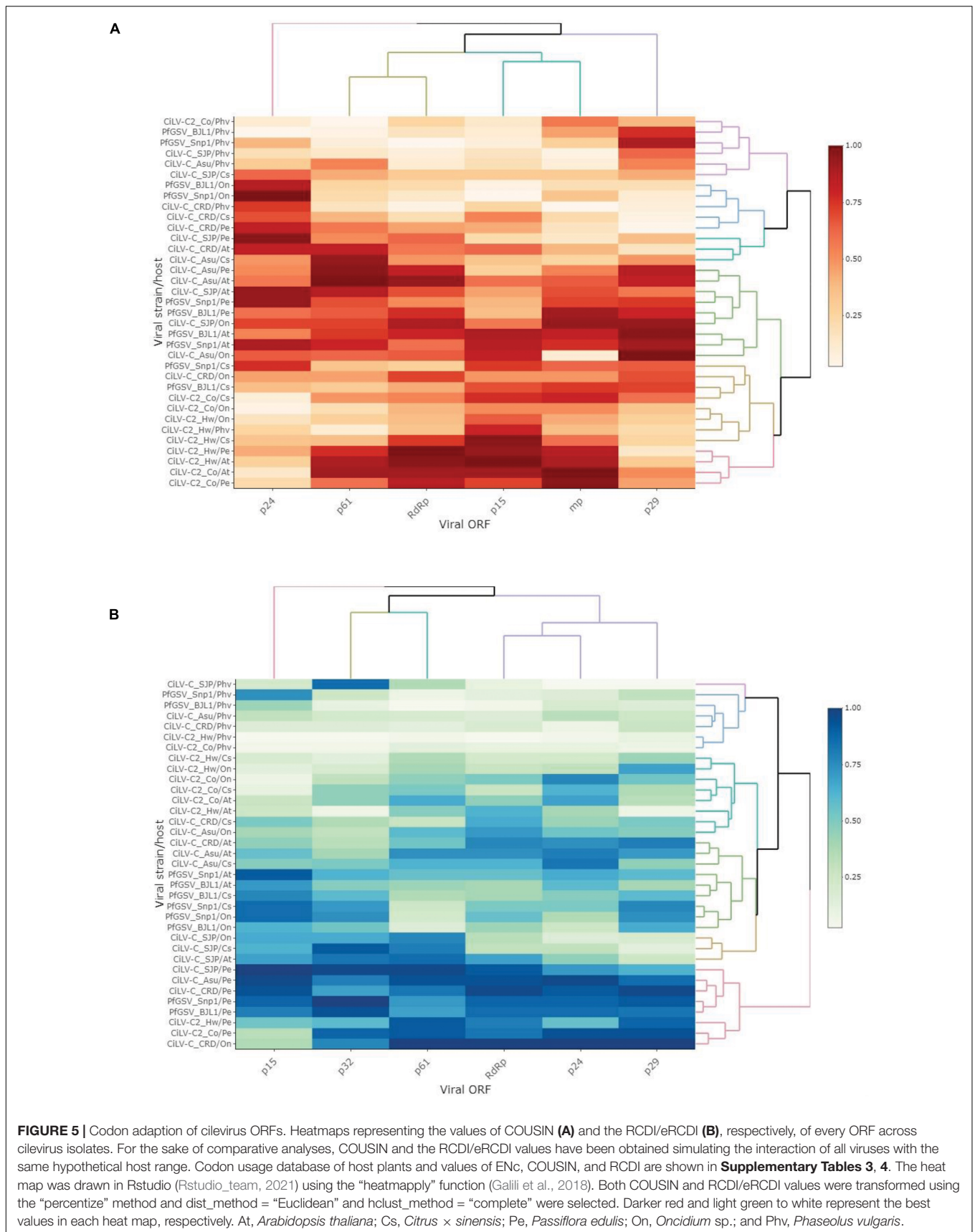
⁹<https://zhanglab.ccmb.med.umich.edu/I-TASSER/>

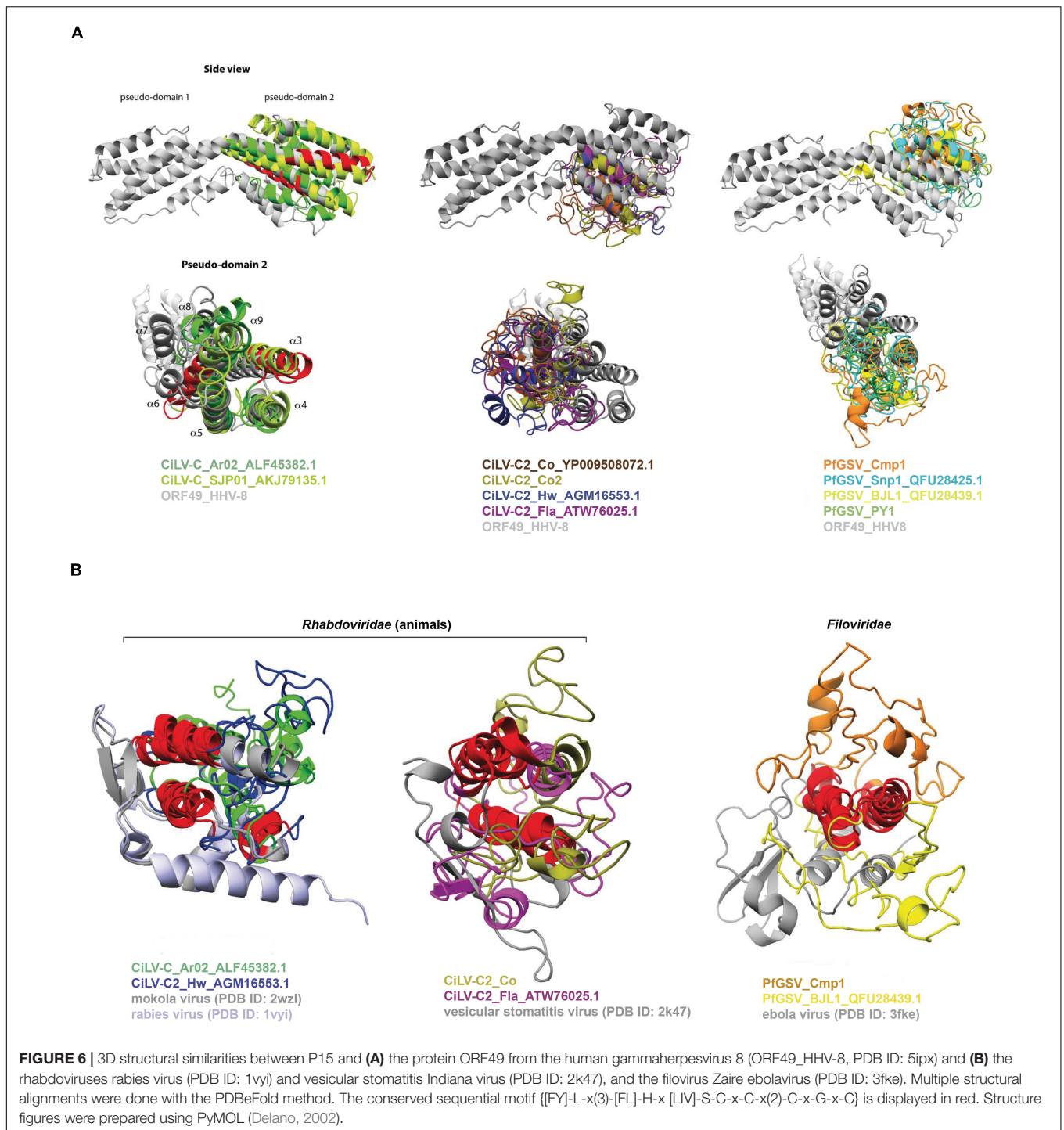
¹⁰<https://www.rcsb.org>

³<http://www.kazusa.or.jp/codon/>

⁴<http://www.kazusa.or.jp/codon/countcodon.html>

⁵<https://genomevolution.org/coge/>





PDBFold server¹¹ (Krissinel and Henrick, 2004). The quality of protein structure templates and predicted full-length models were assessed using the C-score and TM-score implemented in I-TASSER⁹ (Zhang and Skolnick, 2004). C-score is typically in the range [-5,2], where a C-score of higher value signifies a model with high confidence, and vice-versa. TM-score values

range from [0,1], where 1 indicates a perfect match between two structures. A TM-score below 0.17 corresponds to randomly chosen unrelated proteins whereas structures with a score higher than 0.5 assume generally the same fold.

