Microbiology Spectrum

RESEARCH ARTICLE

Genetic Diversity and Characterization of Circular Replication (Rep)-Encoding Single-Stranded (CRESS) DNA Viruses

Perumal Arumugam Desingu,[a] K. Nagarajan[b]

[a]Department of Microbiology and Cell Biology, Indian Institute of Science, Bengaluru, Karnataka, India
[b]Department of Veterinary Pathology, Madras Veterinary College, Veterinary and Animal Sciences University (TANUVAS), Vepery, Chennai, Tamil Nadu, India

ABSTRACT The CRESS-DNA viruses are the ubiquitous virus detected in almost all eukaryotic life trees and play an essential role in the maintaining ecosystem of the globe. Still, their genetic diversity is not fully understood. Here, we bring to light the genetic diversity of replication (Rep) and capsid (Cap) proteins of CRESS-DNA viruses. We divided the Rep protein of the CRESS-DNA virus into 10 clusters using CLANS and phylogenetic analyses. Also, most of the Rep protein in Rep cluster 1 (R1) and R2 (*Circoviridae*, *Smacoviridae*, *Nanoviridae*, and CRESSV1-5) contain the Viral_Rep superfamily and P-loop_NTPase superfamily domains, while the Rep protein of viruses in other clusters has no such characterized functional domain. The *Circoviridae*, *Nanoviridae*, and CRESSV1-3 viruses contain two domains, such as Viral_Rep and P-loop_NTPase; the CRESSV4 and CRESSV5 viruses have only the Viral_Rep domain; most of the sequences in the pCRESS-related group have only P-loop_NTPase; and *Smacoviridae* do not have these two domains. Further, we divided the Cap protein of the CRESS-DNA virus into 20 clusters using CLANS and phylogenetic analyses. The Rep and Cap proteins of *Circoviridae* and *Smacoviridae* are grouped into a specific cluster. Cap protein of CRESS-DNA viruses grouped with one cluster and Rep protein with another cluster. Further, our study reveals that selection pressure plays a significant role in the evolution of CRESS-DNA viruses' Rep and Cap genes rather than mutational pressure. We hope this study will help determine the genetic diversity of CRESS-DNA viruses as more sequences are discovered in the future.

IMPORTANCE The genetic diversity of CRESS-DNA viruses is not fully understood. CRESS-DNA viruses are classified as CRESSV1 to CRESSV6 using only Rep protein. This study revealed that the Rep protein of the CRESS-DNA viruses is classified as CRESSV1 to CRESSV6 groups and the new Smacoviridae-related, CRESSV2-related, pCRESS-related, Circoviridae-related, and 1 to 4 outgroups, according to the Viral_Rep and P-loop_NTPase domain organization, CLANS, and phylogenetic analysis. Furthermore, for the first time in this study, the Cap protein of CRESS-DNA viruses was classified into 20 distinct clusters by CLANS and phylogenetic analysis. Through this classification, the genetic diversity of CRESS-DNA viruses clarifies the possibility of recombinations in Cap and Rep proteins. Finally, it has been shown that selection pressure plays a significant role in the evolution and genetic diversity of Cap and Rep proteins. This study explains the genetic diversity of CRESS-DNA viruses and hopes that it will help classify future detected viruses.

KEYWORDS classification, CRESS DNA virus, Cap gene, Rep gene, evolution, genetic diversity

Circular replication (Rep)-encoding single-stranded (CRESS)-DNA viruses are ubiquitous viruses that are reported to spread worldwide and infect almost all of the eukaryotic tree of life (1 to 3). CRESS-DNA viruses have also been found in environmental samples such as sewage, seawater, lakes, and springs (4 to 11). Recently, ssDNA viruses

Editor Eleanor A. Powell, University of Cincinnati

Address correspondence to Perumal Arumugam Desingu, perumald@iisc.ac.in or padesingu@gmail.com.

The authors declare no conflict of interest.

Received 29 March 2022
Accepted 7 October 2022
Published 8 November 2022

have been classified into 13 families (1); 10 families (*Anelloviridae*, *Bacilladnaviridae*, *Bidnaviridae*, *Circoviridae*, *Geminiviridae*, *Genomoviridae*, *Nanoviridae*, *Parvoviridae*, *Redondoviridae*, and *Smacoviridae*) are reported from the eukaryotes (12). These viruses are commonly found with replication initiation protein (Rep) and structural capsid protein (Cap) (1, 12). Of the 10 ssDNA virus families found in eukaryotes, the *Bidnaviridae* and *Parvoviridae* families have the linear genome topology, and the *Anelloviridae* family has a different Rep protein, with the remaining seven families containing circular ssDNA with Rep protein containing the preserved HUH endonuclease motif and superfamily 3 helicase (S3H) domain (12).

Recently, these characterized seven families of ssDNA viruses infect eukaryotes (*Bacilladnaviridae*, *Circoviridae*, *Geminiviridae*, *Genomoviridae*, *Nanoviridae*, *Redondoviridae*, and *Smacoviridae*), and uncharacterized CRESS-DNA viruses have been classified into separate groups using this characteristic and conserved two-domain Rep protein (12). Thus, unclassified CRESS-DNA viruses are classified as CRESSV1 through CRESSV6 (12). So far, the Rep protein of CRESS-DNA viruses has been characterized to contain the HUH motif and S3H domain (1, 12). It is also not widely known what other domains are present in the rep protein of CRESS-DNA viruses that accumulate day by day through metagenomic sequencing in different environmental samples, and how they help classify CRESS-DNA viruses. Furthermore, the classification of CRESS DNA viruses by capsid proteins is challenging due to the lack of conserved portions of the capsid proteins of the CRESS DNA viruses as found in the Rep protein (12). In particular, the capsid proteins of CRESS DNA viruses are reported to be derived from a number of RNA viruses (13 to 16). It is also largely unknown which of the Cap proteins of the CRESS-DNA viruses that accumulate day by day through metagenomic sequencing in different environmental samples are related to the RNA viruses and the diversity in the Cap proteins of the CRESS-DNA viruses. A recent study found that capsid proteins in cruciviruses (CRESS DNA virus) are highly conserved and possibly acquired from RNA viruses, but the Rep protein is more diversified than Cap protein (17). From these, it is speculated that cruciviruses may have obtained Rep protein from different CRESS-DNA viruses by recombination (17). Therefore, it appears that the genetic variation and recombination of CRESS-DNA viruses can be detected by dividing the capsid proteins of almost identical CRESS-DNA viruses into groups. However, it should be noted that there is no mechanism for classifying the capsid proteins of CRESS-DNA viruses so far.

The present study systematically classified the CRESS-DNA viruses Rep and Cap proteins and reported the presence of different group-specific domain organizations in the Rep protein. Further, it explains the recombination-mediated evolution of the CRESS-DNA virus and reveals that selection pressure plays a significant role in the evolution of CRESS-DNA viruses' Rep and Cap genes rather than mutational pressure.

RESULTS

CLANS-based classification of CRESS-DNA Rep protein. As a first step toward understanding the genetic diversity of the CRESS DNA viruses, we analyzed the interrelationship between the core viral proteins such as Rep and Cap proteins of various isolates of CRESS-DNA viruses. We first chose the Rep protein for our analysis since it shows a high degree of conservation among the CRESS-DNA viruses (2, 18 to 20). To explore the sequence diversity of the Rep protein of CRESS-DNA viruses, we collected 1,160 (sequences details are provided in Data Set S1 in the supplemental material) amino acid sequences of CRESS-DNA viruses from the NCBI database and grouped them based on pairwise sequence similarity using the CLANS (CLuster ANalysis of Sequences) tool (21, 22). The analysis grouped the CRESS-DNA virus Rep protein sequences into 10 different clusters (R1 to R10) (Rep Cluster 1 [R1]) (a minimum of 10 viral sequences to a maximum of 487 sequences per group) (Fig. 1A; the individual sequence details in the different clusters are listed in Data Set S2). The majority of the clusters, except clusters R7 and R8, showed interconnections at a *P*-value threshold of $1e^{-2}$ (Fig. 1A). Further, we also observed three different superclusters (Supercluster 1—clusters R1, R2, R4, and R5; Supercluster 2—clusters R3, R6, and R9; Supercluster 3—clusters R7 and R8) at a *P*-value threshold of $1e^{-5}$ (Fig. S1A).

