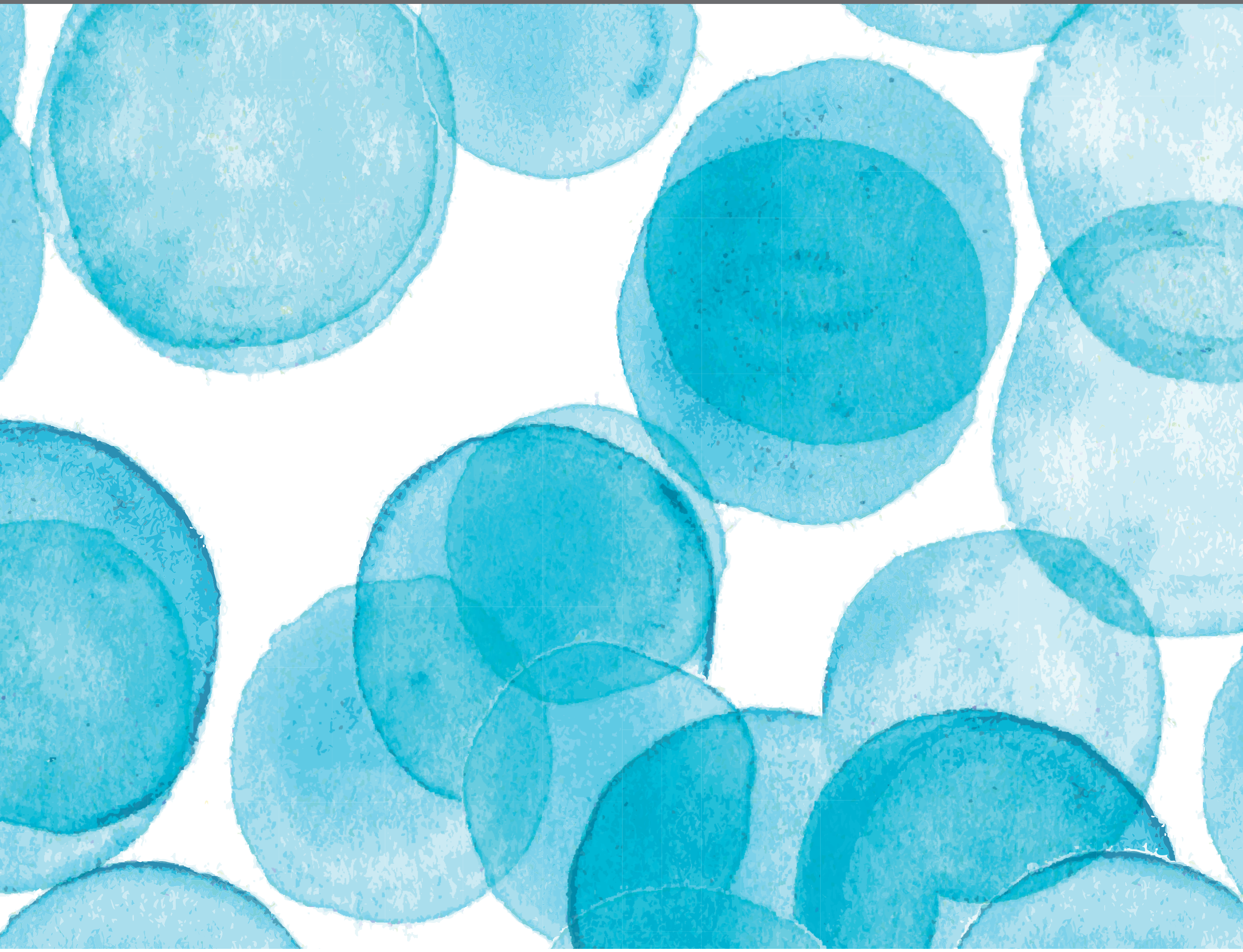
VIROME: DIVERSITY, FUNCTION AND ECOLOGY

EDITED BY: Tao Jin, Mao Ye, Jingzhe Jiang, Pingfeng Yu and
Minh-Thu Nguyen

PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-397-3
DOI 10.3389/978-2-83250-397-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

VIROME: DIVERSITY, FUNCTION AND ECOLOGY

Topic Editors:
Tao Jin, Guangdong Magigene Biotechnology Co.,Ltd, China
Mao Ye, Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences (CAS), China
Jingzhe Jiang, Chinese Academy of Fishery Sciences (CAFS), China
Pingfeng Yu, Zhejiang University, China
Minh-Thu Nguyen, University Hospital Münster, Germany

Citation: Jin, T., Ye, M., Jiang, J., Yu, P., Nguyen, M.-T., eds. (2022).
Virome: Diversity, Function and Ecology. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-397-3

Table of Contents

05 ***Occurrence, Distribution, and Genetic Diversity of Alfalfa*** **(Medicago sativa L.*****) Viruses in Four Major Alfalfa-Producing Provinces of China***

Zhipeng Guo, Tingting Zhang, Zhao Chen, Junpeng Niu, Xuewen Cui, Yue Mao, Mahmood Ul Hassan, Hafiz Abdul Kareem, Nan Xu, Xin Sui, Shuanghong Gao, Momi Roy, Jian Cui and Quanzhen Wang

21 ***RNA Virus Diversity in Birds and Small Mammals From Qinghai–Tibet Plateau of China***

Wentao Zhu, Jing Yang, Shan Lu, Dong Jin, Ji Pu, Shusheng Wu, Xue-Lian Luo, Liyun Liu, Zhenjun Li and Jianguo Xu

32 ***Characterization of Two Novel Toti-Like Viruses Co-infecting the Atlantic Blue Crab,*** **Callinectes sapidus*****, in Its Northern Range of the United States***

Mingli Zhao, Lan Xu, Holly Bowers and Eric J. Schott

45 **Enterococcus faecalis** ***Bacteriophage vB_EfaS_efap05-1 Targets the Surface Polysaccharide and ComEA Protein as the Receptors***

Lingqiong Huang, Wenqiong Guo, Jiahui Lu, Wuliang Pan, Fuqiang Song and Peng Wang

52 ***Picorna-Like Viruses of the Havel River, Germany***

Roland Zell, Marco Groth, Lukas Selinka and Hans-Christoph Selinka

65 ***Distribution Characteristics of Soil Viruses Under Different Precipitation Gradients on the Qinghai-Tibet Plateau***

Miao-Miao Cao, Si-Yi Liu, Li Bi, Shu-Jun Chen, Hua-Yong Wu, Yuan Ge, Bing Han, Li-Mei Zhang, Ji-Zheng He and Li-Li Han

77 ***A Comprehensive Analysis of Citrus Tristeza Variants of Bhutan and Across the World***

Dilip Kumar Ghosh, Amol Kokane, Sunil Kokane, Krishanu Mukherjee, Jigme Tenzin, Datta Surwase, Dhanshree Deshmukh, Mrugendra Gubyad and Kajal Kumar Biswas

94 ***Updating the Phylodynamics of Yellow Fever Virus 2016–2019 Brazilian Outbreak With New 2018 and 2019 São Paulo Genomes***

Ana Paula Moreira Salles, Ana Catharina de Seixas Santos Nastri, Yeh-Li Ho, Luciana Vilas Boas Casadio, Deyvid Emanuel Amgarten, Santiago Justo Arévalo, Michele Soares Gomes-Gouvea, Flair Jose Carrilho, Fernanda de Mello Malta and João Renato Rebello Pinho

103 ***Bacteriophage-Resistant Mutant of*** **Enterococcus faecalis** ***Is Impaired in Biofilm Formation***

Jiazhen Liu, Yanpeng Zhu, Yang Li, Yuwen Lu, Kun Xiong, Qiu Zhong and Jing Wang

113 ***Characterization of*** **Pseudomonas aeruginosa** ***Bacteriophage L5 Which Requires Type IV Pili for Infection***

Lan Yang, Tingting Zhang, Linlin Li, Chao Zheng, Demeng Tan, Nannan Wu, Mingyang Wang and Tongyu Zhu

122 ***Auxiliary Metabolic Gene Functions in Pelagic and Benthic Viruses of the Baltic Sea***

Benedikt Heyerhoff, Bert Engelen and Carina Bunse

138 ***Expanding the Environmental Virome: Infection Profile in a Native Rainforest Tree Species***

Anderson Carvalho Vieira, Ícaro Santos Lopes, Paula Luize Camargos Fonseca, Roenick Proveti Olmo, Flora Bittencourt, Letícia Maróstica de Vasconcelos, Carlos Priminho Pirovani, Fernanda Amato Gaiotto and Eric Roberto Guimarães Rocha Aguiar

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fmicb.2021.771361

Occurrence, Distribution, and Genetic Diversity of Alfalfa (*Medicago sativa* L.) Viruses in Four Major Alfalfa-Producing Provinces of China

Zhipeng Guo[1], Tingting Zhang[2], Zhao Chen[1], Junpeng Niu[1], Xuewen Cui[1], Yue Mao[1], Mahmood Ul Hassan[1], Hafiz Abdul Kareem[1], Nan Xu[1], Xin Sui[1], Shuanghong Gao[1], Momi Roy[1], Jian Cui[3] and Quanzhen Wang[1]*

[1] Department of Grassland Science, College of Grassland Agriculture, Northwest A&F University, Yangling, China, [2] Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, China, [3] Department of Plant Science, College of Life Sciences, Northwest A&F University, Yangling, China

OPEN ACCESS

Edited by:
Mao Ye,
Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences (CAS), China

Reviewed by:
Niu Yanbing,
Shanxi Agricultural University, China
Nicolas Bejerman,
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

****Correspondence:***
Quanzhen Wang
wangquanzhen191@163.com

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *06 September 2021*
Accepted: *22 December 2021*
Published: *13 January 2022*

Citation:
Guo Z, Zhang T, Chen Z, Niu J, Cui X, Mao Y, Hassan MU, Kareem HA, Xu N, Sui X, Gao S, Roy M, Cui J and Wang Q (2022) Occurrence, Distribution, and Genetic Diversity of Alfalfa (Medicago sativa L.) Viruses in Four Major Alfalfa-Producing Provinces of China. Front. Microbiol. 12:771361. doi: 10.3389/fmicb.2021.771361

Alfalfa (*Medicago sativa* L.) is one of the most widely cultivated forage crops in the world. China is the second largest producer of alfalfa in terms of the planting area worldwide, with Gansu, Henan, Inner Mongolia, and Shaanxi provinces being the production hubs. Alfalfa viruses have been reported on a small-scale survey in some of these areas, but they have not been well characterized. In the present study, seven viruses were detected in 12 fields of 10 cities/counties of the four abovementioned provinces by high-throughput sequencing and assembly of small RNA. Their incidence, distribution, and genetic diversity were analyzed by enzyme-linked immunosorbent assay, polymerase chain reaction (PCR)/reverse transcription-PCR and clone sequencing. The results showed that alfalfa mosaic virus (AMV), pea streak virus (PeSV), lucerne transient streak virus (LTSV), alfalfa dwarf virus (ADV), Medicago sativa alphapartitivirus 1 (MsAPV1), MsAPV2, and alfalfa leaf curl virus (ALCV) were the main viruses infecting alfalfa in four examined provinces. AMV and MsAPV1 had the highest incidences in all 4 provinces. SDT analysis of the 7 viruses isolated in China revealed a highly conserved among AMV, LTSV, ADV, MsAPV1, MsAPV2, and ALCV, but the sequence was a high variation between China isolates to abroad isolates in PeSV, ADV, and ALCV. To our knowledge, this is the first report of ADV in Inner Mongolia and Gansu, ALCV in Inner Mongolia, MsAPV1 and MsAPV2 in all 4 provinces, and PeSV and LTSV in China. These findings provide a basis for future research on the genetic evolution of alfalfa viruses in China and on strategies to prevent diseases in alfalfa caused by these viruses.

Keywords: alfalfa, viruses, high-throughput sequencing, small RNA, incidence, genetic diversity, China

INTRODUCTION

Alfalfa (*Medicago sativa* L.), which is referred to as "the queen of forages," is the most important perennial leguminous forage crop worldwide owing to its high nutritional value and value for feeding livestock (Samac et al., 2015). The area of alfalfa cultivation is approximately 32 million hectares globally and is increasing with the development of the animal husbandry market (Samac et al., 2015). In China, alfalfa was planted on about 4.72 million hectares in 2015 (Guo et al., 2020). The major alfalfa-producing provinces include Gansu, Inner Mongolia, Xinjiang, Ningxia, Heilongjiang, Hebei, Shaanxi, Sichuan and Henan, which together produce more than 85% of the total alfalfa production in the country. Gansu and Shaanxi are the leading alfalfa producers in Northwestern China, while the Inner Mongolia Autonomous Region and Henan are the major producers in Northern and Central China, respectively.

To date, about 47 alfalfa viruses have been reported worldwide (**Supplementary Table 1**); of these, 11 have been previously reported in China including alfalfa mosaic virus (AMV; Guo et al., 2019; Li et al., 2021), alfalfa dwarf virus (ADV; Xuehelati et al., 2020), alfalfa leaf curl virus (ALCV; Guo et al., 2020), tomato mosaic virus (ToMV; Wen and Nan, 2015), cowpea mosaic virus (CPMV), bean yellow mosaic virus (BYMV), white clover mosaic virus (WCMV; Zhou et al., 2016), bean leafroll virus (BLRV; Li et al., 2019), Medicago sativa deltapartitivirus 1 (MsDPV1), Medicago sativa amalgavirus 1 (MsAV1; Li et al., 2021), and Medicago sativa alphapartitivirus 1 (MsAPV1; Kim et al., 2018; Nemchinov et al., 2018b; Li et al., 2021). The main symptoms caused by these viruses are macular mosaicism, mottling, ringspot, reddening, etiolation, shrinkage, mosaic shrinkage, and dwarfism (Guo et al., 2020; Samarfard et al., 2020).

High-throughput sequencing (HTS)—which can detect ultra-low levels of plant viruses—and bioinformatics analysis are important tools for the discovery of novel DNA or RNA viruses infecting plants (Mokili et al., 2012). About 10 alfalfa viruses have been identified by HTS technology (Bejerman et al., 2015, 2016; Nemchinov et al., 2017, 2018a; Raza et al., 2017; Kim et al., 2018; Gaafar et al., 2019; Samarfard et al., 2020).

Identifying viruses and analyzing their prevalence and distribution in alfalfa fields are critical for the development of effective disease management strategies. However, very few studies have been conducted in China on alfalfa virus diseases and most were small-scale investigations. In this study we conducted a large-scale survey of viruses in 12 alfalfa fields of 10 cities/counties in Gansu, Henan, Inner Mongolia Autonomous Region, and Shaanxi provinces of China and characterized their prevalence and molecular variability of alfalfa viruses by small (s)RNA HTS, PCR and RT-PCR.

MATERIALS AND METHODS

Plant Material

Tender leaves showing virus-like symptoms, including macular mosaicism (*n* = 104), mottling (*n* = 138), etiolation (*n* = 52), shrinkage (*n* = 328), mosaic shrinkage (*n* = 350), and dwarfism (*n* = 96) (**Figure 1** and **Supplementary Table 2**) were collected in May and June 2020 from alfalfa plants growing in Gansu, Henan, Inner Mongolia Autonomous Region, and Shaanxi provinces (**Figure 2**). The 1,068 samples were collected and included 258 from Jiuquan city, Gansu (G; N39°37′, E98°47′); 80 from Helinger county, Inner Mongolia Autonomous Region (N1; N40°39′, E111°59′); 118 from Tumote Left Banner, Inner Mongolia Autonomous Region (N2; N40°35′, E111°46′); 80 from Yangling district, Xianyang city, Shaanxi (S; N34°18′, E108°0.06′); 120 from Zhengzhou city, Henan (H1; N34°54′, E113°45′); 108 from Yuanyang county-1, Henan (H2; N35°06′, E113°57′); 40 from Yuanyang county-2, Henan (H3; N35°01′, E113°41′); 60 from Lankao county, Henan (H4; N34°49′, E114°59′); 44 from Wenxian county-1, Henan (H5; N34°53′, E113°09′); 40 from Wenxian county-2, Henan (H6; N34°51′, E113°07′); 60 from Yichuan county, Henan (H7; N34°18′, E112°22′); and 60 from Zhenping county, Henan (H8; N32°58′, E112°13′). The samples in each jurisdiction were collected from one field. All samples were immediately transported to the laboratory on dry ice and stored at −80°C until use.

RNA Isolation and Deep Sequencing

Total RNA was extracted from each one of the 1,068 samples using TRIzol reagent (Zhonghuihecai, Shaanxi, China) according to the manufacturer's instructions. For HTS, total RNA was extracted from a pooled sample including 6 leaves exhibiting each of the above-mentioned virus-like symptoms from each location, respectively. And 12 RNA samples from 12 locations were, respectively, subjected to single-end sRNA HTS to detect viruses using the Illumina Hiseq4000 platform at Biomarker Technologies (Beijing, China) and the Illumina Nextseq550 platform at Sangon Biotech (Shanghai, China). The sRNA libraries for deep sequencing were constructed as previously described (Lu et al., 2007). Raw reads were filtered and cleaned by removing those of low quality, with a proportion of unknown bases > 10%, or without adapter sequences along with insert fragments, adapter sequences, and reads > 34 nt and < 15 nt. The Bowtie v1.0.0 software (Langmead et al., 2009) was used to remove rRNA, tRNA, small nuclear RNA, small nucleolar RNA, non-coding RNA and repetitive sequences from the 15- to 34-nt clean reads for alignment to the Silva (Pruesse et al., 2007), GtRNAdb (Chan and Lowe, 2009), Rfam (Griffiths-Jones et al., 2003), Repbase (Jurka et al., 2005) databases. The remaining clean reads were assembled and spliced using SPAdes software (K-mer value = 17; Bankevich et al., 2012). The assembled contigs were compared with GenBank Virus RefSeq nucleotide and protein databases and NCBI Non-redundant protein and nucleotide sequences databases using BLASTn and BLASTx (1e-5).

Virus Detection by PCR and Reverse Transcription (RT)-PCR

To detect viruses in alfalfa samples, total DNA and RNA was extracted from each of the 1,068 samples using the Plant Genomic DNA Extraction Kit (Beijing Solarbio Science and Technology Co., Beijing, China) and the TRIzol reagent (Zhonghuihecai,

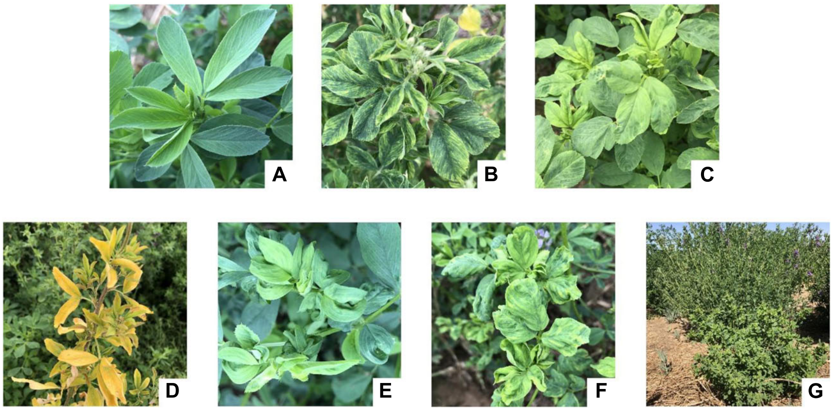


FIGURE 1 | Symptom types of alfalfa virus disease in the field. **(A)** Health; **(B)** macular mosaicism; **(C)** mottling; **(D)** etiolation; **(E)** shrinkage; **(F)** mosaic shrinkage; **(G)** dwarfism.