The best generated P15 3D model corresponded to that for CiLV-C_Ar02 (C-score -2.82, TM-score 0.39 ± 0.13 , **Supplementary Table 8**). The model from this protein suggested the presence of a four-helical bundle-like scaffold with structural

¹¹<https://www.ebi.ac.uk/msd-srv/ssm/>

similarity to the 12 α -helices bundled of protein ORF49 from the human gammaherpesvirus 8 (ORF49_HHV-8, PDB ID: 5ipx). The multiple structural alignments of 3D models of P15 proteins with the crystal structure of ORF49_HHV-8 by using PDBeFold¹¹ revealed the “plasticity” of the helix-bundle protein folds (**Figure 6A**).

PDBeFold provides average root mean square deviation (RMSD) and Q-score values to estimate the quality of matching structures. RMSD distance between corresponding residues is calculated after an optimal rotation of one structure to another. An RMSD of less than about 3–4 Å would generally be considered very close, but there is no absolute rule. The Q-score assesses the number of residues in the matched secondary structural elements and their positions in space. Q-scores reach the value of 1 for identical structures and decline to zero with the lack of similarity, thus, the higher Q, the better.

Among the P15 from CiLV-C isolates, an overall RMSD = 2.47 Å and Q-score = 0.31 over 94 aligned residues were obtained, whereas in those from CiLV-C2 the values were RMSD = 3.83 Å and Q-score = 0.01 over 22 aligned residues. In P15 from the PfGSV isolates, the values were an overall RMSD = 4.1 Å and Q-score = 0.02 over 30 aligned residues. Detailed analysis of the 3D model for the CiLV-C_Ar02 P15 indicated that cysteine residues at positions 85 (Cys85) and 92 (Cys92) could form a disulfide bridge and consequently stabilize the four-helical bundle. Remarkably, this two-cysteine-residues pattern is also present in the conserved sequence motif across the P15 proteins. However, since the spatial localization of the two cysteines is not close enough in some 3D models, the formation of a disulfide bridge is not completely evident. Such discrepancies could ensue from the analysis of low-resolution 3D models.

The resolved crystal structures of ORF49_HHV-8 (PDB ID: 5ipx) and ORF49 from the murine herpesvirus 68 (ORF49_MHV68, PDB ID: 6a4v) (Hew et al., 2017; Chung et al., 2018) show a fold consisting of 12 α -helices bundled into two pseudo-domains (left top and bottom panels in **Figure 6A**). By mapping the identical and similar amino acids onto the helical scaffold, the potentially functional residues in P15 from CiLV-C were predicted. Conserved leucine and valine residues were localized in hydrophobic regions along the interacting surfaces of the α -helices and have the potential to play a structural role. Polar and charged residues were localized on the surface and could be associated with protein fold stability or protein interactions (**Table 1**).

α -helical folds have also been detected in the C-terminal domain of the phosphoprotein (P_{CTD}) of mononegaviruses (Assenberg et al., 2010; Ivanov et al., 2010; Ng et al., 2020). It is of note that cileviruses and mononegaviruses of the genus *Dichorhavirus*, family *Rhabdoviridae* (Dietzgen et al., 2018), share several commonalities, including hosts and vectors, that have led to postulate the likely existence of evolutionary convergence among viruses of these two genera (Freitas-Astúa et al., 2018). The peculiar architecture observed in

TABLE 1 | Summary of the predicted functional residues for the P15 family of proteins.

Amino acid ^a	ORF49 ^b	3D ^c	Description
Cys 18	His 14	$\alpha 1$	Polar, exposed, protein interaction
Cys 20	Phe 16	$\alpha 1$	Polar, exposed, protein interaction
Asp 26	Ser 22	Loop	Charged, exposed, protein interaction
Leu 34	Leu 30	$\alpha 2$	Hydrophobic, buried, helix packing
Val 35	Asp 31	$\alpha 2$	Hydrophobic, buried, helix packing
Leu 50	Ala 46	$\alpha 3$	Hydrophobic, buried, helix packing
Leu 74	Pro 70	$\alpha 3$	Hydrophobic, buried, helix packing
Tyr 76	Leu 72	$\alpha 3$	Polar, exposed, protein interaction
His 79	Ala 75	$\alpha 3$	Charged, exposed, protein interaction
Ser 82	Ala 78	$\alpha 3$	Polar, exposed, protein interaction
Cys 83	Asn 79	$\alpha 3$	Polar, exposed, protein interaction
Asn 84	Leu 80	$\alpha 3$	Polar, exposed, protein interaction
Cys 85	Ala 81	$\alpha 3$	Polar, exposed, disulfide bridge, protein fold stability
Cys 88	Leu 84	$\alpha 3$	Polar, exposed, protein interaction
Gly 90	Gln 86	$\alpha 3$	Polar, exposed, protein fold stability
Cys 92	Tyr 88	$\alpha 3$	Polar, exposed, disulfide bridge, protein fold stability
Gly 98	Lys 94	$\alpha 4$	Polar, exposed, protein fold stability
Gly 114	Phe 113	$\alpha 5$	Polar, exposed, protein fold stability

^aAmino acid position in the CiLV-C_Ar02_ALF45382.1 sequence.

^bEquivalent position in the ORF49 sequence according to the structure superposition between ORF49_HHV-8 (PDB ID: 5ipx) and CiLV-C_Ar02_ALF45382.1 (I-TASSER 3D-model).

^cNomenclature of alpha-helices is based on the 3D structure of ORF49_HHV-8. The structure superposition between ORF49_HHV-8 and CiLV-C 3D-models by PDBeFold is provided in **Supplementary File 1**.

P_{CTD} and P15 prompted us to compare the 3D structure of these proteins.

The P_{CTD} of the cytorhabdovirus lettuce necrotic yellows virus (LNYV) comprises five α -helices that have an overall topology that although with different structural features, seems to be conserved with those in other rhabdoviruses, filoviruses, and paramyxoviruses (Martinez et al., 2013). P_{CTD} crystal structure from plant-infecting rhabdoviruses other than LNYV has not been described. Hence, the predicted 3D models of P15 were aligned with the P_{CTD} of mononegaviruses, regardless of their range hosts. P15 of CiLV-C_Ar02 and CiLV-C2_Hw are structurally related to P_{CTD} of mokola virus (PDB ID: 2wzl) (Delmas et al., 2010) and rabies virus (PDB ID: 1vyi) (Mavrakakis et al., 2004) with overall RMSD = 3.1 Å and Q-score = 0.06 over 42 aligned residues, whereas those of CiLV-C2_Co2 and CiLV-C2_Fla are structurally related to vesicular stomatitis Indiana virus (PDB ID: 2k47) with overall RMSD = 4.0 Å and Q-score = 0.04 over 32 aligned residues (**Figure 6B**). In the case of PfGSV_Cmp1 and PfGSV_BJL1, their modeled P15 are structurally related to Zaire ebolavirus (PDB ID: 3fke) with an overall RMSD = 3.1 Å and Q-score = 0.04 over 36 aligned residues. Based on structural similarities, it could be hypothesized that P15 proteins can perform functions related to those accomplished by ORF49 and P_{CTD} proteins. ORF49 interacts and upregulates the transcriptional activity

of a protein known as RTA (replication and transcription activator) in HHV-8, which drives the expression of all lytic genes (González et al., 2006). Present in a large number of viruses, P protein orthologs are involved in a wide range of functions, e.g., acting as chaperons, regulation of viral transcription, RNA binding, and suppressor of RNA silencing in plants (Masters and Banerjee, 1988; Peluso and Moyer, 1988; Martinez et al., 2013; Bejerman et al., 2016; Mann et al., 2016; Ng et al., 2020).

