AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

GENETIC DIVERSITY AND EVOLUTION

Functional Genomic Analyses Reveal an Open Pan-genome for the Chloroviruses and a Potential for Genetic Innovation in New Isolates

Rodrigo A. L. Rodrigues,[a] Victória F. Queiroz,[a] Jayadri Ghosh,[b] David D. Dunigan,[b,c] James L. Van Etten[b,c]

[a]Virus Laboratory, Department of Microbiology, Federal University of Minas Gerais, Belo Horizonte, Brazil
[b]Nebraska Center for Virology, University of Nebraska—Lincoln, Lincoln, Nebraska, USA
[c]Department of Plant Pathology, University of Nebraska—Lincoln, Lincoln, Nebraska, USA

ABSTRACT Chloroviruses (family *Phycodnaviridae*) are large double-stranded DNA (dsDNA) viruses that infect unicellular green algae present in inland waters. These viruses have been isolated using three main chlorella-like green algal host cells, traditionally called NC64A, SAG, and Pbi, revealing extensive genetic diversity. In this study, we performed a functional genomic analysis on 36 chloroviruses that infected the three different hosts. Phylogenetic reconstruction based on the DNA polymerase B family gene clustered the chloroviruses into three distinct clades. The viral pan-genome consists of 1,345 clusters of orthologous groups of genes (COGs), with 126 COGs conserved in all viruses. Totals of 368, 268, and 265 COGs are found exclusively in viruses that infect NC64A, SAG, and Pbi algal hosts, respectively. Two-thirds of the COGs have no known function, constituting the "dark pan-genome" of chloroviruses, and further studies focusing on these genes may identify important novelties. The proportions of functionally characterized COGs composing the pan-genome and the core-genome are similar, but those related to transcription and RNA processing, protein metabolism, and virion morphogenesis are at least 4-fold more represented in the core genome. Bipartite network construction evidencing the COG sharing among host-specific viruses identified 270 COGs shared by at least one virus from each of the different host groups. Finally, our results reveal an open pan-genome for chloroviruses and a well-established core genome, indicating that the isolation of new chloroviruses can be a valuable source of genetic discovery.

IMPORTANCE Chloroviruses are large dsDNA viruses that infect unicellular green algae distributed worldwide in freshwater environments. They comprise a genetically diverse group of viruses; however, a comprehensive investigation of the genomic evolution of these viruses is still missing. Here, we performed a functional pan-genome analysis comprising 36 chloroviruses associated with three different algal hosts in the family *Chlorellaceae*, referred to as zoochlorellae because of their endosymbiotic lifestyle. We identified a set of 126 highly conserved genes, most of which are related to essential functions in the viral replicative cycle. Several genes are unique to distinct isolates, resulting in an open pan-genome for chloroviruses. This profile is associated with generalist organisms, and new insights into the evolution and ecology of chloroviruses are presented. Ultimately, our results highlight the potential for genetic diversity in new isolates.

KEYWORDS *Phycodnaviridae*, chloroviruses, genomics, pan-genome, evolution

Aquatic viruses are an integral part of Earth's ecosystems, and they play a major role in maintaining global geochemical cycling. It is estimated that 10^{23} virus infections occur each second in ocean waters. These infections are responsible for killing approximately 20% of the marine microbial biomass per day, cycling nutrients, and even changing local

Editor Colin R. Parrish, Cornell University

Address correspondence to Rodrigo A. L. Rodrigues, rodriguesral07@gmail.com, or James L. Van Etten, jvanetten1@unl.edu.

Received 9 August 2021
Accepted 17 October 2021
Accepted manuscript posted online 20 October 2021
Published 26 January 2022

weather patterns (1). Metagenomic studies indicate that after bacteriophages, phycodnaviruses are one of the most abundant viral groups in the oceans (2–4). The *Phycodnaviridae* family currently comprises six genera (*Coccolithovirus*, *Phaeovirus*, *Prasinovirus*, *Prymnesiovirus*, *Raphidovirus*, and *Chlorovirus*) of genetically diverse, large, double-stranded DNA (dsDNA) viruses. Even though the majority of studies have been conducted on marine environments, a plethora of viruses that infect eukaryotic green algae are also present in terrestrial water environments worldwide (5–8). Among the phycodnaviruses identified in inland waters, representatives of the genus *Chlorovirus* have been isolated from different locations across five continents but mainly in North America and Europe, likely due to sampling bias. They infect unicellular eukaryotic algae, commonly referred to as zoochlorellae, that normally exist as symbionts associated with the protozoon *Paramecium bursaria* and other aquatic invertebrates, including *Hydra viridis* and *Acanthocystis turfacea* (9, 10). Chloroviruses possess icosahedral capsids with an internal lipid membrane and linear dsDNA genomes ranging from 290 to 370 kb in length that have up to 400 coding sequences (CDSs). Of the predicted CDSs, approximately 50% resemble proteins of annotated function; however, a large number of putative proteins remain uncharacterized (5, 9).

Phylogenetic analyses indicate that chloroviruses diverged into three monophyletic clades according to the chlorella-like algal host species that they infect, *Chlorella variabilis* NC64A (NC64A viruses), first isolated from a paramecium in the United States; *Micractinium conductrix* Pbi (Pbi viruses), first isolated from a paramecium in Europe; and *Chlorella heliozoae* SAG 3.83 (SAG viruses), isolated from a heliozoon in Europe (9, 11). The core protein family set consists of about 125 members common to all chloroviruses, indicating that they are essential for viral replication; however, many of these core proteins are of unknown function. Besides the conserved genes, some clusters of orthologous groups of genes (COGs) associated with protein families are restricted to each of the three individual clusters in one of the algal hosts; thus, they are not present in the other chloroviruses, and they could be responsible for host recognition and specificity (9). While little is known about their functions and the mechanisms underlying host specificity, it is known that the inability of viruses to attach to nonhost cells is a major factor in limiting chloroviruses' host range (12).

The identification of conserved genes and those specific for each group of viruses is important for the advancement of genomic and evolutionary studies. Furthermore, such studies might provide important insights into the genetic innovation within each group. The *Chlorovirus* conserved genes include a subset that is in clusters (3 to 11 members) of collinear monocistronic genes, referred to as gene gangs (13). There are dozens of chlorovirus isolates described and genetically characterized so far; however, a better comprehension of the genome evolution of these viruses on a large scale is still missing. In this sense, pan-genomic analyses are valuable for understanding the genomic features under selective pressure and the evolutionary history of chloroviruses. In this study, we performed a functional pan-genomic analysis of chloroviruses, including viruses isolated from different places on Earth, infecting the three different algal hosts. We observed a distinct proportion of COG prevalences between the pan-genome and core genome for a few functional categories of genes, and a network construction evidenced COG sharing among host-specific viruses. Finally, our results revealed that chloroviruses have an open pan-genome and a well-established core genome; this observation indicates the potential for genetic innovation in new chlorovirus isolates.

RESULTS

***Phycodnaviridae* phylogeny.** The *Phycodnaviridae* family currently comprises 6 genera, and a total of 33 virus species are officially recognized by the International Committee on Taxonomy of Viruses (ICTV) (14). This family is the only one assigned to the newly created *Algavirales* order. Of the phycodnaviruses, chloroviruses are the most studied, and *Chlorovirus* is the genus with the largest number of sequenced representatives, currently comprising more than 50% of all viruses listed in the *Phycodnaviridae*. The phycodnaviruses are proposed

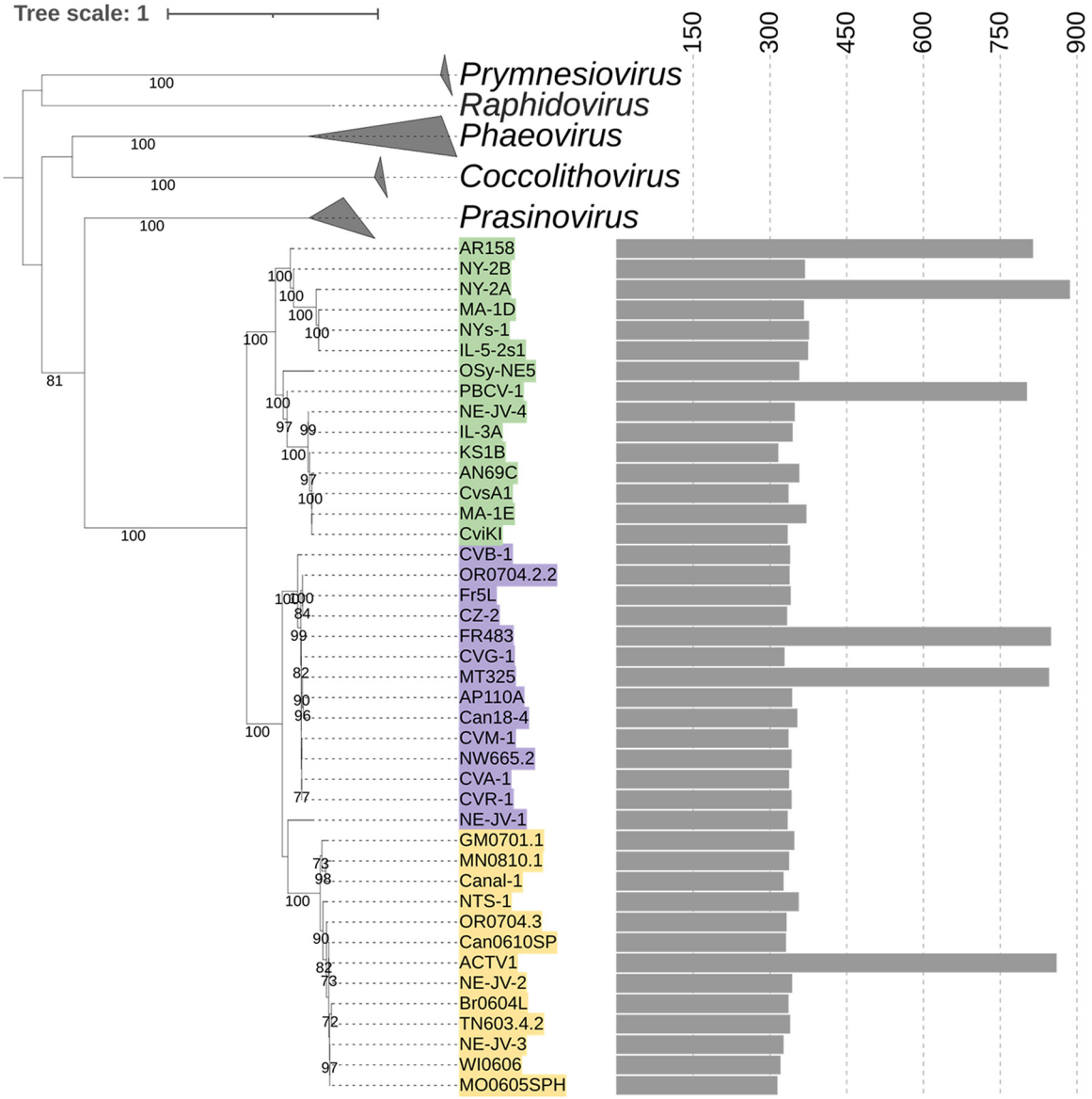

FIG 1 Phylogeny of *Phycodnaviridae*. A phylogenetic reconstruction using amino acid sequences of the DNA polymerase B family from different members of the family *Phycodnaviridae* is shown. Representatives of all six currently established genera were included. Only bootstrap values of >70 are shown. Chloroviruses infecting different host cells are colored as follows: green, *Chlorella variabilis*; purple, *Micractinium conductrix*; yellow, *Chlorella heliozoae*. Gray bars indicate the number of CDSs for each chlorovirus isolate available in GenBank. The bar indicates the rate of evolution.