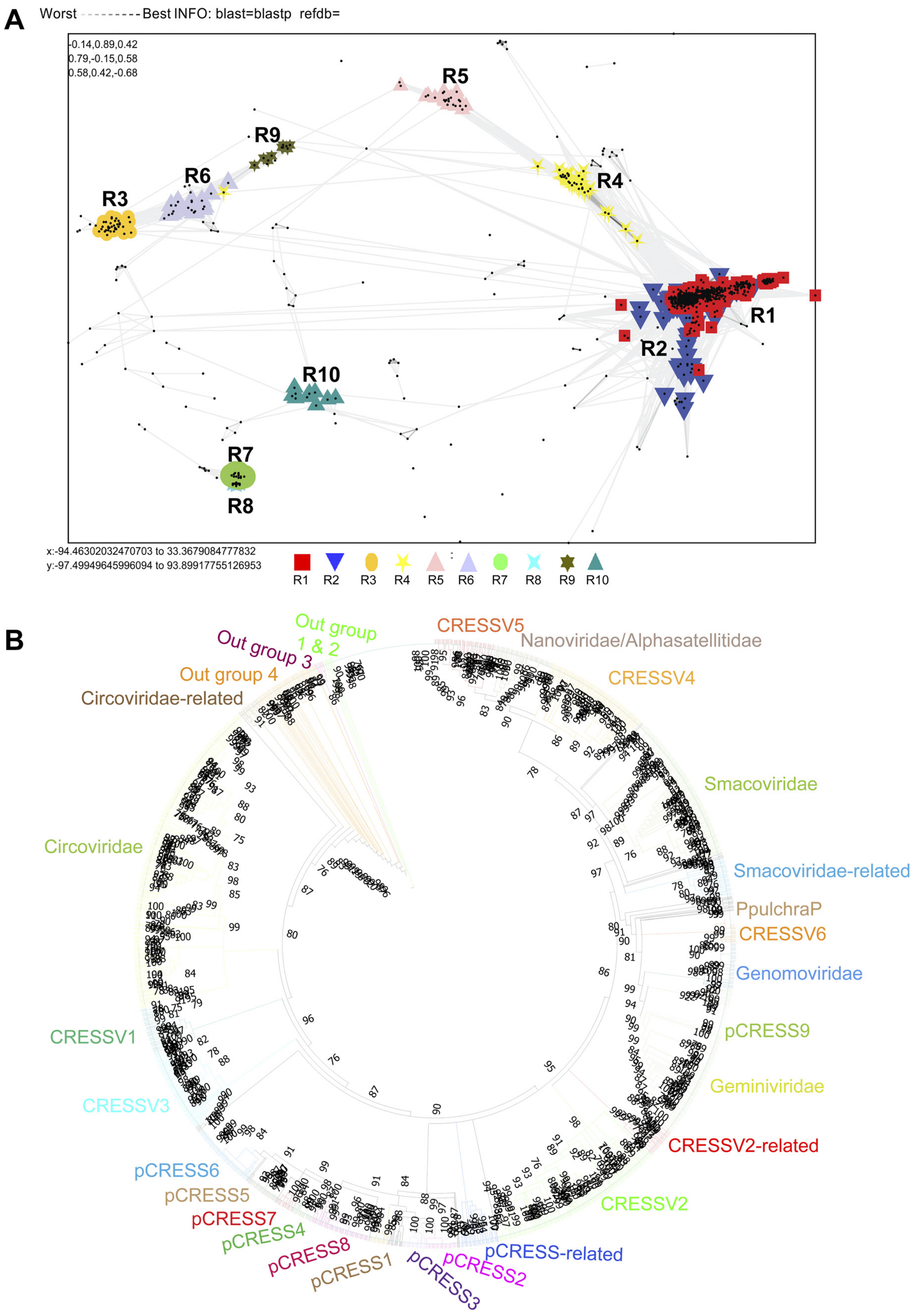

FIG 1 CLANS analysis-based classification of CRESS-DNA virus Rep protein. (A) Representative CRESS-DNA virus Rep protein sequences were clustered using CLANS Toolkit by their pairwise sequence similarity network. A total of 1,160 amino acid sequences of Rep protein

(Continued on next page)

For a better understanding of the genetic diversity of the CRESS-DNA virus, we classified the CRESS-DNA viruses into three broad groups as follows: (i) culturable CRESS-DNA viruses (*Circoviridae*, *Geminiviridae*, *Smacoviridae*, *Cruciviridae*, etc.) (23), which are infective; (ii) replication-competent circular DNA (rccDNA), which includes the bovine meat and milk factors (BMMF) and Sphinx infective DNA molecule (24); and (iii) uncharacterized and uncultivated CRESS-DNA (1), which was detected as a DNA molecule in the viral metagenomic analysis. Interestingly, most of the sequences in clusters R1 and R2 grouped with highly characterized and culturable viral families of *Circoviridae*, *Smacoviridae*, and *Cruciviridae*. Further, cluster R8 sequences exclusively belonged to BMMF of rccDNA, while all other clusters included uncultured CRESS-DNA viruses. The remaining clusters (R3, R4, R5, R6, R7, R9, and R10) were classified as uncharacterized and uncultivated CRESS-DNA.

Domain organization in the CRESS-DNA virus Rep protein. We were interested to find out if these different Rep protein clusters have any significant differences in the organization of functional domains. The Rep genes of CRESS-DNA viruses have been reported to contain two main functional domains/motif, HUH endonuclease motif and superfamily 3 helicase domains (1, 25). In this context, we analyzed the domain organization of the Rep protein of viruses from different clusters using the Conserved Domain search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?) (26 to 29). Interestingly, we found that only the Rep protein of viruses in clusters R1 and R2 displayed functional domains such as Viral_Rep superfamily (Cdd:pfam02407) and P-loop_NTPase superfamily (Cdd:pfam00910). On the other hand, cluster R8 (BMMF) sequences contained Rep_1 superfamily (Cdd:pfam01446), a homologous domain to the Rep1 domain of bacteria involved in plasmid replication. Moreover, we did not find any known putative functional domains in our conserved domain analysis of other clusters consisting of uncultured viruses (Rep clusters R3, R4, R5, R6, R7, R9, and R10). However, it should be noted that R4 and R5 in clusters of Rep protein that do not express these putative functional domains have evolutionary links with R1 and R2 clusters that express functional domains (Fig. 1A). Similarly, cluster R7 has evolutionary links with cluster R8 that holds the Rep_1 domain (Fig. 1A).

We then analyzed the diversity of the Rep protein domains in depth belonging to clusters R1 and R2 to further classify the viruses in these clusters, which are highly related to culturable viruses (sequences details are provided in Data Set S3). To explore the different domain organizations present in the viruses of clusters R1 and R2, we reclustered them into 10 subclusters (clusters a to j) at a *P*-value threshold of $1e^{-38}$ (Fig. S1B). Of these 10 subclusters, we noted that subclusters such as a, b, g, and h formed a single group (group 1), and c, d, i, and j subclusters formed a separate group (group 2) (Fig. S1B). Furthermore, it can be seen that there are some evolutionary links between these two groups (group 1 and group 2), but the subclusters e and f together as a separate group (group 3) (Fig. S1B).

The viruses in the subcluster a majorly contain two main domains: Viral_Rep and P-loop_NTPase domains. Some sequences had one of the following additional domains in between Viral_Rep and P-loop_NTPase domain, such as the AAA ATPase domain, Penta-EF hand, DNA-binding ATP-dependent protease La, Type III secretion system protein PrgH-EprH (PrgH), and parvovirus nonstructural protein NS1 (Data Set S4). Similarly, cluster b viruses also contained the Viral_Rep and P-loop_NTPase domains. In addition, few sequences had a third functional domain between Viral_Rep and P-loop_NTPase domains, such as AAA+-type ATPase, SpoVK/Ycf46/Vps4 family, or Type VI protein secretion system component VasK. Moreover, the cluster b viruses also contained a combination of domains such as (i) incomplete Viral_Rep + P-loop_NTPase

FIG 1 Legend (Continued)

(Supplemental Data 1) of CRESS-DNA viruses were used in this analysis. Classification of clusters was carried out by a network-based method using offset values and global average with maximum rounds 10,000 in CLANS Toolkit analysis. The *P* value of $\leq 1e^{-02}$ was used to show the lines connecting the sequences. (B) Phylogenetic relationship of Rep protein of CRESS-DNA viruses. The maximum-likelihood method inferred the evolutionary history using the Subtree-Pruning-Regrafting algorithm in PhyML 3.3_1. A total of 1,509 amino acid sequences of Rep protein (Supplemental Data 5) of CRESS-DNA viruses were used in this analysis.

and (ii) Viral_Rep + incomplete P-loop_NTPase (Data Set S4). Also, most sequences in subcluster g contain Viral_Rep + P-loop_NTPase domains, and some sequences are incomplete with these domains or possess only one of the two domains (Data Set S4). Significantly, most sequences in the subcluster h contain only the P-loop_NTPase domains (Data Set S4). More interestingly, it was revealed that most of the sequences in subcluster c and d have only Viral_Rep domains (Data Set S4). Also, subcluster i, which is grouped with subclusters c and d, contains the Viral_Rep+P-loop_NTPase domains, and subcluster j contains the Viral_Rep domain+incomplete P-loop_NTPase domains (Data Set S4). Finally, it is essential to note that the sequences in subclusters e and f have no known putative functional domains (Data Set S4). Collectively, our analyses reveal a vast diversity of domains in the viral Rep protein of CRESS-DNA viruses ranging from lack of any known functional domains to the combination of multiple functional domains.