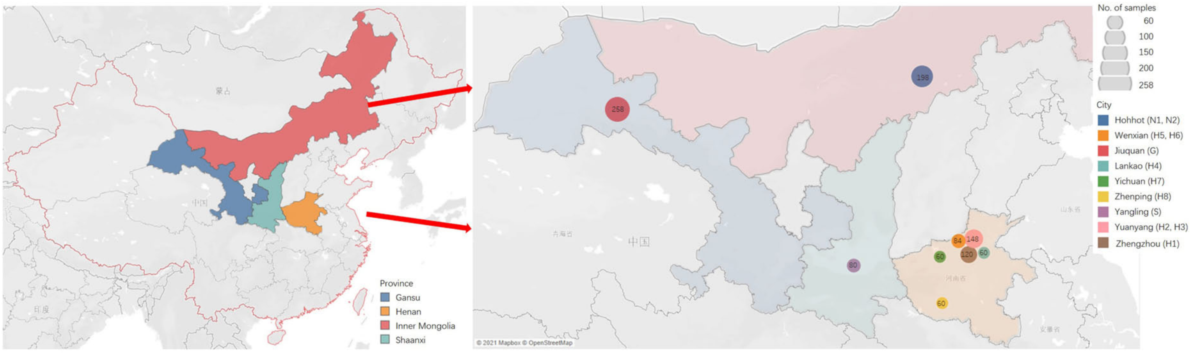


FIGURE 2 | The sites of samples with virus-induced-symptoms in Inner Mongolia, Gansu, Shaanxi and Henan provinces of China.

Shaanxi, China), respectively, according to the manufacturer's recommendations. For RNA samples, the cDNA was synthesized using HiScript II reverse transcriptase and oligo (dT)23VN primer (Vazyme Biotech, Nanjing, China) according to the manufacturer's instructions. PCR (used for DNA virus detection) and RT-PCR (used for RNA virus detection) were performed using the Taq PCR Master Mix Kit (Jiangsu CoWin Biosciences, Taizhou, China) with the following programs: predenaturation at 94°C for 2 min; 35 cycles of denaturation at 94°C for 30 s, annealing at 56°C for 30 s, and extension at 72°C for 2 min; and final extension at 72°C for 2 min. Primers used to amplify specific sequences of AMV, pea streak virus (PeSV), lucerne transient streak virus (LTSV), ADV, MsAPV1, Medicago sativa alphapartitivirus 2 (MsAPV2), and ALCV are listed in **Supplementary Table 3**. The PCR products were resolved by gel electrophoresis on a 1.2% agarose gels and purified using SanPrep Column DNA Gel Extraction Kit (Sangon Biotech). The purified PCR products were verified by Sanger sequencing by Sangon Biotech.

Cloning and Sequencing Analysis

The purified PCR products consisting of the sequence fragments of virus capsid protein were cloned into the pUCm-T vector (Sangon Biotech) and transformed into *Escherichia coli* DH5α competent cells (Sangon Biotech). Positive clones were verified by PCR and at least 3 independently derived clones were sequenced by Sangon Biotech. By cloning and sequencing, we obtained the sequences of the complete coat protein (CP) gene sequences of PeSV, LTSV, MsAPV1, and MsAPV2, and the complete Nucleocapsid (N) gene of ADV, and the complete genome sequence of ALCV. Almost the whole genomic sequence was obtained by splicing the assembled contigs sequences and cloning sequencing sequences with DNAMAN v6 (Lynnon Biosoft, QC, Canada) software (**Supplementary Tables 4–10**).

Viral reference sequences were downloaded from GenBank (**Supplementary Tables 4–10**) and aligned using Muscle with default parameters (Kumar et al., 2016). Phylogenetic and molecular evolutionary analyses were performed using MEGA v7.0 (Kumar et al., 2016) with 1,000 bootstrap replicates. A phylogenetic tree was constructed by the maximum likelihood (ML) method using MEGA v7.0 software. The other parameters were as follows: substitution type = nucleotide; model/method = Jukes–Cantor model; rates among sites = uniform; gaps/missing data treatment = complete deletion; ML = heuristic (nearest neighbor interchange); initial tree for ML = automatically constructed (maximum parsimony).

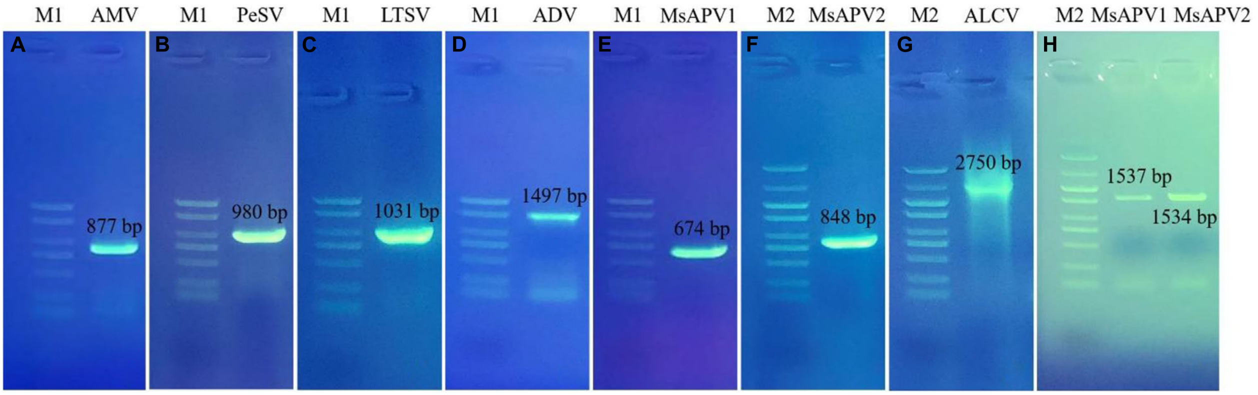


FIGURE 3 | PCR validation for the presence of seven viruses. **(A)** AMV; **(B)** PeSV; **(C)** LTSV; **(D)** ADV; **(E)** MsAPV1; **(F)** MsAPV2; **(G)** ALCV, **(H)** MsAPV1 (left) and MsAPV2 (right). M1 and M2 represent DL 2000 DNA marker and DL 5000 DNA marker, respectively. The specific bands in A-H were amplified by corresponding primers from **Supplementary Table 3**, respectively.

Pairwise sequence alignment of viral reference sequences (**Supplementary Tables 4–10**) was performed using SDT software (Muhire et al., 2014). The parameters were as follows: alignment programs = Muscle; Cluster sequences using a neighbor joining tree.

Recombination analysis was performed by SimPlot 3.5 software (Lole et al., 1999). The nucleotide sequence of the isolate from different host and country were used as the reference sequence (**Supplementary Tables 4–10**), then Similarity plot and Bootscanning analysis were performed using SimPlot 3.5 (Lole et al., 1999). Genetic Algorithm Recombination Detection (GARD) was used to detect the Recombination sites and evaluate their reliability (Pond et al., 2006a,b).

AMV, PeSV, LTSV, and ALCV Detection by Enzyme-Linked Immunosorbent Assay

Five samples of each virus (AMV, PeSV, LTSV, and ALCV)-positive responses to RT-PCR were used for enzyme-linked immunosorbent assay (ELISA), respectively. AMV, PeSV, LTSV, and ALCV using ELISA kits obtained from Shanghai Yuanxin Biotechnology Co., Ltd., to identify the virus following the procedure provided by the supplier.

RESULTS

Detection of Viruses Infecting Alfalfa by sRNA High-Throughput Sequencing

Each RNA sample was extracted from a pooled sample of 6 leaves exhibiting each of the above-mentioned virus-like symptoms from each location and pooled, fragmented into libraries, and sequenced on an Illumina platform (San Diego, CA, United States), respectively. From the 12 HTS data, 6,933,449–15,668,658 clean reads were selected from 9,788,810–16,082,935 raw reads (**Supplementary Table 11**); 23–111 contigs were mapped to 22 viral reference sequences (**Supplementary Table 12**). The contigs of sample G were aligned to nucleotide sequences of AMV, PeSV, ADV, MsAPV1, BLRV, and raspberry vein chlorosis virus. The contigs of sample N1 were aligned to nucleotide sequences of AMV, PeSV, LTSV, ADV, MsAPV1, ALCV, BLRV, alfalfa latent virus (ALV), Allium fistulosum carlavirus, Ilex cornuta carlavirus, garlic common latent virus, birch carlavirus, cowpea mild mottle virus, cherry twisted leaf-associated virus, and cherry green ring mottle virus. The contigs of sample N2 were aligned to nucleotide sequences of AMV, PeSV, ADV, ALCV, BLRV, and ALV. The contigs of sample S were aligned to nucleotide sequences of AMV, MsAPV1, and MsAPV2. The contigs of sample H1 were aligned to nucleotide sequences of AMV, ADV, MsAPV1, ALCV, grapevine cabernet sauvignon reovirus, raspberry latent virus, and cassava frogskin virus. The contigs of sample H2 were aligned to nucleotide sequences of AMV, ADV, MsAPV1, and ALCV. The contigs of sample H3 were aligned to nucleotide sequences of AMV, MsAPV1, MsAPV2, ALCV, and MsAV1. The contigs of sample H4 were aligned to nucleotide sequences of AMV, MsAPV1, and ALCV. The contigs of sample H5 were aligned to nucleotide sequences of AMV, ADV, MsAPV1, and ALCV. The contigs of sample H6 were aligned to nucleotide sequences of AMV and ALCV. The contigs of sample H7 were aligned to nucleotide sequences of AMV, MsAPV1, and ALCV. The contigs of sample H8 were aligned to nucleotide sequences of AMV, MsAPV1, and capsicum chlorosis virus (**Supplementary Table 12**). Nearly complete genomic sequences of PeSV, ALCV, and AMV were assembled from samples G, H1, and H8, respectively (**Supplementary Table 12**). Those assembly sequences were derived from each mixed sample; thus, could potentially be chimeric. Hence a further molecular validation should be performed by PCR and RT-PCR.

Detection of Viruses Infecting Alfalfa Detected by PCR and RT-PCR

Twenty-five primer pairs were used to detect 22 different viruses by PCR and RT-PCR. Fifteen of the viruses tested negative, but 7 of the viruses (AMV, PeSV, LTSV, ADV, MsAPV1, MsAPV2, and ALCV) tested positive in alfalfa pooled samples. After that, we detected the above seven viruses in 1,068 samples, separately. Of these, AMV, MsAPV1, and MsAPV2 were detected in all 4 provinces; PeSV was detected in Gansu and Inner Mongolia; LTSV was detected in Inner Mongolia; ADV was detected in Gansu, Henan, and Inner Mongolia, and ALCV was detected in Henan and Inner Mongolia. The PCR products of ALCV and RT-PCR products of AMV, PeSV, LTSV, ADV, MsAPV1, and MsAPV2 showed distinct bands by agarose gel electrophoresis, with sizes of 2,750, 877, 980, 1031, 1497, 674 bp (MsAPV1); 848 bp (MsAPV2); 1,537 bp (MsAPV1 CP gene); and 1,534 bp (MsAPV1 CP gene) (**Figure 3**), confirming the presence of the 7 viruses in the alfalfa samples.

AMV, PeSV, LTSV, and ALCV Detection by Enzyme-Linked Immunosorbent Assay

The results of ELISA for four kinds of viruses (AMV, PeSV, LTSV, and ALCV) detection were shown in **Figure 4**. As shown in **Figure 4**, five samples of each virus that were detected positive for the above viruses by RT-PCR were also detected positive by ELISA, respectively. While the negative control of alfalfa without showing symptoms typical of virus infection were not detected for these four kinds of viruses by ELISA.

Virus Prevalence and Distribution

The viruses had the highest prevalence among samples from all surveyed cities/counties in Henan province (100%), followed by Tumote Left Banner, Inner Mongolia Autonomous Region (89.83%); Jiuquan, Gansu (84.50%); Helinger county, Inner Mongolia Autonomous Region (80.00%); and Yangling, Shaanxi (65.00%) (**Supplementary Table 13**). AMV was detected at the highest rate, followed by MsAPV1, PeSV, ADV, ALCV, MsAPV2, and LTSV. AMV and MsAPV1 were detected in all 12 alfalfa-growing regions of the 4 provinces (**Figure 5A**). The main symptoms of samples single-infected with AMV were etiolation (10/52) and macular mosaicism (14/104), while samples only infected with MsAPV1 showed shrinkage (18/328) and mosaic shrinkage (18/350) as the major symptoms (**Supplementary Table 2**). In samples infected with multiple viruses, the most frequent combinations were AMV + MsAPV1, AMV + PeSV, AMV + PeSV + MsAPV1, AMV + ADV + MsAPV1, and AMV + MsAPV1 + ALCV (**Figure 5C**). The main symptoms of samples infected with virus combinations were as follows: AMV + MsAPV1, shrinkage (94/328), macular mosaicism (28/104), and mosaic shrinkage (86/350); AMV + PeSV, macular mosaicism (12/104) and etiolation (4/52); AMV + PeSV + MsAPV1, mottling (26/138) and shrinkage (40/328); AMV + ADV + MsAPV1, dwarfism (18/96) and mosaic shrinkage (36/350); and AMV + MsAPV1 + ALCV, dwarfism (12/96), mosaic shrinkage (28/350), and shrinkage (26/328) (**Supplementary Table 2**). The detection rates of AMV, MsAPV1, ADV, ALCV, and LTSV were lower in samples planted before 2012 than in those planted after 2012. On the contrary, PeSV and MsAPV2 detection rates were higher among samples planted before 2012 than among those planted after 2012 (**Figure 5B**).

The incidences of single and multiple infections varied across 12 alfalfa-growing locations in the 4 provinces (**Supplementary Tables 14–25**). AMV and MsAPV1 were main viruses involved in single infections in all of the fields (**Supplementary Tables 14–25**). The sites with the highest single-infection rates were Yuanyang-2, Yuanyang-1, Helinger, Tumote Left Banner, Lankao and Zhenping (AMV); and Wenxian-1, Zhenping, Yangling, and Zhengzhou (MsAPV1) (**Figure 6A**). The most common 2-virus combinations were AMV + MsAPV1 and AMV + PeSV; the sites with the highest dual infection rates were Lankao, Zhenping, Yuanyang-2, Yangling, Yichuan, Yuanyang-1 and Zhengzhou (AMV + MsAPV1); and Helinger and Jiuquan (AMV + PeSV) (**Figure 6B**). The most frequent combinations of multiple viruses were AMV + PeSV + MsAPV1, AMV + ADV + MsAPV1, and AMV + MsAPV1 + ALCV; the sites with the highest multiple infection rates were Jiuquan (AMV + PeSV + MsAPV1); Zhengzhou and Yuanyang-1 (AMV + ADV + MsAPV1); and Wenxian-2 and Yichuan (AMV + MsAPV1 + ALCV) (**Figure 6C**).

Recombination Analysis of Alfalfa Viruses CP Gene or N Gene

The Simplot analysis and GARD found no recombination evidence of MsAPV1, and MsAPV2 (**Figures 7G–J**). Simplot analysis did not detect recombination signals in PeSV isolates (**Figure 7A**), but GARD found evidence of recombination with up to 8 breakpoints (**Figure 7B**). On the contrary, recombination signals of LTSV and ADV isolates from 4 provinces were detected by using Simplot (**Figures 7C,E**), and the recombination sites of these isolates were further confirmed by GARD (**Figures 7D,F**). For LTSV, GARD found 3 recombination sites, which were located at 587, 635, and 794 sites of CP gene, respectively (**Figure 7D**), with an average model approval rate of 26.41%, 51.17%, and 28.58%, respectively (**Figure 7D**). For ADV, GARD found 4 recombination sites, which were located at 286 (approval rate 30.80%), 1,064 (46.51%),

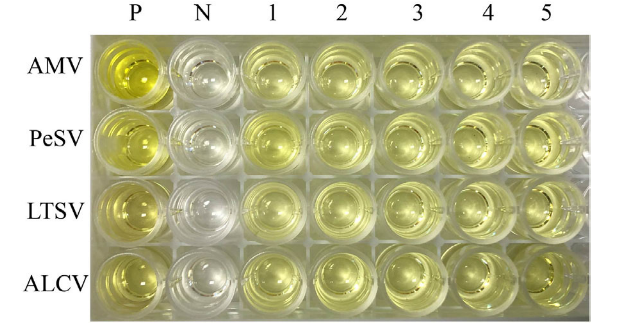


FIGURE 4 | The results of ELISA for AMV, PeSV, LTSV, and ALCV detection. P: Positive control (*Medicago sativa* leaves that were infected by AMV, PeSV, LTSV, and ALCV, respectively), N: Negative control (virus-free alfalfa leaves), 1–5: Five samples of each virus that were detected positive for RT-PCR were also detected positive by ELISA. Line 1–4 represent AMV, PeSV, LTSV, and ALCV, respectively.

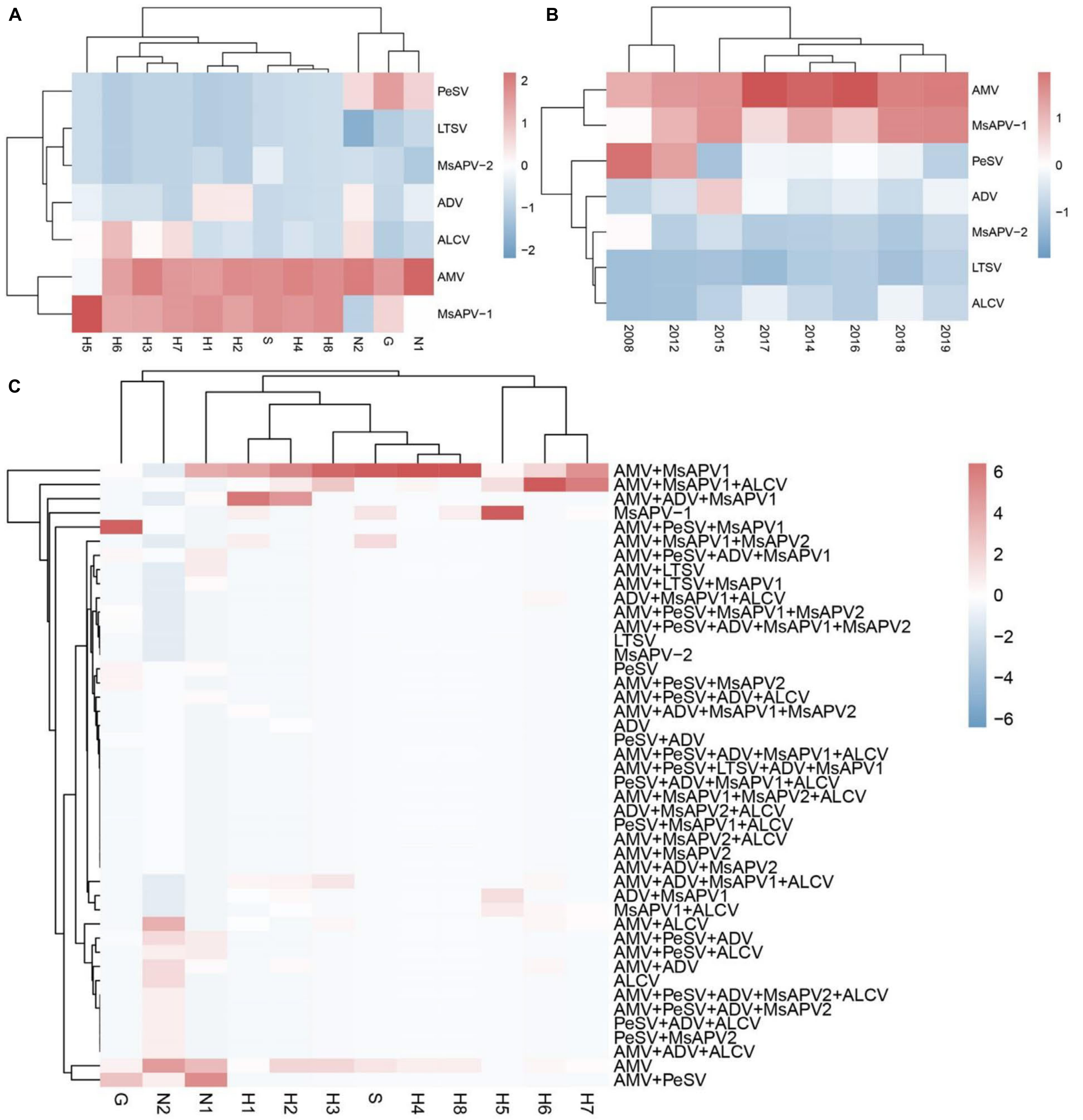


FIGURE 5 | The colored heat map of detection rate of AMV, PeSV, LTSV, ADV, MsAPV1, MsAPV2 and ALCV of samples, in different locations **(A)**, planted in different year **(B)**, and incidence of various viruses with both single and multiple infection in different alfalfa-growing locations **(C)**. Jiuquan (G), Helinger (N1), Tumote Left Banner (N2), Yangling (S), Zhengzhou (H1), Yuanyang-1 (H2), Yuanyang-2 (H3), Lankao (H4), Wenxian-1 (H5), Wenxian-2 (H6), Yichuan (H7) and Zhenping (H8). The red color represents a high detection rate of virus, while the blue color represents a low detection rate of virus.