to have a common origin with the whole group of nucleocytoplasmic large DNA viruses (NCLDVs) (15–17). Phylogenetic reconstruction of phycodnaviruses based on the DNA polymerase B gene, a well-conserved gene among members of the phylum *Nucleocytoviricota*, shows a clear separation of the six genera with high bootstrap support, with *Chlorovirus* forming a sister clade to *Prasinovirus* (Fig. 1). It is important to note that despite being monophyletic groups, viruses in different genera have high genetic diversity, as noted by the long branches of each clade, suggesting a particular evolutionary history for each group of viruses after its origin. A deeper analysis of the complete genome and DNA polymerase B gene shows intragenus conservation but great intergenus diversity (Fig. 2 and 3).

Within the genus *Chlorovirus*, it was possible to observe high bootstrap support for the formation of three distinct clades (Fig. 1). These clades grouped viruses that infect the same host, namely, *Chlorella variabilis* NC64A, *Micractinium conductrix* Pbi, and *Chlorella heliozoae* SAG 3.83. These hosts were recently reclassified, and before that,

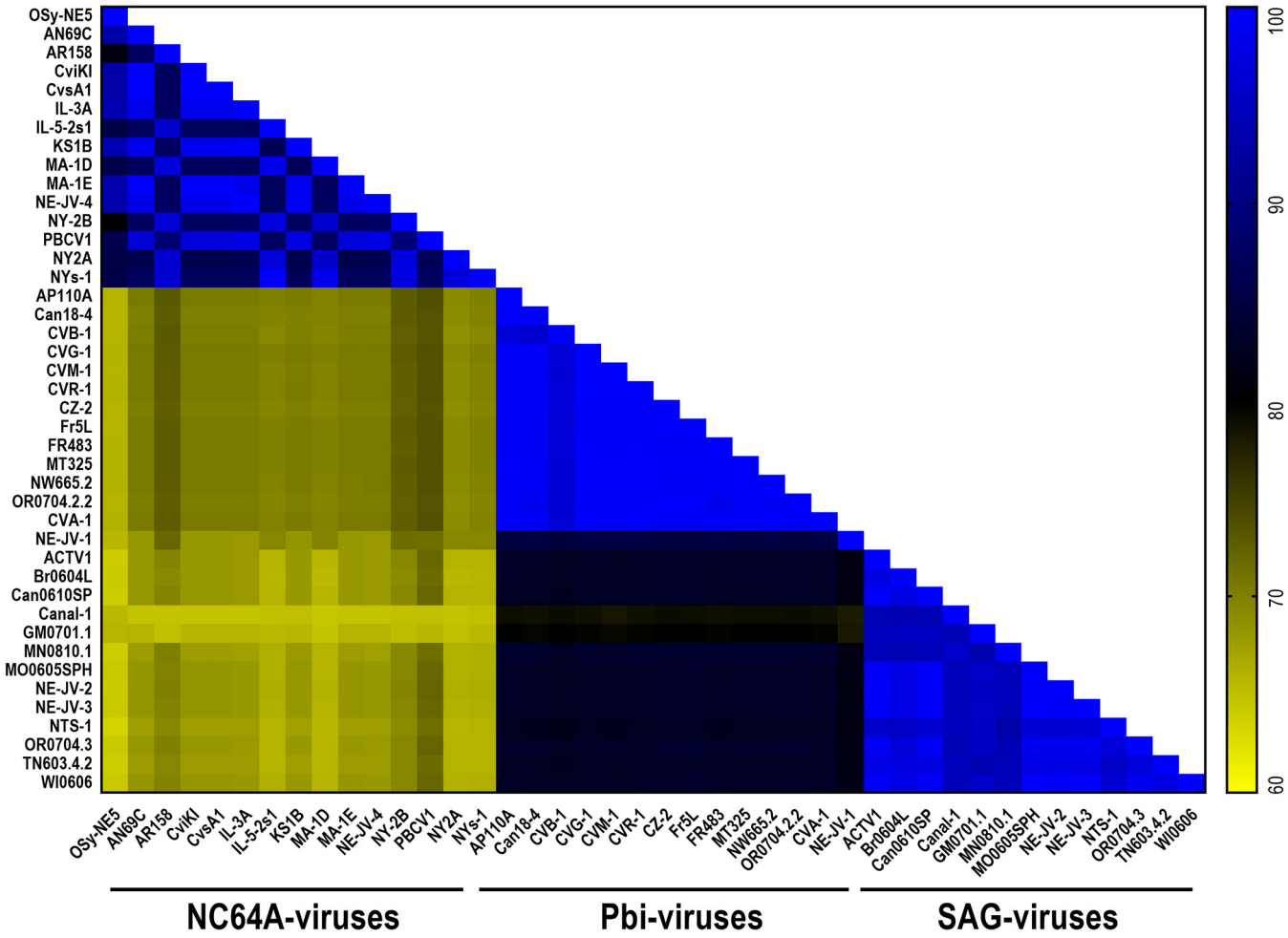

FIG 2 Similarity range of chlorovirus genomes. Shown is a heat map of identity analysis of the complete genomes of chloroviruses included for phylogenetic reconstruction. Identity ranges from 60% (yellow) to 100% (blue).

they were commonly called NC64A, Pbi, and SAG, respectively, and are named as such throughout this work to allow for a more adequate comparison with other data in the literature. Interestingly, one of the viruses included in this analysis (NE-JV-1) appeared in a separate branch within the SAG clade, although it was isolated from Pbi (9). Finally, there was high gene diversity among the NC64A viruses because they separated into two clades, compared to the Pbi viruses and SAG viruses, which had an apparently lower evolution rate (Fig. 1); however, this may be biased due to a narrow sampling range of space and time.

Evolution of the chlorovirus pan-genome. As mentioned above, chloroviruses have a large dsDNA genome, harboring hundreds of genes. A total of 42 chloroviruses have had their genomes sequenced and annotated and are available in public databases (last checked in May 2021). These viruses were isolated beginning in 1983 and have been collected from at least 11 countries (Fig. 4 and Table 1); however, the majority have been isolated from different sites in the United States. These viruses have genomes ranging from 287 to 369 kb and contain between 314 and 886 coding sequences (CDSs); they also contain 3 to 16 genes encoding tRNAs (Table 1). These data suggest large genomic diversity among chloroviruses. To gain a better understanding of which genes are well conserved in this group of viruses as well as a dimension on the evolution of viral genomic diversity, we gathered 36 viruses for a pan-genome analysis. Six of the 42 viruses with available genomes were excluded from the analysis because they have a much higher number of CDSs (Fig. 1 and Table 1), which would distort the analysis and may not adequately reflect the evolution of the viral pan-genome.

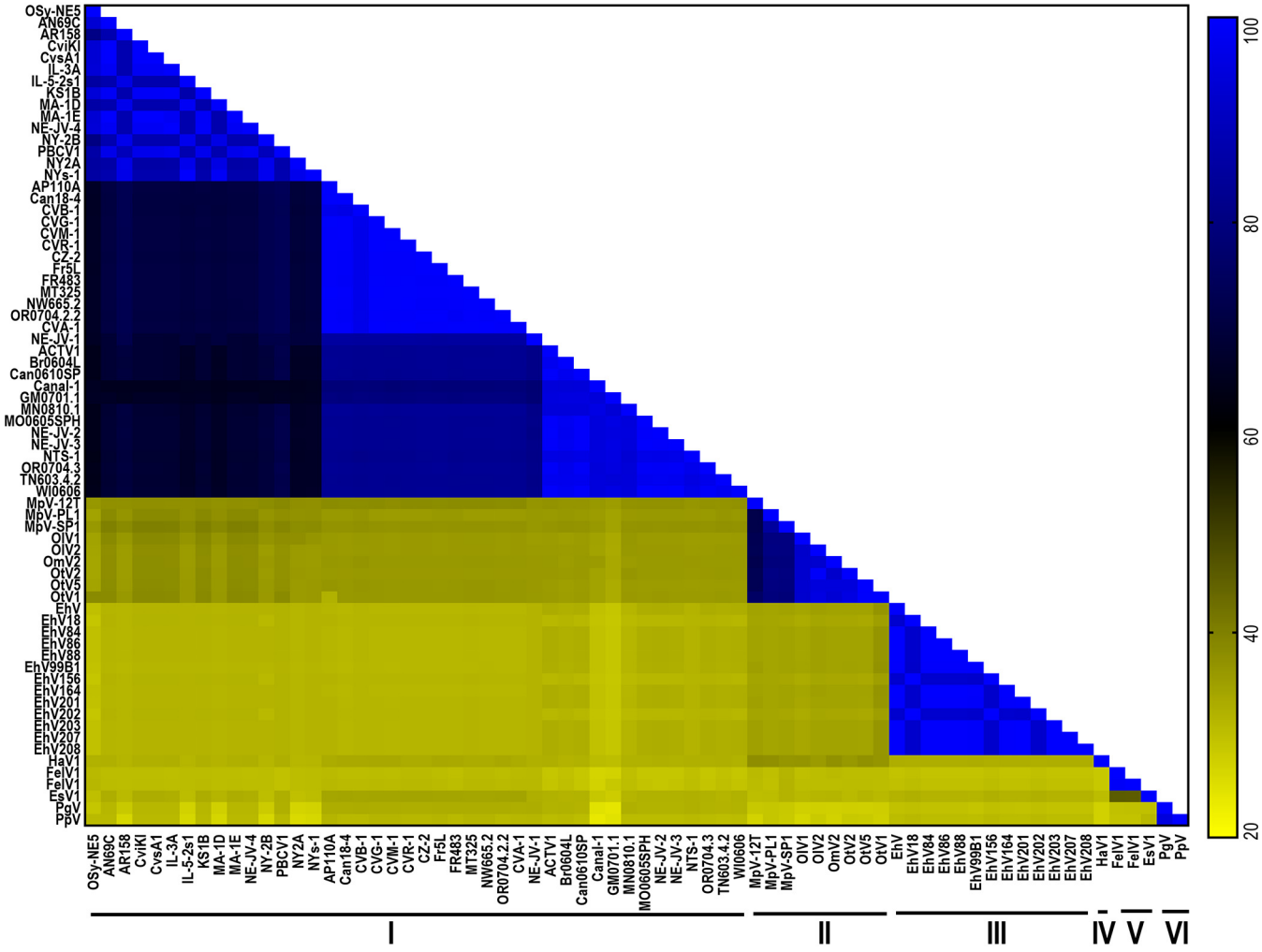

FIG 3 Similarity range of chlorovirus DNA polymerase B genes. Shown is a heat map of identity analysis of DNA polymerase B genes of all *Phycodnaviridae* members included for phylogenetic reconstruction. Genera are indicated in roman numerals. I, *Chlorovirus*; II, *Prasinovirus*; III, *Coccolithovirus*; IV, *Raphidovirus*; V, *Phaeovirus*; VI, *Prymnesiovirus*. Identity ranges from 20% (yellow) to 100% (blue).