Phylogenetic tree-based classification of CRESS-DNA virus Rep protein and group-specific domain organization. Recently, CRESS-DNA viruses have been classified into different groups CRESSV1 to CRESSV6 using Rep protein (1, 12). Therefore, we are interested in finding out which CRESSV groups the clusters of Rep protein with various organizations of domain identified in this current study belong to. To find out, we performed a phylogenetic analysis of the sequences of the Rep protein used in this present study with the sequences used to classify the CRESS-DNA viruses in the previous study (1) (Data Set S5). In this phylogenetic analysis, we observed that CRESSV6, *P. pulchra*, pCRESS9, *Genomoviridae*, and *Geminiviridae* were grouped together, and CRESSV4, CRESSV5, and *Nanoviridae* have formed another group (Fig. 1B), as in the previous study (1, 12). As in the previous study (1), in plasmid CRESS sequences (pCRESS), CRESS1, pCRESS2, and pCRESS3 formed a separate group, and CRESS4, pCRESS5, CRESS6, pCRESS7, and pCRESS8 formed another group (Fig. 1B). Furthermore, CRESSV1 and CRESSV3 revealed a close association with *Circoviridae* (Fig. 1B). In addition, the group that showed a relationship with Smacoviridae was called Smacoviridae-related; the group that showed contact with the CRESSV2 sequences was called CRESSV2-related; the group that showed a relationship with the pCRESS sequences was called pCRESS-related; the group that showed contact with *Circoviridae* was called Circoviridae-related; and also, the groups formed an outgroup were named as outgroups 1 to 4 (Fig. 1B).

We first explored the subclusters a to j created by the clusters R1 and R2 with domain organizations. Notably, we observed the subcluster a and b sequences that revealed the domain organization Viral_Rep+P-loop_NTPase grouped into the CRESSV1, CRESSV2, CRESSV3, *Circoviridae*, and Circoviridae-related groups (Fig. 1B; Data Set S4). Interestingly, subclusters c and d, which contain only Viral_Rep domains, are grouped with CRESSV4 and CRESSV5, respectively (Fig. 1B; Data Set S4). We observed that subclusters e and f grouped with *Smacoviridae* without any known putative functional domains (Fig. 1B; Data Set S4). Significantly, subcluster g, which displays mostly Viral_Rep +P-loop_NTPase domains and some sequences with these domains incomplete or with only one of the two domains, formed the CRESSV2-related group (Fig. 1B; Data Set S4). Similarly, it should be noted that the subcluster h, which contains, in most of the sequences, only P-loop_NTPase domains, formed the pCRESS-related group (Fig. 1B; Data Set S4). Also, subcluster i often has Viral_Rep+P-loop_NTPase domains, and subclusters j has Viral_Rep domain+incomplete P-loop_NTPase domains grouped with Nanoviridae (Fig. 1B; Data Set S4).

Next, we explored clusters R3, R4, R5, R6, and R9 without any known putative functional domains. Of these clusters, R4 and R5 combined with clusters R1 and R2 to form Supercluster 1 (Fig. S1A). Note that cluster R4 forms the Smacoviridae-related group, and cluster R5 forms the outgroup 1 (Fig. 1B; Data Set S4). We observed that the R3, R6, and R9 clusters formed Supercluster 2 and created outgroup 4, outgroup 3, and outgroup 2, respectively (Fig. 1B; Data Set S4). These results show that CRESS-DNA virus Rep proteins group into the phylogenetic tree, as is the case with CLANS clustering and domain organizations.

Classification of CRESS-DNA virus Cap protein using CLANS. While the Rep protein of CRESS-DNA viruses is evolutionarily conserved, the Cap protein is highly diverse (2, 18 to 20). Therefore, previous studies analyzed the evolution of capsid proteins primarily by structural fold comparisons rather than sequence comparisons (23, 30 to 32). However, we took advantage of the recent explosion in the metagenomic data from CRESS-DNA viruses. We employed a sequence comparison method to classify and identify the genetic diversity of CRESS-DNA virus Cap protein. We collected 1,823 amino acid sequences of CRESS-DNA viruses from the NCBI database and grouped them based on pairwise similarity (CLANS analysis) (sequences details are provided in Data Set S6). The analysis classified the CRESS-DNA virus Cap gene sequences into 20 different clusters (minimum of 10 sequences per group was considered to classify them as an individual cluster) (the individual sequence details in the different clusters are listed in Data Set S7). Most of the clusters show interconnections with other clusters, except the clusters C3 (Cap cluster 3), C4, C19, and C20, which were isolated from other clusters (orphan clusters) in a pairwise similarity network (Fig. S2) (*P*-value threshold of $1e^{-02}$). Cluster C1 of CRESS-DNA virus sequences clustered with *Circoviridae* viruses, while cluster C2 showed a relationship with *Geminiviridae* viruses, cluster C3 sequences clustered with *Smacoviridae* viruses, and cluster C6 sequences clustered with *Cruciviridae* virus sequences (Data Set S7). Among the 20 clusters identified for the Cap protein of the CRESS-DNA viruses (Fig. 2A), the clusters C2, C7, C8, C11, C12, C15, and C18 form a supercluster (Fig. 2A) in the sequence similarity network analysis at a *P*-value threshold of $>1e^{-04}$. Similarly, the supercluster consists of clusters C1, C14, and C17 in CLANS analysis (Fig. 2A; Data Set S7). In addition, clusters C3, C4, C5, C6, C9, C10, C13, C16, C19, and C20 were isolated from other clusters (orphan clusters) in a pairwise similarity network (Fig. 2A; Data Set S7) (*P*-value threshold of $1e^{-04}$). Collectively, CRESS-DNA virus Cap proteins also split into separate groups in CLANS analysis and are thought to support the classification of Cap proteins.

Only *Cruciviridae* Cap proteins related to RNA viruses. In previous studies, it has been reported that the cap protein of the CRESS-DNA virus is related to the RNA virus (13 to 16), so we were interested to find out which of these 20 clusters is related to the RNA virus. To do this, we retrieved the RNA virus sequences associated with the Cap protein of the CRESS-DNA virus from the NCBI database and performed CLANS analysis (sequences details are provided in Data Set S8). This analysis noted that RNA viruses revealed association only with *Cruciviridae* virus sequences belonging to cluster C6 at a *P*-value threshold of $1e\text{-}^{02}$ (Fig. 2B; Data Set S9).

Phylogenetic tree-based classification of CRESS-DNA virus Cap protein. We examined whether CLANS analysis-based clustering of CRESS-DNA virus cap protein sequences also grouped into the phylogenetic tree. Because cap proteins do not have common domains as seen in CRESS-DNA virus Rep proteins, and the sequence alignments are low from most genetic variants, we performed separate phylogenetic analysis for (i) superclusters C1, C14, and C17; (ii) superclusters C2, C7, C8, C11, C12, C15, and C18; and (iii) orphan clusters such as C3, C4, C5, C6, C9, C10, C13, C16, C19, and C20. To do this, we first performed phylogenetic analysis using sequences from the C1, C14, and C17 clusters that formed the Cap protein supercluster. These clusters C1, C14, and C17 are well aligned (Data Set S10) and split into separate groups for the phylogenetic tree (Fig. 3A). Similarly, C2, C7, C8, C11, C12, C15, and C18 clusters are well aligned (Data Set S11) and split into separate groups for the phylogenetic tree (Fig. 3B). In particular, C8, C11, and C12 formed an outgroup, C8 was somewhat detached, and C11 and C12 grouped slightly closer together into the phylogenetic tree (Fig. 3B), as seen in the CLANS analysis (Fig. 2A). Similarly, the C7 and C15 clusters grouped in the phylogenetic tree (Fig. 3B), as seen in the CLANS analysis (Fig. 2A), and the C18 and some C2 sequences grouped together with this (C7 and C15) group (Fig. 3B). Also, although the cluster C2 sequences are majorly grouped together, it is noteworthy that some sequences are grouped together with a group formed by C8, C11, and C12 and a group created by C7, C15, and C18 (Fig. 3B). We then performed phylogenetic analysis separately for the orphan clusters C3, C4, C5, C6, C9, C10, C13,