Zn-FINGER-LIKE PROTEINS IN ARTHROPOD-INFECTING VIRUSES CLOSELY RELATED TO KITAVIRUSES ARE NOT RELATED WITH THE CILEVIRUS P15 PROTEINS

Based on the predicted isoelectric point and the presence of a cysteine-rich pattern compatible with a potential Zn-finger domain, an acidic cilevirus P15-like protein was previously recognized in the genome of Wuhan insect virus 8 (WhIV-8) (Vinokurov and Koloniuk, 2019). The protein is encoded by the putative ORF *p11* (Figure 1) that was neither identified during the virus genome description (Shi et al., 2016) nor annotated at its corresponding GenBank locus KX883812.

Our search for supposed *p15* homologs based on the presence of a zinc-finger-like motif was expanded to nege/kita-like viruses and other related arthropod-infecting viruses. It also included previously ignored or poorly characterized putative ORFs. In members of the proposed subgroup Centivirus of the nege/kita-like viruses, a homolog of WhIV-8 *p11* was identified in the uncharacterized *Glycine dolichocarpa* virus (GDV, Transcriptomic shotgun assembly TSA GGIW01009300) (Kondo et al., 2020). Proteins encoded by the ORF *p11* in GDV and WhIV-8 have 41.4% amino acid sequence identity. The region comprising the cysteine-rich pattern in these two proteins shows a marginal identity (*E*-value < 1 according to MOTIF Search¹²) with the Pfam motif C4-type zinc-finger of DNA polymerase delta (PF14260). Similarly, the ORF *p19* in the putative centivirus Wuhan house centipede virus 1 (WHCV-1) and ORF *p9* of Hubei virga-like virus 9 (HVLV-9) encode proteins with signatures of a putative zinc-finger domain (Vinokurov and Koloniuk, 2019), which in the case of HVLV-9 also displays a subtle identity [*E*-value < 1 according to MOTIF Search¹¹] with the Pfam motif SLBP (stem-loop binding protein) of the RNA binding superfamily (PF15247). The N-terminal of the ORF3-encoded protein of *Pyrrhocoris apterus* virus slightly resembles the canonical Pfam DSRM (double-stranded RNA binding motif, PF00035) architecture, whereas, although with low sequence similarity, the C-terminal end of this protein displays a conservative basic RNA-interacting motif. However, neither nucleotide sequences of these ORFs nor their predicted polypeptides show significant sequence identity to any element across cileviruses.

¹²<https://www.genome.jp/tools/motif/>

To get further insight into the putative relationships between P15 and P11 proteins of WhIV-8 and GDV, their 3D structures were predicted using the I-TASSER server⁹. Although P11 proteins show α helix scaffolds, the arrangement of helices is different from that observed in P15. Multiple structural alignments of 3D models by PDBeFold¹¹ indicated that only one α -helix in P11 proteins shows structural similarity to α -helices in the P15 3D models (overall RMSD = 2.68 Å and Q-score = 0.01 over 17 aligned residues). The lack of overall sequence or 3D structure similarity between both groups of proteins drastically reduces the probability of a common origin.

THE 5'-END OF THE RNA2 IN CILEVIRUSES, BESIDES THE ORF *p15*, HARBORS ORFs ENCODING SMALL HYDROPHOBIC PROTEINS

Downstream the ORF *p15*, the genomic stretch extended up to *p61* in the 5'-end of the RNA2 of cileviruses is known as the intergenic region (IR). However, this is not a proper description because rather than lacking ORFs, the IR lacks ORFs greater than 300 nucleotides. *In silico* tools have traditionally considered a minimum of 100 amino acids for an ORF to be annotated as a putative coding protein sequence, which has led to the inaccurate annotation of certain genomes. However, despite the detection of those small ORFs, the IRs of cileviruses show a very low density of predicted coding sequences, only 20–30%, with large non-coding stretches (Figure 1).

In congruence with ORF *p15*, the interspecies comparisons of the IRs of cileviruses show low levels of nucleotide sequence identity ($\leq 45\%$) (Ramos-González et al., 2020). With variable length, the IR is larger in CiLV-C isolates ($1,095 \pm 56$ nucleotides) and shorter in PfGSV isolates (850 ± 21 nucleotides). The region harbors small ORFs encoding hydrophobic proteins with predicted TM domains, but which are unrelated to the SP24 motif (Melzer et al., 2013; Roy et al., 2013; Ramos-González et al., 2016, 2020). ORF *p13* in PfGSV_Snp1, the largest detected so far (333 nucleotides), presents four regularly distributed cysteine residues and low sequence identity with motifs identified in the cytokine-induced anti-apoptosis inhibitor 1 (CIAPIN1, PF05093), and PGC7/Stella/Dppa3-like domain (PGC7_Stella, PF15549) (Ramos-González et al., 2020).

Further inspection in the 5'-end of the RNA2 in cileviruses, allowed us to detect the ORF *p8* overlapping the ORF *p15* of CiLV-C. ORF *p8* is present in every sequenced virus of the three lineages of CiLV-C (Chabi-Jesus et al., 2021), but it is absent in CiLV-C2 and PfGSV. Previously overlooked, the ORF *p8* spans 54% of *p15* and has 213 nucleotides in the frame +2 (Figure 1). ORF *p8* is significantly longer than any other ORF expected in all reading frames within the ORF *p15*, which is a fair indicator of its real existence according to an *in silico* method for detecting overlapping genes (Schlub et al., 2018) performed in this study (Supplementary Table 9). The presence of the overlapping ORF *p8* in CiLV-C may limit the variability of *p15* and suggests a

different evolutionary trajectory in the formation of this genomic region among cileviruses.

The predicted P8 protein seems to be hydrophobic with a TM domain according to MEMSAT-SVM¹³ and TMHMM Server v. 2.0¹⁴. This TM domain resembles the predicted properties of the small P7 proteins encoded downstream the ORF *p15*, in the IR of CiLV-C2 (Roy et al., 2013), CiLV-C_SJP, and PfGSV (Figure 1). Small hydrophobic proteins by themselves are outside the scope of this work but what is particularly pertinent is their constant presence in the 5'-end of the RNA2 of cileviruses, and the association with the small Zn-finger-like or nucleic acid binding-like proteins in some nege/kita-like viruses. The overlapping gene array observed in the pair P15/P8 in CiLV-C was also observed in the pairs P11/P6 and P19/P8 of the centiviruses WhIV-8 and WHCV-1, respectively (Figure 1). Overlapping genes *p6* and *p8* of centiviruses were detected following the same procedure described for the identification of *p8* in CiLV-C (Supplementary Table 9). Small hydrophobic proteins have been described in several taxonomic groups of plant RNA viruses where they can be involved in coupling viral replication and cell-to-cell movement showing properties similar to those of reticulons (Solovyev and Morozov, 2017; Morozov and Solovyev, 2020; Lazareva et al., 2021).

DISCUSSION

This study was driven by the search for commonalities between the ORFs *p15*, their encoded proteins, and the questioning of whether these polypeptides can be considered orthologs. Finally, these analyses led us to a critical appraisal of the 5'-end of the RNA2 of cileviruses under an evolutionary perspective.

Absent in kitavirids beyond the genus *Cilevirus* and with no apparent homology with any sequence available in public databases, ORF *p15* was considered an orphan gene (Arena et al., 2016). P15 proteins from different species show a remarkably reduced amino-acid sequence identity between them, with values as low as those shown by the global alignment of unrelated proteins ($\approx 16\%$) (Konaté et al., 2019; Ramos-González et al., 2020). As well as P13 from PfGSV, it has been speculated that P15 proteins are viral auxiliary factors that act as regulators of host physiology during the plant-virus interplay (Ramos-González et al., 2020). Particularly, the putative role of P15 from CiLV-C of the strain CRD as a viral suppressor of the plant gene silencing antiviral defense was analyzed (Leastro et al., 2020), but further studies are needed to reach sound conclusions.

Together with the hallmark ORF *RdRp* (Koonin et al., 2006), the genomes of the typical cileviruses comprise the ORFs *p24*, *p32*, and the taxonomically restricted *p61*. Their RNA2 molecules also accommodate at their 5'-ends the *p15* and the IR, which in addition to *p11* and *p13* in PfGSV strains, harbor some putative small ORFs encoding predicted TM small polypeptides.