It should be noted that all open reading frames (ORFs) with 64 or more codons were counted in the six viruses AR158, NY-2A, PBCV-1, FR483, MT325, and ATCV-1. These viruses have more than 800 CDSs, which is due to different strategies used for gene prediction and annotation (18–20). In contrast, the viruses with lower numbers of CDSs were selected by removing any ORFs that were inside larger ORFs because they were unlikely to be real protein-encoding genes. When this was done with the 6 viruses mentioned above, the number of likely authentic CDSs resembled those of the 36 viruses used in this study.

Our analysis of the 36 chloroviruses identified a total of 1,345 clusters of orthologous groups of genes (COGs) as components of the chlorovirus pan-genome (Fig. 5). With the addition of new viruses, new unique genes were identified, contributing to a constant increase in the viral pan-genome. At the same time, well-conserved genes were identified, making up the core genome of chloroviruses. A total of 543 COGs were singletons; that is, they had only one gene and were unique to a specific virus. Three hundred COGs were shared by 2 to 5 viruses, and 131 COGs were shared by between 6 and 10 viruses. Only 126 (9.36%) COGs were shared by all 36 chloroviruses, constituting the core genome of this group of viruses (Fig. 5; see also Data Set S1 in the supplemental material). Interestingly, the addition of genes identified in specific host viruses did not result in a significant increase or decrease in the pan-genome or core genome, respectively. A slightly greater slope in the pangenome curve was

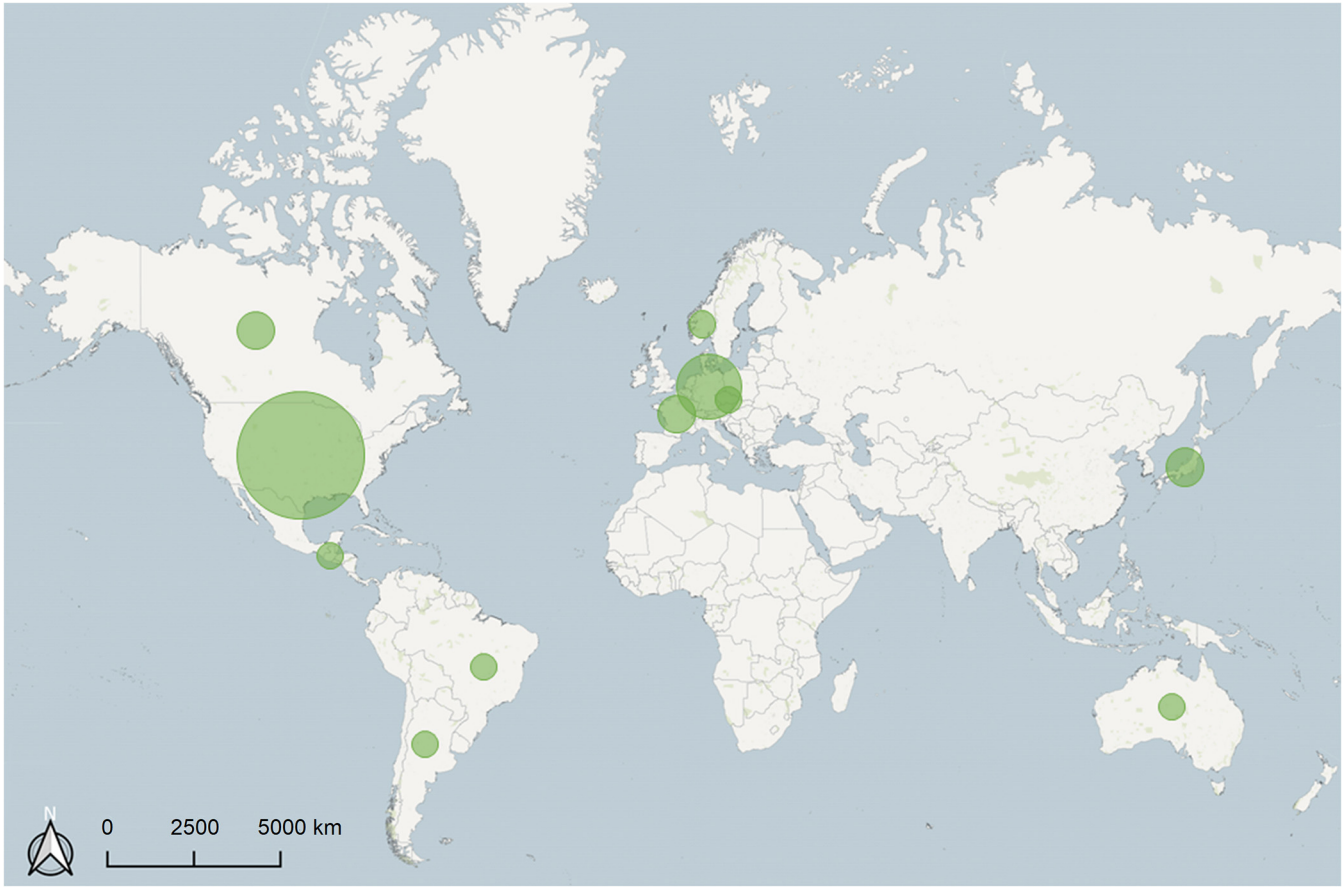

FIG 4 Worldwide chlorovirus isolation sites. The size of the areas of the green circles corresponds to the number of chlorovirus isolates in each country, as identified in Table 1.

observed with the inclusion of Only-Syngen Nebraska virus 5 (Osy-NE5). Although this virus infects the same host cell at the species level, i.e., *Chlorella variabilis*, only the Syngen 2-3 strain supports virus replication, thus having a different virus-host relationship than the other NC64A viruses (21). Such specific interactions can be attributed to a greater genetic diversity of this virus, and this hypothesis will be tested in future studies. Our analysis showed an open pan-genome for chloroviruses and a well-conserved core genome for this group of viruses.

Functional characterization of the pan-/core genome of chloroviruses. Among the 1,345 COGs, 891 (66.2%) had no defined function, being considered hypothetical proteins with no known conserved domains (Fig. 6A). A total of 115 COGs do not currently have a well-established function (listed as miscellaneous), but conserved domains were identified in proteins from genes belonging to these clusters, such as domains containing glycoprotein repeats and ankyrin repeats, chitin binding domains, and PBCV-specific basic adaptor domain-containing proteins, among others. Thus, currently, approximately 75% of chlorovirus COGs do not have a well-defined function, constituting a vast field for future investigations. Among the COGs with known functions, most were related to DNA replication, recombination, and repair processes (123; 9.1%), including, among others, DNA polymerase B, DNA primase, DNA topoisomerase type II, DNA ligase, and various endonucleases containing GIY-YIG catalytic domains (Fig. 6A; Data Set S1). Forty-three COGs were associated with carbohydrate metabolism, including enzymes synthesizing extracellular matrix polysaccharides such as hyaluronan and chitin, enzymes synthesizing sugar nucleotides, enzymes responsible for the synthesis of the glycans attached to the virus major capsid proteins, and enzymes involved in algal cell wall degradation such as chitinase and chitosanase activities (22, 23). The other COGs were distributed into functions such as nucleotide (21;

TABLE 1 General information on chloroviruses with available genomes in a public database