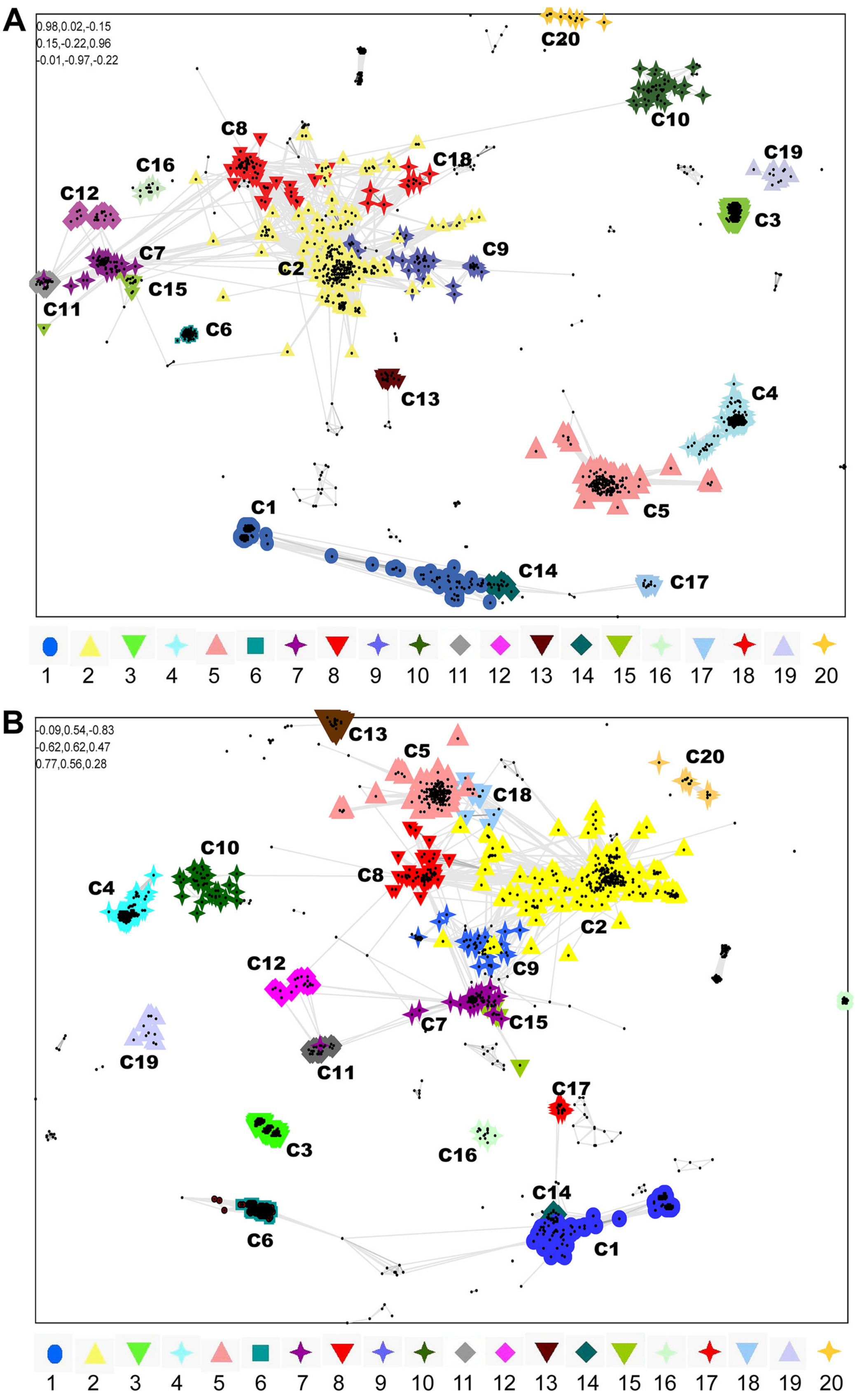

FIG 2 Sequence similarities (CLANS) analysis-based CRESS-DNA virus capsid protein clustering. (A) A total of 1,823 amino acid sequences of Cap protein of CRESS-DNA viruses (Supplemental Data 6) were used and classified by their

(Continued on next page)

C16, C19, and C20. Thus, the sequences in these clusters are well-aligned C3 (Data Set S12), C4 (Data Set S13), C5 (Data Set S14), C6 (Data Set S15), C9 (Data Set S16), C10 (Data Set S17), C13, C16, C19, and C20 (Data Set S18), to form the phylogenetic tree C3 (Fig. S3A), C4 (Fig. S3B), C5 (Fig. S4A), C6 (Fig. S4B), C9 (Fig. S5A), C10 (Fig. S4B), C13, C16, C19, and C20 (Fig. S5C).

Recombination mediated evolution of CRESS-DNA viruses. Recently, it has been reported that the cap protein of cruciviruses is very similar, but the Rep protein may be derived from different sources with greater diversity (17); we examined whether the sequences in the cluster of these 20 Cap proteins received the Rep protein from the same group or from different groups. To do this, we took the representative sequences in each Cap cluster and identified the phylogenetic tree groups that contain its Rep protein (Data Set S19). In this analysis, it appears that the sequences in the same Cap cluster have different groups of rep proteins (Data Set S19). From these, it can be inferred that the CRESS-DNA virus has the potential to acquire genetic diversity through recombination in the Cap and Rep genes.

Role of host codon usage selection pressure on Rep gene evolution. Since we observe homology at amino acid levels between the Rep gene of CRESS-DNA viruses but not any significant identity at the nucleotide sequence level, we suspected this might be due to this virus's host codon usage bias-based selection pressure. To explore this, we first analyzed the base composition of 1,115 nucleotide sequences of CRESS-DNA viruses' Rep genes (the details of nucleotide sequences used in the analysis are presented in Data Set S20) as AT to GC ratio can affect codon usage in microbes (33, 34). Our study revealed that the Rep gene of CRESS-DNA viruses contains A > T > G > C with AT% > GC% (average GC content is 43.6 ± SD 7.06) (Fig. 4A; Data Set S21). We next analyzed the codon usage bias using the effective number of codon usage (ENc) analysis. ENc values of <35 indicate high codon bias, and values of >50 show general random codon usage (35, 36). The Rep gene of CRESS-DNA viruses has ENc values ranging from 31 to 61, while most of the ENc values fall between 40 and 60 (average ENc 51.004 ± SD 5.73) (Fig. 4B; Data Set S21), indicating weak to strong codon usage bias. In addition, we calculated the relative synonymous codon usage (RSCU) value, which is the ratio between the observed to the expected value of synonymous codons for a given amino acid. A RSCU value of 1 indicates that there is no bias for that codon. In contrast, RSCU values of >1.0 have positive codon usage bias (defined as abundant codons), and RSCU values of <1.0 have negative codon usage bias (defined as less-abundant codons) (36, 37). Our analysis revealed that the RSCU values of 28 codons were >1, and 31 codons were <1 in the Rep gene of all the CRESS-DNA viruses (Fig. 4C; Data Set S21), clearly indicating a codon usage bias (both positive and negative).

Next, we performed ENc-GC3s plot analysis where the ENc values are plotted against the GC3s values (GC content at the third position in the codon) to determine the significant factors such as selection or mutation pressure affecting the codon usage bias (38). In this analysis, genes whose codon bias is affected by mutations will lie on or around the expected curve. In contrast, genes whose codon bias is affected by selection and other factors will lie beneath the expected curve (36, 38). Interestingly, we observed that most of the points fall below the expected curve in the ENc-GC3s plot analysis (Fig. 4D; Data Set S21), indicating the strong presence of selection pressure rather than mutation pressure. Similarly, neutrality plot analysis where GC12 values (av-

FIG 2 Legend (Continued)
pairwise sequence similarity network using CLANS. The clusters were classified using the network-based method using offset values and global average with a maximum of 10,000 in CLANS Toolkit analysis. The *P* value of $\leq 1e^{-05}$ was used to show the lines connecting the sequences. (B) Pairwise sequence similarity based on CRESS-DNA virus capsid protein and +RNA viruses relationship. Representative CRESS-DNA virus capsid protein sequences and their relationship with RNA viruses using CLANS Toolkit. A total of 1,967 amino acid sequences of Cap protein of CRESS-DNA viruses and +RNA viruses were used in this analysis (Supplemental Data 8). The clusters were classified using the network-based method using offset values and global average with a maximum of 10,000 in CLANS Toolkit analysis. The *P* value of $\leq 1e^{-02}$ was used to show the lines connecting the sequences.