Interestingly, hibiscus yellow blotch virus (HYBV) (Olmedo-Velarde et al., 2021), a specimen considered a cilevirus with an unusual genomic organization and phylogenetically intermediate among typical cileviruses and hibiscus green spot virus 2 (HGSV-2), genus *Higrevirus* (Melzer et al., 2012), poses an additional puzzle on the composition of the RNA2 segments of typical cileviruses and the role of the ORFs in their 5'-ends, e.g., *p15*. While the RNA1 in all these viruses encodes two proteins, the RdRp and P29 or P10 in typical cileviruses and HYBV, respectively, the RNA2 of HYBV resembles a truncated version of the RNA2 of typical cileviruses, lacking the 5'-end stretch upstream the homologs of the threesome *p61-p32-p24* (Figure 1). RNA2 of HYBV also includes an additional ORF, homolog to the cilevirus RNA1 *p29*, at its 3'-end. The very conserved synteny and composition among the genomes of HYBV and typical cileviruses suggest, under a reductionist approach, the rather dispensable role of the 5'-end of the RNA2 of cileviruses. Perhaps, this region only provides proteins and/or unknown regulatory non-coding RNAs that albeit non-essential could expand the viral fitness (Frias-Lasserre, 2012; Miller et al., 2016). Alternatively, we could speculate, for instance, that the putative functions encoded by this genomic region in typical cileviruses are taken over by other ORFs in HYBV, likely involving a compensatory evolutionary process (Rokyta et al., 2002; Harcombe et al., 2009). However, while the lack of function of the 5' end of RNA2 in typical cileviruses does not appear to be a rational choice, in the absence of compelling evidence suggesting explicit roles, the analysis of *p15* across the genus, intrinsically, may shed some light on the 5'-end of cilevirus RNA2.

In this study, we described that *p15* in each of the three cilevirus species is under purifying selection, suggesting an ongoing action to conserve its nucleotide sequence, most likely precluding the change of amino acid residues and preserving function at directly selected sites (Cvijović et al., 2018). Furthermore, we detected that in some viral strains the synonymous codon usage of this ORF is biased showing the same codon patterns adopted by other viral genes and displaying COUSIN and RCDI values that suggest a certain level of adaptation to its plant hosts. We also revealed that although with extremely divergent amino acid sequences, P15 proteins show a conserved sequence motif, and their 3D models have preserved a backbone topology that resembles a helical bundle-like scaffold. Altogether, these two aspects suggest elements of common ancestry and conserved functionality of the P15 proteins (Arendsee et al., 2019; Ng et al., 2020; Ravantti et al., 2020).

The superimposition of the 3D models of P15 and the resolved crystal structures of ORF49_HHV-8 and ORF49_MHV68 (Hew et al., 2017; Chung et al., 2018) and the P_{CTD} of negative-stranded RNA viruses of the order *Mononegavirales* (Martinez et al., 2013) indicated the presence of an α -helical bundle-like scaffold in the cilevirus protein. Bundles of four α -helices are frequently found in proteins with a wide variety of functions (Rhys et al., 2019). ORF49 shows a fold consisting of 12 helices bundled into two pseudo-domains, with charged patches on its surface likely acting in protein-protein interaction sites. The protein is stabilized by double-stranded oligonucleotides suggesting its interaction and

¹³<http://bioinf.cs.ucl.ac.uk/psipred/>

¹⁴<http://www.cbs.dtu.dk/services/TMHMM/>

binding with DNA molecules (Hew et al., 2017). The P_{CTD} of negative-stranded RNA viruses of the order *Mononegavirales* (Martinez et al., 2013) has been shown to regulate viral transcription, RNA binding, and immunomodulatory functions (Ng et al., 2020). The P protein of cytorhabdoviruses has weak local RNA silencing suppressor activity, strongly suppresses systemic RNA silencing, and the P_{CTD} is essential for both local RNA silencing suppression and the interaction with AGO1, AGO2, AGO4, RDR6, and SGS3, proteins of the main core of the gene silencing mechanism (Bejerman et al., 2016; Mann et al., 2016). Therefore, the presence of an α -helical bundle-like scaffold in P15 could provide a structural base by mean it could interact with other proteins and likely mediate the plant-virus interplay. P15 from CiLV-C of the strain CRD forms homo- and heterodimers with other viral proteins (Leastro et al., 2018), and considering the diversity of its interactions, it appears that the protein could act modulating a wide spectrum of viral and cellular pathways, as observed in other viral proteins (Deblasio et al., 2018; Dao et al., 2020; Li et al., 2020).

The Origin and Evolution of the 5'-End of Cileviruses: A Theory

Phylogenetic inferences suggest a common ancestor between kitaviruses and a variety of arthropod-infecting viruses including negevirus (Vasilakis et al., 2013; Kondo et al., 2019; Ramos-González et al., 2020; Quito-Avila et al., 2021). Having regards to the presence of the orthologs *RdRp* and *p24* in these viruses (Kuchibhatla et al., 2014), a synteny-based analysis indicates a parsimonious evolutionary scenario in which cileviruses, and kitaviruses in general, derived from a monopartite ss(+) RNA through genome segmentation. In brief, the process could have started amid assortments of defective molecules resulting from error-prone replication and recombination events (Bangham and Kirkwood, 1993; García-Arriaza et al., 2004; Vignuzzi and López, 2019) and sub-genomic RNA transcripts derived from the ancestor. Those molecules with self-replication activity could have recruited segments supplying viral complementary functions. *Cis*- and *trans*-acting sequence motifs likely evolved providing these novel genomic segments with the capacity to coordinate essential processes for the persistence of split viral genomes, e.g., multiplication and packaging (Newburn and White, 2015). Further modifications of kitavirus genomes likely included the acquisition of auxiliary genes, as seems to be a general pathway in the formation of plant virome (Dolja et al., 2020). The clearest examples of horizontal gene transfer in the genome of current kitaviruses are epitomized by the ORF *p32* in cileviruses, and ORFs *BMB1* and 2, which encode the binary movement block proteins in HGSV-2 (Quito-Avila et al., 2013; Solovyev and Morozov, 2017; Hao et al., 2018).

Multiplication of unnecessary viral sequences appears to have fitness costs and the size of virus genomes shows a natural trend to shrinkage (Belshaw et al., 2007; Zwart et al., 2014). The occurrence of IRs as long as observed in the RNA2 of cileviruses is atypical. Whether large IRs existed in the ancestors of kita/nege-like viruses, they were most likely purged by selection. The genomic organization of known negevirus and kita/nege-like

viruses are rather compact, with short stretches of intergenic sequences. Therefore, a relevant interrogation in this context is whether the IRs of cileviruses deviate from the selection rules. In a sense, most likely not, and some features might suggest it. The IR of PfGSV isolates are shorter and contain more predicted coding regions than those in CiLV-C isolates, whereas the presence of the overlapping ORF *p8* in CiLV-C may also indicate genomic compression.

As defined in this study, the 5'-end of the RNA2 genomic segments in typical cileviruses comprises the ORF *p15* and the IR. There is no obvious evidence allowing for a direct connection between the 5'-end of the cilevirus RNA2 molecules and any genetic element in the genome of their closely related negevirus and nege/kita-like viruses. Yet, the relationship between the 5'-end of the RNA2 molecules from cileviruses would be difficult to establish, except for the presence of a conserved amino acid motif across the P15 family and the 3D structure similarities among these proteins. These two features, moreover, may be pertinent clues suggesting the common ancestry of the ORF *p15* present in the cileviruses CiLV-C, CiLV-C2, and PfGSV, giving support to a theory about their common origin.

The origin of the 5'-end of the RNA2 in typical cileviruses could have involved the following plausible scenarios: (i) a reminiscent ancient ORF from the monopartite ancestor or (ii) its acquisition from a heterologous source by recombination, i.e., horizontal gene transfer. Although we lack evidence to exclude any of these possibilities, our approach using prediction and computational analysis of 3D protein structures has provided information on a possible external origin of P15 proteins. The scaffold of α -helices detected in P15 shows structural resemblance with those detected in well-characterized proteins from herpesvirids and mononegavirids. Importantly, the odds of the 3D structure of P15 is the result of convergence are very low considering the small number of protein folds available (Malik et al., 2017).