Virus	Host species	Genome size (kb)	GC content (%)	No. of CDSs	No. of tRNAs	GenBank accession no.	Source of isolate	Yr of isolation
AN69C	*Chlorella variabilis* (NC64A)	332	40	362	10	JX997153	Canberra, Australia	1995
CviKI	*Chlorella variabilis* (NC64A)	308	40	336	14	JX997162	Kyoto, Japan	1990
CvsA1	*Chlorella variabilis* (NC64A)	310	40	342	14	JX997165	Sawara, Japan	1992
IL-3A	*Chlorella variabilis* (NC64A)	323	40	349	12	JX997169	Illinois, USA	1983
IL-5-2s1	*Chlorella variabilis* (NC64A)	344	41	379	8	JX997170	Illinois, USA	1986
KS1B	*Chlorella variabilis* (NC64A)	287	40	319	13	JX997171	Kansas, USA	2003
MA-1D	*Chlorella variabilis* (NC64A)	339	41	371	11	JX997172	Massachusetts, USA	1984
MA-1E	*Chlorella variabilis* (NC64A)	336	40	376	14	JX997173	Massachusetts, USA	1984
NE-JV-4	*Chlorella variabilis* (NC64A)	328	40	352	11	JX997179	Nebraska, USA	2008
NY-2B	*Chlorella variabilis* (NC64A)	344	41	371	8	JX997182	New York, USA	1986
NYs-1	*Chlorella variabilis* (NC64A)	348	41	381	7	JX997183	New York, USA	1984
AR158	*Chlorella variabilis* (NC64A)	344	41	814	6	NC_009899.1	Buenos Aires, Argentina	1997
NY-2A	*Chlorella variabilis* (NC64A)	369	41	886	7	NC_009898.1	New York, USA	1984
PBCV-1	*Chlorella variabilis* (NC64A)	330	40	802	11	NC_000852.5	New York, USA	1984
OSy-NE5	*Chlorella variabilis* (Syngen 2-3)	327	42	357	13	NC_032001.1	New York, USA	2012
AP110A	*Micractinium conductrix* (Pbi)	327	44	348	9	JX997154		
Can18-4	*Micractinium conductrix* (Pbi)	329	45	357	10	JX997157	Canada	1995
CVA-1	*Micractinium conductrix* (Pbi)	326	45	346	9	JX997159	Amonau, Germany	1984
CVB-1	*Micractinium conductrix* (Pbi)	319	44	346	10	JX997160	Berlin, Germany	1984
CVG-1	*Micractinium conductrix* (Pbi)	318	45	333	9	JX997161	Gottingen, Germany	1984
CVM-1	*Micractinium conductrix* (Pbi)	327	44	341	9	JX997163	Marburg, Germany	1984
CVR-1	*Micractinium conductrix* (Pbi)	329	45	351	9	JX997164	Rauschenberg, Germany	1984
CZ-2	*Micractinium conductrix* (Pbi)	305	45	340	10	JX997166	Czech Republic	1995
Fr5L	*Micractinium conductrix* (Pbi)	302	45	345	11	JX997167	France	1995
NE-JV-1	*Micractinium conductrix* (Pbi)	326	45	337	3	JX997176	Nebraska, USA	2008
NW665.2	*Micractinium conductrix* (Pbi)	325	44	350	8	JX997181	Norway	1984
OR0704.2.2	*Micractinium conductrix* (Pbi)	313	45	344	7	JX997184	Oregon, USA	2007
FR483	*Micractinium conductrix* (Pbi)	321	44	849	9	NC_008603.1	France	1997
MT325	*Micractinium conductrix* (Pbi)	314	45	845	10	DQ491001.1	Montana, USA	1996
Br0604L	*Chlorella heliozoae* (SAG 3.83)	295	49	346	9	JX997155	Sao Paulo, Brazil	2006
Can0610SP	*Chlorella heliozoae* (SAG 3.83)	307	49	341	13	JX997156	British Columbia, Canada	2006
Canal-1	*Chlorella heliozoae* (SAG 3.83)	293	51	336	10	JX997158	Nebraska, USA	2008
GM0701.1	*Chlorella heliozoae* (SAG 3.83)	315	48	362	10	JX997168	Guatemala	2007
MN0810.1	*Chlorella heliozoae* (SAG 3.83)	327	52	343	9	JX997174	Minnesota, USA	2008
MO0605SPH	*Chlorella heliozoae* (SAG 3.83)	289	49	323	11	JX997175	Missouri, USA	2006
NE-JV-2	*Chlorella heliozoae* (SAG 3.83)	319	48	346	13	JX997177	Nebraska, USA	2008
NE-JV-3	*Chlorella heliozoae* (SAG 3.83)	298	49	334	12	JX997178	Nebraska, USA	2008
NTS-1	*Chlorella heliozoae* (SAG 3.83)	323	48	364	7	JX997180	Nebraska, USA	2008
OR0704.3	*Chlorella heliozoae* (SAG 3.83)	311	49	342	13	JX997185	Oregon, USA	2007
TN603.4.2	*Chlorella heliozoae* (SAG 3.83)	321	49	351	9	JX997186	Tennessee, USA	2006
WI0606	*Chlorella heliozoae* (SAG 3.83)	289	50	329	11	JX997187	Wisconsin, USA	2006
ACTV-1	*Chlorella heliozoae* (SAG 3.83)	288	49	860	11	NC_008724.1	Stuttgart, Germany	2002

1.6%), protein (16; 1.2%), and lipid (6; 0.4%) metabolism; transcription and RNA processing (20; 1.5%); signal transduction (16; 1.2%); integration and transposition (9; 0.7%); virus-host interaction (8; 0.6%); and virion structure and morphogenesis (25; 1.8%) (Fig. 6A).

Regarding the core genome of chloroviruses composed of 126 COGs, almost half of these had no known function (60; 47.5%), evidencing a large number of genes with homologues present in all known chloroviruses that may have important functions for viral biology (Fig. 6B). Among the COGs with known function, 11 were related to DNA replication, recombination, and repair, including 2 genes conserved in other NCLDVs, such as DNA primase and DNA polymerase, and another 11 related to RNA transcription and processing, including, among others, the well-conserved VLTF3 (Fig. 6B; Data Set S1). Six COGs were related to nucleotide and protein metabolism. Sixteen COGs were associated with virion structure and morphogenesis, with genes corresponding to the major capsid protein, the A32-like ATPase, postulated to be involved in the encapsidation of the viral genome, and also the penton protein and 13 other minor capsid

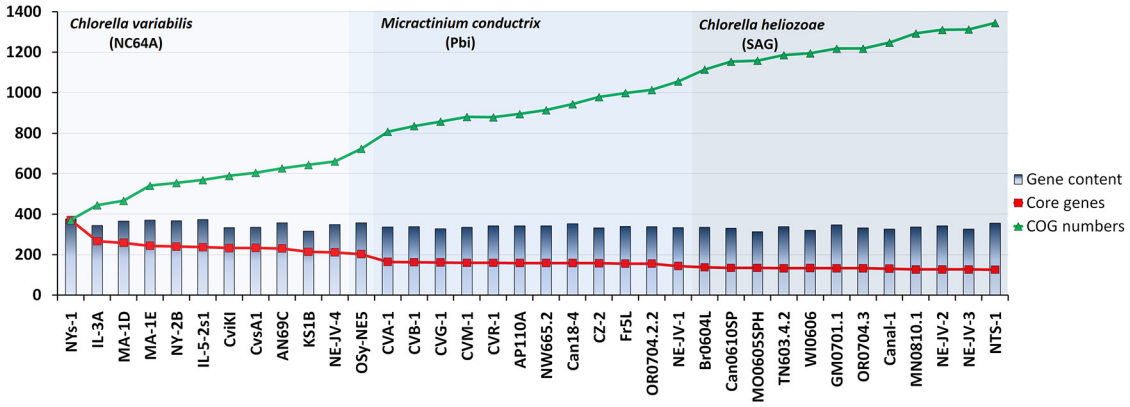

FIG 5 Evolution of the pan-genome of chloroviruses. The stepwise inclusion of chloroviruses indicates an "open" pan-genome and a core genome of this group of viruses. Blue bars indicate the number of CDSs for each virus. The green line indicates the evolution of the gene content (pan-genome), and the red line indicates the evolution of conserved genes (core genome). Viruses are separated according to the host with which they are associated, indicated by shades of blue. The *y* axis shows the number of COGs, and the *x* axis shows the chlorovirus isolates.

proteins, previously identified as P2 to P14 in PBCV-1 (24) (Fig. 6B; Data Set S1). Only two COGs related to carbohydrate metabolism were part of the core genome of chloroviruses, a chitinase and a multidomain glycosyltransferase, a protein with three distinct domains, each with a specific activity (25). Furthermore, a gene encoding a lipid hydrolase was present in all chloroviruses, being the only COG related to lipid metabolism composing the core genome of the viruses (Fig. 6B; Data Set S1). No genes related to integration and transposition or virus-host interaction were conserved in all chloroviruses, suggesting a high specificity of these genes for their respective viruses.

In general, there was a proportional correspondence in the numbers of COGs in the different functional categories between the pan-genome and the core genome (Fig. 6C). Interestingly, COGs related to RNA transcription and processing, protein metabolism, and virion structure and morphogenesis were at least 4-fold more represented in the core genome than in the pan-genome, suggesting that these genes and processes are essential for the chloroviruses (Fig. 6C). Finally, we identified 390 genes in the prototype chlorovirus PBCV-1 with homologues in at least one of the 36 chloroviruses included in this analysis. A total of 141 of the 148 virus-encoded proteins incorporated into the PBCV-1 particle were assigned to a COG (Data Set S1). Among these proteins, 54 (35.5%) fell into COGs shared by all 36 chloroviruses, including the major capsid protein (*A430L*), the penton protein (*A310L*), and 13 minor capsid proteins, named P2 to P14, identified by cryo-electron microscopy (cryo-EM) analysis (24) (Data Set S1).

Nearly 60% (754 of 1,345 COGs) of the chloroviruses' pan-genome is unique to this group of viruses (i.e., the proteins have no homologues outside chloroviruses), while the other 40% possibly originated from *Bacteria* (293; 21.8%), *Eukarya* (160; 11.9%), viruses (128; 9.5%), and *Archaea* (10; 0.7%) (Fig. 7A). Most notably, more than 50% of the COGs with best hits with eukaryotic organisms have homologues in green algae from the *Chlorellaceae* and *Chlorophyceae* families (Fig. 7A). Among COGs of putative viral origins, most are likely from other members of the *Phycodnaviridae* or *Mimiviridae* family (Fig. 7A). When considering the core genome, 30.2% (38 of 126 COGs) is unique to chloroviruses, and most COGs have best hits with *Bacteria* (57; 45.2%), followed by viruses (16; 12.7%), *Eukarya* (13; 10.3%), and *Archaea* (2; 1.6%) (Fig. 7B). Only a small fraction of the most conserved COGs was inferred to originate from green algae (9), and more than 90% of the proteins with best hits with viruses have homologues in members of the *Phycodnaviridae* and *Mimiviridae* families (Fig. 7B).

COG sharing among host-specific viruses. The 1,345 COGs were evenly shared among the 36 chloroviruses included in this analysis (Fig. 8). In our network graph, the presence of COGs shared among viruses from each of the three host groups was evident, and there were specific COGs present in only a single isolate (singletons), which,