1,073 (8.94%), and 1,163 (42.08%) positions of the N gene, respectively (**Figure 7F**).

Recombination Analysis of Whole Genome Sequence of Alfalfa Viruses

The Simplot analysis and GARD found recombination evidence of AMV, and ALCV (**Figure 8**). Simplot analysis detected recombination signals in AMV and ALCV isolates (**Figures 8A,C,E,G**), and GARD found evidence of recombination with up to 7, 6, 4, and 6 breakpoints in AMV-RNA1, AMV-RNA2, AMV-RNA3, and ALCV, respectively (**Figures 8B,D,F,H**). For AMV-RNA1, GARD found 7 recombination sites, which were located at 99, 321, 942, 1,441, 2,142, 2,670, and 3,500 sites of AMV-RNA1 genome, respectively (**Figure 8B**), with an average model approval rate of

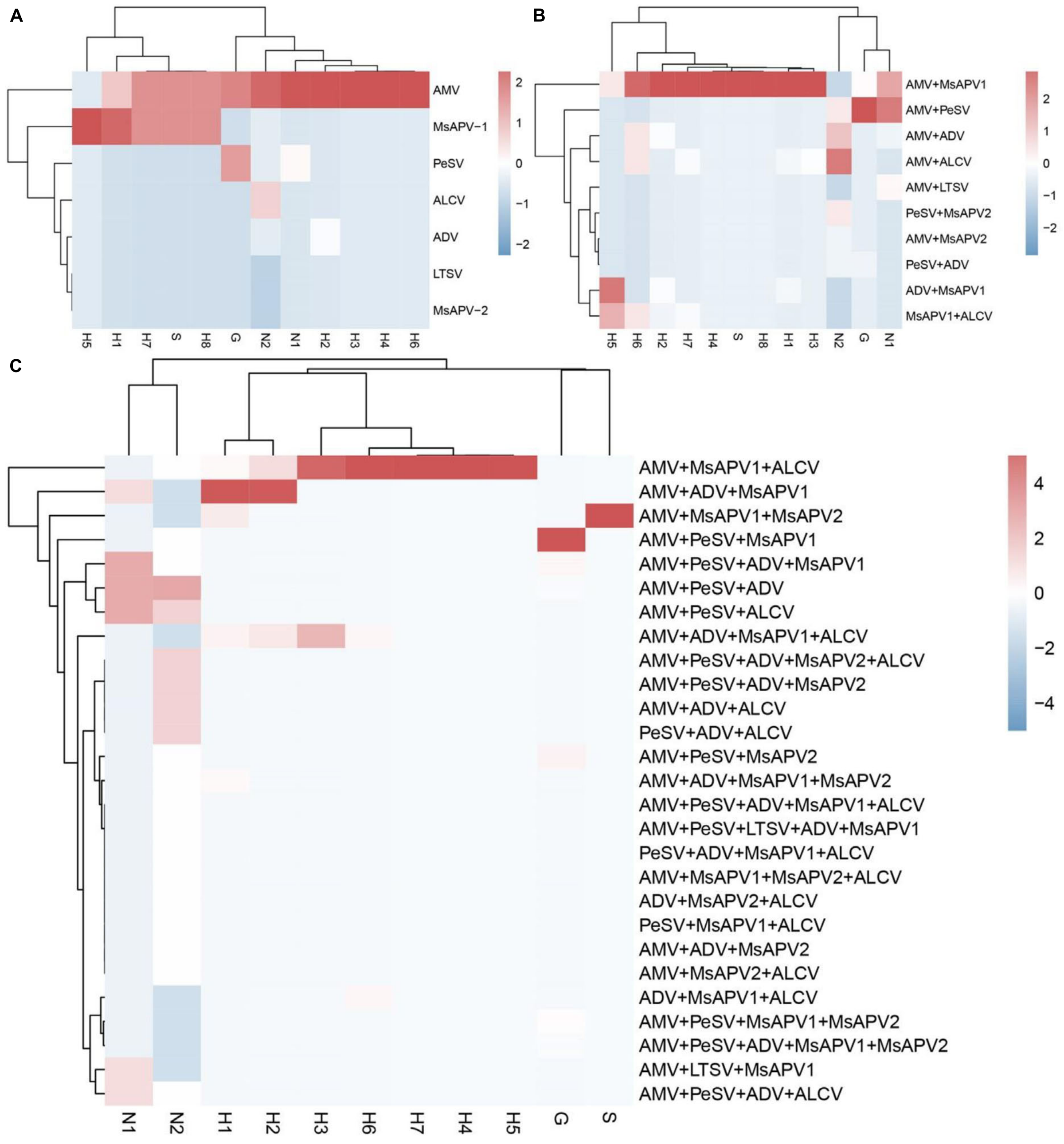


FIGURE 6 | The colored heat map of Incidence of single infection **(A)**, dual infection **(B)** and multiple infection **(C)** in different alfalfa-growing locations. Jiuquan (G), Helinger (N1), Tumote Left Banner (N2), Yangling (S), Zhengzhou (H1), Yuanyang-1 (H2), Yuanyang-2 (H3), Lankao (H4), Wenxian-1 (H5), Wenxian-2 (H6), Yichuan (H7) and Zhenping (H8). The red color represents a high detection rate of virus, while the blue color represents a low detection rate of virus.

74.23, 42.53, 95.98, 86.04, 95.72, 95.42, and 99.99%, respectively (**Figure 8B**). For AMV-RNA2, six recombination sites were detected at 202 (99.48%), 681 (95.47%), 1,004 (99.98%), 1,961 (92.51%), 2,225 (99.08%), and 2,474 (78.41%) positions of the AMV-RNA2 genome, respectively (**Figure 8D**). For AMV-RNA3, four recombination sites were located at 138 (99.70%), 631 (57.69%), 1,165 (64.67%), and 1,939 (98.76%) positions of the AMV-RNA3 genome, respectively (**Figure 8F**). For ALCV, GARD found 6 recombination sites, which were detected at 285 (89.69%), 625 (99.51%), 1,373 (99.99%), 1,832 (99.75%), 2,104 (44.81%), and 2,595 (81.17%) positions of the ALCV genomic sequence, respectively (**Figure 8H**).

Evolution Analysis of Alfalfa Viruses

The complete CP gene sequences of 4 alfalfa viruses (PeSV, LTSV, MsAPV1, and MsAPV2) and the complete N gene sequences of ADV identified in this study were deposited in GenBank (**Supplementary Tables 5–9**). For PeSV, three complete PeSV CP

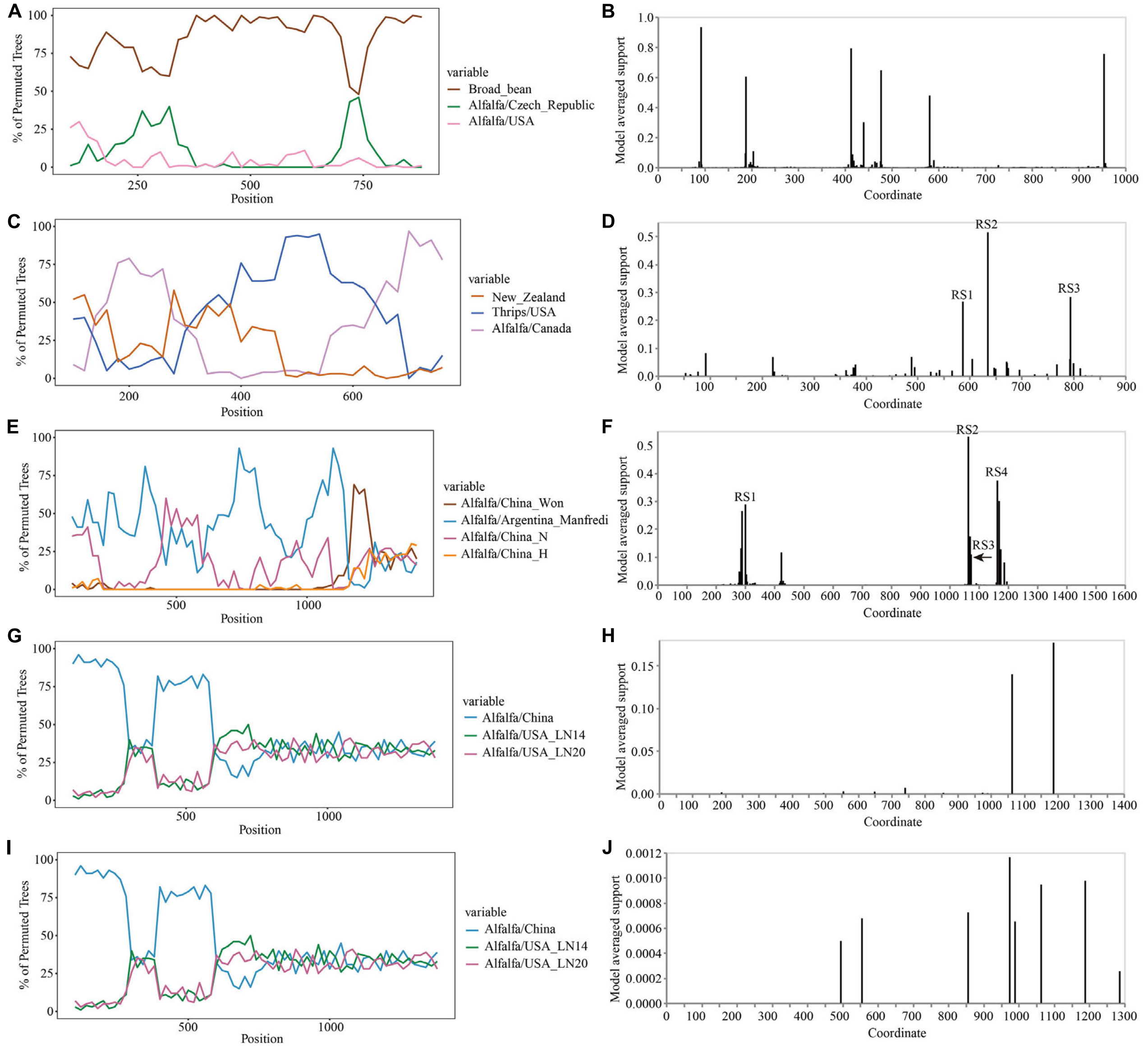


FIGURE 7 | Recombination analysis of CP gene (PeSV, LTSV, MsAPV1, and MsAPV2) and N gene (ADV). Bootscanning validation of recombinant isolates [**(A)** PeSV; **(C)** LTSV; **(E)** ADV; **(G)** MsAPV1; and **(I)** MsAPV2]. The nucleotide sequence of the isolate from different host and country were used as the reference sequence, and the nucleotide sequence of the isolate from 4 provinces in this study was as the test sequence. GARD detection of recombinant sites of alfalfa viruses [**(B)** PeSV; **(D)** LTSV; **(F)** ADV; **(H)** MsAPV1; and **(J)** MsAPV2]. RS represents the recombination site.

gene sequences and 1 complete poplar mosaic virus (PopMV, as outgroup sequence) CP gene sequence downloaded from GenBank, and 3 complete PeSV CP gene sequences in this study (**Supplementary Table 5**) were used to build a phylogenetic tree (**Figure 9A**). The phylogenetic tree of 7 complete CP gene sequences showed that PeSV isolates formed 2 groups (**Figure 9A**). The Jiuquan G isolate was clustered in group IA and was most closely related to isolate VRS541. Helinger N1 and Tumote Left Banner N2 isolates were placed in group IB, while the PopMV ATCC PV257 isolate was an outgroup (**Figure 9A**). There was significant variation between the PeSV isolates from China and those from other countries (**Figure 9B**), and the nucleotide identity between these isolates was from 79.8 to 100.0% (**Figure 9C**).

For LTSV, three complete LTSV CP gene sequences and 1 complete subterranean clover mottle virus (SCMoV, as outgroup sequence) CP gene sequence downloaded from GenBank, and 2 complete LTSV CP gene sequences in this study (**Supplementary Table 6**) were used to build a phylogenetic tree (**Figure 9D**). The phylogenetic tree of the 6 complete CP sequences revealed that the Canada, New Zealand, and United States isolates of LTSV were clustered in group I (**Figure 9D**). The isolates from Helinger and Tumote Left Banner of Inner Mongolia Autonomous Region were outliers, while the SCMoV MJ isolate was an outgroup

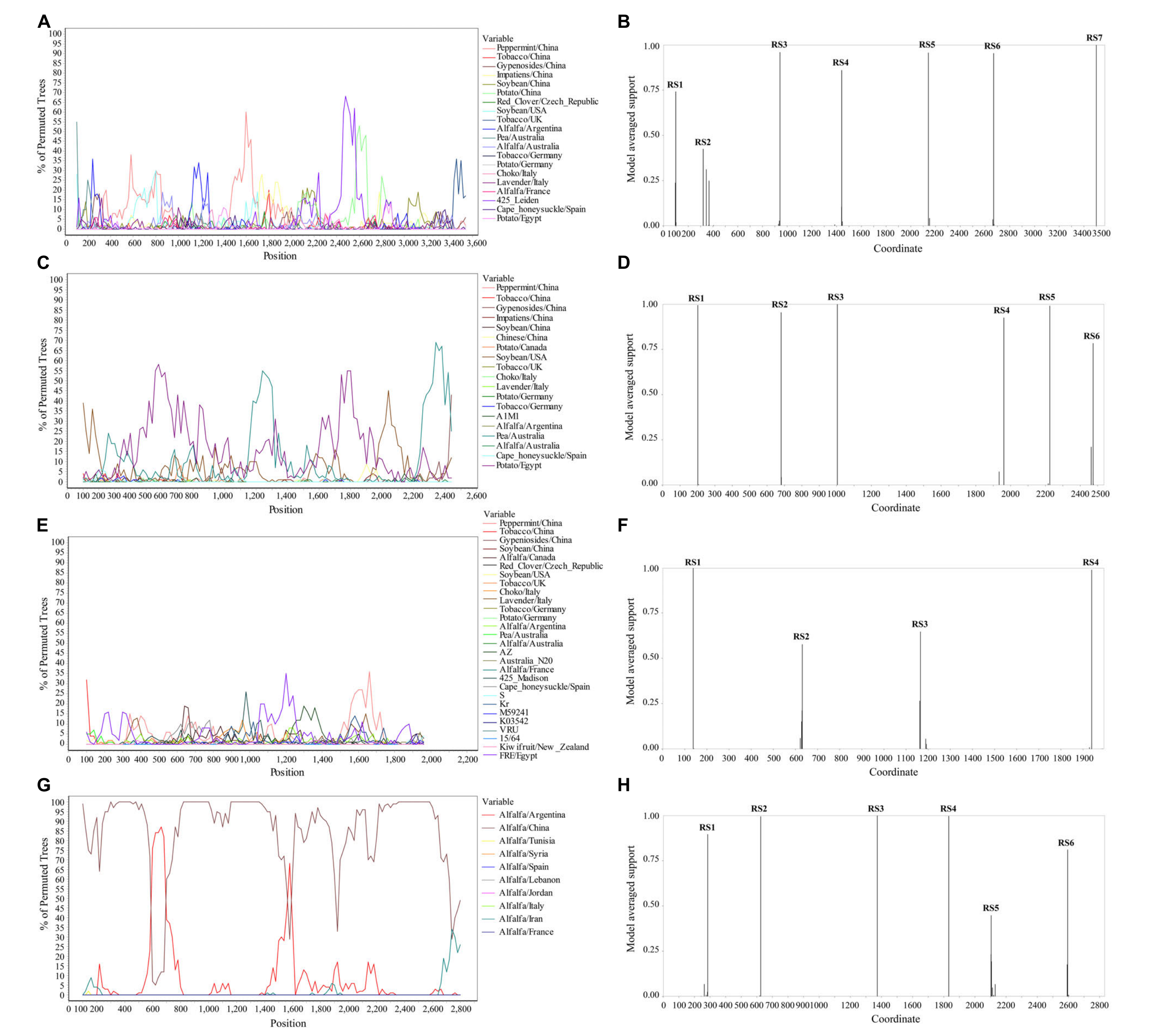


FIGURE 8 | Recombination analysis of whole genome sequence (AMV and ALCV). Bootscanning validation of recombinant isolates [**(A)** AMV-RNA1; **(C)** AMV-RNA2; **(E)** AMV-RNA3; and **(G)** ALCV]. The nucleotide sequence of the isolate from different host and country were used as the reference sequence, and the nucleotide sequence of the isolate from 4 provinces in this study was as the test sequence. GARD detection of recombinant sites of alfalfa viruses [**(B)** AMV-RNA1; **(D)** AMV-RNA2; **(F)** AMV-RNA3; and **(H)** ALCV]. RS represents the recombination site.

(**Figure 9D**). There was low variation between the LTSV isolates from China and those from other countries (**Figure 9E**), and the nucleotide identity between these isolates was higher than 96.5% (**Figure 9F**).