The external origin of the ORF *p15* could be considered part of the horizontal gene transfer events that likely happened in the course of kitavirus evolution, but what is peculiar in this case is the participation of a mononegavirid, probably an ancestor of the current plant rhabdovirus, as its putative donor. Cileviruses and plant-infecting rhabdoviruses have a partially shared plant host range, and, in the case of dichorhavirus, they also share mite vectors (Dietzgen et al., 2018, 2020; Freitas-Astúa et al., 2018). Likely, *Brevipalpus* mites, where dichorhavirus also multiply, have underpinned a primordial and essential role in the segmentation of both cileviruses and dichorhavirus (Kondo et al., 2017).

However, whatever the scenario for the origin of *p15* in cileviruses, the very conservation of a sequence motif and fold in the encoded proteins indicate that the "ancestor" ORF was present before the radiation of, at least, CiLV-C, CiLV-C2, and PfGSV. Since then, it has been likely undergone a transformation process under selective pressure, acquiring or losing protein secondary structure elements and involving host-specific patterns to provide evolutionarily competitiveness to these viruses under new ecological conditions, e.g., the change of natural host range from arthropods to plants. Indeed, if this would be the case,

p15 likely became a fitness factor whose fixation might have indirectly contributed to the conservation of the IR. This large chunk of genomic sequence with apparently an unclear role is preserved between functional ORFs, that away of inactivity, it may be a zone of gain and loss of start and stop codons gradually incorporating novel ORFs.

Two mechanisms explaining the origin of orphan genes are currently accepted: (i) divergence from the coding sequence of a preexisting gene (Tautz and Domazet-Lošo, 2011; Vakirlis et al., 2020), and (ii) *de novo* appearance from intergenic or non-genic regions (Carvunis et al., 2012; Van Oss and Carvunis, 2019). The irregular organization and lack of identity among the small ORFs other than *p15* found in the 5'-end of the RNA 2 of cileviruses support the idea that this genomic region might act as “the cradle of new genes” (Forterre and Gaïa, 2016). Several of the predicted ORFs found in the IR are small, of ≈ 200 nucleotides, and they encode putative proteins where TM domains are detected by *in silico* analyses. Viral miniproteins with TM domains are encoded by a wide range of DNA and RNA viruses and they can bind and modulate cellular transmembrane proteins, providing flexible mechanisms to regulate several cellular activities (Dimaio, 2014; Finkel et al., 2018). Moreover, small ORFs can act as seeds for larger coding sequences (CDS) through a mechanism known as “CDS elongation” when their stop codons undergo mutations (Couso and Patraquim, 2017). Thus, small ORFs could provide an extra coding capacity that in the course of evolution could eventually become an additional source of fitness (Guerra-Almeida and Nunes-da-Fonseca, 2020).

CONCLUDING REMARKS

In this study, we have addressed one of the most intriguing features of the molecular biology of cileviruses: the 5' end of the RNA2 molecule and its two most salient elements, the ORF *p15*, and the IR. We presented pieces of evidence on the putative biochemical activity of the protein P15 over the genus and elucidate on its putative contribution to viral fitness. The results are supported by the analyses of 58 nucleotide sequences of isolates of the three known species whose P15 amino acid sequences cover a spectrum of identity ranging from 14 to 100%. The main predictions have been obtained by integrating the results of different computational methods because of the premise that each of these elements taken in isolation does not allow to draw solid conclusions about the activity and function of a given protein (Omer et al., 2017; Ponting, 2017;

REFERENCES

- Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi: 10.1093/nar/gky379
- Arena, G. D., Ramos-González, P. L., Falk, B. W., Casteel, C. L., Freitas-Astúa, J., and Machado, M. A. (2020). Plant Immune System Activation Upon Citrus Leprosis Virus C Infection Is Mimicked by the Ectopic Expression of the P61 Viral Protein. *Front. Plant Sci.* 11:1188. doi: 10.3389/fpls.2020.01188
- Ravantti et al., 2020). Taken them together, we posit a theory that partially describes the evolution of the cilevirus genomes that involves elements denoting the likely close interplay during the coevolutionary history of *Brevipalpus*-transmitted viruses. In light of the plausible scenario described by the theory, two aspects of the biology of cileviruses gain new distinctions: (i) since *p15* across cileviruses can be considered alleles descendant from a common ancestor likely co-opted from another virus, this ORF should not be longer considered an orphan gene, and (ii) the markedly chimeric origin of the RNA2 genome segment, which adds further support to what appears to be a common feature of members of the family *Kitaviridae* (Morozov et al., 2020).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

PR-G and TP: conceptualization. PR-G, TP, CC-J, and GA: formal analysis, investigation, and methodology. JF-A: funding acquisition. TP and JF-A: resources. PR-G and TP: writing—original draft. PR-G, TP, CC-J, GA, and JF-A: writing—review and editing. All authors approved the submitted version.

ACKNOWLEDGMENTS

The authors are grateful to CAPES, CNPq, and FAPESP for scholarships and research grants associated with this work (CAPES PNP20132154 – 33141010001P4 – PNP2 – IBSP, 88882.157041/2017-01; CNPq 312528/2020-5; FAPESP 2017/50222-0, 2019/25078-9, 2019/02137-0, and 2020/15413-2). The authors also thank the reviewers and editor for their useful and constructive comments on our work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2021.771983/full#supplementary-material>

- Arena, G. D., Ramos-González, P. L., Nunes, M. A., Alves, M. R., Camargo, L. E. A., Kitajima, E. W., et al. (2016). Citrus leprosis virus C infection results in hypersensitive-like response, suppression of the JA/ET plant defense pathway and promotion of the colonization of its mite vector. *Front. Plant Sci.* 7:1757. doi: 10.3389/fpls.2016.01757
- Arena, G. D., Ramos-González, P. L., Nunes, M. A., Jesus, C. C., Calegario, R. F., Kitajima, E. W., et al. (2017). *Arabidopsis thaliana* as a model host for *Brevipalpus* mite-transmitted viruses. *Sci. Agric.* 74, 85–89. doi: 10.1590/1678-992x-2015-0380

- Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: Orphan genes in plants. *Trends Plant Sci.* 19, 698–708. doi: 10.1016/j.tplants.2014.07.003
- Arendsee, Z., Li, J., Singh, U., Bhandary, P., Seetharam, A., and Wurtele, E. S. (2019). Fagin: Synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics* 20:440. doi: 10.1186/s12859-019-3023-y
- Assenberg, R., Delmas, O., Ren, J., Vidalain, P.-O., Verma, A., Larrous, F., et al. (2010). Structure of the Nucleoprotein Binding Domain of Mokola Virus Phosphoprotein. *J. Virol.* 84, 1089–1096. doi: 10.1128/jvi.01520-09
- Bangham, C. R. M., and Kirkwood, T. B. L. (1993). Defective interfering particles and virus evolution. *Trends Microbiol.* 1, 260–264. doi: 10.1016/0966-842X(93)90048-V
- Behra, S. K., and Severson, D. W. (2013). Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biol. Rev.* 88, 49–61. doi: 10.1111/J.1469-185X.2012.00242.X
- Bejerman, N., Mann, K. S., and Dietzgen, R. G. (2016). Alfalfa dwarf cytorhabdovirus P protein is a local and systemic RNA silencing suppressor which inhibits programmed RISC activity and prevents transitive amplification of RNA silencing. *Virus Res.* 224, 19–28. doi: 10.1016/j.virusres.2016.08.008
- Belshaw, R., Pybus, O. G., and Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.* 17, 1496–1504. doi: 10.1101/gr.6305707
- Biswas, K. K., Palchoudhury, S., Chakraborty, P., Bhattacharyya, U. K., Ghosh, D. K., Debnath, P., et al. (2019). Codon usage bias analysis of citrus tristeza virus: Higher codon adaptation to citrus reticulata host. *Viruses* 11:v11040331. doi: 10.3390/v11040331
- Bourret, J., Alizon, S., and Bravo, I. G. (2019). COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biol. Evol.* 11, 3523–3528. doi: 10.1093/gbe/evz262
- Butt, A. M., Nasrullah, I., Qamar, R., and Tong, Y. (2016). Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerg. Microbes Infect.* 5, 1–14. doi: 10.1038/emi.2016.106
- Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., et al. (2012). Proto-genes and de novo gene birth. *Nature* 487, 370–374. doi: 10.1038/nature11184
- Chabi-Jesus, C., Ramos-González, P. L., Postclam-Barro, M., Fontenele, R. S., Harakava, R., Bassanezi, R. B., et al. (2021). Molecular Epidemiology of Citrus Leprosis Virus C: A New Viral Lineage and Phylodynamic of the Main Viral Subpopulations in the Americas. *Front. Microbiol.* 12:986. doi: 10.3389/fmicb.2021.641252
- Chung, W. C., Kim, J., Kim, B. C., Kang, H. R., Son, J., Ki, H., et al. (2018). Structure-based mechanism of action of a viral poly(ADP-ribose) polymerase 1-interacting protein facilitating virus replication. *IUCr J* 5, 866–879. doi: 10.1107/S2052252518013854
- Colson, P., La Scola, B., Levasseur, A., Caetano-Anollés, G., and Raoult, D. (2017). Mimivirus: leading the way in the discovery of giant viruses of amoebae. *Nat. Rev. Microbiol.* 15, 243–254. doi: 10.1038/nrmicro.2016.197
- Comeron, J. M., and Aguadé, M. (1998). An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.* 47, 268–274. doi: 10.1007/PL00006384
- Couso, J. P., and Patraquim, P. (2017). Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* 18, 575–589. doi: 10.1038/nrm.2017.58
- Cvijović, I., Good, B. H., and Desai, M. M. (2018). The effect of strong purifying selection on genetic diversity. *Genetics* 209, 1235–1278. doi: 10.1534/genetics.118.301058
- Dao, T. N. M., Kang, S.-H., Bak, A., and Folimonova, S. Y. (2020). A Non-Conserved p33 Protein of Citrus Tristeza Virus Interacts with Multiple Viral Partners. *Mol. Plant Microbe Interact.* 33, 859–870. doi: 10.1094/MPMI-11-19-0328-FI
- de Castro, E., Sigrist, C., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P., Gastegger, E., et al. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 34, 362–365. doi: 10.1093/nar/gkl124
- Deblasio, S. L., Xu, Y., Johnson, R. S., Rebelo, A. R., Maccoss, M. J., Gray, S. M., et al. (2018). The interaction dynamics of two potato leafroll virus movement proteins affects their localization to the outer membranes of mitochondria and plastids. *Viruses* 10:v10110585. doi: 10.3390/v10110585
- Delano, W. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* 40, 82–92.
- Delmas, O., Assenberg, R., Grimes, J. M., and Bourhy, H. (2010). The structure of the nucleoprotein binding domain of lyssavirus phosphoprotein reveals a structural relationship between the N-RNA binding domains of Rhabdoviridae and Paramyxoviridae. *RNA Biol.* 7, 322–327. doi: 10.4161/rna.7.3.11931
- Dietzgen, R. G., Bejerman, N. E., Goodin, M. M., Higgins, C. M., Huot, O. B., Kondo, H., et al. (2020). Diversity and epidemiology of plant rhabdoviruses. *Virus Res.* 281:197942. doi: 10.1016/j.virusres.2020.197942
- Dietzgen, R. G., Freitas-Astúa, J., Chabi-Jesus, C., Ramos-González, P. L., Goodin, M. M., Kondo, H., et al. (2018). “Dichorhavirus in their Host Plants and Mite Vectors,” in *Advances in Virus Research*, (Cambridge, MA: Academic Press Inc), 119–148. doi: 10.1016/bs.aivir.2018.06.001
- Dimaio, D. (2014). Viral miniproteins. *Annu. Rev. Microbiol.* 68, 21–43. doi: 10.1146/annurev-micro-091313-103727
- Dolja, V. V., Krupovic, M., and Koonin, E. V. (2020). Deep Roots and Splendid Boughs of the Global Plant Virome. *Annu. Rev. Phytopathol.* 58, 23–53. doi: 10.1146/ANNUREV-PHYTO-030320-041346
- Entwistle, S., Li, X., and Yin, Y. (2019). Orphan Genes Shared by Pathogenic Genomes Are More Associated with Bacterial Pathogenicity. *mSystems* 4:18. doi: 10.1128/msystems.00290-18
- Finkel, Y., Stern-Ginossar, N., and Schwartz, M. (2018). Viral Short ORFs and Their Possible Functions. *Proteomics* 18:201700255. doi: 10.1002/pmic.201700255
- Fischer, D., and Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics* 15, 759–762.
- Forterre, P., and Gaïa, M. (2016). Giant viruses and the origin of modern eukaryotes. *Curr. Opin. Microbiol.* 31, 44–49. doi: 10.1016/j.mib.2016.02.001
- Freitas-Astúa, J., Ramos-González, P. L., Arena, G. D., Tassi, A. D., and Kitajima, E. W. (2018). Brevipalpus-transmitted viruses: parallelism beyond a common vector or convergent evolution of distantly related pathogens? *Curr. Opin. Virol.* 33, 66–73. doi: 10.1016/j.COVIRO.2018.07.010
- Frias-Lasserre, D. (2012). Non coding RNAs and viruses in the framework of the phylogeny of the genes, epigenesis and heredity. *Int. J. Mol. Sci.* 13, 477–490. doi: 10.3390/ijms13010477
- Galili, T., O’Callaghan, A., Sidi, J., and Sievert, C. (2018). Heatmaply: An R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* 34, 1600–1602. doi: 10.1093/bioinformatics/btx657
- García-Arriaza, J., Manrubia, S. C., Toja, M., Domingo, E., and Escarmis, C. (2004). Evolutionary Transition toward Defective RNAs That Are Infectious by Complementation. *J. Virol.* 78, 11678–11685. doi: 10.1128/jvi.78.21.11678-11685.2004
- Garita, L., Tassi, A. D., Calegario, R. F., Kitajima, E. W., Carbonell, S. A. M., and Freitas-Astúa, J. (2013). Common bean (*Phaseolus vulgaris* L.): Experimental local lesion host for citrus leprosis virus C (CiLV-C) and some other cytoplasmic-type Brevipalpus-transmitted viruses. *Plant Dis.* 97, 1346–1351. doi: 10.1094/PDIS-12-12-1143-RE
- González, C. M., Wong, E. L., Bowser, B. S., Hong, G. K., Kenney, S., and Damania, B. (2006). Identification and Characterization of the Orf49 Protein of Kaposi’s Sarcoma-Associated Herpesvirus. *J. Virol.* 80, 3062–3070. doi: 10.1128/jvi.80.6.3062-3070.2006
- Guerra-Almeida, D., and Nunes-da-Fonseca, R. (2020). Small Open Reading Frames: How Important Are They for Molecular Evolution? *Front. Genet.* 11:574737. doi: 10.3389/fgene.2020.574737
- Hao, X., Zhang, W., Zhao, F., Liu, Y., Qian, W., Wang, Y., et al. (2018). Discovery of Plant Viruses From Tea Plant (*Camellia sinensis* (L.) O. Kuntze) by Metagenomic Sequencing. *Front. Microbiol.* 9:2175. doi: 10.3389/fmicb.2018.02175
- Harcombe, W. R., Springman, R., and Bull, J. J. (2009). Compensatory evolution for a gene deletion is not limited to its immediate functional network. *BMC Evol. Biol.* 9:106. doi: 10.1186/1471-2148-9-106
- Hasegawa, M., Kishino, H., and Yano aki, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174. doi: 10.1007/BF02101694
- He, Z., Gan, H., and Liang, X. (2019). Analysis of synonymous codon usage bias in potato virus M and its adaption to hosts. *Viruses* 11:v11080752. doi: 10.3390/v11080752