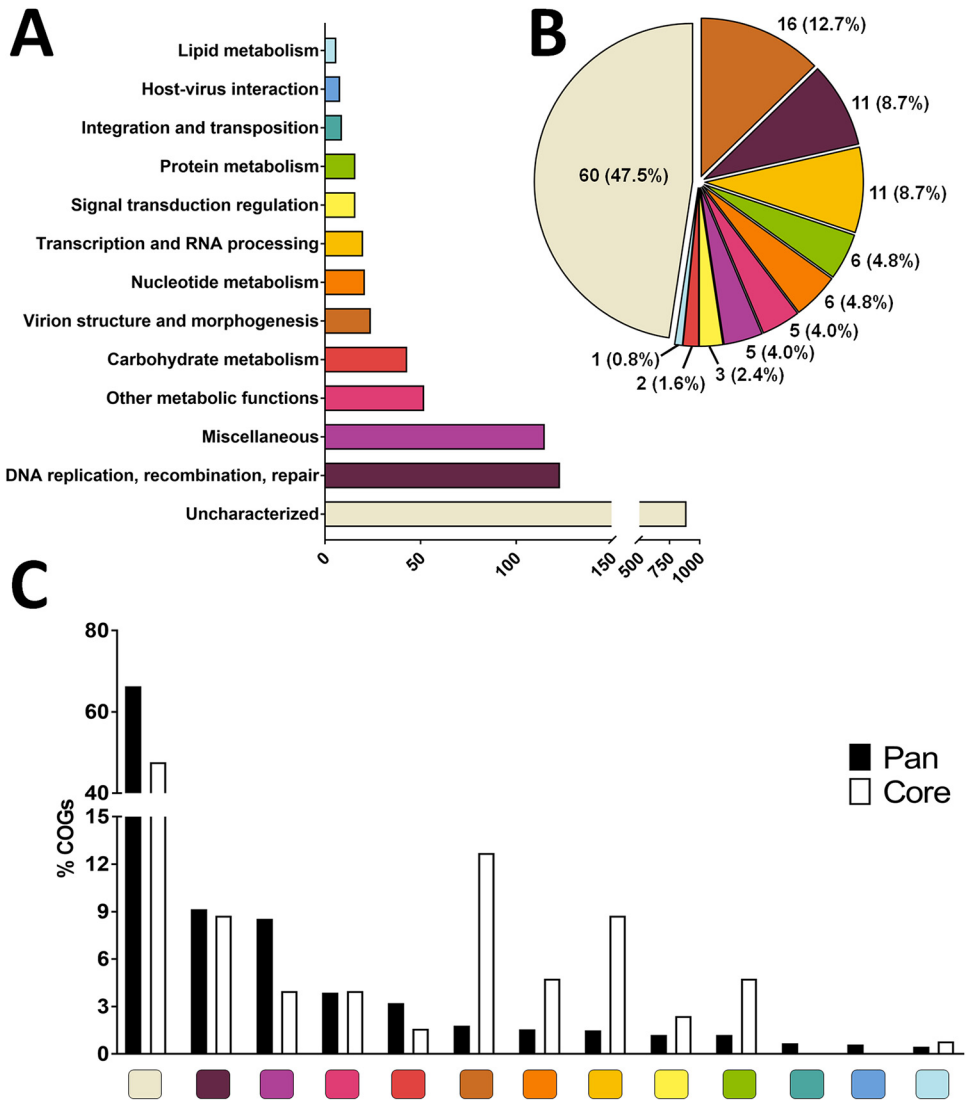

FIG 6 Functional characterization of the pan-genome and core genome of chloroviruses. (A) Distribution of all COGs according to different functional categories of genes; (B) distribution of COGs from the core genome (126 COGs) into functional categories, evidencing both raw numbers and percentages; (C) comparison of COG distributions in the pan-genome and core genome considering the different functional categories. Colors in panels B and C correspond to those depicted in panel A.

for the most part, had no known function (Fig. 8). A total of 270 COGs were shared by at least one specific virus isolated from each host species. Also, some COGs were found specifically in viruses that infect only one type of host, possibly being important in the direct interaction of these viruses with their hosts. NC64A viruses shared 73 COGs exclusively with Pbi viruses and shared only 32 with SAG viruses, while 69 COGs were shared exclusively between Pbi viruses and SAG viruses (Fig. 8). Totals of 368, 268, and 265 COGs were found exclusively in virus isolates that infect NC64A, SAG, and Pbi, respectively. Although the same number of isolates from each host was included in the analysis (12 isolates for each host), there were considerable differences in exclusive COGs among them, with a higher number for NC64A viruses, around 40% more, than for the viruses specific for the other two hosts. In total, 743 COGs were shared among NC64A viruses, 677 COGs were shared among Pbi viruses, and 639 were shared among SAG viruses.

Of the 891 uncharacterized COGs, 138 were present in at least one virus that infects each of the 3 host types, while 253, 208, and 200 COGs were shared exclusively among

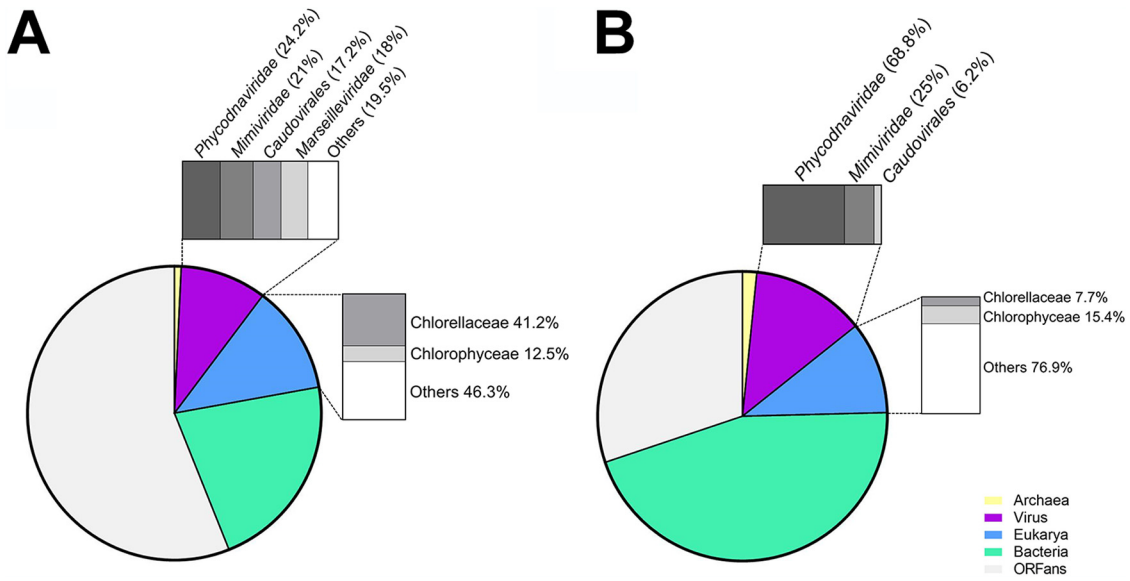

FIG 7 Putative origin of genes composing the pan-genome of chloroviruses. (A) Distribution of all COGs according to the best-hit analysis against the NCBI nr database excluding the chloroviruses; (B) distribution of COGs from the core genome (126 COGs) according to the best-hit analysis. The contributions of different groups of viruses and green algae to the evolution of the pan-genome are indicated. COGs with no hits in the analysis are referred to as "ORFans."

NC64A viruses, Pbi viruses, and SAG viruses, respectively (Fig. 9A). One-third of the COGs related to carbohydrate metabolism were shared among viruses of the 3 host species, including chitinase, chitosanase, and glycosyltransferase clusters, and it is particularly notable that around 30% of COGs were unique to NC64A viruses (Fig. 9B). Thirty-two COGs related to DNA replication, recombination, and repair were shared among viruses that infect the three types of hosts, while 61 were unique to viruses infecting a given host, with 32 shared exclusively among NC64A viruses, 15 shared among SAG viruses, and 14 shared among Pbi viruses (Fig. 5C). Unlike the functional categories of COGs mentioned above, for those related to RNA transcription and processing (20 COGs), 65% were shared among viruses of all hosts, 2 were unique to NC64A viruses, 3 were unique to SAG viruses, and there were no unique COGs shared among the Pbi viruses (Fig. 9D). Regarding the other functional categories, the sharing of COGs was slightly more homogeneous among the specific viruses for each host (Fig. 10). An exception would be the set of genes whose functions are not well elucidated, but there was the presence of conserved domains (miscellaneous), in which there was a greater proportion of these exclusive genes in NC64A viruses than in the others.

DISCUSSION

The *Phycodnaviridae* family was officially created by the ICTV in 1990 to group the newly discovered chloroviruses, large dsDNA viruses associated with chlorella-like green algae (26). Currently, the *Phycodnaviridae* family is composed of 6 distinct genera, among which the genus *Chlorovirus* is the one with the largest number of recognized species and by far the most studied of the family. Recently, in a notable effort by the ICTV to create a viral megataxonomy (27), phycodnaviruses were added to a new order, *Algavirales*, which is included in the class *Megaviricetes*, phylum *Nucleocytoviricota*. Members of this phylum include other families of large DNA viruses that replicate in the cytoplasm and/or nucleus of their host cells, commonly called NCLDVs (15, 16). The NCLDVs have some well-conserved genes (e.g., DNA polymerase B), which point to a common origin, possibly prior to the emergence of eukaryotic organisms (28). Our phylogenetic analysis based on the DNA polymerase B gene, including distinct members of the *Phycodnaviridae* family, showed a clear separation between the different genera, with high statistical support. It is interesting to

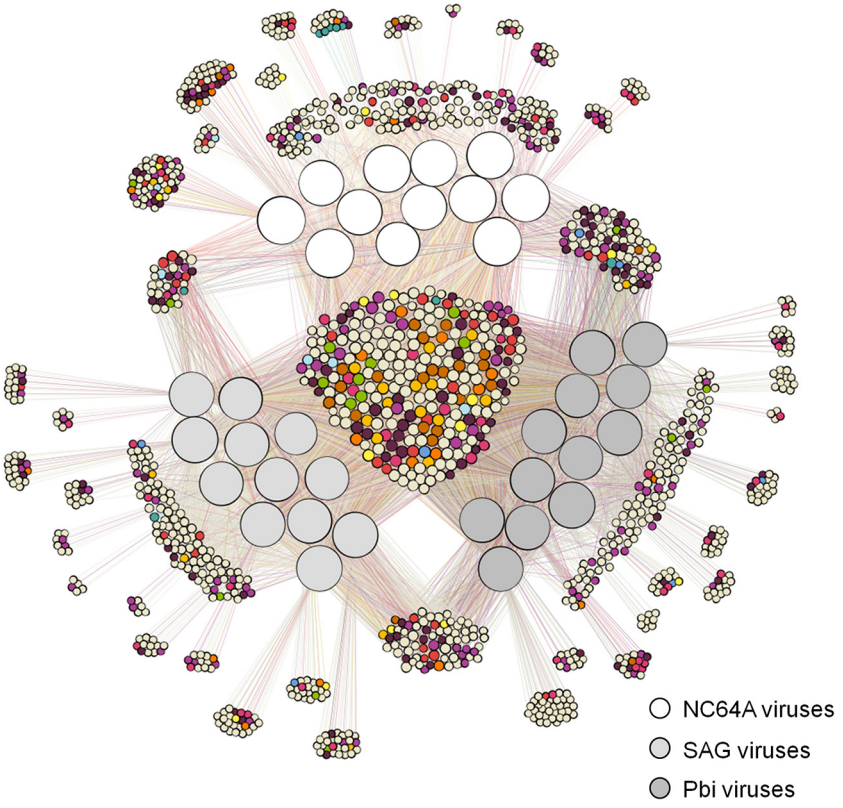

FIG 8 COG sharing among the group of chloroviruses. A bipartite network graph connects the 1,345 COGs to the 36 chloroviruses included in the pan-genome analysis. Larger nodes correspond to the viruses, with different colors corresponding to the specific host with which they are associated. Smaller nodes correspond to the COGs, and colors refer to different functional categories, as indicated in the legend of Fig. 3. The graph was generated using a force-based algorithm, and the nodes were then manually organized to allow better visualization of the connections, especially the singletons. Nodes in the middle of the graph are the COGs shared by at least one virus infecting one of the three distinct hosts.

note, however, that there was a large evolutionary distance between viruses of different genera, which suggests a unique evolutionary path for each group of viruses. Identity analyses, both at the individual-gene level and at the complete-genome level, showed major differences between the different genera (below 60%). Such differences could be enough for a taxonomic restructuring of *Phycodnaviridae*, a point that has recently been analyzed for different members of the *Nucleocytoviricota* (29). Even within the genus *Chlorovirus*, it was possible to observe a separation into at least three distinct clades, which could be classified into different subgenera. Each clade corresponds to viruses infecting a specific host, which indicates that the most recent ancestor of each group already infected its respective host lineage before divergence (9).