For ADV, two complete ADV N gene sequences and 1 complete persimmon virus A (PeVA, as outgroup sequence) N gene sequences downloaded from GenBank, and 8 complete ADV N gene sequences in this study (**Supplementary Table 7**) were used to build a phylogenetic tree (**Figure 9G**). The phylogenetic tree of the 11 complete N gene sequences showed that ADV isolates were divided into 3 groups (**Figure 9G**). Zhengzhou H1, Wenxian-1 H5, Yuanyang-1 H2, and Wenxian-2 H6 clustered together in group I and showed the closest relationship to the Won isolate identified in China. Isolates Helinger N1 and Tumote Left Banner N2 were placed in group II. Isolates Jiuquan G and Yuanyang-2 H3 were clustered in group III. The Manfredi Isolate from Argentina was an outlier, while PeVA was an outgroup (**Figure 9G**). All the ADV isolates from China showed a high degree of homogeneity (**Figure 9H**) with > 93.4% nucleotide identity between these isolates (**Figure 9I**), but a significant variation between China isolates and Argentina isolates (**Figure 9H**) and the nucleotide

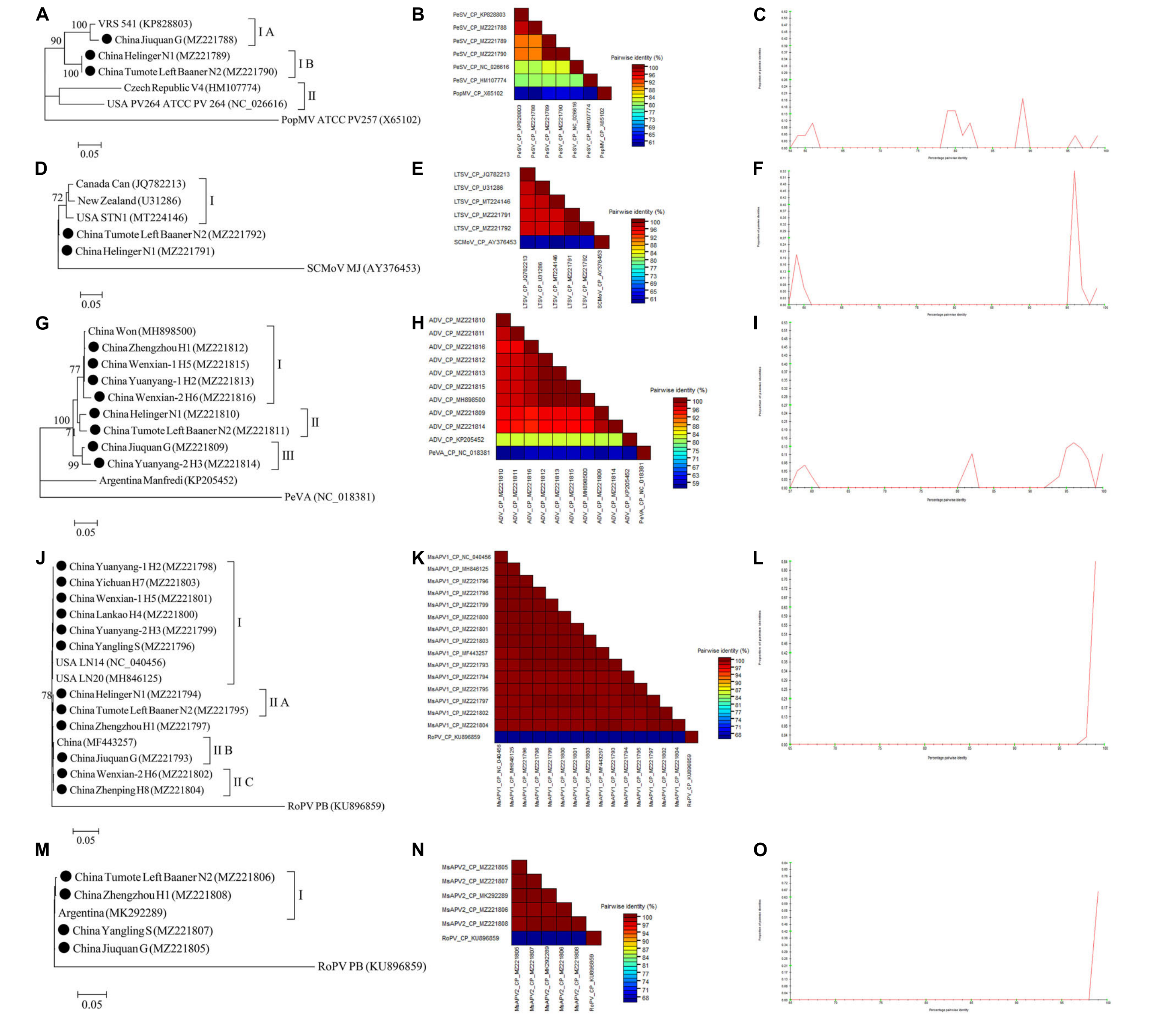


FIGURE 9 | Rooted Maximum Likelihood phylogenetic trees basing on complete CP gene nucleotide sequences [**(A)** PeSV; **(D)** LTSV; **(J)** MsAPV1, and **(M)** MsAPV2), and complete N gene nucleotide sequences [**(G)** ADV] respectively. The black dot marks the sequence obtained in this study. Jiuquan denote location in Gansu province, China. Helinger and Tumote Left Banner denote location in Inner Mongolia Autonomous Region, China. Yangling denote location in Shaanxi province, China. Zhengzhou, Yuanyang-1, Yuanyang-2, Lankao, Wenxian-1, Wenxian-2, Yichuan and Zhenping denote location in Henan province, China. Bootstrap values below 70% were not shown. The SDT interface **(B,C,E,F,H,I,K,L,N,O)**: Color-coded pairwise identity matrix generated from PeSV sequences **(B)**, LTSV sequences **(E)**, ADV sequences **(H)**, MsAPV1 sequences **(K)**, and MsAPV2 sequences **(N)**. Each colored cell represents a percentage identity score between two sequences (one indicated horizontally to the left and the other vertically at the bottom). Pairwise identity frequency distribution plot of PeSV **(C)**, LTSV **(F)**, ADV **(I)**, MsAPV1 **(L)**, and MsAPV2 **(O)**. The horizontal axis indicates percentage pairwise identities, and the vertical axis indicates proportions of these identities within the distribution.

identity of isolates from both countries was as low as 81.2–81.8% (**Figure 9I**).

For MsAPV1, three complete MsAPV1 CP gene sequences and 1 complete rose partitivirus (RoPV, as outgroup sequence) CP gene sequence downloaded from GenBank, and 12 complete MsAPV1 CP gene sequences in this study (**Supplementary Table 8**) were used to build a phylogenetic tree (**Figure 9J**). The phylogenetic tree of the 16 complete CP sequences showed that MsAPV1 isolates were divided into 2 groups (**Figure 9J**). The Yuanyang-1 H2, Yichuan H7, Wenxian-1 H5, Lankao H4, Yuanyang-2 H3 isolates from Henan, and Yangling S from Shaanxi, and LN20, and LN14 from United States were clustered in group I. The Helinger N1 and Tumote Left Banner N2 isolate from Inner Mongolia were clustered in group IIA. The

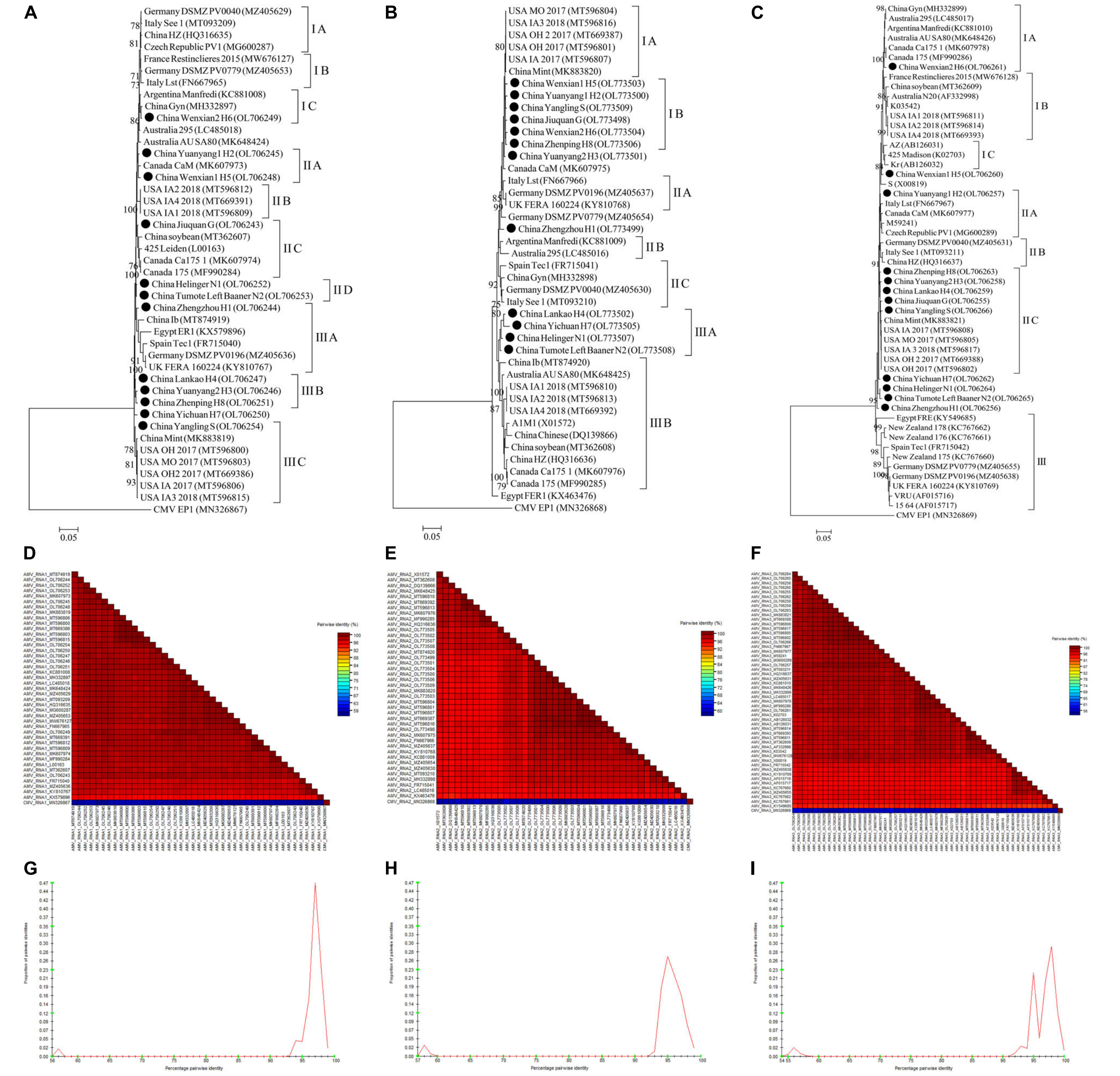


FIGURE 10 | Rooted Maximum Likelihood phylogenetic trees basing on complete AMV-RNA nucleotide sequences of AMV-RNA1 **(A)**, AMV-RNA2 **(B)**, and AMV-RNA3 **(C)** respectively. The black dot marks the sequence obtained in this study. Jiuquan denote location in Gansu province, China. Helinger and Tumote Left Banner denote location in Inner Mongolia Autonomous Region, China. Yangling denote location in Shaanxi province, China. Zhengzhou, Yuanyang-1, Yuanyang-2, Lankao, Wenxian-1, Wenxian-2, Yichuan and Zhenping denote location in Henan province, China. Bootstrap values below 70% were not shown. The SDT interface **(D–F,G–I)**: Color-coded pairwise identity matrix generated from AMV-RNA1 sequences **(D)**, AMV-RNA2 sequences **(E)** and AMV-RNA3 sequences **(F)**. Each colored cell represents a percentage identity score between two sequences (one indicated horizontally to the left and the other vertically at the bottom). Pairwise identity frequency distribution plot of AMV-RNA1 **(G)**, AMV-RNA2 **(H)** and AMV-RNA3 **(I)**. The horizontal axis indicates percentage pairwise identities, and the vertical axis indicates proportions of these identities within the distribution.

isolate Jiuquan G, which was most closely related to an isolate from China, was placed in group IIB. The Wenxian-2 H6 and Zhenping H8 isolates were placed in group IIC, while the PoPV PB isolate was an outgroup (**Figure 9J**). There was a highly conserved between the MsAPV1 isolates in this study (**Figure 9K**) with > 99.4% nucleotide identity between these isolates (**Figure 9L**).

For MsAPV2, one complete MsAPV2 CP gene sequence and 1 complete RoPV CP gene sequence (as outgroup sequence) downloaded from GenBank, and 4 complete MsAPV2 CP gene sequences in this study (**Supplementary Table 9**) were used to build a phylogenetic tree (**Figure 9M**). The phylogenetic tree of the 6 complete CP sequences showed that the Tumote Left Banner N2 and Zhengzhou H1 isolates, which were most closely related to the isolate from Argentina, clustered together in group I (**Figure 9M**). The isolates Yangling S and Jiuquan G were outliers, while the PoPV PB isolate was an outgroup (**Figure 9M**). All the MsAPV2 isolates in this study showed a high degree of homogeneity (**Figure 9N**) with > 99.5% nucleotide identity between these isolates (**Figure 9O**).

For AMV, almost whole genomic sequence of AMV identified in this study were deposited in GenBank (**Supplementary Table 4**). For AMV-RNA1, thirty-one AMV-RNA1 sequences and 1 complete cucumber mosaic virus (CMV, as outgroup sequence) RNA1 sequence downloaded from GenBank, and 12 AMV-RNA1 sequences in this study (**Supplementary Table 4**) were used to build a phylogenetic tree (**Figure 10A**). The phylogenetic tree of 44 RNA1 sequences showed that the isolates were divided into 3 groups (**Figure 10A**). Isolates Wenxian-2 H6 was placed in group IC and was most closely related to the isolate Gyn from China. The isolates Yuanyang-1 H2 and Wenxian-1 H5 clustered together in group IIA and were most closely related to the isolate CaM from Canada. Isolates Jiuquan G was placed in group IIC and was most closely related to the isolate from China. The isolates Helinger N1 and Tumote Left Banner N2 were clustered in group IID. Isolate Zhengzhou H1 was placed in group IIIA and was most closely related to isolate Ib from China. The isolates Lankao H4, Yuanyang-2 H3, and Zhenping H8 clustered in group IIIB. Yangling S isolate was placed in group IIIC and was most closely related to isolate Mint from China, while the CMV EP15 isolate was an outgroup (**Figure 10A**). All the AMV isolates in this study showed a high degree of homogeneity (**Figure 10D**) with > 94.3% nucleotide identity between these isolates (**Figure 10G**).

For AMV-RNA2, twenty-nine AMV-RNA2 sequences and 1 complete cucumber mosaic virus (CMV, as outgroup sequence) RNA2 sequence downloaded from GenBank, and 12 AMV-RNA2 sequences in this study (**Supplementary Table 4**) were used to build a phylogenetic tree (**Figure 10B**). The phylogenetic tree of 44 RNA2 sequences showed that the isolates were divided into 3 groups (**Figure 10B**). Isolates Wenxian-1 H5, Yuanyang-1 H2, Yangling S, Jiuquan G, Wenxian-2 H6, and Zhenping H8 were placed in group IB and were most closely related to the isolate Mint from China. Lankao H4, Yichuan H7, Helinger N1, and Tumote Left Banner N2 isolates clustered together in group IIIA and were most closely related to the isolate Ib from China. Isolates FER1 from Egypt was an outlier, while the CMV EP1 isolate was an outgroup (**Figure 10B**). There was a highly conserved between the AMV isolates from China (**Figure 10E**) with > 96.15% nucleotide identity between these isolates (**Figure 10H**) and a low variation between the AMV isolates from China and those from other countries (**Figure 10E**), and the nucleotide identity between these isolates was higher than 93.6% (**Figure 10H**).

For AMV-RNA3, forty AMV-RNA3 sequences and 1 complete cucumber mosaic virus (CMV, as outgroup sequence) RNA3 sequence downloaded from GenBank, and 12 AMV-RNA3 sequences in this study (**Supplementary Table 4**) were used to build a phylogenetic tree (**Figure 10C**). The phylogenetic tree of 53 RNA3 sequences showed that the isolates were divided into 3 groups (**Figure 10C**). Isolates Wenxian-2 H6 was placed in group IA and was most closely related to the isolate from Canada. Yuanyang-1 H2 isolate was clustered in group IIA and was most closely related to the isolate Lst from Italy. Zhenping H8, Yuanyang-2 H3, Lankao H4, Jiuquan G, and Yangling S clustered together in group IIC and were most closely related to the isolate Mint from China, while the CMV EP1 isolate was an outgroup (**Figure 10C**). The AMV isolates in this study showed a high degree of homogeneity (**Figure 10F**) with > 97.3% nucleotide identity between these isolates (**Figure 10I**). All the AMV isolates showed a high degree of homogeneity (**Figure 10F**) which nucleotide identity was higher than 91.6% (**Figure 10I**).

For ALCV, nineteen ALCV whole genome sequences and 1 Euphorbia caput-medusae latent virus (EcmLV, as outgroup sequence) downloaded from GenBank, and 9 ALCV whole genome sequences in this study (**Supplementary Table 10**) were used to build phylogenetic trees (**Figure 11A**). The phylogenetic trees of 29 complete genomic sequences revealed that the isolates formed 2 groups (**Figure 11A**). All isolates from China and Argentina were placed in one group and were closely related. All of the isolates from Henan province in this study were clustered in 1 group and showed the closest relationship to the isolate SLSC410-1 isolate detected in alfalfa from Henan. The EcmLV A14 isolate was an outgroup (**Figure 11A**). All the ALCV isolates from China showed highly conserved (**Figure 11B**) with > 97.2% nucleotide identity between these isolates (**Figure 11C**), but a significant variation between China isolates and abroad isolates (**Figure 11B**) and the nucleotide identity of isolates between China and other countries was from 80.3 to 97.6% (**Figure 11C**).

DISCUSSION

This is the first large-scale survey of alfalfa viruses in 4 provinces of China (Gansu, Henan, Inner Mongolia, and Shaanxi). Seven viruses—namely, AMV, PeSV, LTSV, ADV, MsAPV1, MsAPV2, and ALCV—were detected in alfalfa plant samples as individual virus infections or as dual or multiple infections. Samples with virus-induced symptoms were frequently found to be infected with multiple viruses (**Figure 5C**). The main symptoms of samples infected with AMV only were etiolation and macular mosaicism (**Supplementary Table 2**), while those of samples infected with MsAPV1 or ALCV only were shrinkage and mosaic shrinkage (**Supplementary Table 2**). These symptoms may serve as biological indicators for diagnosing alfalfa virus infection, although additional studies are needed to establish the typical symptoms triggered by a single virus.