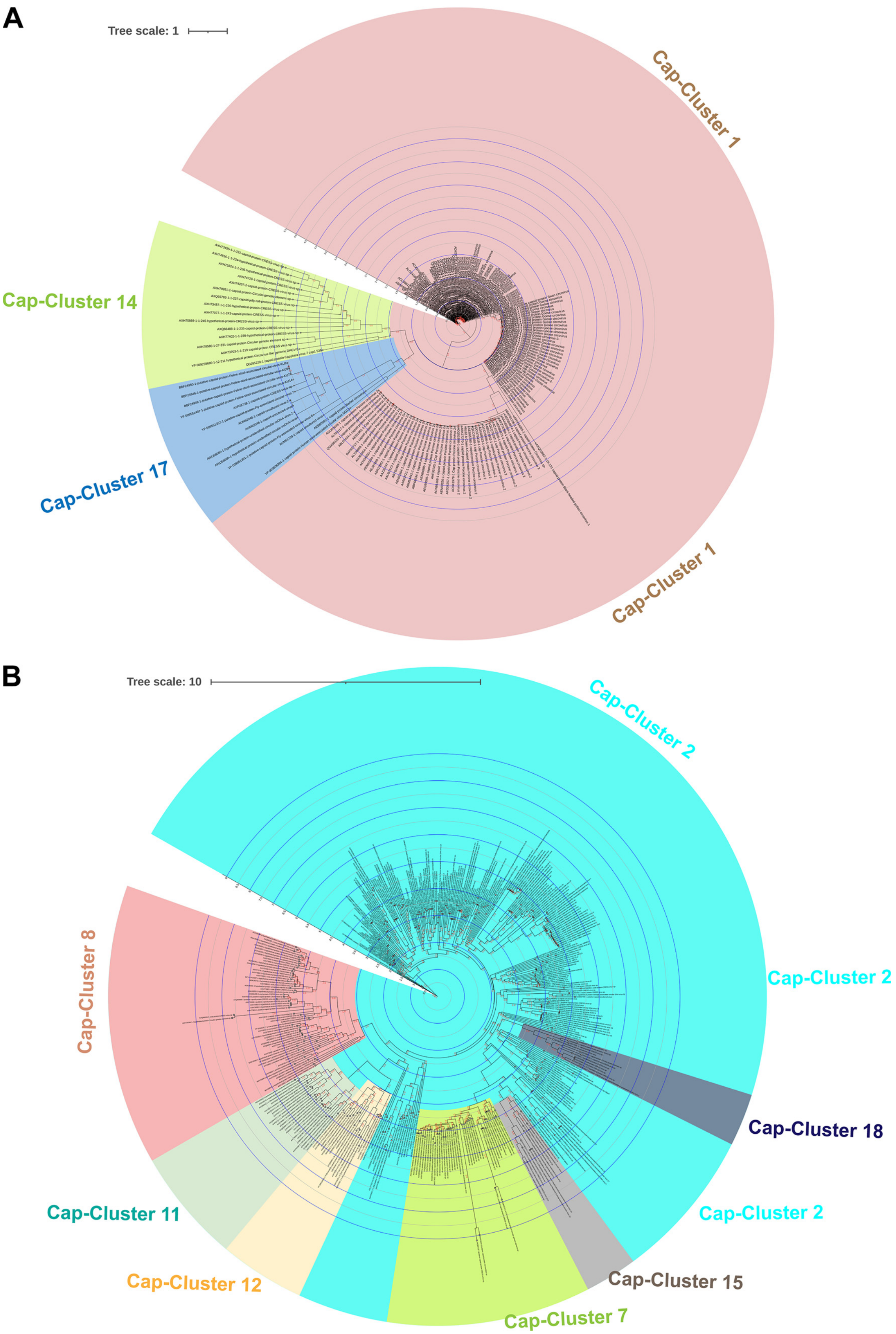

FIG 3 Phylogenetic relationship of CRESS-DNA virus Cap protein superclusters. (A) Phylogenetic tree depicting the genetic relationship between the CRESS-DNA virus Cap protein supercluster formed by the clusters C1, C14, and C17. The details of

(Continued on next page)

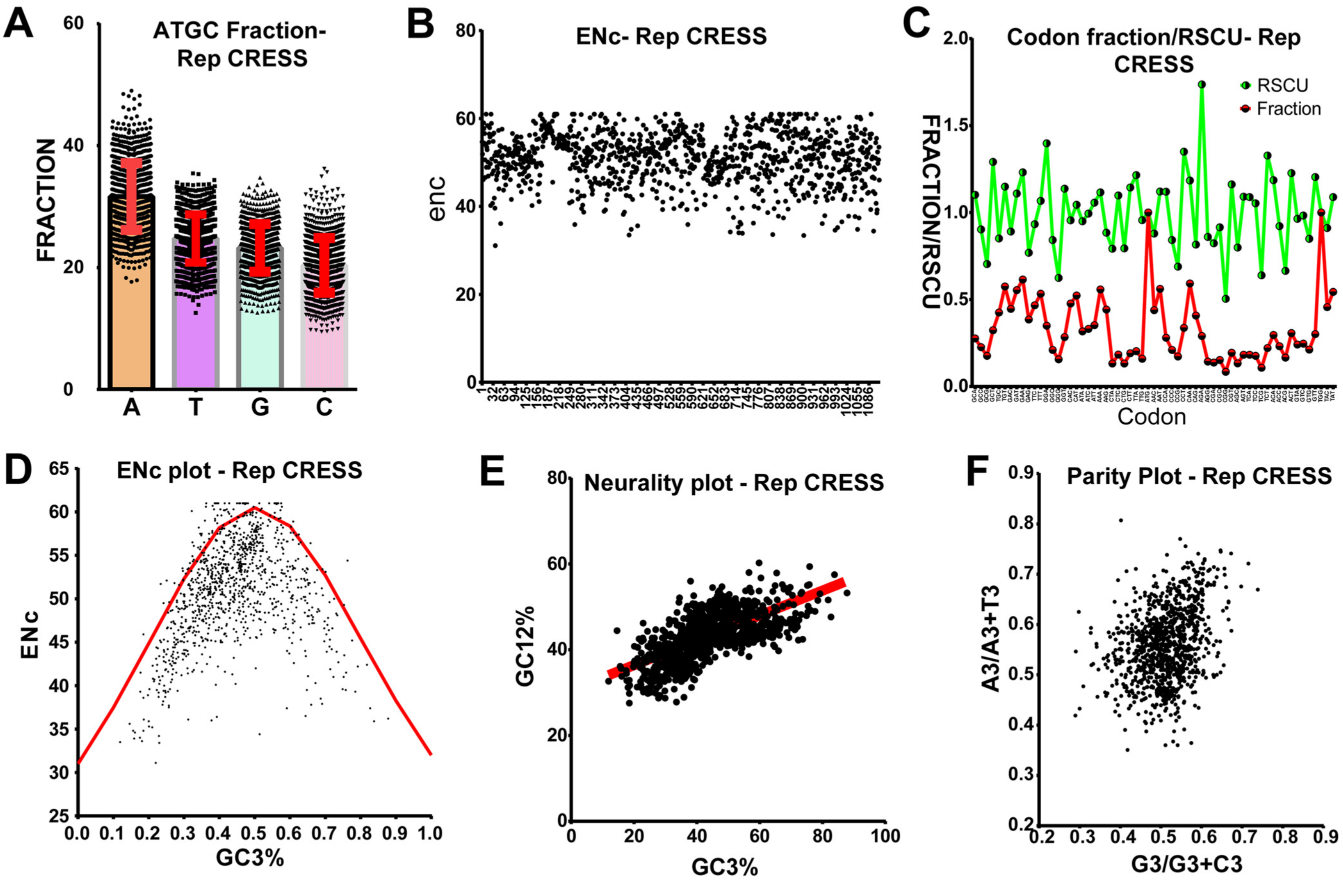

FIG 4 Host codon usage selection pressure on Rep gene of CRESS-DNA virus evolution. (A) Representing the A, T, G, and C fraction; (B) representing the ENc values; (C) representing the codon usage fraction and RSCU values; (D) representing ENc plotted against GC3s; (E) neutrality plot analysis of the GC12 and that of the GC3; and (F) parity rule 2 (PR2)-bias plot. (Total of 1,115 nucleotide sequences of Rep gene of CRESS-DNA viruses were used in this analysis.

erage of the GC content percentage at the first and second position in the codon) are plotted against GC3 values to evaluate the degree of influence of mutation pressure and natural selection on the codon usage patterns, displayed a slope of 0.2899 (Y = 0.2899*X + 30.61, *r* = 0.662; *P* < 0.0001) (Fig. 4E; Data Set S21), indicating that the mutation pressure and natural selection were 28.9% and 71.1%, respectively. Moreover, we performed parity rule 2 bias analysis, where the AT bias [A3/(A3+ T3)] is plotted against GC bias [G3/(G3 + C3)] to determine whether mutation pressure and natural selection affect the codon usage bias (38). If A = T and G = C, it indicates no mutation pressure and natural selection, while any discrepancies indicate mutation pressure and natural selection. Our analysis of CRESS-DNA Rep gene sequences shows unequal A to T and G to C numbers, indicating the presence of mutation and selection pressure (Fig. 4F, Data Set S21). Taken together, these results suggest that CRESS-DNA has wide host-range adaptation, maintaining better codon usage pattern with bacteria, and further selection pressure has played a significant role in the evolution of the CRESS-DNA viruses Rep gene rather than mutational pressure.

Role of host codon usage selection pressure on Cap gene evolution. Similar to the Rep gene, our NCBI nucleotide BLAST analysis of the Cap gene also showed limited

FIG 3 Legend (Continued)

sequences in each cluster (Supplemental Data 7) and alignment are provided in Supplemental Data 10. (B) The phylogenetic tree depicting the genetic relationship between the CRESS-DNA virus Cap protein supercluster created by the clusters C2, C7, C8, C11, C12, C15, and C18. The details of sequences in each cluster (Supplemental Data 7) and alignment are provided in Supplemental Data 11. The maximum-likelihood method inferred the evolutionary history using the Subtree-Pruning-Regrafting algorithm and bootstrap values in PhyML 3.3_1.