- Hew, K., Veerappan, S., Sim, D., Cornvik, T., Nordlund, P., and Dahlroth, S.-L. (2017). Structure of the Open Reading Frame 49 Protein Encoded by Kaposi's Sarcoma-Associated Herpesvirus. *J. Virol.* 91:16. doi: 10.1128/jvi.01947-16
- Ivanov, I., Crépin, T., Jamin, M., and Ruigrok, R. W. H. (2010). Structure of the Dimerization Domain of the Rabies Virus Phosphoprotein. *J. Virol.* 84, 3707–3710. doi: 10.1128/jvi.02557-09
- Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010
- Kaur Saini, H., and Fischer, D. (2007). Structural and functional insights into Mimivirus ORFans. *BMC Genomics* 8:115. doi: 10.1186/1471-2164-8-115
- Khandia, R., Singhal, S., Kumar, U., Ansari, A., Tiwari, R., Dhama, K., et al. (2019). Analysis of nipah virus codon usage and adaptation to hosts. *Front. Microbiol.* 10:886. doi: 10.3389/fmicb.2019.00886
- Kitajima, E. W., and Alberti, G. (2014). Anatomy and fine structure of *Brevipalpus* mites (Tenuipalpidae) – Economically important plant virus vectors. Part 7. Ultrastructural detection of cytoplasmic and nuclear types of *Brevipalpus* transmitted viruses. *Zoologica* 160, 174–192.
- Kitajima, E. W., Rodrigues, J. C. V., and Freitas-Astua, J. (2010). An annotated list of ornamentals naturally found infected by *Brevipalpus* mite-transmitted viruses. *Sci. Agric.* 67, 348–371. doi: 10.1590/S0103-901620100003000014
- Konaté, M. M., Plata, G., Park, J., Usmanova, D. R., Wang, H., and Vitkup, D. (2019). Molecular function limits divergent protein evolution on planetary timescales. *Life* 8:39705. doi: 10.7554/eLife.39705
- Kondo, H., Chiba, S., Maruyama, K., Andika, I. B., and Suzuki, N. (2019). A novel insect-infecting virga/nege-like virus group and its pervasive endogenization into insect genomes. *Virus Res.* 262, 37–47. doi: 10.1016/j.virusres.2017.11.020
- Kondo, H., Fujita, M., Hisano, H., Hyodo, K., Andika, I. B., and Suzuki, N. (2020). Virome Analysis of Aphid Populations That Infest the Barley Field: The Discovery of Two Novel Groups of Nege/Kita-Like Viruses and Other Novel RNA Viruses. *Front. Microbiol.* 11:509. doi: 10.3389/fmicb.2020.00509
- Kondo, H., Hirota, K., Maruyama, K., Andika, I. B., and Suzuki, N. (2017). A possible occurrence of genome recombination among bipartite rhabdoviruses. *Virology* 508, 18–25. doi: 10.1016/j.virol.2017.04.027
- Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient virus world and evolution of cells. *Biol. Direct* 1:29. doi: 10.1186/1745-6150-1-29
- Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 60, 2256–2268. doi: 10.1107/S0907444904026460
- Kuchibhatla, D. B., Sherman, W. A., Chung, B. Y. W., Cook, S., Schneider, G., Eisenhaber, B., et al. (2014). Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently “orphan” viral proteins. *J. Virol.* 88, 10–20. doi: 10.1128/JVI.02595-13
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096
- Lazareva, E. A., Lezzhov, A. A., Chergintsev, D. A., Golyshv, S. A., Dolja, V. V., Morozov, S. Y., et al. (2021). Reticulon-like properties of a plant virus-encoded movement protein. *New Phytol.* 229, 1052–1066. doi: 10.1111/nph.16905
- Leastro, M. O., Castro, D. Y. O., Freitas-Astúa, J., Kitajima, E. W., Pallás, V., and Sánchez-Navarro, J. Á. (2020). Citrus Leprosis Virus C Encodes Three Proteins With Gene Silencing Suppression Activity. *Front. Microbiol.* 11:1231. doi: 10.3389/fmicb.2020.01231
- Leastro, M. O., Kitajima, E. W., Silva, M. S., Resende, R. O., and Freitas-Astúa, J. (2018). Dissecting the Subcellular Localization, Intracellular Trafficking, Interactions, Membrane Association, and Topology of Citrus Leprosis Virus C Proteins. *Front. Plant Sci.* 9:1299. doi: 10.3389/fpls.2018.01299
- Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239
- Li, H., Li, F., Zhang, M., Gong, P., and Zhou, X. (2020). Dynamic Subcellular Localization, Accumulation, and Interactions of Proteins From Tomato Yellow Leaf Curl China Virus and Its Associated Betasatellite. *Front. Plant Sci.* 11:840. doi: 10.3389/fpls.2020.00840
- Locali-Fabris, E. C., Freitas-Astúa, J., Souza, A. A., Takita, M. A., Astúa-Monge, G., Antonioli-Luizon, R., et al. (2006). Complete nucleotide sequence, genomic organization and phylogenetic analysis of citrus leprosis virus cytoplasmic type. *J. Gen. Virol.* 87, 2721–2729. doi: 10.1099/vir.0.82038-0
- Malik, S. S., Azem-e-Zahra, S., Kim, K. M., Caetano-Anollés, G., and Nasir, A. (2017). Do viruses exchange genes across superkingdoms of life? *Front. Microbiol.* 8:02110. doi: 10.3389/fmicb.2017.02110
- Mann, K. S., Johnson, K. N., Carroll, B. J., and Dietzgen, R. G. (2016). Cytorhabdovirus P protein suppresses RISC-mediated cleavage and RNA silencing amplification in planta. *Virology* 490, 27–40. doi: 10.1016/j.virol.2016.01.003
- Martinez, N., Ribeiro, E. A., Leyrat, C., Tarbouriech, N., Ruigrok, R. W. H., and Jamin, M. (2013). Structure of the C-Terminal Domain of Lettuce Necrotic Yellows Virus Phosphoprotein. *J. Virol.* 87, 9569–9578. doi: 10.1128/jvi.00999-13
- Masters, P. S., and Banerjee, A. K. (1988). Resolution of multiple complexes of phosphoprotein NS with nucleocapsid protein N of vesicular stomatitis virus. *J. Virol.* 62, 2651–2657. doi: 10.1128/JVI.62.8.2651-2657.1988
- Mavrakis, M., McCarthy, A. A., Roche, S., Blondel, D., and Ruigrok, R. W. H. (2004). Structure and function of the C-terminal domain of the polymerase cofactor of rabies virus. *J. Mol. Biol.* 343, 819–831. doi: 10.1016/j.jmb.2004.08.071
- Melzer, M. J., Sether, D. M., Borth, W. B., and Hu, J. S. (2012). Characterization of a virus infecting Citrus volkameriana with citrus leprosis-like symptoms. *Phytopathology* 102, 122–127. doi: 10.1094/PHYTO-01-11-0013
- Melzer, M. J., Simbajon, N., Carillo, J., Borth, W. B., Freitas-Astúa, J., Kitajima, E. W., et al. (2013). A cilevirus infects ornamental hibiscus in Hawaii. *Arch. Virol.* 2013, 1745–1740. doi: 10.1007/s00705-013-1745-0
- Miller, W. A., Shen, R., Staplin, W., and Kanodia, P. (2016). Noncoding RNAs of plant viruses and viroids: Sponges of host translation and RNA interference machinery. *Mol. Plant Microbe Interact.* 29, 156–164. doi: 10.1094/MPMI-10-15-0226-FI
- Morozov, S. Y., and Solovyev, A. G. (2020). Small hydrophobic viral proteins involved in intercellular movement of diverse plant virus genomes. *AIMS Microbiol.* 6, 305–329. doi: 10.3934/microbiol.2020019
- Morozov, S. Y., Lazareva, E. A., and Solovyev, A. G. (2020). Sequence Relationships of RNA Helicases and Other Proteins Encoded by Blunervirus RNAs Highlight Recombinant Evolutionary Origin of Kitaviral Genomes. *Front. Microbiol.* 11:561092. doi: 10.3389/fmicb.2020.561092
- Mueller, S., Papamichail, D., Coleman, J. R., Skiena, S., and Wimmer, E. (2006). Reduction of the Rate of Poliovirus Protein Synthesis through Large-Scale Codon Deoptimization Causes Attenuation of Viral Virulence by Lowering Specific Infectivity. *J. Virol.* 80, 9687–9696. doi: 10.1128/jvi.00738-06
- Newburn, L. R., and White, K. A. (2015). Cis-acting RNA elements in positive-strand RNA plant virus genomes. *Virology* 479–480, 434–443. doi: 10.1016/j.virol.2015.02.032
- Ng, W. M., Stelfox, A. J., and Bowden, T. A. (2020). Unraveling virus relationships by structure-based phylogenetic classification. *Virus Evol.* 6:veaa003. doi: 10.1093/ve/veaa003
- Olmedo-Velarde, A., Hu, J., and Melzer, M. J. (2021). A Virus Infecting Hibiscus rosa-sinensis Represents an Evolutionary Link Between Cileviruses and Higreviruses. *Front. Microbiol.* 12:988. doi: 10.3389/fmicb.2021.660237
- Omer, S., Harlow, T. J., and Gogarten, J. P. (2017). Does Sequence Conservation Provide Evidence for Biological Function? *Trends Microbiol.* 25, 11–18. doi: 10.1016/j.tim.2016.09.010
- Ouzounis, C., Sander, C., Scharf, M., and Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.* 232, 805–825. doi: 10.1006/jmbi.1993.1433
- Peluso, R., and Moyer, S. (1988). Viral proteins required for the *in vitro* replication of vesicular stomatitis virus defective interfering particle genome RNA. *Virology* 162, 369–376. doi: 10.1016/0042-6822(88)90477-1
- Ponting, C. P. (2017). Biological function in the twilight zone of sequence conservation. *BMC Biol.* 15:71. doi: 10.1186/s12915-017-0411-5
- Puigbò, P., Aragonès, L., and Garcia-Vallvé, S. (2010). RCDI/eRCDI: A web-server to estimate codon usage deoptimization. *BMC Res. Notes* 3:87. doi: 10.1186/1756-0500-3-87