Interestingly, our analysis placed the chlorovirus NE-JV-1 isolate in the SAG virus clade, although it was isolated from *Micractinium conductrix* Pbi (9). However, a phylogenetic reconstruction based on a concatenated alignment of 32 conserved chlorovirus genes places this isolate next to the Pbi virus clade (9). This suggests that the evolutionary history of a gene may not fully reflect the evolutionary history of a virus, something also observed for other biological entities (30).

Our analysis revealed an open pan-genome for chloroviruses, and unlike those observed for other virus groups such as *Cedratvirus* and *Marseillevirus* (31, 32), the addition of viruses from different genetic groups (possibly subgenera) did not result in an abrupt increase of the pan-genome or a significant decrease in the core genome, suggesting a more constant evolution for chloroviruses. This general theory assumes that the evolution of the pan-genome is shaped mainly by gene gain and loss by horizontal gene transfer

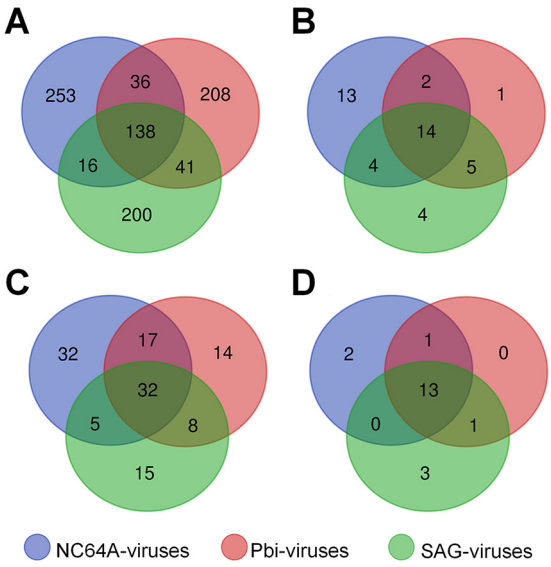

FIG 9 Functional diversity of chlorovirus COGs according to specific host species. Venn diagrams evidence the number of COGs shared by viruses infecting different hosts according to distinct functional categories: uncharacterized (A); carbohydrate metabolism (B); DNA replication, recombination, and repair (C); and transcription and RNA processing (D).

(HGT) events (33, 34). Considering this scenario, one can assume that a considerable fraction of the chloroviruses' pan-genome came from cellular organisms, especially *Bacteria* and *Eukarya*, based on best-hit analysis. Surprisingly, a considerable fraction of genes with best hits with eukaryotic organisms is related to green algae, a result in contrast to those previously reported (9). This could be due to differences in the parameters used in the analysis or in the database. The former is unlikely since similar strategies were used (i.e., a BLASTP search against the NCBI nr database with an E value of $<10^{-5}$) (9). The latter is a better explanation, especially considering that new green alga genomes have been available recently. In fact, most of the best hits with green algae were related to *Chlorella sorokiniana*, *Chlorella desiccata*, and *Micractinium conductrix*, all organisms whose genomes became available after 2018. This update to the genomic database highlights the importance of new studies involving whole-genome sequencing of new green algae, which will profoundly benefit the limnology field but will also have a great impact on the virology field, allowing for an improvement in the understanding of chlorovirus evolution and virus-host interactions.

It is notable that nearly 60% of the genes that constitute the pan-genome are unique to chloroviruses, suggesting that these genes were possibly created *de novo* in the last common ancestor of chloroviruses (or in specific groups within *Chlorovirus*). *De novo* gene creation was proposed to make an important contribution to the evolution of the genomes of giant pandoraviruses, a group closely related to *Phycodnaviridae* (35). A more robust investigation in this regard based on the phylogenetic reconstruction of each of these genes could confirm this hypothesis and provide new insights into the origin and evolution of the chlorovirus pan-genome.

Interestingly, previous studies suggest that the evolution of large and giant DNA viruses occurs following an accordion-like model, wherein their ancestors were small viruses that evolved by gaining and losing genes throughout history through HGT events from their hosts and sympatric organisms (36–38). In this sense, our data supported this hypothesis for chloroviruses, wherein coevolution with their algal hosts would be critical for viral genome evolution. Moreover, the pan-genomic analysis assumes that each gene is separate (or interacts separately) from one another. However, certain chlorovirus genes exist in conserved clusters, referred to as "gene gangs," so the entire cluster appears to be acting as a single unit (13). Therefore, the pan-genome evolution of chloroviruses may have been shaped by additional forces other than HGT, with some genes being conserved among distinct

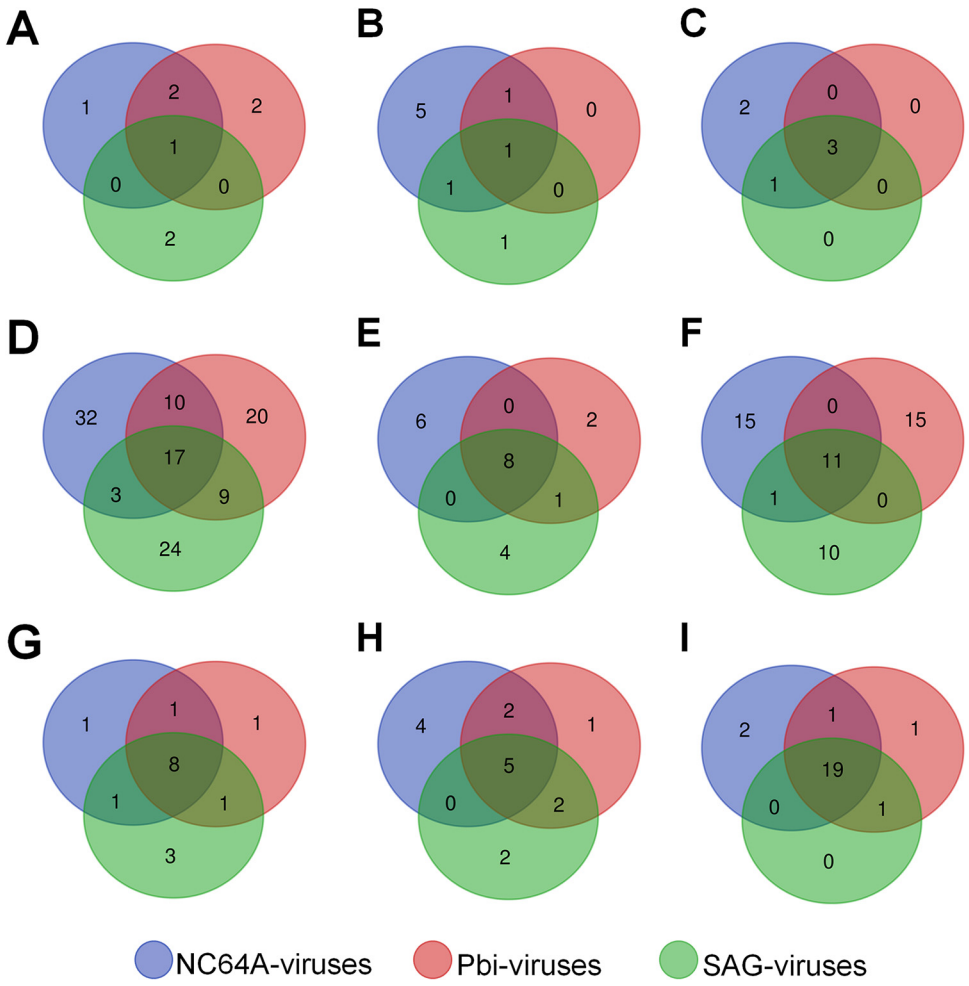

FIG 10 Venn diagrams evidencing the number of COGs shared by viruses infecting different hosts according to distinct functional categories. (A) Host-virus interaction; (B) integration and transposition; (C) lipid metabolism; (D) miscellaneous; (E) nucleotide metabolism; (F) other metabolic functions; (G) protein metabolism; (H) signal transduction regulation; (I) virion structure and morphogenesis.

chloroviruses as gangs, given their pivotal importance to virus replicative success. In addition, organisms with an open pan-genome are usually characterized by having large population sizes, having diverse community interactions, and being niche generalists (34). Although it appears that the chloroviruses are strictly related to chlorella-like green algae, it is possible that these viruses are capable of infecting a broader host range, both now and in the past. This would allow diverse interactions with a plethora of organisms, thus contributing to the increase of the pan-genome, until the establishment of an optimal system for development, in this case, the modern chlorella-like green algal cells.

Chloroviruses have a large dsDNA genome encoding hundreds of proteins (5). On average, chloroviruses have around 350 protein-coding sequences. However, 6 of the 42 chloroviruses whose genomes are annotated and deposited in the databases have a far higher number of CDSs than this, ranging from 802 to 886 CDSs, which results in a major distortion in the gene annotation for some isolates. Although we did not include the prototype chlorovirus PBCV-1 in the pan-genome construction because of the inflated number of CDSs, we searched for homologues and associated them with each COG. Of the 802 CDSs predicted for PBCV-1, 390 were associated with a COG, some of them included in the same cluster, being considered paralogues. Most of the remaining CDSs in PBCV-1 probably do not encode proteins. It should be noted that the PBCV-1 genome was hand-curated after resequencing and comparing the virion proteome (39), allowing an appropriate adjustment for the initial gene prediction and

annotation for this virus. However, the same approach has not been applied to the other virus genomes.

It is worth noting that of the 148 virus-encoded proteins present in the PBCV-1 virion (39), 141 had homologues in other chloroviruses and were assigned to a COG. Interestingly, among the remaining seven proteins, four were products of ORFan genes (i.e., genes found exclusively in PBCV-1). Among them is *A256/257L*, a 2,514-bp gene whose product is a 97-kDa protein (39). This unique protein must have an important role for the virus and should be investigated in further studies.

Much of the chlorovirus pan-genome can be considered a "dark pan-genome," as nearly 70% of all COGs were composed of uncharacterized genes. This constitutes a major unexplored research field that could reveal important novelties of interest for the scientific community. A possible way to infer the functions of these genes would be through the "guilt by association" strategy (13). In that case, conserved collinearity of genes, or gene gangs, could provide useful insights for a functional classification of a gene that belongs to a functionally characterized gang, even though the exact function of that gene would require an in-depth experimental investigation. Considering that the biological function of this hypothetical gene could not be confidently assigned by this strategy, we decided to maintain the "hypothetical" status in our analysis. Even without a defined function, these genes must play an important role in the virus's biology since they are expressed during the viral replication cycle, as evidenced by the transcriptomic data for PBCV-1, where 99% of the virus genome was covered by poly(A)$^+$ mRNA reads (40). The exact mechanisms in the regulation of chlorovirus transcription are currently unknown, but viral promoter motifs are likely to have an influence, at least, for early gene expression (41, 42). These promoter motifs seem to be conserved among the different groups of chloroviruses, a characteristic that may have existed prior to the diversification of this group of viruses.