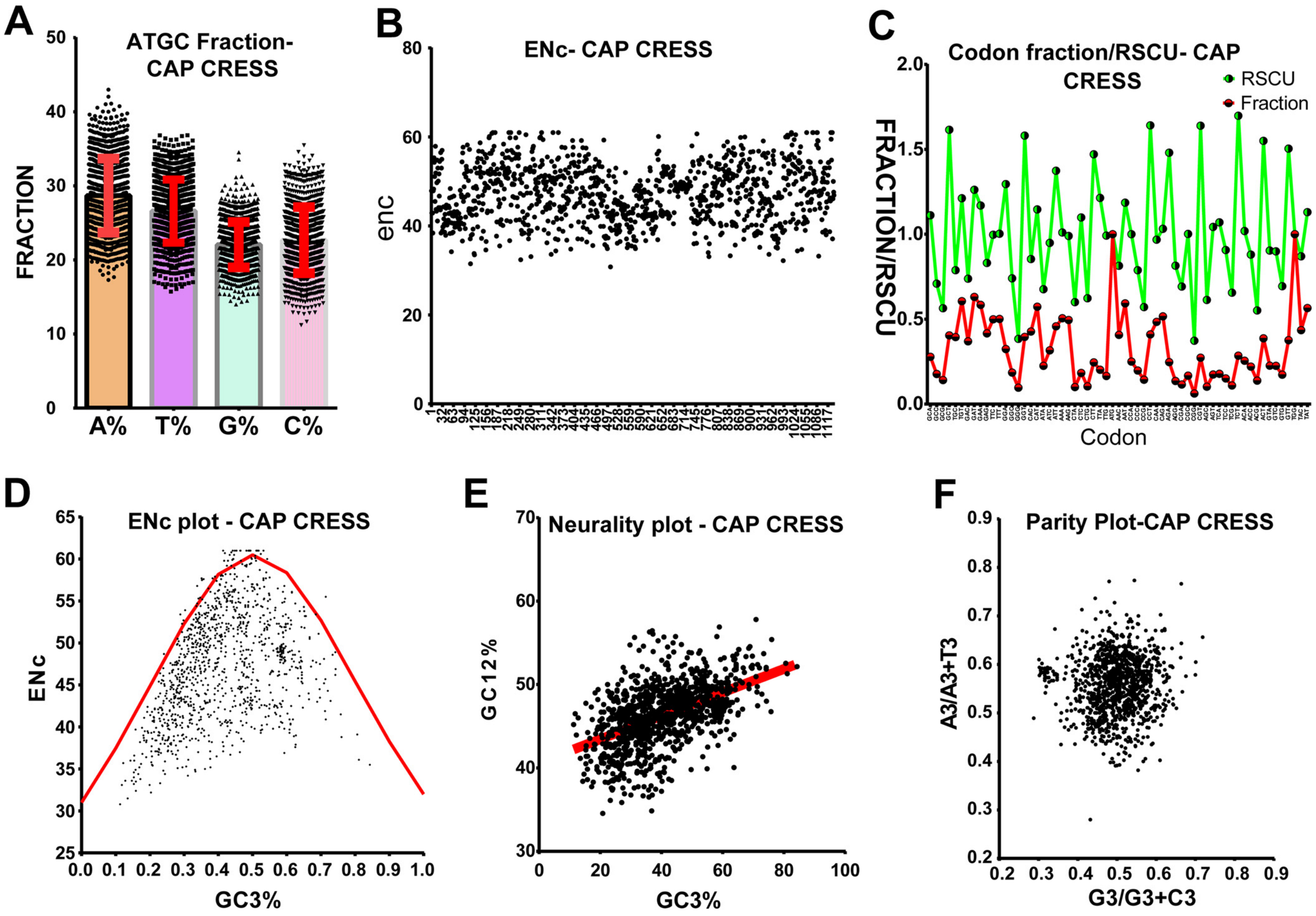

FIG 5 Host codon usage selection pressure on Cap gene of CRESS-DNA virus evolution. (A) A, T, G, and C fraction; (B) ENc values; (C) codon usage fraction and RSCU values; (D) ENc plotted against GC3s; (E) neutrality plot analysis of the GC12 and that of the GC3; and (F) parity rule 2 (PR2)-bias plot. (Total of 1,134 nucleotide sequences of Cap gene of CRESS-DNA viruses used in this analysis.)

homology between the Cap gene of CRESS-DNA viruses. Since we observed a strong codon-bias-based evolution in the Rep gene of CRESS-DNA viruses (Fig. 4A to F), we tested whether the Cap gene of the CRESS-DNA viruses also shows codon-bias-based evolution to explore whether the evolution of the Cap gene was influenced by mutation pressure or selection pressure. We retrieved 1,134 nucleotide sequences of Cap genes of CRESS-DNA viruses (Data Set S22) from the NCBI public database. Our analysis of the nucleotide base composition of the Cap gene revealed that the Cap gene contains AT% > GC% (average GC content is 44.72 ± SD 5.96) (Fig. 5A; Data Set S23). Further, the Cap protein of the CRESS-DNA virus has ENc value ranging from 33 to 61, while most of the sequence ENc values fall between 40 to 60 (average ENc 51.54 ± SD 5.36) (Fig. 5B; Data Set S23). Similarly, the RSCU values of 27 codons were >1, and 31 codons were <1 in all the Cap genes of CRESS-DNA viruses (Fig. 5C; Data Set S23). Also, nine codons showed RSCU values of <0.7, and 6 codons showed RSCU values of >1.5, indicating the presence of underrepresented and overrepresented codon bias in the Cap gene, respectively (Data Set S23). Moreover, we performed ENc-GC3s plot analysis and found that most points fall below the expected curve in the ENC-GC3s plot (Fig. 5D; Data Set S23). In line with this, the neutrality plot displayed a slope of 0.1404 (Y = 0.1404*X + 40.64; *r* = 0.512; *P* < 0.0001) (Fig. 5E; Data Set S23), indicating 14% of mutation pressure and 86% of selection pressure in this gene, and parity rule 2 bias analysis showed discrepancies in the A to T and G to C numbers in the third position of the codon (Fig. 5F; Data Set S23). Taken together, these results indicate that the selection pressure played a more significant role in the Cap gene than the Rep genes of CRESS-DNA viruses.

DISCUSSION

The genetic diversity of CRESS-DNA viruses so far is known only to be the tip of the iceberg. Many novel CRESS-DNA viruses have recently been detected by metagenomic sequencing (8, 39 to 41). The rapid development of metagenomic sequencing suggests that in the future, most CRESS-DNA viruses will be detected from different sources and that these CRESS-DNA viruses will be divided into different virus families. Therefore, it is hoped that identifying and classifying genetic diversity in CRESS-DNA viruses will help determine their importance in transmission and pathogenesis and design antivirals and vaccines for appropriate control and prevention. However, the classification of CRESS-DNA viruses has been determined using only the Rep protein (1, 12). This is because Rep protein contains conserved HUH motif and S3H domains, while Cap protein is unclassified because it has high genetic diversity without being conserved (1, 12). However, of the cruciviruses that classify Cap protein well, the report that Cap proteins are nearly identical and that the highly diverse Rep protein may be derived from different CRESS-DNA virus sources is critical here (17). Therefore, it can be expected that the genetic diversity and genetic recombination events of CRESS-DNA viruses can be determined by detecting and classifying the diversity in both Rep and Cap proteins.

It is noteworthy that recently, unclassified CRESS-DNA viruses using the Rep protein of CRESS-DNA viruses were grouped into six groups called CRESSV1 to CRESSV6 (1, 12). The present study reveals that there is not only the presence of CRESSV1-CRESSV6 groups in the CRESS-DNA viruses but also reveals the presence of groups such as Smacoviridae-related, CRESSV2-related, pCRESS-related, Circoviridae-related, and 1 to 4 outgroups. So far, it has been reported that the Rep protein of the CRESS-DNA virus contains the HUH motif and S3H domains (1, 12). In this study, we report the presence of domains such as the Viral_Rep superfamily (Cdd: pfam02407) and P-loop_NTPase superfamily (Cdd: pfam00910) in the Rep protein of most CRESS-DNA viruses. Furthermore, this present study revealed the presence of these two domains in the CRESSV1, CRESSV2, CRESSV3, *Circoviridae*, and Circoviridae-related groups and the *Nanoviridae* group. However, CLANS and phylogenetic analyses clarify that the viral_Rep and P-loop_NTPase domains in the CRESSV1, CRESSV2, CRESSV3, Circoviridae, and Circoviridae-related groups are very close and distinct from the *Nanoviridae* group. It is noteworthy that CRESSV1, CRESSV2, CRESSV3, Circoviridae, and Circoviridae-related groups together formed the Rep subcluster a and b and the sequences in the *Nanoviridae* group Rep subcluster i and j (Fig. S1B). Our phylogenetic tree (Fig. 1B) and previous study (12) reflect this diversity. In particular, some sequences in the rep subcluster a and b appear to have an additional domain (Penta-EF hand, DNA-binding ATP-dependent protease La, Type III secretion system protein PrgH-EprH [PrgH], etc.) between the Viral_Rep and P-loop_NTPase domains. From the acquisition of such additional functional domains, it is clear that these viruses are stepping into the next stage of evolution, and when more sequences are found later, it is possible to speculate that they are likely to be classified as separate virus families. However, since these sequences are detected by metagenomic sequencing from uncultured viruses and maybe sequence alignment error, it may be imperative to isolate the viruses and identify the significance of these additional functional domains.