- Quito-Avila, D. F., Brannen, P. M., Cline, W. O., Harmon, P. F., and Martin, R. R. (2013). Genetic characterization of Blueberry necrotic ring blotch virus, a novel RNA virus with unique genetic features. *J. Gen. Virol.* 94, 1426–1434. doi: 10.1099/vir.0.050393-0
- Quito-Avila, D. F., Freitas-Astúa, J., and Melzer, M. J. (2021). Bluner-, Cile-, and Higreviruses (Kitaviridae). *Encycl. Virol.* 2021, 247–251. doi: 10.1016/B978-0-12-809633-8.21248-X
- Ramos-González, P. L., Chabi-Jesus, C., Arena, G. D., Tassi, A. D., Kitajima, E. W., and Freitas-Astúa, J. (2018). Citrus leprosis: a unique multietiological disease. *Citrus Am.* 1, 4–19. doi: 10.6084/m9.figshare.13513857.v2
- Ramos-González, P. L., Chabi-Jesus, C., Guerra-Peraza, O., Breton, M. C., Arena, G. D., Nunes, M. A., et al. (2016). Phylogenetic and molecular variability studies reveal a new genetic clade of Citrus leprosis virus C. *Viruses* 8:153. doi: 10.3390/v8060153
- Ramos-González, P. L., Santos, G. F., dos, Chabi-Jesus, C., Harakava, R., Kitajima, E. W., et al. (2020). Passion Fruit Green Spot Virus Genome Harbors a New Orphan ORF and Highlights the Flexibility of the 5'-End of the RNA2 Segment Across Cileviruses. *Front. Microbiol.* 11:00206. doi: 10.3389/fmicb.2020.0206
- Ravanti, J. J., Martínez-Castillo, A., and Abrescia, N. G. A. (2020). Superimposition of viral protein structures: A means to decipher the phylogenies of viruses. *Viruses* 12:v12101146. doi: 10.3390/v12101146
- Rhys, G. G., Wood, C. W., Beesley, J. L., Zaccai, N. R., Burton, A. J., Brady, R. L., et al. (2019). Navigating the Structural Landscape of de Novo α -Helical Bundles. *J. Am. Chem. Soc.* 141, 8787–8797. doi: 10.1021/jacs.8b13354
- Rokyta, D., Badgett, M. R., Molineux, I. J., and Bull, J. J. (2002). Experimental Genomic Evolution: Extensive Compensation for Loss of DNA Ligase Activity in a Virus. *Mol. Biol. Evol.* 19, 230–238. doi: 10.1093/oxfordjournals.molbev.a004076
- Roy, A., Choudhary, N., Guillermo, L. M., Shao, J., Govindarajulu, A., Achour, D., et al. (2013). A novel virus of the genus *Cilevirus* causing symptoms similar to citrus leprosis. *Phytopathology* 103, 488–500. doi: 10.1094/PHYTO-07-12-0177-R
- Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248
- Rstudio_team, (2021). *RStudio | Open source & professional software for data science teams - RStudio*. Available online at: <https://www.rstudio.com/> (accessed June 9, 2021)
- Schlub, T. E., Buchmann, J. P., and Holmes, E. C. (2018). A simple method to detect candidate overlapping genes in viruses using single genome sequences. *Mol. Biol. Evol.* 35, 2572–2581. doi: 10.1093/molbev/msy155
- Sharp, P., and Li, W. (1987). The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. doi: 10.1093/NAR/15.3.1281
- Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature* 540, 539–543. doi: 10.1038/nature20167
- Sievers, F., and Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* 1079, 105–116. doi: 10.1007/978-1-62703-646-7_6
- Singh, U., and Wurtele, E. S. (2020). Genetic novelty: How new genes are born. *Elife* 9:55136. doi: 10.7554/eLife.55136
- Solovyev, A. G., and Morozov, S. Y. (2017). Non-replicative Integral Membrane Proteins Encoded by Plant Alpha-Like Viruses: Emergence of Diverse Orphan ORFs and Movement Protein Genes. *Front. Plant Sci.* 8:1820. doi: 10.3389/fpls.2017.01820
- Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692–702. doi: 10.1038/nrg3053
- Vakirlis, N., Carvunis, A.-R., and McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife* 9:53500. doi: 10.7554/eLife.53500
- Van Oss, S. B., and Carvunis, A. R. (2019). De novo gene birth. *PLoS Genet.* 15:1008160. doi: 10.1371/journal.pgen.1008160
- Vasilakis, N., Forrester, N. L., Palacios, G., Nasar, F., Savji, N., Rossi, S. L., et al. (2013). Negevirus: a Proposed New Taxon of Insect-Specific Viruses with Wide Geographic Distribution. *J. Virol.* 87, 2475–2488. doi: 10.1128/JVI.00776-12
- Vignuzzi, M., and López, C. B. (2019). Defective viral genomes are key drivers of the virus–host interaction. *Nat. Microbiol.* 4, 1075–1087. doi: 10.1038/s41564-019-0465-y
- Vinokurov, K. S., and Koloniuk, I. (2019). Discovery and characterization of a novel alphavirus-like RNA virus from the red firebug *Pyrrhocoris apterus* L. (Heteroptera). *J. Invertebr. Pathol.* 166:107213. doi: 10.1016/j.jip.2019.107213
- Wright, F. (1990). The “effective number of codons” used in a gene. *Gene* 87, 23–29. doi: 10.1016/0378-1119(90)90491-9
- Yang, J., and Zhang, Y. (2015). I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* 43, W174–W181. doi: 10.1093/nar/gkv342
- Yin, Y., and Fischer, D. (2008). Identification and investigation of ORFans in the viral world. *BMC Genomics* 9:24. doi: 10.1186/1471-2164-9-24
- Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinforma.* 57, 702–710. doi: 10.1002/prot.20264
- Zwart, M. P., Willemsen, A., Daròs, J. A., and Elena, S. F. (2014). Experimental evolution of pseudogenization and gene loss in a plant RNA virus. *Mol. Biol. Evol.* 31, 121–134. doi: 10.1093/molbev/mst175

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Ramos-González, Pons, Chabi-Jesus, Arena and Freitas-Astua. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.