Eleven alfalfa viruses have been detected in specific alfalfa-growing provinces of China including ADV and BLRV in Xinjiang (Li et al., 2019; Xuehelati et al., 2020); AMV in Beijing, Gansu, Henan, and Xinjiang (Wen and Nan, 2015; Zhang, 2016; Guo et al., 2019; Li et al., 2021); ALCV in Henan (Guo et al., 2020); ToMV, WCMV, CPMV, and BYMV in Gansu (Wen and Nan, 2015; Zhou et al., 2016); and MsAPV1 in Beijing (Kim et al., 2018; Li et al., 2021). We report here for the first time

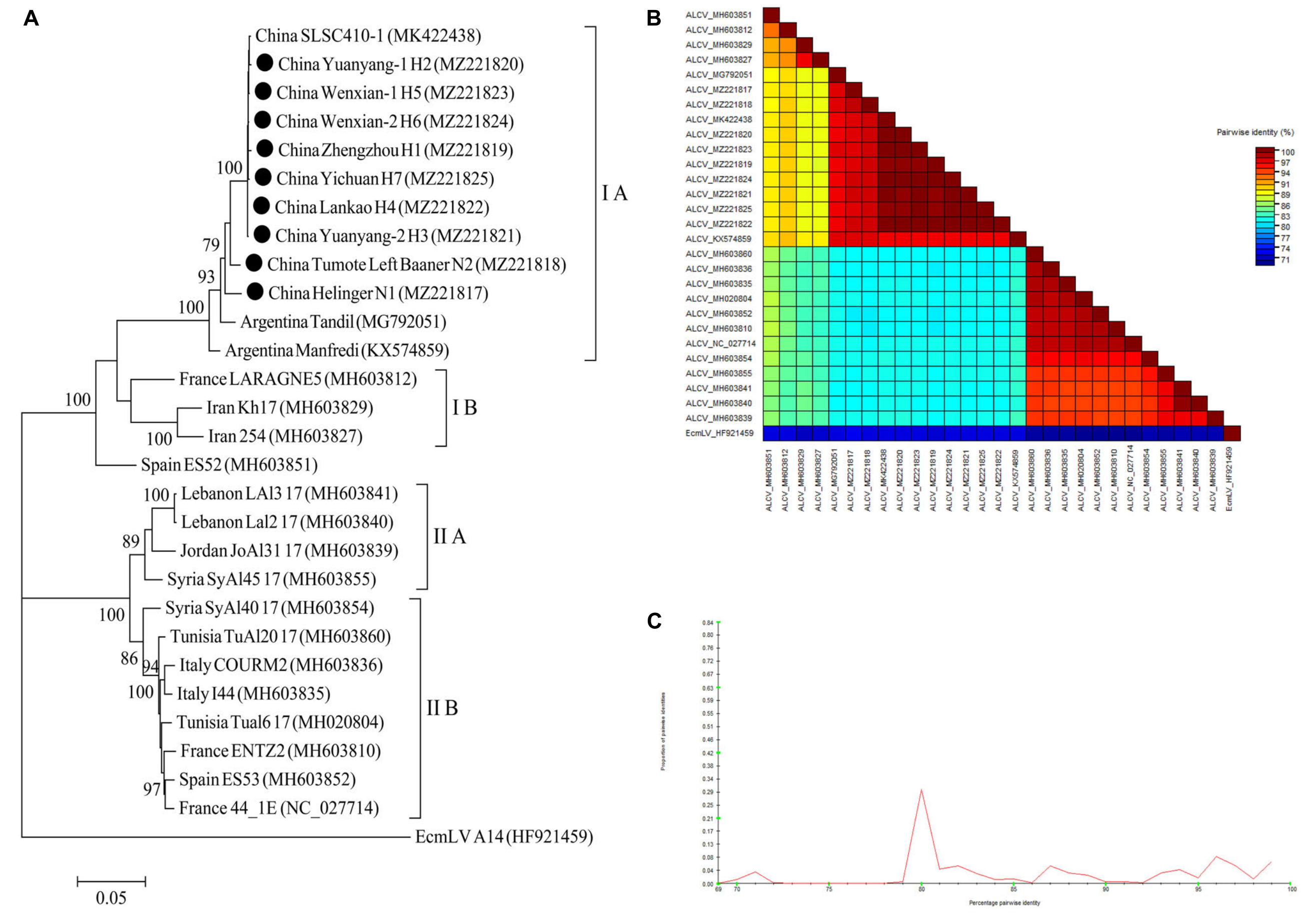


FIGURE 11 | Rooted Maximum Likelihood phylogenetic trees basing on nucleotide sequences of ALCV whole genome sequence **(A)**. The black dot marks the sequence obtained in this study. Helinger and Tumote Left Banner denote location in Inner Mongolia Autonomous Region, China. Zhengzhou, Yuanyang-1, Yuanyang-2, Lankao, Wenxian-1, Wenxian-2 and Yichuan denote location in Henan province, China. Bootstrap values below 70% were not shown. The SDT interface **(B,C)**: Color-coded pairwise identity matrix generated from ALCV whole genome sequences **(B)**. Each colored cell represents a percentage identity score between two sequences (one indicated horizontally to the left and the other vertically at the bottom). Pairwise identity frequency distribution plot of ALCV whole genome sequences **(C)**. The horizontal axis indicates percentage pairwise identities, and the vertical axis indicates proportions of these identities within the distribution.

the infection of alfalfa with the following viruses in the surveyed provinces: PeSV in Gansu and Inner Mongolia; LTSV in Inner Mongolia; ADV in Gansu, Henan, and Inner Mongolia; MsAPV1 and MsAPV2 in Gansu, Henan, Inner Mongolia, and Shaanxi; and ALCV in Inner Mongolia. Importantly, the incidence of MsAPV1 (65.36%) was almost as high as that of AMV (79.96%) in China (**Figure 5A**). AMV and MsAPV1 were detected at similar rates in Yichuan and Zhenping (Henan; **Figure 5A**). On the other hand, the incidence of MsAPV1 was higher than that of AMV in Zhengzhou and Wenxian (Henan; **Figure 5A**). AMV and MsAPV1 are the main RNA viruses infecting alfalfa in Beijing (Li et al., 2021); our results indicate that MsAPV1 is also the predominant virus in Henan.

The rate of detection of AMV in the 12 locations surveyed in this study ranged from 18.18% in Wenxian-1 to 100% in Yuanyang-2 and Lankao regions. This is in accordance with rates reported in other countries; for example, in a survey carried out in the Saudi Arabia, AMV was detected in alfalfa at rates ranging from 41.9 to 82.5% (Abdalla et al., 2020). In this study, the incidences of AMV in alfalfa grown for more than 2 years in China is > 90% (**Figure 5B**). AMV—the most widespread virus species infecting alfalfa globally—can infect up to 80% of plants in an alfalfa stand more than 2 years old and 100% of plants in 3-year-old stands (Samac et al., 2015); it can also reduce the alfalfa yield by 9%–82%, plant height by 7%–57% (Guo et al., 2019), and crude protein content by 42.70% (Han et al., 2019). AMV has a wide host range that includes more than 700 species in over 70 families. Alfalfa is an important perennial host of AMV and a reservoir for AMV strains, which can be transmitted by aphids to other host crops such as pea, chickpea, and tomato (Samac et al., 2015). Our data suggest that AMV is a significant threat to alfalfa production in China.

High variability is one of the typical characteristics of RNA viruses, mainly due to the lack of correction function of RNA-dependent RNA polymerase (RdRp) or replicase. RNA viruses have a very high mutation rate (about 10^{-4} nucleotides per replication cycle), which is also an evolutionary strategy of RNA viruses (Malpica et al., 2002). The ability to transfer the hereditary information encased inside capsids—the protective proteinaceous shells that include the centers of infection particles (virions)—is unique to bona fide viruses and recognizes them from other sorts of egotistical hereditary elements such as transposons and plasmids (Krupovič and Bamford, 2009). Recombination can have a significant effect on the evolutionary process and is of intriguing in its own right (Pond et al., 2006a). GARD has not required a non-recombinant reference arrangement and recombination between sequences is also accommodated, which can be run in parallel on a cluster of computers, and so is well suited to screen for recombination in big datasets (Pond et al., 2006a). In this study, recombination signals of LTSV, ADV, AMV, and ALCV isolates from 4 provinces were detected by using Simplot (**Figures 7C,E**, **8A,C,E,G**), and GARD found the recombination sites of LTSV located at 635 sites of CP gene with an average model approval rate of 51.17% (**Figure 7D**). The recombination sites of ADV, which were located at 1,064 (46.51%) position of the N gene, respectively (**Figure 7F**). Most the recombination sites of AMV and ALCV were higher than 80.00 (**Figures 8B,D,F,H**). These results indicated that the confidence of the recombinant site is high, and the variation of these viruses is mainly caused by base site mutation and gene recombination. And these two factors play an important role in the evolution of these viruses and are the main factors for the formation of new strains of alfalfa viruses. Analyzing the whole genome sequence of the virus can get more mutation sites and sufficient time, which can help us understand the evolutionary relationship of the virus more comprehensively.

Taxonomic classification approaches which are based on pairwise genomic identity measures are potentially highly automatable and are progressively popular with the International Committee on Taxonomy of Viruses (ICTV). SDT, a virus classification tool based on pairwise sequence alignment and identity calculation, can produce publication-quality color-coded distance matrices and pairwise identity plots to further assist the classification of sequences according to the taxonomic demarcation criteria approved by ICTV (Muhire et al., 2014). LTSV and AMV isolate from alfalfa plants in China showed a high degree of homogeneity (**Figures 9E**, **10D–F**). AMV isolates from Gansu and Shaanxi were closely related, and both were distantly related to isolates from Inner Mongolia (**Figures 10A–C**). ALCV sequences from the 4 surveyed provinces were highly conserved (**Figures 11B,C**) and most closely related to the isolates from Argentina (**Figure 11A**), but a significant variation to the isolates from other countries (**Figures 11B,C**). The results suggested that ALCV isolates in China originated from a single ALCV isolate similar to what has been reported for ALCV in Argentina (Davoodi et al., 2018). These authors also suggested that the virus most likely originated in Iran (Davoodi et al., 2018). For PeSV and ADV, there was significant variation between the isolates from China and those from other countries (**Figures 9B,H**). Chinese isolates MsAPV1 and MsAPV2 showed minor variations in the CP gene sequences (**Figures 9K,N**). MsAPV1 was first identified through an analysis of a public transcriptome dataset (Kim et al., 2018). The complete genome sequences of MsAPV1 and MsAPV2 were obtained based on that of MsAPV (Bejerman et al., 2019). In our study, we readily detected the RNA-dependent RNA polymerase (RdRp) gene of MsAPV1 and MsAPV2 and confirmed that these are different virus species (Bejerman et al., 2019).

High-throughput sequencing is widely used for the detection of plant viruses as it allows a comprehensive, large-scale, and unbiased analysis of the genome (Villamor et al., 2019; Bejerman et al., 2020). In the present study, most of the fragments in the high-quality assembly were mapped to the viral nucleic acid sequences of 7 viruses with high homology, and the viruses were verified by PCR/RT-PCR, cloning, and sequencing. Notably, we did not detect any unknown viruses.

CONCLUSION

This study identified the main virus species infecting alfalfa in Gansu, Henan, Inner Mongolia, and Shaanxi provinces of China and analyzed their incidence, distribution, and genetic diversity. To our knowledge, as the virus host is alfalfa, this is the first report of PeSV and LTSV in China; ADV in Gansu and Inner Mongolia; ALCV in Inner Mongolia; and MsAPV1 and MsAPV2 in all 4 surveyed provinces. The incidence of MsAPV1 was high and close to that of AMV in China. SDT analysis of the 7 viruses isolated in China revealed a highly conserved among AMV, LTSV, ADV, MsAPV1, MsAPV2, and ALCV, but the sequence was a high variation between China isolates to abroad isolates in PeSV, ADV, and ALCV. These results provide a basis for the studies on the genetic evolution of alfalfa viruses, particularly the 7 species identified in this work, and can guide the development of strategies for preventing diseases in alfalfa caused by these viruses.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI SRA BioProject, accession no: PRJNA761453.

AUTHOR CONTRIBUTIONS

ZG and QW conceived the ideas designed the methodology. ZG and JN collected alfalfa samples. ZG, TZ, JN, XC, YM, MH, HK, NX, XS, SG, and MR conducted the experiments. ZG, TZ, and ZC analyzed the data and wrote the manuscript. ZC, MH, JC, and QW edited the language of the manuscript. QW supervised the project and provided the constructive suggestions for revisions. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Key R&D Program of China (Grant No. 2017YFE0111000).

ACKNOWLEDGMENTS

We would like to thank Quan Zhang (Jiuquan Daye Seed Industry Co., Ltd., Jiuquan, China); Fu-Gui Mi (College of Grassland, Resources and Environment, Inner Mongolia Agricultural University, Hohhot, China); Zhao-Lan Wang (Institute of Grassland research of Chinese Academy of Agricultural Sciences, Hohhot, China); Cheng-Zhang Wang and Yu-Xia Guo (College of Animal Science and Technology, Henan Agricultural University, Zhengzhou, China); and Chang-Song Feng and Hong-Qi Du (Institute of Animal Husbandry and Veterinary Science, Henan Academy of Agricultural Sciences, Zhengzhou, China) for their assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.771361/full#supplementary-material

REFERENCES

Abdalla, O. A., Al-Shahwan, I. M., Al-Saleh, M. A., and Amer, M. A. (2020). Molecular characterization of alfalfa mosaic virus (AMV) isolates in alfalfa and other plant species in different regions in Saudi Arabia. *Eur. J. Plant Pathol.* 156, 603–613. doi: 10.1007/s10658-019-01910-z

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bejerman, N., Debat, H., Nome, C., Cabrera-Mederos, D., Trucco, V., de Breuil, S., et al. (2019). Redefining the Medicago sativa alphapartitiviruses genome sequences. *Virus Res.* 265, 156–161. doi: 10.1016/j.virusres.2019.03.021

Bejerman, N., Giolitti, F., Breuil, S. D., Trucco, V., Nome, C., Lenardon, S., et al. (2015). Complete genome sequence and integrated protein localization and interaction map for alfalfa dwarf virus, which combines properties of both cytoplasmic and nuclear plant *Rhabdoviruses*. *Virology* 483, 275–283. doi: 10.1016/j.virol.2015.05.001

Bejerman, N., Giolitti, F., Trucco, V., de Breuil, S., Dietzgen, R. G., and Lenardon, S. (2016). Complete genome sequence of a new enamovirus from Argentina infecting alfalfa plants showing dwarfism symptoms. *Arch. Virol.* 161, 2029–2032. doi: 10.1007/s00705-016-2854-3

Bejerman, N., Roumagnac, P., and Nenchinov, L. G. (2020). High-throughput sequencing for deciphering the virome of alfalfa (*Medicago sativa L.*). *Front. Microbiol.* 11:553109. doi: 10.3389/fmicb.2020.553109

Chan, P. P., and Lowe, T. M. (2009). Gtrnadb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 37, D93–D97. doi: 10.1093/nar/gkn787

Davoodi, Z., Bejerman, N., Richet, C., Filloux, D., Kumari, S. G., Chatzivassiliou, E. K., et al. (2018). The westward journey of alfalfa leaf curl virus. *Viruses* 10:542. doi: 10.3390/v10100542

Gaafar, Y. Z. A., Richert-Pöggeler, K. R., Maaß, C., Vetten, H. J., and Ziebell, H. (2019). Characterization of a novel nucleorhabdovirus infecting alfalfa (Medicago sativa). *Virol. J.* 16:55. doi: 10.1186/s12985-019-1147-3

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Res.* 31, 439–441. doi: 10.1093/nar/gkg006

Guo, Z. P., Feng, C. S., Zhang, J. X., Wang, M. L., Qu, G., Liu, J. Y., et al. (2019). Field resistance to alfalfa mosaic virus among 30 alfalfa varieties. *Acta Pratacul. Sin.* 28, 157–167. doi: 10.11686/cyxb2018259

Guo, Z. P., Zhang, J. X., Wang, M. L., Guan, Y. Z., Qu, G., Liu, J. Y., et al. (2020). First report of alfalfa leaf curl virus infecting alfalfa (*Medicago sativa* L.) in China. *Plant Dis.* 104:1001. doi: 10.1094/PDIS-02-19-0318-PDN

Han, Y. Z., Hu, H. Q., Yu, Y. X., Zhang, C. P., and Fan, Z. W. (2019). Effects of alfalfa mosaic disease on photosynthetic performance, growth, and forage quality of Medicago sativa. *Pratacult. Sci.* 36, 2061–2068. doi: 10.11829/j.issn.1001-0629.2018-0705

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kim, H., Park, D., and Hahn, Y. (2018). Identification a novel RNA viruses in alfalfa (*Medicago sativa*): an Alphapartitivirus, Deltapartitivirus, and a Marafivirus. *Gene* 638, 7–12. doi: 10.1016/j.gene.2017.09.069

Krupovič, M., and Bamford, D. H. (2009). Does the evolution of viral polymerases reflect the origin and evolution of viruses? *Nat. Rev. Microbiol.* 7, 250–250. doi: 10.1038/nrmicro2030-c1

Kumar, S., Stecher, G., and Tamura, K. (2016). Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25

Li, J., Gu, H., Liu, Y., Wei, S., Hu, G., Wang, X., et al. (2021). RNA-seq reveals plant virus composition and diversity in alfalfa, thrips, and aphids in Beijing, China. *Arch. Virol.* 166, 1711–1722. doi: 10.1007/s00705-021-05067-1

Li, K. M., Muhanmaiti, A., Ge, R. Y., Liu, X. X., Li, B. Y., and Xuehelati, R. (2019). Identification of a new virus on alfalfa in Xinjiang. *Pratacult. Sci.* 36, 2319–2324. doi: 10.11829/j.issn.1001-0629.2019-0120

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160. doi: 10.1128/JVI.73.1.152-160.1999

Lu, C., Meyers, B. C., and Green, P. J. (2007). Construction of small RNA cDNA libraries for deep sequencing. *Methods* 43, 110–117. doi: 10.1016/j.ymeth.2007.05.002

Malpica, J. M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W., and García-Arenal, F. (2002). The rate and character of spontaneous mutation in an RNA virus. *Genetics* 162, 1505–1511. doi: 10.1093/genetics/162.4.1505

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

Muhire, B. M., Varsani, A., and Martin, D. P. (2014). SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One* 9:e108277. doi: 10.1371/journal.pone.0108277

Nemchinov, L. G., Grinstead, S. C., and Mollov, D. S. (2017). Alfalfa virus S, a new species in the family *Alphaflexiviridae*. *PLoS One* 12:e0178222. doi: 10.1371/journal.pone.0178222

Nemchinov, L. G., Lee, M. N., and Shao, J. (2018b). First report of alphapartitiviruses infecting alfalfa (*Medicago sativa* L.) in the United States. *Microbiol. Resour. Announc.* 7, e1266–e1218. doi: 10.1128/MRA.01266-18

Nemchinov, L. G., Francois, S., Roumagnac, P., Ogliastro, M., Hammond, R. W., Mollov, D. S., et al. (2018a). Characterization of alfalfa virus F, a new member of the genus *Marafivirus*. *PLoS One* 13:e0203477. doi: 10.1371/journal.pone.0203477

Pond, S. L. K., Posadab, D., Gravenorc, M. B., Woelka, C. H., and Frosta, S. D. (2006a). GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22, 3096–3098. doi: 10.1093/bioinformatics/btl474

Pond, S. L. K., Posadab, D., Gravenorc, M. B., Woelka, C. H., and Frosta, S. D. (2006b). Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* 23, 1891–1901. doi: 10.1093/molbev/msl051

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196. doi: 10.1093/nar/gkm864

Raza, A., Al-Shahwan, I. M., Abdalla, O. A., Al-Saleh, M. A., and Amer, M. A. (2017). Lucerne transient streak virus; a recently detected virus infecting alfalfa (*Medicago sativa*) in central Saudi Arabia. *Plant Pathol. J.* 33, 43–52. doi: 10.5423/PPJ.OA.06.2016.0143

Samac, D. A., Rhodes, L. H., and Lamp, W. O. (2015). *Compendium of Alfalfa Diseases and Pests*, 3nd Edn. St. Paul, MN: APS Press.

Samarfard, S., McTaggart, A. R., Sharman, M., Bejerman, N. E., and Dietzgen, R. G. (2020). Viromes of ten alfalfa plants in Australia reveal diverse known viruses and a novel RNA virus. *Pathogens* 9:214. doi: 10.3390/pathogens9030214

Villamor, D. E. V., Ho, T., Al-Rwahnih, M., Martin, R. R., and Tzanetakis, I. E. (2019). High throughput sequencing for plant virus detection and discovery. *Phytopathology* 109, 716–725. doi: 10.1094/PHYTO-07-18-0257-RVW

Wen, Z. H., and Nan, Z. B. (2015). Detection of pathogenic organisms in *Medicago sativa* in Zhangye, Gansu province. *Acta Pratacult. Sin.* 24, 121–126. doi: 10.11686/cyxb20150414

Xuehelati, R., Fan, J. X., Wang, L. L., Ge, R. Y., and Li, K. M. (2020). Identification and detection of alfalfa dwarf virus (ADV) isolates in China. *J. Xinjiang Agric. Univ.* 43, 177–181.

Zhang, X. W. (2016). *Molecular Identification and Detection of Lucerne Witches' Broom and Mosaic Disease in Xinjiang.* Master Thesis. Urumqi: Xinjiang Agricultural University.