Furthermore, in CLANS analysis, the subclusters c and d were grouped with the subclusters i and j reacting with *Nanoviridae* (Fig. S2B), of which the subcluster c was CRESSV4 and the subcluster d was CRESSV5 viruses (Data Set S4), as reflected in our phylogenetic tree (Fig. 1B) and previous study (12). In particular, the CRESSV4 and CRESSV5 viruses have only the Viral_Rep domain, the sequences in the subcluster j related to Nanoviridae are the Viral_Rep+incomplete P-loop_NTPase, and the sequences in the subcluster i are the Viral_Rep+P-loop_NTPase domains. Of these, it can be speculated that the CRESSV4 and CRESSV5 viruses, which have only the Viral_Rep domain, may have appeared first, followed by the subcluster j with the viral_Rep+incomplete P-loop_NTPase, and finally the subcluster i virus with the Viral_Rep+P-loop_NTPase domains. Similarly, subcluster g (CRESSV2-related group), which is often Viral_Rep+P-loop_NTPase domains and some sequences where these domains are incomplete or show only one of the two domains, may have led to the emergence of CRESSV2 viruses with Viral_Rep+P-loop_NTPase domains. Furthermore, it is essential to note that subcluster h, which usually contains only the P-loop_NTPase domain, formed the pCRESS-related group. Interestingly, no functional domains were found in Smacoviridae's Rep protein in the Conserved Domain search tool, which revealed links between *Circoviridae* and *Nanoviridae* in CLANS and biogenetic analyzes. Similarly, no functional domains were found in the sequences in group R5 (CLANS) or Smacoviridae-related group (phylogenetic tree). It is noteworthy that the sequences of Rep protein that formed the outgroups in this phylogenetic analysis are the clusters of CLANS analysis, R3, R5, R6, and R9, forming separate groups. R3, R5, R6, and R9 clusters formed Supercluster 2 and created outgroup 4, outgroup 1, outgroup 3, and outgroup 2, respectively. Remarkably, no functional domains are found in the sequences in the R3, R5, R6, and R9 clusters that make up the outgroups. However, the sequences that make up the outgroups are detected from uncultured viruses by metagenomic sequencing, and it can be expected that the functional significance will be revealed by isolating these viruses and characterizing the Rep protein.

Cap protein was high in genetic diversity, making it challenging to align and phylogenetically classify correctly. Therefore, in this study, we subdivided the closest sequences into clusters using CLANS analysis and then did phylogenetic classification by aligning them well using the corresponding clusters. First, in this study, the Rep proteins were divided into clusters in the CLANS analysis and then phylogenetically classified using the related clusters, which is consistent with phylogenetic classification in the previous studies (1, 12). Accordingly, we divide the cap protein into 20 clusters using CLANS analysis, and (i) superclusters C1, C14, C17; (ii) superclusters C2, C7, C8, C11, C12, C15, and C18; and (iii) orphan clusters such as C3, C4, C5, C6, C9, C10, C13, C16, C19, and C20 became well aligned and led to phylogenetic classification. Furthermore, only the Cruciviridae virus sequences in Cap-cluster C6 revealed evolutionary relationships with RNA viruses, but future studies need to determine the evolutionary origins of the sequences in other Cap clusters. Remarkably, this study revealed that viruses in the same Cap cluster derive their Rep protein from groups of different Rep proteins, which can be speculated to be generated by genetic recombination. These can be believed to underscore the importance of classifying Cap protein. Finally, this study makes it clear that selection pressure plays a more significant role than mutational pressure in the genetic diversity and evolution of CRESS-DNA virus Cap and Rep protein. Therefore, it can be expected that there will be more opportunities to detect CRESS-DNA viruses with greater genetic diversity and/or recombination in the future. We hope this study will help determine the genetic diversity/recombination of CRESS-DNA viruses as more sequences are discovered in the future.

In conclusion, to the best of our knowledge, this is the first report on the CRESS-DNA virus Rep protein classification using a different domain organization pattern, and CLANS and phylogenetic analysis based on the classification of Cap protein. Furthermore, this study also clarifies the genetic diversity in CRESS-DNA viruses formed by recombination and selection pressures in Cap and Rep proteins. It is widely expected that CRESS-DNA viruses, which have tremendous genetic diversity in the future, will be able to be detected from different sources in different parts of the world through rapidly growing metagenomic sequences. We hope this study will help you determine and accurately classify using CLANS, phylogenetic groups, the domain organization pattern, genetic diversity, and recombination of those CRESS-DNA viruses.

MATERIALS AND METHODS

Database search, collection, and curation. Complete genome sequences of CRESS-DNA viruses were retrieved from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/). Rep and Cap genes' characterized protein-coding sequence (CDS) region and their corresponding amino acid sequences were retrieved from the database in the available complete genome sequence of CRESS-DNA viruses. The uncharacterized CDS of CRESS-DNA viruses were classified as a Cap/Rep protein using NCBI protein BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Protein) analysis, and the sequences were retrieved. Further, complete genome sequences that contain Cap/Rep of every CRESS-DNA were individually used to perform separate NCBI protein BLAST analysis (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Protein). Their BLAST-aligned sequences of other ssDNA viruses (example: Circoviridae, Smacoviridae, Cruciviridae, etc.) were retrieved.

CLANS (CLuster ANalysis of Sequences) analysis. The CLANS analysis was performed in the online Toolkit software (https://toolkit.tuebingen.mpg.de/tools/clans). The protein sequences retrieved from the NCBI database were subjected to the pairwise sequence similarity calculation using the online CLANS analysis in the Toolkit (21) with a scoring matrix of BLOSUM45 and BLAST HSP (High Scoring Pair) up to an E value of $1e^{-2}$. Next, the CLANS files obtained from the Toolkit were visualized in a Java application (clans.jar) (22). A minimum of 100,000 rounds were used to show the sequences connection and clusters in the clans.jar application. The clusters were classified based on the network method using offset values and global average with maximum rounds of 10,000 in clans.jar analysis.

Analysis of functional domain organization in the protein. We determined the domain organizations in the Rep protein of the CRESS-DNA virus using the Conserved Domain search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). For this, we used the CDD v3.19-58235 PSSms database, Expect Value threshold line 0.01, and composition-based statistics adjustment applied, and performed by the maximum number of hits to 500 in the Conserved Domain search tool (26 to 29).

Phylogenetic analyses. The phylogenetic analysis was performed in PhyML 3.3_1 using the amino acid sequences of the Rep/Cap protein of the CRESS-DNA virus clustered into clusters in the CLANS analysis, which was retrieved from the NCBI public database. The phylogenetic analysis is in PhyML 3.3_1, evolutionary model LG, equilibrium frequencies Empirical ML- Model, discrete gamma model (number of categories [$n = 4$]), tree topology search with SPR (Subtree Pruning and Regraphing), tree topology, and branch length, model parameters are optimizing parameters, and SH-like statistics are used to test the branch support (42 to 44). Further, the phylogenetic trees were visualized through the interactive tree of life (iTOL) v5 (45).

Codon usage bias analysis. (i) Nucleotide sequence composition analysis. The nucleotide composition of CDSs, specifically the A%, T%, G%, and C% composition of the Rep/Cap genes of CRESS-DNA viruses, was analyzed using Automated Codon Usage Analysis (ACUA) software (46).

(ii) Relative synonymous codon usage (RSCU) analysis. RSCU value is the ratio between the observed and the expected value of synonymous codons for a given amino acid. When the RSCU value is one, it indicates that there is no bias for that codon (36, 37). This study determined the RSCU values using the ACUA software (46). The nucleotide sequences of Rep/Cap genes of CRESS-DNA viruses obtained from the NCBI nucleotide public database were used for this analysis.

(iii) Effective number of codons (ENc). The effective number of codon usage from 61 codons for the 20 amino acids is one method that determines the codon usage bias and may range from 20 to 61. ENc values of <35 indicate high codon bias, and values of >50 show general random codon usage (35, 36). In this study, the ENc values were determined on the online server (http://ppuigbo.me/programs/CAIcal/) (47), and the input nucleotide sequences used in the codon adaptation index (CAI) calculation were used in this analysis.

Determining the selection and mutation pressure. (i) ENc-GC3s plot. In this analysis, the ENc values are plotted against the third position of GC3s of codon values to determine the significant factors such as selection or mutation pressure affecting the codon usage bias (38). The expected curve was determined by estimating the expected ENc values for each GC3s as recommended in previous publications (36, 38). The ENc and GC3s for every gene were obtained from an online CAI analysis server (http://ppuigbo.me/programs/CAIcal/) (47). The genes would lie on or around the expected curve when mutation pressure only affects codon bias. In contrast, they would fall considerably below the expected curve if codon bias is influenced by selection and other factors (36, 38).

(ii) Neutrality plot analysis. In a neutrality plot, GC12 values of the codon are plotted against GC3 values to evaluate the degree of influence of mutation pressure and natural selection on the codon usage patterns. The GC12 and GC3 values for the nucleotide sequences of Rep/Cap genes of CRESS-DNA viruses were obtained from an online CAI analysis server (http://ppuigbo.me/programs/CAIcal/) (47).

(iii) Parity rule 2 (PR2)-bias plot. The PR2-bias, the AT bias [A3/(A3+ T3)] is plotted against GC-bias [G3/(G3 + C3)] to mutation pressure and natural selection affecting the codon usage bias (38). The A3, T3, G3, and C3 values of nucleotide sequences of Rep/Cap genes of CRESS-DNA viruses were obtained using the ACUA Software (46).

Data availability. We have retrieved the nucleotide sequences from publicly available NCBI databases. Further, all the nucleotide sequence accession numbers and names are indicated in the respective figures and supplemental data.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLSX file, 0.7 MB.

SUPPLEMENTAL FILE 2, XLSX file, 1.4 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.3 MB.

SUPPLEMENTAL FILE 4, XLSX file, 0.2 MB.

SUPPLEMENTAL FILE 5, XLSX file, 0.04 MB.

SUPPLEMENTAL FILE 6, XLSX file, 0.02 MB.

SUPPLEMENTAL FILE 7, XLSX file, 1 MB.

SUPPLEMENTAL FILE 8, XLSX file, 0.8 MB.
SUPPLEMENTAL FILE 9, XLSX file, 0.7 MB.
SUPPLEMENTAL FILE 10, PDF file, 34.4 MB.