For the functional classification of COGs, we assigned only one major function to each COG, but we know that some chlorovirus genes have multidomains with specific functions. For example, the PBCV-1 A064R protein, initially described as a glycosyltransferase, has three functional domains: domain 1 is a β-L-rhamnosyltransferase, domain 2 is an α-L-rhamnosyltransferase, and domain 3 is a methyltransferase that methylates the C-2 hydroxyl group of an L-rhamnose unit (43). In these cases, we classified the proteins based on the initial functional annotation. These proteins attest to the great coding potential of chloroviruses, being a good example of genetic economy among giant viruses.

Regarding the COGs that have defined functions, we did not observe a large proportional difference in quantity when considering the pan-genome or just the core genome, with a few exceptions. Of these, it is interesting to note a greater presence of conserved genes related to transcription and RNA processing, suggesting an important ability to manipulate the process once it is triggered in the host cell. Interestingly, no chlorovirus has an RNA polymerase homologue, which makes it completely dependent on the cellular enzyme to produce its transcripts (40). It is interesting to note, however, that all chloroviruses encode two enzymes involved in mRNA capping. One enzyme is responsible for capping the 5′ end of the primary transcripts (44, 45), and the other is an RNA triphosphatase that is part of the capping apparatus (45, 46). The third component of the capping apparatus, methylation of the GpppN cap by *S*-adenosylmethionine:RNA (guanine-N7) methyltransferase, is presumed to be provided by the algal host. Furthermore, four transcription factors are present in all chloroviruses and are possibly used to initiate the gene transcription process, recruiting the host RNA polymerase.

Like transcription-related COGs, those belonging to other functional groups were present in viruses that infect all three host species. A slightly higher number of COGs was exclusively associated with NC64A viruses (368 COGs). It is worth noting that one of the viruses included in the analysis was the isolate named Osy-NE5, a virus isolated from the *Chlorella variabilis* Syngen 2-3 strain (21). Osy-NE5 can infect *C. variabilis* NC64A but does not complete replication in the alga (susceptible but not competent). A total of 348 COGs from NC64A viruses were shared with OSy-NE5, of which 51 were unique to it. It is possible

that with the isolation and characterization of new specific viruses for this strain of *C. variabilis*, we will have new information that will allow a more detailed comparative genomic analysis between the groups. This isolate as well as those closest to it are considered Osy viruses, a new group (possibly a new subgenus) within *Chlorovirus*, and further studies should be carried out to better understand the genomics and biology of Osy viruses.

Conclusions. Our analyses revealed an open pan-genome for chloroviruses, suggesting a great potential for genetic innovation for this group of viruses, once new isolates are obtained from unexplored sites on the planet. Similar to other organisms, the pan-genome evolution of chloroviruses may have been shaped by HGT events, but other forces are likely to have an influence, considering the conserved structures of gene gangs in these viruses (13). Despite the great genetic diversity, the chlorovirus core genome appears to be well delimited, with approximately 1/3 of the genome of these viruses being well conserved, regardless of the subgroup to which it belongs (related to the host). Furthermore, our analyses reinforce the monophyletism of chloroviruses but call attention to the need for a restructuring at the taxonomic level of this important group of aquatic viruses, ubiquitous on the planet. The study of chloroviruses has revealed unique characteristics of these viruses, including an emerging potential for relevant biotechnological applications (47–52). Thus, exploring new sites to isolate and characterize chloroviruses could reveal great surprises, further expanding the pan-genome of these viruses, being a valuable source of genetic diversity and biotechnological innovation.

MATERIALS AND METHODS

Data set. Fasta files containing the complete genomes as well as the amino acid sequences (CDSs) of different chlorovirus isolates were downloaded from GenBank, constituting the raw data set. The search was done using the words "chlorovirus" and "chlorella virus" as bait. Furthermore, some genomes have been identified based on previously described deposits (9). The search was carried out in February 2021, and all sequences available at the time were obtained (42 isolated viruses). Of the 42 isolates, 14 were isolated from *Chlorella variabilis* NC64A, 1 was from *Chlorella variabilis* Syngen 2-3, 14 were from *Micractinium conductrix* Pbi, and 13 were from *Chlorella heliozoae* SAG 3.83. Of the 42 viruses, 6 had a number of CDSs much higher than the average number of CDSs (average of 350 genes) and were excluded from the pan-genome construction in order to avoid distortions in the analysis. Thus, the final data set for pan-genome analysis was composed of CDSs from 36 virus isolates, with 12 representatives of each host species in which these viral genomes have been annotated.

Phylogenetic analysis. The phylogenetic reconstruction was based on the DNA polymerase B gene, as it is a well-conserved gene in all members of the *Phycodnaviridae* family (28). Amino acid sequences from all 42 chloroviruses were included, along with sequences from other representatives of the family, distributed among the other five genera. Sequence alignment was performed using Muscle software with default parameters (53). Noninformative positions were removed using TrimAl with a gap threshold of 0.2 (54). The tree was built using the maximum likelihood method with IQ-TREE, with the evolutionary model being selected by the program (55). Bootstrap analysis was performed with 1,000 replicates, and the tree was visualized with iTOL (56).

Pan-genome construction. To estimate the size of the pan-genome, the chlorovirus gene sequences were clustered using the Proteinortho tool (57), based on the reciprocal best-hit strategy, using an amino acid sequence identity of 30% and a sequence coverage of 60% as thresholds and an E value of $<10^{-5}$. The genes were clustered based on homology, constituting clusters of orthologous genes (COGs). For analysis of the evolution of the pan-genome and core genome, we performed a stepwise inclusion of each virus annotation in the pairwise comparisons of the gene contents of the 36 chlorovirus genome sequences included in the final data set.

Functional analysis and network construction. Once the COGs were established, a representative gene for each of them was used for functional annotation and classification. This was made possible considering the homology of the genes included in each COG; thus, the functions of the genes included in the same cluster should be similar. The gene annotation was performed using BLASTP with an E value of $<10^{-5}$ against the NCBI nr database (58). Conserved domains in the genes were searched using InterProScan (59). After function annotation, the COGs were classified into distinct functional groups based on NCVOG (Nucleo-Cytoplasmic Virus Orthologous Groups) classification and additional categories previously established for chloroviruses (39, 60). To obtain insights on the putative origin of the COGs, a BLASTP search was also performed excluding the *Chlorovirus* (taxid 181083) and unclassified *Chlorovirus* (taxid 346674) taxa and categorized into ORFans (genes with no homologues outside *Chlorovirus*) and those having the best hits with *Bacteria*, *Eukarya*, *Archaea*, and viruses. This analysis was performed using the DIAMOND tool in very-sensitive mode (61). To obtain a general picture of COG sharing among chloroviruses, a bipartite network was built using Gephi (62). The layout was generated using a force-based algorithm, followed by manual rearrangement of the nodes to better visualize the connections and unique groups of genes shared among each virus group based on the host with which they are associated.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLS file, 0.1 MB.

ACKNOWLEDGMENTS

We thank the colleagues of the Virus Laboratory (UFMG) for their constant support during the development of this work.

This work was funded in part by the National Science Foundation under grant no. 1736030 (J.L.V.E.), the University of Nebraska—Lincoln Agricultural Research Division and Office of Research and Economic Development (D.D.D.), and the University of Nebraska Collaborative Initiative Seed Grant Program (D.D.D.). R.A.L.R. and V.F.Q. were CNPq and/or CAPES stipend recipients of the Post-graduation Program in Microbiology/UFMG.

R.A.L.R. designed the study, performed functional pan-genome analyses, and wrote the first draft of the manuscript; V.F.Q. performed phylogenetic and genomic similarity analyses; J.G. analyzed the data and revised the manuscript; and D.D.D. and J.L.V.E. analyzed the data, revised the manuscript, and obtained grants. All authors read and approved the final version of the manuscript.