Zhou, Q. Y., Liang, Q. L., and Han, L. (2016). Symptoms and pathogen detection of alfalfa virus disease. *Pratacult. Sci.* 33, 1297–1305. doi: 10.11829/j.issn.1001-0629.2015-0652

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

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

frontiers in Microbiology

ORIGINAL RESEARCH
published: 18 February 2022
doi: 10.3389/fmicb.2022.780651

RNA Virus Diversity in Birds and Small Mammals From Qinghai–Tibet Plateau of China

*Wentao Zhu[1], Jing Yang[1,2,3], Shan Lu[1,2,3], Dong Jin[1,2,3], Ji Pu[1], Shusheng Wu[4], Xue-Lian Luo[1], Liyun Liu[1], Zhenjun Li[1] and Jianguo Xu[1,2,3,5]**

[1] State Key Laboratory of Infectious Disease Prevention and Control, Chinese Center for Disease Control and Prevention, National Institute for Communicable Disease Control and Prevention, Beijing, China, [2] Shanghai Public Health Clinical Center, Fudan University, Shanghai, China, [3] Research Units of Discovery of Unknown Bacteria and Function, Chinese Academy of Medical Sciences, Beijing, China, [4] Yushu Prefecture Center for Disease Control and Prevention, Yushu, China, [5] Research Institute of Public Heath, Nankai University, Tianjin, China

Most emerging and re-emerging viruses causing infectious diseases in humans and domestic animals have originated from wildlife. However, current knowledge of the spectrum of RNA viruses in the Qinghai-Tibet Plateau in China is still limited. Here, we performed metatranscriptomic sequencing on fecal samples from 56 birds and 91 small mammals in Tibet and Qinghai Provinces, China, to delineate their viromes and focused on vertebrate RNA viruses. A total of 184 nearly complete genome RNA viruses belonging to 28 families were identified. Among these, 173 new viruses shared <90% amino acid identity with previously known viral sequences. Several of these viruses, such as those belonging to genera *Orthonairovirus* and *Hepatovirus*, could be zoonotic viruses. In addition, host taxonomy and geographical location of these viruses showed new hosts and distribution of several previously discovered viruses. Moreover, 12 invertebrate RNA viruses were identified with <40% amino acid identity to known viruses, indicating that they belong to potentially new taxa. The detection and characterization of RNA viruses from wildlife will broaden our knowledge of virus biodiversity and possible viral diseases in the Qinghai–Tibet Plateau.

Keywords: virome, bird, rodent, small mammals, Qinghai–Tibet Plateau, diversity, fecal sample

OPEN ACCESS

Edited by:
*Tao Jin,
Guangdong Magigene Biotechnology Co., Ltd., China*

Reviewed by:
*John-Sebastian Eden,
The University of Sydney, Australia
Laurent Dacheux,
Institut Pasteur, France*

****Correspondence:***
*Jianguo Xu
xujianguo@icdc.cn*

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *21 September 2021*
Accepted: *12 January 2022*
Published: *18 February 2022*

Citation:
Zhu W, Yang J, Lu S, Jin D, Pu J, Wu S, Luo X-L, Liu L, Li Z and Xu J (2022) RNA Virus Diversity in Birds and Small Mammals From Qinghai–Tibet Plateau of China. Front. Microbiol. 13:780651. doi: 10.3389/fmicb.2022.780651

INTRODUCTION

The majority of emerging and re-emerging viral infectious diseases in humans have originated from wildlife, including rodents and birds, and are increasing with time (Jones et al., 2008; Wu et al., 2017; He et al., 2021). The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has reminded us of the pathogenetic potential of viruses and their capacity to cause deadly outbreaks on a global scale. RNA viruses are likely to be present in all cellular life (Koonin et al., 2006) and constitute the vast majority of the global virome (Mu et al., 2017; Wolf et al., 2018; Kondo et al., 2020). Most studies primarily focus on identifying RNA viruses that are pathogenic to humans and animals. However, emerging viruses appear to be well adapted to thrive in their reservoir host with little or no obvious evidence of clinical features (Wu et al., 2018). Besides, our knowledge of the viral population and its ecological diversity harbored by wildlife is largely obscure (Paez-Espino et al., 2016; Carroll et al., 2018; Wu et al., 2018). Therefore, an in-depth understanding of the spectrum of viruses existing in wildlife, in addition to their prevalence and distribution, will contribute to the prevention and control of emerging viral infectious diseases of wildlife origin (Olival et al., 2017).

More than 10,000 living species of birds, the richest lineage of extant tetrapod vertebrates, are distributed all over the world and show a broad diversity in morphology, and ecology (Gill et al., 2015; Prum et al., 2015). Birds exhibit a flocking behavior and can fly over great distances, thereby effectively spreading emerging and re-emerging RNA viruses, such as avian influenza viruses, Usutu virus, West Nile virus, and coronaviruses, among themselves and to humans and other animals, Reed et al. (2003), Olsen et al. (2006), Woo et al. (2012), Lühken et al. (2017), Zhu et al. (2021c). However, studies on avian viruses have mainly focused on influenza viruses, and little attention has been given to virus biodiversity.

Small mammals, such as shrews, rats, hamsters, and pikas, vary widely in preferred food, habits, habitat use, and lifestyle (Animal and Veterinary Service, 2021). Rodentia is the most diverse order of class Mammalia. It includes 33 families with more than 2,200 species and accounts for about 43% of all mammalian species (Blanga-Kanfi et al., 2009). Rodents are hosts and reservoirs of several important emerging and re-emerging zoonotic viruses, such as Rift Valley fever virus, bornavirus, Lassa virus, tick-borne encephalitis virus, lymphocytic choriomeningitis virus, and hantaviruses, which cause severe diseases in humans (Woo et al., 2012; Rabiee et al., 2018). The transmission and prevalence of rodent-borne viral diseases vary in different regions, and some of these viruses exhibit a global distribution pattern (Mills, 2005; Rabiee et al., 2018). Rodents live in close proximity with humans and play an important role in the interaction between human and arthropod vectors and other wildlife (Blanga-Kanfi et al., 2009; Meerburg et al., 2009). Further studies of the viral spectrum in the wild rodent populations could help in understanding viral evolution, emergence, and biodiversity.

The Qinghai–Tibet Plateau is a global biodiversity hotspot with diverse geographical and topographic characteristics. It is the highest and largest plateau (average elevation >4,500 m) in the world with expansive planation surfaces, mountain ranges, and basins (Zhou et al., 2006).

Knowledge of virus biodiversity is limited, and zoonotic viruses are still poorly understood. Our preliminary exploration of pikas and marmots (Luo et al., 2018; Zhu et al., 2021b) was used to establish the baseline of RNA viruses in the Qinghai–Tibet plateau. Here, we report RNA viruses from birds and small mammals in Tibet and Qinghai Provinces, China, to outline their viral spectrum, including evolutionary, genetic, and distributional characteristics.

MATERIALS AND METHODS

Sample Collection

In April 2018, fecal samples of 31 birds and 41 rats were collected from various locations in the Cona County, Tibet Province (4,360 m above sea level; **Supplementary Figure 1** and **Supplementary Table 1**). In July 2019, fecal samples of 25 birds and 50 small mammals were collected from various locations in Yushu and Nangqian Counties of Qinghai Province (3,890 m above sea level; **Supplementary Figure 1** and **Supplementary Table 1**). Small mammals were captured in their natural habitat using mousetraps, and birds were coincidentally captured while catching pikas near pika holes (Zhu et al., 2021b). The wild animals were euthanized and dissected. Their fecal samples were collected and preserved in maintenance medium consisting of Hank's balanced salt solution (pH 7.4) with penicillin G (100 U/ml) and streptomycin (50 μg/ml), and kept at −20°C while transfer to the laboratory, and stored at −80°C in the laboratory. The species of animals were identified using the mitochondrial cytochrome b (*Cyt b*) gene (Sorenson et al., 1999) or by morphological observation by experts. The sampling process was conducted by the local center for disease control and prevention (CDC) as part of the National Surveillance Program for Plague in Wildlife, and authorized by the Ethics Committee of National Institute for Communicable Disease Control and Prevention, China CDC (ICDC-2019012).

RNA Extraction

The process was mainly followed as previously reported (Wu et al., 2018). Briefly, each specimen was homogenized in phosphate buffer saline (PBS). Clear suspensions were obtained by centrifugation at 15,000 × *g* for 20 min and were filtered using a 0.22-μm polyvinylidene difluoride filter. The filtered supernatant was centrifuged at 300,000 × *g* for 2 h at 4°C. Pellets were re-suspended in PBS and digested using the RNase-Free DNase I Kit (Qiagen) at 37°C for 1 h. RNA was extracted using the QIAamp Viral RNA Mini Kit (Qiagen). RNA concentration and quality of each sample were determined using Qubit (Thermo Fisher) and 2100 Bioanalyzer (Agilent). The same animal classes from the same sampling county were divided into the same group (**Supplementary Table 1**), resulting in four groups. RNA of each sample of the same group was pooled in equal quantity.

RNA Library Construction and Next-Generation Sequencing

To facilitate virus discovery, rRNA of each library was removed as previously descripted (Zhu et al., 2021a) using the Ribo-Zero Gold rRNA Removal Kit (Illumina). Libraries were constructed using the TruSeq Stranded Total RNA Library Prep Gold Kit (Illumina) according to manufacturer's instructions. RNA was fragmented, and random hexamers were used to transcribe RNA into cDNA. The second strand of cDNA was obtained using the DNA polymerase I large fragment. The next steps included end repair, adapter ligation, purification, and fragment selection. The constructed libraries were sequenced using the Illumina HiSeq 2000 platform with 150 bp paired-end reads.

Virus Discovery

RNA viruses were detected according to an established metatranscriptomic pipeline (Shi et al., 2018). Raw reads were trimmed to remove the adapter and low-quality reads (<Q20) using Trimmomatic v0.32 (Bolger et al., 2014). The obtained high-quality reads were assembled *de novo* per pool using both

Trinity v2.4.0 (Grabherr et al., 2011) and Megahit v1.1.2 (Li et al., 2015). The resulting contigs were first annotated using the database including all reference virus proteins downloaded from NCBI[1] by Diamond BLASTx with the e-values to $1e^{-5}$ (Buchfink et al., 2015). In an attempt to identify highly divergent viral sequences, the assembled contigs were BLAST searched against the conserved domain database (CDD) v3.14 (Lu et al., 2020). Subsequently, the obtained viral contigs were verified and BLAST searched against both non-redundant protein and nucleotide databases, and contigs showing similarity to the host, plant, bacterial, and fungal sequences were eliminated. The resulting viral contigs were compared with their closely related members, and those with all viral proteins of the corresponding genus or family were retained for further analyses. The RNA-Seq by expectation–maximization algorithm (Li and Dewey, 2011) was used to quantify the abundance of contigs. Finally, viral contigs sharing <90% RNA-dependent RNA polymerase (RdRp) amino acid identity with any previously known virus were identified as new viruses (Shi et al., 2016).

Genomic and Phylogenetic Analyses

Putative open reading frames (ORFs) of viral genomes were predicted using NCBI ORF Finder.[2] The putative function of viral protein was annotated using CDD.[3] Genetic distance (p-dist) was estimated using MEGA X software (Kumar et al., 2018), and amino acid identity was calculated using BioAider v1.314 (Zhou et al., 2020). Codon usage preferences were estimated using the codon usage similarity index (COUSIN[4]).

Potential viral contigs and their closely related members were aligned by multiple alignment program using MAFFT v7 (Katoh and Standley, 2013). The best-fit substitution models were estimated using ModelFinder in IQ-TREE v2 (Minh et al., 2020). Phylogenetic trees were constructed using PhyML v3.0 based on the maximum-likelihood method (Guindon et al., 2010) with corresponding substitution models from ModelFinder and 1,000 bootstrap replicates. Trees were finally edited and visualized in interactive Tree of Life v1.0 (Letunic and Bork, 2016).

Inferring Zoonotic Potential

The probability of being able to infect humans was ranked using machine learning models with a 0.303 of the cutoff value following a recent report (Mollentze et al., 2021). Briefly, viral genomes were merged into a file in the FASTA sequence format, and the PredictNovel.R script was run.

Confirmation and Prevalence Screening

Gaps between viral contigs with unassembled overlaps were filled by RT-PCR, which was performed using the PrimeScript^{TM} One-Step RT-PCR Kit with specific primers (**Supplementary Table 2**) based on assembled sequences, and Sanger DNA sequencing. To confirm the assembly results, reads were mapped back to the viral sequences and aligned using Bowtie 2 (Langmead and Salzberg, 2012). To exclude the contigs belonging to expressed endogenous virus elements (EVEs), DNA was extracted from the corresponding samples and used for PCR amplification (**Supplementary Table 2**) using *Taq* DNA polymerase (TaKaRa). The sequence was eliminated if the PCR results were positive. In addition, these vial sequence (including host hits) were removed when performing BLASTn searches against the non-redundant nucleotide database.

To confirm the prevalence of vertebrate viruses, we designed specific primers (**Supplementary Table 2**) for viruses based on assembled genomes, and performed RT-PCR to screen corresponding viral sequence in individual samples. PCR products were subjected to gel purification and Sanger DNA sequencing. We did not submit the sequences from Sanger DNA sequencing to public databases.

RESULTS

Metatranscriptomic sequencing was performed on fecal samples collected from 31 birds (library XZNCD) and 41 rats (library XZSCD) from Tibet Province in April 2018, and 25 birds (library YSNCD) and 50 small mammals (library YSSCD) from Qinghai Province in July 2019. The samples were organized into four pools for high-throughput RNA sequencing according to animal species and sampling location (**Supplementary Table 1**). The four rRNA-depleted libraries resulted in 474,166,740 paired-end reads with 65,810,192–182,565,940 reads per pool (**Supplementary Figure 2A**), which were deposited in the NCBI Sequence Reads Archive under accession numbers SRR13847367, SRR13847389, SRR13847390, and SRR13857276.

Virome Overview

A total of 184 complete or near-complete viral genomes that contain the complete RdRp domain were obtained (**Figure 1** and **Supplementary Table 3**). Sequence comparisons indicated that 173 of them were divergent from previously known viruses, sharing <90% amino acid identity with known viruses. Coronaviruses detected in fecal samples have already been reported (Zhu et al., 2021c). Viruses of the family *Picobirnaviridae* accounted for >50% of the total number of viruses in the XZSCD and YSSCD libraries (**Supplementary Figure 2B**). All four libraries contained viruses from families *Astroviridae*, *Iflaviridae*, *Partitiviridae*, and *Solemoviridae* (**Supplementary Figure 3**).

To assess the amount of each viral RNA read, reads were mapped back to viral genomes. The viral family abundance among the four libraries showed marked differences (**Supplementary Figure 2C**). Viral families that revealed relatively high abundances were *Nodaviridae*, *Chuviridae*, *Partitiviridae*, *Dicistroviridae* and *Solemoviridae* in the XZNCD library; *Solemoviridae*, *Chuviridae*, *Picornaviridae* and *Mitoviridae* in the XZSCD library; *Totiviridae* in the YSNCD library; *Nairoviridae* in the YSSCD library (**Supplementary Figure 2**).

[1] https://ftp.ncbi.nlm.nih.gov/genomes/Viruses/

[2] https://www.ncbi.nlm.nih.gov/orffinder/

[3] https://www.ncbi.nlm.nih.gov/cdd

[4] https://cousin.ird.fr/

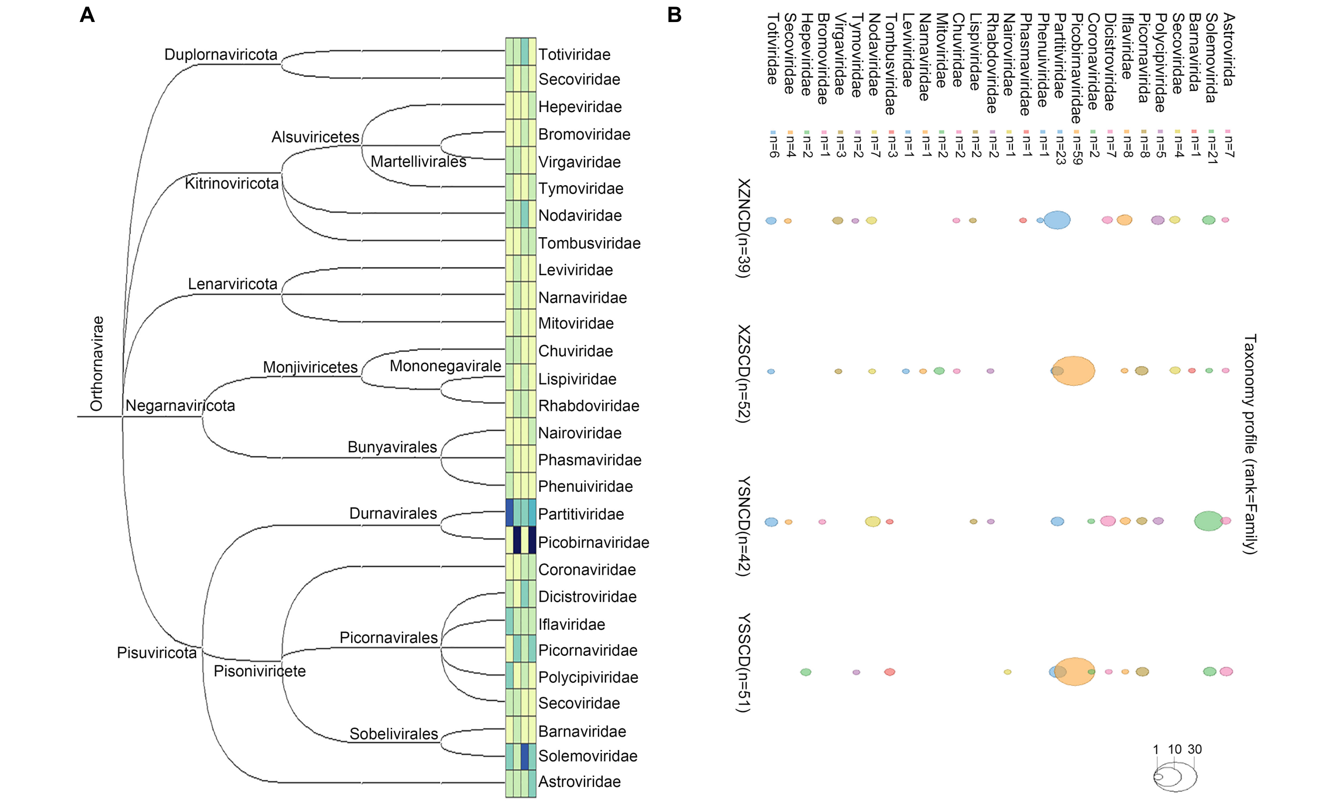


FIGURE 1 | Overview of RNA viruses identified in this study. **(A)** Phylogenetic tree of the 28 virus families. The tree was constructed by loading the names of virus families and its corresponding numbers to STAMP v2 (Parks et al., 2014). **(B)** Bubble map showing the number of viruses identified in the corresponding family in each library. Same color squares and circles represent corresponding virus families before the squares. The total number of each family is labeled after corresponding squares. The viral number of corresponding families in each pool is indicated by circles of different sizes. The no-filled circles of three sizes represent 1, 10, 30 viruses, respectively.

Genomic Characterization and Phylogenetic Analysis of Vertebrate Viruses

Despite the large number of viruses discovered, we mainly focused on the characteristics of vertebrate viruses, which signified birds or mammals as virus hosts. In addition to these insect, plant, and fungal virus families (Luo et al., 2018; Wu et al., 2018), viruses of families *Rhabdoviridae*, *Phasmaviridae* and *Phenuiviridae* appeared to be insect viruses, because they were grouped with invertebrate RNA viruses in the phylogenetic trees (**Supplementary Figure 4**). Thus, we mainly describe the characteristics and phylogenetic relationships of viruses within families *Astroviridae*, *Hepeviridae*, *Nairoviridae*, *Picornaviridae* and *Picobirnaviridae*.