ACKNOWLEDGMENTS

P.A.D. is a DST-INSPIRE faculty member and is supported by research funding from the Department of Science and Technology (DST/INSPIRE/04/2016/001067), the Government of India, and a Core grant from the Science and Engineering Research Board (SERB) (CRG/2018/002192), Department of Science and Technology (DST), Government of India.

We declare no potential conflicts of interest.

REFERENCES

1. Kazlauskas D, Varsani A, Koonin EV, Krupovic M. 2019. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat Commun 10:3425. https://doi.org/10.1038/s41467-019-11433-0.
2. Zhao L, Rosario K, Breitbart M, Duffy S. 2019. Eukaryotic circular rep-encoding single-stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. Adv Virus Res 103:71–133. https://doi.org/10.1016/bs.aivir.2018.10.001.
3. Krupovic M. 2013. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol 3:578–586. https://doi.org/10.1016/j.coviro.2013.06.010.
4. Chow CE, Suttle CA. 2015. Biogeography of viruses in the sea. Annu Rev Virol 2:41–66. https://doi.org/10.1146/annurev-virology-031413-085540.
5. Labonte JM, Suttle CA. 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. ISME J 7:2169–2177. https://doi.org/10.1038/ismej.2013.110.
6. Ng TFF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E. 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J Virol 86:12161–12175. https://doi.org/10.1128/JVI.00869-12.
7. Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, Harding JS, Roumagnac P, Martin DP, Lefeuvre P, Varsani A. 2016. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect Genet Evol 39:304–316. https://doi.org/10.1016/j.meegid.2016.02.011.
8. Rosario K, Schenck RO, Harbeitner RC, Lawler SN, Breitbart M. 2015. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. Front Microbiol 6:696. https://doi.org/10.3389/fmicb.2015.00696.
9. Dayaram A, Goldstien S, Argüello-Astorga GR, Zawar-Reza P, Gomez C, Harding JS, Varsani A. 2015. Diverse small circular DNA viruses circulating amongst estuarine molluscs. Infect Genet Evol 31:284–295. https://doi.org/10.1016/j.meegid.2015.02.010.
10. Bistolas KSI, Rudstam LG, Hewson I. 2017. Gene expression of benthic amphipods (genus: Diporeia) in relation to a circular ssDNA virus across two Laurentian Great Lakes. PeerJ 5:e3810. https://doi.org/10.7717/peerj.3810.
11. Blinkova O, Rosario K, Li L, Kapoor A, Slikas B, Bernardin F, Breitbart M, Delwart E. 2009. Frequent detection of highly diverse variants of cardiovirus, cosavirus, bocavirus, and circovirus in sewage samples collected in the United States. J Clin Microbiol 47:3507–3513. https://doi.org/10.1128/JCM.01062-09.
12. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Rosario K, Yutin N, Wolf YI, Harrach B, Zerbini FM, Dolja VV, Kuhn JH, Koonin EV. 2020. *Cressdnaviricota*: a virus phylum unifying seven families of Rep-encoding viruses with single-stranded, circular DNA genomes. J Virol 94:e00582-20. https://doi.org/10.1128/JVI.00582-20.
13. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. 2017. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology 504:114–121. https://doi.org/10.1016/j.virol.2017.02.001.
14. Diemer GS, Stedman KM. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct 7:13. https://doi.org/10.1186/1745-6150-7-13.
15. Roux S, Enault F, Bronner G, Vaulot D, Forterre P, Krupovic M. 2013. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat Commun 4:2700. https://doi.org/10.1038/ncomms3700.
16. Krupovic M, Ravantti JJ, Bamford DH. 2009. Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol Biol 9:112. https://doi.org/10.1186/1471-2148-9-112.
17. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, Bisseux M, Ravet V, Dayaram A, Stainton D, Kraberger S, Zawar-Reza P, Goldstien S, Briskie JV, White R, Taylor H, Gomez C, Ainley DG, Harding JS, Fontenele RS, Schreck J, Ribeiro SG, Oswald SA, Arnold JM, Enault F, Varsani A, Stedman KM. 2020. Unveiling crucivirus diversity by mining metagenomic data. mBio 11:e01410-20. https://doi.org/10.1128/mBio.01410-20.
18. Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D. 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332:714–717. https://doi.org/10.1126/science.1203163.
19. Kazlauskas D, Varsani A, Krupovic M. 2018. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses 10:187. https://doi.org/10.3390/v10040187.
20. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM. 2017. Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15:161–168. https://doi.org/10.1038/nrmicro.2016.177.
21. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at its core. J Mol Biol 430:2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.
22. Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20:3702–3704. https://doi.org/10.1093/bioinformatics/bth444.
23. Krupovic M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A 114:E2401–E2410. https://doi.org/10.1073/pnas.1621061114.
24. Whitley C, Gunst K, Müller H, Funk M, zur Hausen H, de Villiers E-M. 2014. Novel replication-competent circular DNA molecules from healthy cattle serum and milk and multiple sclerosis-affected human brain tissue. Genome Announc 2:e00849-14. https://doi.org/10.1128/genomeA.00849-14.
25. Gorbalenya AE, Koonin EV, Wolf YI. 1990. A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. FEBS Lett 262:145–148. https://doi.org/10.1016/0014-5793(90)80175-I.
26. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265–D268. https://doi.org/10.1093/nar/gkz991.
27. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 45:D200–D203. https://doi.org/10.1093/nar/gkw1129.
28. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015.

CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. https://doi.org/10.1093/nar/gku1221.

29. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39: D225–D229. https://doi.org/10.1093/nar/gkq1189.
30. Abrescia NG, Bamford DH, Grimes JM, Stuart DI. 2012. Structure unifies the viral universe. Annu Rev Biochem 81:795–822. https://doi.org/10.1146/annurev-biochem-060910-095130.
31. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA. 2007. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291–D297. https://doi.org/10.1093/nar/gkl959.
32. Krupovic M, Bamford DH. 2011. Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. Curr Opin Virol 1:118–124. https://doi.org/10.1016/j.coviro.2011.06.001.
33. Du M-Z, Zhang C, Wang H, Liu S, Wei W, Guo F-B. 2018. The GC content as a main factor shaping the amino acid usage during bacterial evolution process. Front Microbiol 9:2948. https://doi.org/10.3389/fmicb.2018.02948.
34. Bohlin J, Brynildsrud O, Vesth T, Skjerve E, Ussery DW. 2013. Amino acid usage is asymmetrically biased in AT- and GC-rich microbial genomes. PLoS One 8:e69878. https://doi.org/10.1371/journal.pone.0069878.
35. Zhao Y, Zheng H, Xu A, Yan D, Jiang Z, Qi Q, Sun J. 2016. Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution. BMC Genomics 17:677. https://doi.org/10.1186/s12864-016-3021-7.
36. Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, Feng W, Zhang G, Song X, Sun X. 2018. Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS One 13:e0194372. https://doi.org/10.1371/journal.pone.0194372.
37. Gun L, Yumiao R, Haixian P, Liang Z. 2018. Comprehensive analysis and comparison on the codon usage pattern of whole *Mycobacterium tuberculosis* coding genome from different area. Biomed Res Int 2018:3574976. https://doi.org/10.1155/2018/3574976.
38. Tian H-F, Hu Q-M, Xiao H-B, Zeng L-B, Meng Y, Li Z. 2020. Genetic and codon usage bias analyses of major capsid protein gene in Ranavirus. Infect Genet Evol 84:104379. https://doi.org/10.1016/j.meegid.2020.104379.
39. Guo Z, He Q, Tang C, Zhang B, Yue H. 2018. Identification and genomic characterization of a novel CRESS DNA virus from a calf with severe hemorrhagic enteritis in China. Virus Res 255:141–146. https://doi.org/10.1016/j.virusres.2018.07.015.
40. Moens MAJ, Perez-Tris J, Cortey M, Benitez L. 2018. Identification of two novel CRESS DNA viruses associated with an *Avipoxvirus* lesion of a blue-and-gray Tanager (*Thraupis episcopus*). Infect Genet Evol 60:89–96. https://doi.org/10.1016/j.meegid.2018.02.015.
41. Liu Q, Wang H, Ling Y, Yang S-X, Wang X-C, Zhou R, Xiao Y-Q, Chen X, Yang J, Fu W-G, Zhang W, Qi G-L. 2020. Viral metagenomics revealed diverse CRESS-DNA virus genomes in faeces of forest musk deer. Virol J 17:61. https://doi.org/10.1186/s12985-020-01332-y.
42. Lemoine F, Domelevo Entfellner J-B, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature 556:452–456. https://doi.org/10.1038/s41586-018-0043-0.
43. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321. https://doi.org/10.1093/sysbio/syq010.
44. Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O. 2019. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. Nucleic Acids Res 47:W260–W265. https://doi.org/10.1093/nar/gkz303.
45. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301.
46. Vetrivel U, Arunkumar V, Dorairaj S. 2007. ACUA: a software tool for automated codon usage analysis. Bioinformation 2:62–63. https://doi.org/10.6026/97320630002062.
47. Puigbo P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct 3:38. https://doi.org/10.1186/1745-6150-3-38.