REFERENCES

1. Suttle CA. 2007. Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5:801–812. https://doi.org/10.1038/nrmicro1750.
2. Mihara T, Koyano H, Hingamp P, Grimsley N, Goto S, Ogata H. 2018. Taxon richness of "Megaviridae" exceeds those of bacteria and archaea in the ocean. Microbes Environ 33:162–171. https://doi.org/10.1264/jsme2.ME17203.
3. Ghedin E, Claverie JM. 2005. Mimivirus relatives in the Sargasso Sea. Virol J 2:62. https://doi.org/10.1186/1743-422X-2-62.
4. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, de Vargas C, Sullivan MB, Bowler C, Wincker P, Karp-Boss L, Sunagawa S, Ogata H. 2020. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. Nat Ecol Evol 4:1639–1649. https://doi.org/10.1038/s41559-020-01288-w.
5. Van Etten JL, Agarkova IV, Dunigan DD. 2019. Chloroviruses. Viruses 12:20. https://doi.org/10.3390/v12010020.
6. Van Etten JL, Dunigan DD, Nagasaki K, Schroeder DC, Grimsley N, Brussaard CPD, Nissimov JI. 2021. Phycodnaviruses (Phycodnaviridae), p 687–695. *In* Bamford DH, Zuckerman M (ed), Encyclopedia of virology, 4th ed. Elsevier, New York, NY.
7. Van Etten JL, Lane LC, Meints RH. 1991. Viruses and virus-like particles of eukaryotic algae. Microbiol Rev 55:586–620. https://doi.org/10.1128/MMBR.55.4.586-620.1991.
8. Van Etten JL, Graves MV, Müller DG, Boland W, Delaroque N. 2002. Phycodnaviridae—large DNA algal viruses. Arch Virol 147:1479–1516. https://doi.org/10.1007/s00705-002-0822-6.
9. Jeanniard A, Dunigan DD, Gurnon JR, Agarkova IV, Kang M, Vitek J, Duncan G, McClung OW, Larsen M, Claverie JM, Van Etten JL, Blanc G. 2013. Towards defining the chloroviruses: a genomic journey through a genus of large DNA viruses. BMC Genomics 14:158. https://doi.org/10.1186/1471-2164-14-158.
10. Van Etten JL, Dunigan DD. 2012. Chloroviruses: not your everyday plant virus. Trends Plant Sci 17:1–8. https://doi.org/10.1016/j.tplants.2011.10.005.
11. Pröschold T, Darienko T, Silva PC, Reisser W, Krienitz L. 2011. The systematics of Zoochlorella revisited employing an integrative approach. Environ Microbiol 13:350–364. https://doi.org/10.1111/j.1462-2920.2010.02333.x.
12. Meints RH, Lee K, Burbank DE, Van Etten JL. 1984. Infection of a Chlorella-like alga with the virus, PBCV-1: ultrastructural studies. Virology 138:341–346. https://doi.org/10.1016/0042-6822(84)90358-1.
13. Seitzer P, Jeanniard A, Ma F, Van Etten JL, Facciotti MT, Dunigan DD. 2018. Gene gangs of the chloroviruses: conserved clusters of collinear monocistronic genes. Viruses 10:576. https://doi.org/10.3390/v10100576.
14. ICTV. 2021. ICTV—taxonomy. https://talk.ictvonline.org/taxonomy/.
15. Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 75:11720–11734. https://doi.org/10.1128/JVI.75.23.11720-11734.2001.
16. Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117:156–184. https://doi.org/10.1016/j.virusres.2006.01.009.
17. Dunigan DD, Fitzgerald LA, Van Etten JL. 2006. Phycodnaviruses: a peek at genetic diversity. Virus Res 117:119–132. https://doi.org/10.1016/j.virusres.2006.01.024.
18. Fitzgerald LA, Graves MV, Li X, Feldblyum T, Nierman WC, Van Etten JL. 2007. Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect Chlorella NC64A. Virology 358:472–484. https://doi.org/10.1016/j.virol.2006.08.033.
19. Fitzgerald LA, Graves MV, Li X, Feldblyum T, Hartigan J, Van Etten JL. 2007. Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect Chlorella Pbi. Virology 358:459–471. https://doi.org/10.1016/j.virol.2006.08.034.
20. Fitzgerald LA, Graves MV, Li X, Hartigan J, Pfitzner AJP, Hoffart E, Van Etten JL. 2007. Sequence and annotation of the 288-kb ATCV-1 virus that infects an endosymbiotic chlorella strain of the heliozoon Acanthocystis turfacea. Virology 362:350–361. https://doi.org/10.1016/j.virol.2006.12.028.
21. Quispe CF, Esmael A, Sonderman O, McQuinn M, Agarkova I, Battah M, Duncan GA, Dunigan DD, Smith TPL, De Castro C, Speciale I, Ma F, Van Etten JL. 2017. Characterization of a new chlorovirus type with permissive and non-permissive features on phylogenetically related algal strains. Virology 500:103–113. https://doi.org/10.1016/j.virol.2016.10.013.
22. Van Etten JL, Agarkova I, Dunigan DD, Tonetti M, De Castro C, Duncan GA. 2017. Chloroviruses have a sweet tooth. Viruses 9:88. https://doi.org/10.3390/v9040088.
23. Van Etten JL, Gurnon JR, Yanai-Balser GM, Dunigan DD, Graves MV. 2010. Chlorella viruses encode most, if not all, of the machinery to glycosylate their glycoproteins independent of the endoplasmic reticulum and Golgi. Biochim Biophys Acta 1800:152–159. https://doi.org/10.1016/j.bbagen.2009.07.024.
24. Fang Q, Zhu D, Agarkova I, Adhikari J, Klose T, Liu Y, Chen Z, Sun Y, Gross ML, Van Etten JL, Zhang X, Rossmann MG. 2019. Near-atomic structure of a giant virus. Nat Commun 10:388. https://doi.org/10.1038/s41467-019-08319-6.
25. Noel E, Notaro A, Speciale I, Duncan GA, De Castro C, Van Etten JL. 2021. Chlorovirus PBCV-1 multidomain protein A111/114R has three glycosyltransferase functions involved in the synthesis of atypical N-glycans. Viruses 13:87. https://doi.org/10.3390/v13010087.
26. ICTV. 1990. Minutes of the 8th plenary meeting of the ICTV, Berlin, 29 August 1990.
27. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. 2020. Global organization and proposed megataxonomy of the virus world. Microbiol Mol Biol Rev 84:e00061-19. https://doi.org/10.1128/MMBR.00061-19.
28. Guglielmini J, Woo A, Krupovic M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. Proc Natl Acad Sci U S A 116:19585–19592. https://doi.org/10.1073/pnas.1912006116.
29. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. 2021. A phylogenomic framework for charting the diversity and evolution of giant viruses. bioRxiv https://doi.org/10.1101/2021.05.05.442809.

30. Futuyma DJ, Kirkpatrick M. 2017. Evolution, 4th ed, chapter 14: The evolution of genes and genomes. Sinauer Associates, Sunderland, MA.
31. Rodrigues RAL, Andreani J, Andrade ACDSP, Machado TB, Abdi S, Levasseur A, Abrahão JS, La Scola B. 2018. Morphologic and genomic analyses of new isolates reveal a second lineage of cedratviruses. J Virol 92:e00372-18. https://doi.org/10.1128/JVI.00372-18.
32. Dornas FP, Assis FL, Aherfi S, Arantes T, Abrahão JS, Colson P, La Scola B. 2016. A Brazilian Marseillevirus is the founding member of a lineage in family Marseilleviridae. Viruses 8:76. https://doi.org/10.3390/v8030076.
33. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr Opin Genet Dev 15:589–594. https://doi.org/10.1016/j.gde.2005.09.006.
34. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The ecology and evolution of pangenomes. Curr Biol 29:R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012.
35. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, Beucher L, Philippe N, Bertaux L, Christo-Foroux E, Labadie K, Couté Y, Abergel C, Claverie J-M. 2018. Diversity and evolution of the emerging Pandoraviridae family. Nat Commun 9:2285. https://doi.org/10.1038/s41467-018-04698-4.
36. Filée J. 2013. Route of NCLDV evolution: the genomic accordion. Curr Opin Virol 3:595–599. https://doi.org/10.1016/j.coviro.2013.07.003.
37. Filée J. 2015. Genomic comparison of closely related giant viruses supports an accordion-like model of evolution. Front Microbiol 6:593. https://doi.org/10.3389/fmicb.2015.00593.
38. Koonin EV, Yutin N. 2019. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. Adv Virus Res 103:167–202. https://doi.org/10.1016/bs.aivir.2018.09.002.
39. Dunigan DD, Cerny RL, Bauman AT, Roach JC, Lane LC, Agarkova IV, Wulser K, Yanai-Balser GM, Gurnon JR, Vitek JC, Kronschnabel BJ, Jeanniard A, Blanc G, Upton C, Duncan GA, McClung OW, Ma F, Van Etten JL. 2012. Paramecium bursaria chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. J Virol 86:8821–8834. https://doi.org/10.1128/JVI.00907-12.
40. Blanc G, Mozar M, Agarkova IV, Gurnon JR, Yanai-Balser G, Rowe JM, Xia Y, Riethoven JJ, Dunigan DD, Van Etten JL. 2014. Deep RNA sequencing reveals hidden features and dynamics of early gene transcription in Paramecium bursaria chlorella virus 1. PLoS One 9:e90989. https://doi.org/10.1371/journal.pone.0090989.
41. Fitzgerald LA, Boucher PT, Yanai-Balser G, Suhre K, Graves MV, Van Etten JL. 2008. Putative gene promoter sequences in the chlorella viruses. Virology 380:388–393. https://doi.org/10.1016/j.virol.2008.07.025.
42. Kawasaki T, Tanaka M, Fujie M, Usami S, Yamada T. 2004. Immediate early genes expressed in chlorovirus infections. Virology 318:214–223. https://doi.org/10.1016/j.virol.2003.09.015.
43. Speciale I, Laugieri ME, Noel E, Lin S, Lowary TL, Molinaro A, Duncan GA, Agarkova IV, Garozzo D, Tonetti MG, Van Etten JL, De Castro C. 2020. Chlorovirus PBCV-1 protein A064R has three of the transferase activities necessary to synthesize its capsid protein N-linked glycans. Proc Natl Acad Sci U S A 117:28735–28742. https://doi.org/10.1073/pnas.2016626117.
44. Håkansson K, Doherty AJ, Shuman S, Wigley DB. 1997. X-ray crystallography reveals a large conformational change during guanyl transfer by mRNA capping enzymes. Cell 89:545–553. https://doi.org/10.1016/s0092-8674(00)80236-6.
45. Ho CK, Van Etten JL, Shuman S. 1996. Expression and characterization of an RNA capping enzyme encoded by Chlorella virus PBCV-1. J Virol 70:6658–6664. https://doi.org/10.1128/JVI.70.10.6658-6664.1996.
46. Ho CK, Gong C, Shuman S. 2001. RNA triphosphatase component of the mRNA capping apparatus of Paramecium bursaria chlorella virus 1. J Virol 75:1744–1750. https://doi.org/10.1128/JVI.75.4.1744-1750.2001.
47. Lohman GJ, Zhang Y, Zhelkovsky AM, Cantor EJ, Evans TC, Jr. 2014. Efficient DNA ligation in DNA-RNA hybrid helices by Chlorella virus DNA ligase. Nucleic Acids Res 42:1831–1844. https://doi.org/10.1093/nar/gkt1032.
48. Krzywkowski T, Nilsson M. 2017. Fidelity of RNA templated end-joining by chlorella virus DNA ligase and a novel iLock assay with improved direct RNA detection accuracy. Nucleic Acids Res 45:e161. https://doi.org/10.1093/nar/gkx708.
49. Dickey JS, Choi TJ, Van Etten JL, Osheroff N. 2005. Chlorella virus Marburg topoisomerase II: high DNA cleavage activity as a characteristic of Chlorella virus type II enzymes. Biochemistry 44:3899–3908. https://doi.org/10.1021/bi047777f.
50. Rakkhumkaew N, Kawasaki T, Fujie M, Yamada T. 2018. Chitin synthesis by Chlorella cells infected by chloroviruses: enhancement by adopting a slow-growing virus and treatment with aphidicolin. J Biosci Bioeng 125:311–315. https://doi.org/10.1016/j.jbiosc.2017.10.002.
51. Nelson M, Burbank DE, Van Etten JL. 1998. Chlorella viruses encode multiple DNA methyltransferases. Biol Chem 379:423–428. https://doi.org/10.1515/bchm.1998.379.4-5.423.
52. Chan SH, Zhu Z, Dunigan DD, Van Etten JL, Xu SY. 2006. Cloning of Nt.CviQII nicking endonuclease and its cognate methyltransferase: M.CviQII methylates AG sequences. Protein Expr Purif 49:138–150. https://doi.org/10.1016/j.pep.2006.04.002.
53. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113.
54. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
55. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
56. 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.
57. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124. https://doi.org/10.1186/1471-2105-12-124.
58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
59. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD. 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49:D344–D354. https://doi.org/10.1093/nar/gkaa977.
60. Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223. https://doi.org/10.1186/1743-422X-6-223.
61. Buchfink B, Reuter K, Drost HG. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. https://doi.org/10.1038/s41592-021-01101-x.
62. Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks, p 361–362. *In* Proceedings of the Third International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence, Menlo Park, CA.