Astroviridae

Seven nearly complete astrovirus genomes (length ranging from 5,779 to 7,066 bp) were detected and assembled from the four libraries. Avastrovirus YSN01 and YSN02 were identified in one bird (*Montifringilla taczanowskii*) fecal sample from Qinghai Province. Mamastrovirus YSS01–YSS03 were detected in one (*Apodemus peninsulae*), two (*A. peninsulae* and *Cricetulus kamensis*), and two (*A. peninsulae* and *Crocidura* sp.) fecal samples from Qinghai Province, respectively. Mamastrovirus XZS01 and avastrovirus XZN01 were detected from six rat fecal samples (all positive samples were of *Phaiomys leucurus*) and one bird (*Leucosticte brandti*) fecal sample, respectively (**Supplementary Table 4**). Each of these detected astroviruses showed typical genome organization comprising a single-stranded positive RNA containing three ORFs. ORF1a and ORF1b encoded non-structural polyproteins (a protease and an RdRp), and ORF2 encoded the viral capsid precursor (**Figure 2A**). According to the phylogenetic analysis (**Figure 2B**) based on ORF2 protein sequences, avastrovirus YSN01, YSN02, and XZN01 were assigned to species *Avastrovirus* 5 of the genus *Avastrovirus* (Fernández-Correa et al., 2019), whereas, mamastrovirus YSS01 and YSS02 formed an independent clade. The genetic distances (p-dist) and amino acid identity of ORF2 between mamastrovirus YSS01, YSS02, and their closely related members were greater than the threshold value 0.741 and less than 75%, respectively, indicating that mamastrovirus YSS01 and YSS02 may represent a new species of genus *Mamastrovirus*. Mamastrovirus YSS03 and XZS01 were closely related to members of species *Mamastrovirus* 6 (Alves et al., 2018), but their p-dist (>0.741) and amino acid identity (<75%)

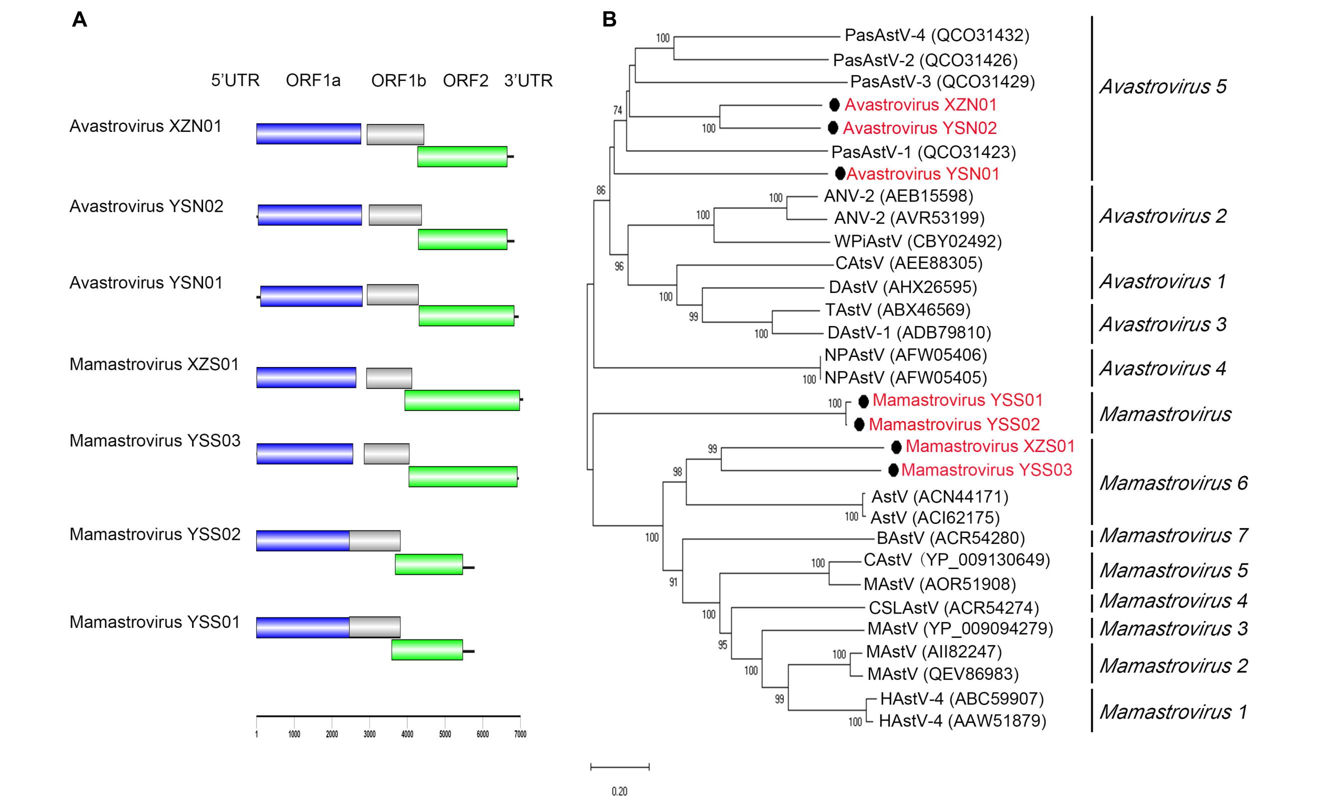


FIGURE 2 | Genomic characterization and phylogenetic analysis of astroviruses. **(A)** Genomic characterization of astroviruses. **(B)** Phylogenetic analysis based on capsid amino acid sequences in the family *Astroviridae*. Phylogenetic tree was constructed using the maximum-likelihood method with 1,000 bootstrap replicates. The best-fit substitution model was Dayhoff. Only bootstrap values >70% are shown. Scale bar indicates nucleotide substitutions per site. Viruses in this study are indicated by solid black circles and in red font.

values shared with closely related members indicate that they represent new species belonging to genus *Mamastrovirus*.

Hepeviridae

In the YSSCD library, one highly divergent virus (7,934 bp; MW826539) of the family *Hepeviridae* was identified that showed only 37.2% amino acid identity to known *Hepeviridae* sp. muf159hep1. Another nearly complete genome (6,940 bp; MW391926) was obtained and named *Orthohepevirus C* strain YS19. The virus was detected in two fecal samples of *Ochotona curzoniae* and *A. peninsulae* from Qinghai Province. Complete genome analysis showed that *Orthohepevirus C* strain YS19 was closely related to *Orthohepevirus C* isolates RdHEVAc86 (77.6% nucleotide identity) and RdHEVAc14 (77.3% nucleotide identity). The genome organization of YS19 shows four ORFs similar to those of RdHEVAc86 and RdHEVAc14 (**Figure 3A** and **Supplementary Figure 5**). The amino acid identity between YS19 and RdHEVAc86 and RdHEVAc14 ranged from 70 to 91% (**Supplementary Table 5**). A genome-based maximum-likelihood phylogenetic tree revealed that *Orthohepevirus C* strain YS19 formed an independent cluster along with RdHEVAc86 and RdHEVAc14 within the group of the species *Orthohepevirus C* (**Figure 3B**). Values of pairwise distances between YS19 and RdHEVAc86 and RdHEVAc14 were 0.068 and 0.069, respectively, which were below the threshold of 0.088 (Smith et al., 2016). These results indicate that YS19 belongs to the HEV-C3, which includes RdHEVAc86 and RdHEVAc14.

Nairoviridae

We identified an orthonairovirus in the family *Nairoviridae* from the YSSCD library, which was confirmed by PCR amplification using three paired primers (**Supplementary Table 2**). All three segments were found in only one shrew fecal sample from the Qinghai Province. The virus was named orthonairovirus YSS19 (YSV) under accession numbers MW391927–MW391929. The amino acid identity analysis indicated a close relationship of YSV with *Erve orthonairovirus* (ERVEV, 52.3–59.7%), *Thiafora orthonairovirus* (TFAV, 51.2–59.4%), and *Crimean-Congo hemorrhagic fever orthonairovirus* (CCHFV, 47.6–48.8%). Phylogenetic analysis based on the putative protein sequence of the L segment showed that YSV grouped with ERVEV and TFAV, which were closely related to CCHFV (**Figure 3C**). In addition, phylogenetic analysis based on putative protein sequences of M and S segments also showed similar results (**Supplementary Figures 6, 7**).

Codon usage preferences of YSV, ERVEV, TFAV, and CCHFV showed a highly similar pattern (**Supplementary Figure 8**). The glycoprotein precursor of YSV includes one unique proteolytic

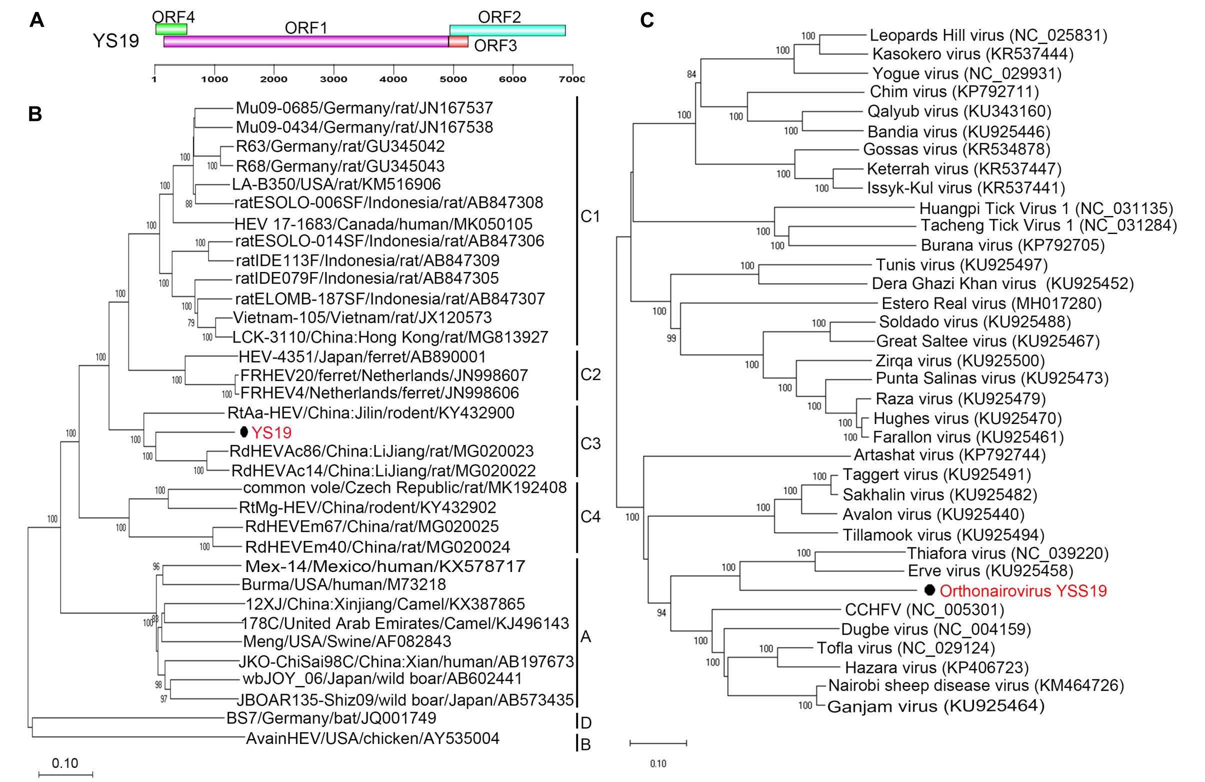

A
YS19
ORF4
ORF1
ORF2
ORF3
1
1000
2000
3000
4000
5000
6000
7000
B
Mu09-0685/Germany/rat/JN167537
Mu09-0434/Germany/rat/JN167538
R63/Germany/rat/GU345042
R68/Germany/rat/GU345043
LA-B350/USA/rat/KM516906
ratESOLO-006SF/Indonesia/rat/AB847308
HEV 17-1683/Canada/human/MK050105
ratESOLO-014SF/Indonesia/rat/AB847306
ratIDE113F/Indonesia/rat/AB847309
ratIDE079F/Indonesia/rat/AB847305
ratELOMB-187SF/Indonesia/rat/AB847307
Vietnam-105/Vietnam/rat/JX120573
LCK-3110/China:Hong Kong/rat/MG813927
C1
HEV-4351/Japan/ferret/AB890001
FRHEV20/ferret/Netherlands/JN998607
FRHEV4/Netherlands/ferret/JN998606
C2
RtAa-HEV/China:Jilin/rodent/KY432900
● YS19
RdHEVAc86/China:LiJiang/rat/MG020023
RdHEVAc14/China:LiJiang/rat/MG020022
C3
common vole/Czech Republic/rat/MK192408
RtMg-HEV/China/rodent/KY432902
RdHEVEm67/China/rat/MG020025
RdHEVEm40/China/rat/MG020024
C4
Mex-14/Mexico/human/KX578717
Burma/USA/human/M73218
12XJ/China:Xinjiang/Camel/KX387865
178C/United Arab Emirates/Camel/KJ496143
Meng/USA/Swine/AF082843
JKO-ChiSai98C/China:Xian/human/AB197673
wbJOY_06/Japan/wild boar/AB602441
JBOAR135-Shiz09/wild boar/Japan/AB573435
A
BS7/Germany/bat/JQ001749
D
AvainHEV/USA/chicken/AY535004
B
0.10
C
Leopards Hill virus (NC_025831)
Kasokero virus (KR537444)
Yogue virus (NC_029931)
Chim virus (KP792711)
Qalyub virus (KU343160)
Bandia virus (KU925446)
Gossas virus (KR534878)
Keterrah virus (KR537447)
Issyk-Kul virus (KR537441)
Huangpi Tick Virus 1 (NC_031135)
Tacheng Tick Virus 1 (NC_031284)
Burana virus (KP792705)
Tunis virus (KU925497)
Dera Ghazi Khan virus (KU925452)
Estero Real virus (MH017280)
Soldado virus (KU925488)
Great Saltee virus (KU925467)
Zirqa virus (KU925500)
Punta Salinas virus (KU925473)
Raza virus (KU925479)
Hughes virus (KU925470)
Farallon virus (KU925461)
Artashat virus (KP792744)
Taggert virus (KU925491)
Sakhalin virus (KU925482)
Avalon virus (KU925440)
Tillamook virus (KU925494)
Thiafora virus (NC_039220)
Erve virus (KU925458)
● Orthonairovirus YSS19
CCHFV (NC_005301)
Dugbe virus (NC_004159)
Tofla virus (NC_029124)
Hazara virus (KP406723)
Nairobi sheep disease virus (KM464726)
Ganjam virus (KU925464)
0.10

FIGURE 3 | Genomic characterization and phylogenetic analysis of orthohepeviruses and orthonairoviruses. **(A)** Genomic characterization of *Orthohepevirus C* strain YS19. **(B)** Phylogenetic analysis using genomes of representatives within genus *Orthohepevirus*. Only bootstrap values >80% are shown. **(C)** Phylogenetic analysis based on amino acid sequences of the L segment from all species in genus *Orthonairovirus*. Bootstrap values (≥ 90%) are shown along branches. Each phylogenetic tree was constructed using the maximum-likelihood method with 1000 bootstrap replicates. The best-fit substitution models were JC and Dayhoff, respectively. Scale bar indicates nucleotide substitutions per site. Viruses in this study are indicated by solid black circles and in red font. CCHFV, *Crimean-Congo hemorrhagic fever orthonairovirus*.

site and another identical to the proteolytic site of CCHFV and Nairobi sheep disease virus (NSDV) groups. The ovarian tumor domain (Frias-Staheli et al., 2007; Dilcher et al., 2012) of YSV (116 amino acids) shared 60.2%, 55.9%, and 45.8% amino acid identity with domains of ERVEV, TFAV, and CCHFV, respectively (**Supplementary Figure 9**).

Picornaviridae

Eight nearly complete picorna-like virus genomes were obtained (**Figure 4A**). Three viruses [*Picornaviridae* sp. XZS01 and XZS02, and rat sapelovirus (SPV) XZS03] were detected in rat fecal samples from Tibet Province, three viruses [rat mosavirus YSS01, rat SPV YSS02, and hepatovirus A (HAV) isolate YSS03] were identified in rat fecal samples from Qinghai Province, and two viruses (bird SPV YSN01 and *Picornaviridae* sp. YSN02) were found in bird fecal samples from Qinghai Province. A virus with <33% and 36% amino acid identity in P1 protein and 2C + 3 CD proteins may belong to a new genus of the family *Picornaviridae* (Zell et al., 2017). Rat SPV YSS02 and XZS03 shared 55.0% and 61.4% polyprotein amino acid identity with known members of genus *Sapelovirus*, indicating that they represent putative new species of that genus (**Figure 4B**). Polyprotein of rat mosavirus YSS01 shared 61.4% amino acid identity to that of *Mosavirus A2*. *Picornaviridae* sp. YSN02 shared 50.9% polyprotein amino acid identity with known *Picornaviridae* sp. sru051pic1. Bird SPV YSN01, *Picornaviridae* sp. XZS01, and *Picornaviridae* sp. XZS02 shared <40.5%, <38.7%, and <31.4% polyprotein amino acid identity with all known viruses of the family *Picornaviridae*, respectively, indicating that these three viruses may belong to putative new species or genera in family *Picornaviridae*.

HAV YSS03 (7,575 bp; MW391925) was identified in two samples from *A. peninsulae*; its genome included a 719-nucleotide 5′-untranslated region, 6756-nucleotide ORF-encoding polyprotein precursor (2,251 amino acids), and 100-nucleotide 3′-untranslated region. Sequence comparison of the HAV YSS03 genome showed 98.1% and 97.8% nucleotide similarities with Marmota himalayana hepatovirus (MHHAV) 2ID and 3ID, respectively. Interestingly, HAV YSS03 also shared 99.6% and 99.3% amino acid identity with MHHAV 2ID and 3ID, respectively. The potential polyprotein cleavage sites were identical between HAV YSS03 and MHHAV (Yu et al., 2016). Phylogenetic analysis based on genomic sequences (**Supplementary Figure 10**) revealed that HAV YSS03 clustered with MHHAV forming an independent branch within genus *Hepatovirus*. These results indicated that a nearly

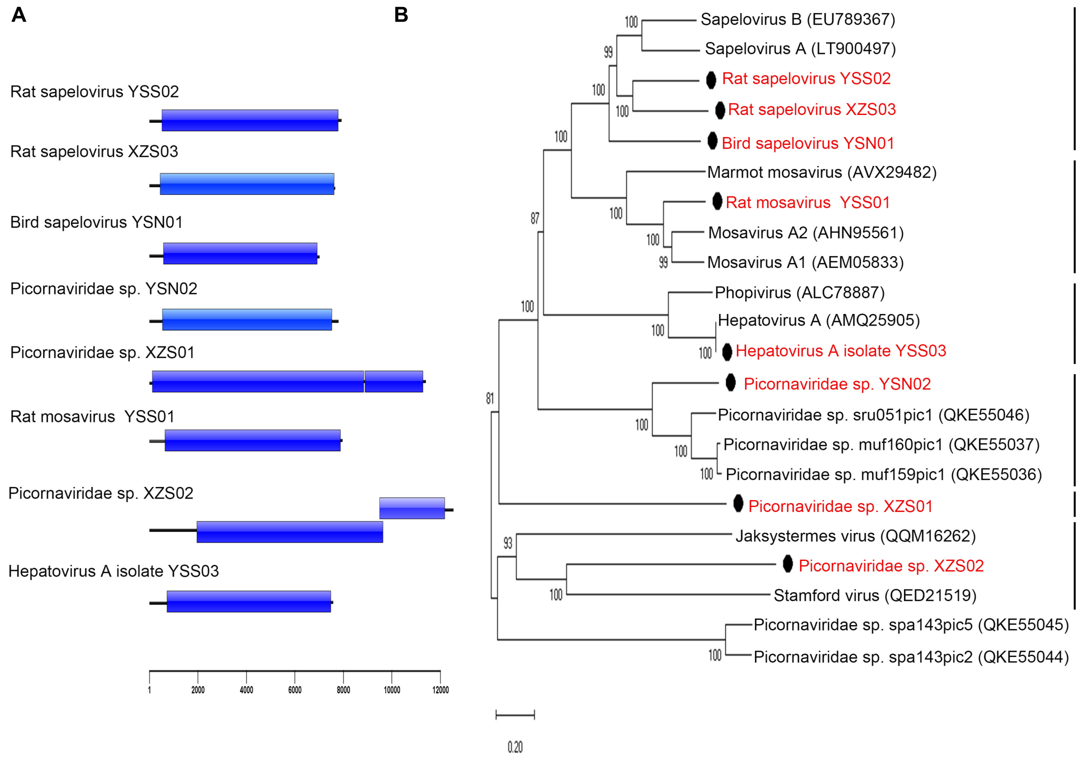


FIGURE 4 | Genomic characterization and phylogenetic analysis of viruses belonging to *Picornaviridae*. **(A)** Genomic characterization of picorna-like viruses in this study. **(B)** Phylogenetic analysis based on amino acid sequences of the polyprotein of the family *Picornaviridae*. The phylogenetic tree was constructed using the maximum-likelihood method with 1000 bootstrap replicates. The best-fit substitution model was Dayhoff. Only bootstrap values >80% are shown. Scale bar indicates nucleotide substitutions per site. Viruses in this study are indicated by solid black circles and in red font.

identical hepatovirus was present in sympatric yet genetically different hosts.

Picobirnaviridae

The family *Picobirnaviridae* includes only one genus, *Picobirnavirus*, which is usually found in vertebrate fecal samples (Luo et al., 2018). Viruses possess bi-segmented double-stranded RNA genomes (i.e., segments 1 and 2) and are rarely unsegmented (Luo et al., 2018; Delmas et al., 2019). A total of 29 segment 1 (containing the RdRp domain) and 39 segment 2 (capsid) sequences were identified in rat fecal samples from Tibet Province with the addition of two unsegmented picobirnaviruses (PBVs). Moreover, a total of 28 segment 1 and nine segment 2 sequences were identified in rat fecal samples from Qinghai Province. The amino acid identity of RdRp with that of known PBVs was <83.6%, including only two RdRp sequences showing >80% identity. To investigate the evolutionary position of these PBVs, phylogenetic trees based on RdRp and capsid amino acid sequences were constructed (**Figure 5A**), and they revealed that newly discovered capsid sequences belong to eight clusters (C1–C8). In addition to known genotypes (GI–GVI), these newly discovered RdRp sequences also formed two new genogroups (GVII and GVIII). The six members of the GVII genogroup (**Figure 5A**) were detected in fecal samples from Tibet Province, sharing 23.6–43.1% amino acid identity with known PBVs. The GVIII genogroup contained only one virus (Rat PBV XZ01) with 48.1% amino acid identity with its closest relative. Interestingly, two unsegmented PBVs, named rat PBV XZ03 (3,494 bp) and rat PBV XZ04 (4,494 bp), were identified and had complete ORFs encoding the RdRp and capsid regions. The assortment types (**Figure 5B**) of rat PBV XZ03 (C2: GI) and rat PBV XZ04 (C8: GV) were different from those previously reported (Luo et al., 2018). Rat PBV XZ03 was detected in two rat (*Phaiomys leucurus*) fecal samples, and rat PBV XZ04 was detected in one fecal sample.

Zoonotic Prediction

Zoonotic risk (i.e., the probability of being able to infect humans) of newly discovered vertebrate viruses was evaluated as previously reported (Mollentze et al., 2021). Ten viruses (avastrovirus XZN01, avastrovirus YSN02, mamastrovirus YSS02, mamastrovirus YSS01, rat SPV YSS02, bird SPV YSN01, *Picornaviridae* sp. XZS02, HAV YSS03, YSV, and *Orthohepevirus C* strain YS19) were high priority (**Figure 6**). Mamastrovirus XZS01, rat SPV XZS03, and Tibet bird virus 2 were ranked as low priority (**Figure 6**).

Invertebrate RNA Viruses

In addition to the viruses described above, a large number of invertebrate RNA viruses were detected. These viruses belonged

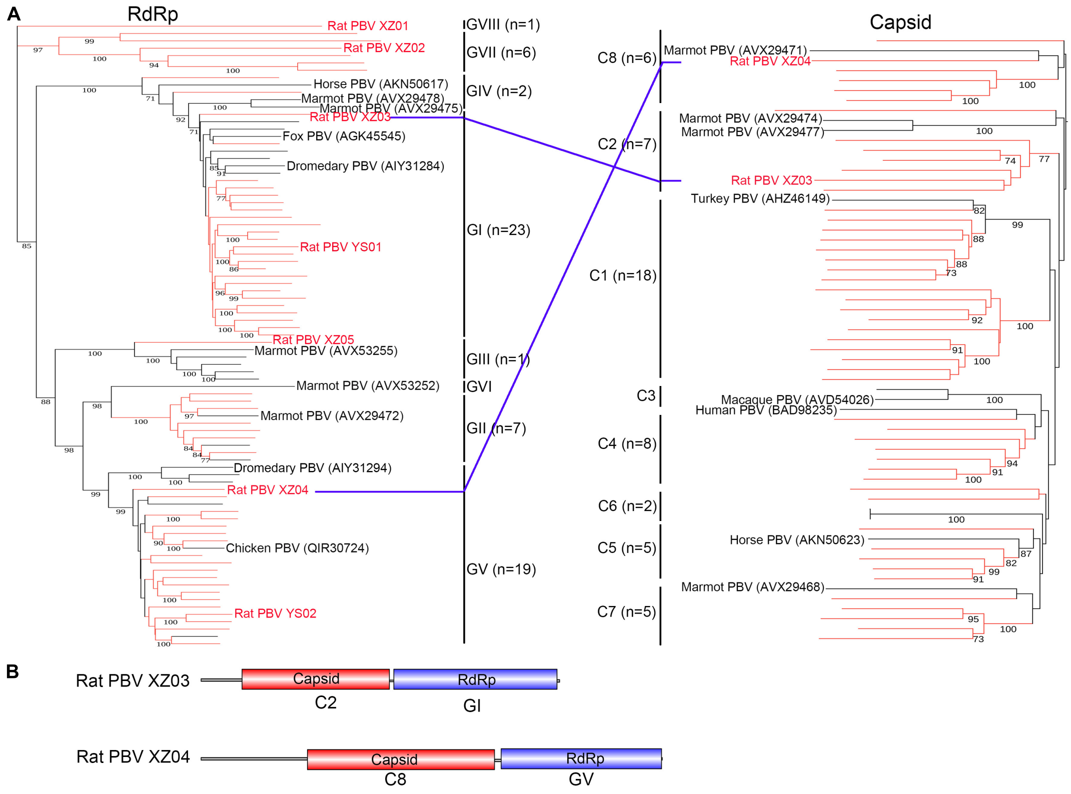


FIGURE 5 | Phylogenetic analysis and genomic characterization of *Picobirnavirus*. **(A)** Phylogenetic analysis based on amino acid sequences of RNA-dependent RNA polymerase and capsid of genus *Picobirnavirus*. Each phylogenetic tree was constructed using the maximum-likelihood method with 1000 bootstrap replicates. The best-fit substitution model was Dayhoff. Only bootstrap values >70% are shown. Viruses in this study are indicated by red font and/or red branches. The two unsegmented picobirnaviruses are also indicated by blue lines. **(B)** Genomic characterization of the two unsegmented picobirnaviruses. PBV, *Picobirnavirus*.

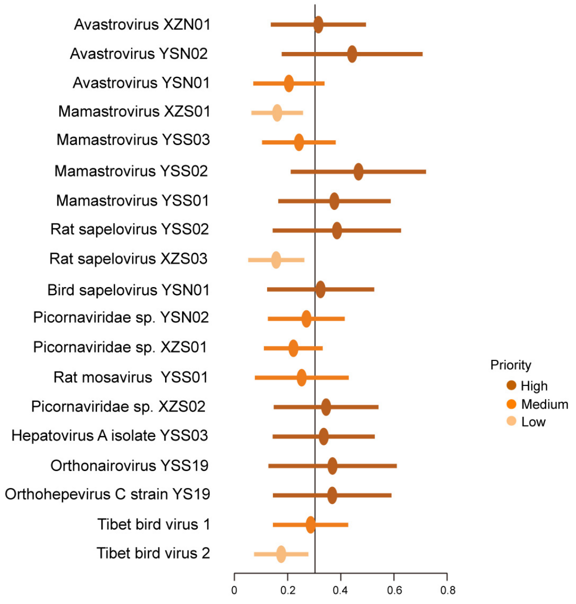


FIGURE 6 | Probability of human infection based on viral genomes. Points reveal the mean calibrated score, with lines indicating 95% confidence intervals. The black line indicates a cutoff at 0.303.

to the families *Partitiviridae* (n = 23), *Solemoviridae* (n = 21), *Iflaviridae* (n = 8), *Dicistroviridae* (n = 7), *Nodaviridae* (n = 7) and *Totiviridae* (n = 6) (**Figure 1**). Also, 12 of these viruses were distantly related to previously known viruses and each other. The RdRp domain of these viruses shared 26.3–37.6% amino acid identity with corresponding sequences of their closest relatives (**Table 1**), indicating that these 12 viruses may belong to potentially new genera or families.

DISCUSSION

The SARS-CoV-2 pandemic highlighted the need to investigate previously unknown viruses from wild animals in an unbiased manner (Albery et al., 2021; Grange et al., 2021; Hu et al., 2021). Information on viral diversity in animals, especially in birds and small mammals from the Qinghai–Tibet Plateau (average altitude >4000 m), is still limited. Here, we identified 184 RNA viruses with low prevalence rate from birds and small mammals, which belong to 28 virus families, suggesting a higher viral diversity in the Qinghai–Tibet Plateau.

These new viruses extend our understanding of RNA virus diversity in wild animals from the highest and largest plateau in the world. We could identify new genera of *Picornaviridae*,

TABLE 1 | Contig length, amino acid identity, and closest relative of the divergent viruses detected in this study.

Virus name	Accession numbers	Length of RdRp or polyprotein (amino acids)	Samples (pool)	Amino acid identity (%) [closest relative]
Arlivirus sp. YSN1024	MW826497	2160	bird (YSNCD)	32.2 [Hemipteran arli-related virus OKIAV95]
Arlivirus sp. XZN142933	MW864073	1195	bird (XZNCD)	34.1 [Lishi spider virus 2]
Comovirus sp. 143027	MW930299	1829	bird (XZNCD)	31.6 [Phaseolus vulgaris severe mosaic virus]
Dicistroviridae sp. XZN128099	MW826400	1800	bird (XZNCD)	35.7 [*Dicistroviridae* sp.]
Iflaviridae sp. XZN178790	MW826394	2998	bird (XZNCD)	29.6 [Vespa velutina associated ifla-like virus]
Leviviridae sp. XZS180134	MW826429	615	rat (XZSCD)	33.8 [*Leviviridae* sp.]
Mitovirus sp. XZS182170	MW826428	830	rat (XZSCD)	27.8 [Plasmopara viticola lesion associated mitovirus 39]
Mitovirus sp. XZS182324	MW826426	872	rat (XZSCD)	29.2 [Gergich narna-like virus]
Nodaviridae sp. XZS178253	MW826427	504	rat (XZSCD)	37.0 [Nodaviridae sp.]
Nodaviridae sp. YSN11758	MW826486	524	rat (YSNCD)	37.6 [*Nodaviridae* sp.]
Polycipiviridae sp. XZN137140	MW826414	2559	bird (XZNCD)	30.0 [Nuksystermes virus]
Polycipiviridae sp. XZN136291	MW826415	2813	bird (XZNCD)	26.3 [Yongsan picorna-like virus 3]

new species of genera *Mamastrovirus* and *Orthonairovirus*, and new genogroups of *Picobirnavirus* in these samples. We also discovered a new variant (orthohepevirus strain YS19) belonging to the HEV-C3 genotype that was clustered with RdHEVAc86, RdHEVAc14, and RtAa-HEV/JL2014. In addition, invertebrate RNA viruses that were significantly distant from known genera or families increase the number of viral taxa, providing a basis for further virus identification.

This study also broadens our knowledge in hosts of known RNA virus. First, we identified a hepatovirus (hepatovirus A isolate YSS03) that was identical to previously reported MHHAV (99.3–99.6% amino acid identity) causing fever in *M. himalayana* (Yu et al., 2016). The nearly identical hepatovirus was detected in the samples from the same region but genetically different hosts (Yu et al., 2016), suggesting a possible cross-host transmission and circulation between *M. himalayana* and *A. peninsulae* in the region. Second, a virus of genus *Tobamovirus* in the family *Virgaviridae* (accession number MW826405) was identified in a bird fecal sample from Tibet Province. It shared 100% RdRp amino acid identity with pepper mild mottle virus (accession number QIM41079) from *Nicotiana occidentalis* in Slovenia. Third, a large number of new PBVs were identified in rat fecal samples from Tibet and Qinghai Provinces, which indicates that the rat is an important wildlife host for PBV.

Although zoonotic potential is estimated based only on genomic signatures, genome-based ranking may shed new light on further studies. The training datasets contained genome sequences of 861 virus species within 36 families, which were known to infect humans as previously reported (Olival et al., 2017; Woolhouse and Brierley, 2018; Mollentze et al., 2021). The predicted probability was trained and calculated on genomic signatures (146 measures), such as dinucleotide biases, amino acid biases and relative frequency of each codon (Mollentze et al., 2021). The cut-off value (0.303) is the optimal balance between sensitivity and specificity in the dataset (Bergner et al., 2021; Mollentze et al., 2021). The best model is built based on generalizable signatures of virus genomes and combined information across viral families, and evaluated using genomes of 758 other virus species, which predicts more accurately than those based on relatedness (e.g., taxonomy) based models (Ladner, 2021). In addition, although compositional similarity may influence predictions, no individual or specific characteristic of viral genomes is primarily responsible for ranking zoonotic potential, with complex and non-linear relationships among characteristics (Mollentze et al., 2021). Thus, the genome-based ranking could be stable and reliable. However, we should acknowledge that the zoonotic potentials of these viruses are still preliminary and need additional confirmatory testing.

CONCLUSION

In conclusion, the RNA viromes of birds and small mammals were characterized, providing a fresh perspective on the viral diversity in the Qinghai–Tibet Plateau. Because of the global diversity and distribution of rodent and bird viruses, it is crucial to pay more attention to their role in viral diseases.

DATA AVAILABILITY STATEMENT

All genome sequences were submitted to GenBank under accession numbers MW391925-MW391929, MW826371-MW 826560, MW864073-MW864077, and MW930234-MW930300, respectively.

ETHICS STATEMENT

The animal study was reviewed and approved by the Ethics Committee of National Institute for Communicable Disease Control and Prevention, China CDC.

AUTHOR CONTRIBUTIONS

JX, WZ, and ZL contributed to conception and designed of the study. WZ, SL, JY, and SW collected the samples. WZ performed the experiments. WZ, JP, DJ, and X-LL performed the statistical analysis. WZ and JY wrote the first draft of the manuscript. JX, JY, and LL acquired the funding. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from National Key R&D Program of China (2019YFC1200501 and 2019YFC1200505), and Research Units of Discovery of Unknown Bacteria and Function (2018RU010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.780651/full#supplementary-material

REFERENCES

Animal and Veterinary Service (2021) Available online at: https://www.nparks.gov.sg/avs, Accessed September 10, 2021.

Albery, G. F., Becker, D. J., Brierley, L., Brook, C. E., Christofferson, R. C., Cohen, L. E., et al. (2021). The science of the host-virus network. *Nat. Microbiol.* 6, 1483–1492. doi: 10.1038/s41564-021-00999-5

Alves, C. D. B. T., Budaszewski, R. F., Torikachvili, M., Streck, A. F., Weber, M. N., Cibulski, S. P., et al. (2018). Detection and genetic characterization of Mamastrovirus 5 from Brazilian dogs. *Braz. J. Microbiol.* 49, 575–583. doi: 10.1016/j.bjm.2017.09.008

Bergner, L. M., Mollentze, N., Orton, R. J., Tello, C., Broos, A., Biek, R., et al. (2021). Characterizing and Evaluating the Zoonotic Potential of Novel Viruses Discovered in Vampire Bats. *Viruses* 13:252. doi: 10.3390/v13020252

Blanga-Kanfi, S., Miranda, H., Penn, O., Pupko, T., DeBry, R. W., and Huchon, D. (2009). Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evol. Biol.* 9:71. doi: 10.1186/1471-2148-9-71

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., et al. (2018). The Global Virome Project. *Science* 359, 872–874. doi: 10.1126/science.aap7463

Delmas, B., Attoui, H., Ghosh, S., Malik, Y. S., Mundt, E., Vakharia, V. N., et al. (2019). ICTV virus taxonomy profile: picobirnaviridae. *J. Gen. Virol.* 100, 133–134. doi: 10.1099/jgv.0.001186

Dilcher, M., Koch, A., Hasib, L., Dobler, G., Hufert, F. T., and Weidmann, M. (2012). Genetic characterization of Erve virus, a European Nairovirus distantly related to Crimean-Congo hemorrhagic fever virus. *Virus Genes* 45, 426–432. doi: 10.1007/s11262-012-0796-8

Fernández-Correa, I., Truchado, D. A., Gomez-Lucia, E., Doménech, A., Pérez-Tris, J., Schmidt-Chanasit, J., et al. (2019). A novel group of avian astroviruses from Neotropical passerine birds broaden the diversity and host range of Astroviridae. *Sci. Rep.* 9:9513. doi: 10.1038/s41598-019-45889-3

Frias-Staheli, N., Giannakopoulos, N. V., Kikkert, M., Taylor, S. L., Bridgen, A., Paragas, J., et al. (2007). Ovarian tumor domain-containing viral proteases evade ubiquitin- and ISG15-dependent innate immune responses. *Cell Host Microbe* 2, 404–416. doi: 10.1016/j.chom.2007.09.014

Gill, F., Donsker, D., and Rasmussen, P. (eds) (2015). *IOC World Bird List (v11.1)*. New Delhi: IOC. doi: 10.14344/IOC.ML.11.1

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Grange, Z. L., Goldstein, T., Johnson, C. K., Anthony, S., Gilardi, K., Daszak, P., et al. (2021). Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2002324118. doi: 10.1073/pnas.2002324118

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

He, W., Gao, Y., Wen, Y., Ke, X., Ou, Z., Li, Y., et al. (2021). Detection of Virus-Related Sequences Associated With Potential Etiologies of Hepatitis in Liver Tissue Samples From Rats, Mice, Shrews, and Bats. *Front. Microbiol.* 12:653873. doi: 10.3389/fmicb.2021.653873

Hu, B., Guo, H., Zhou, P., and Shi, Z. L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* 19, 141–154. doi: 10.1038/s41579-020-00459-7

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., et al. (2008). Global trends in emerging infectious diseases. *Nature* 451, 990–993. doi: 10.1038/nature06536

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kondo, H., Fujita, M., Hisano, H., Hyodo, K., Andika, I. B., and Suzuki, N. (2020). Virome Analysis of Aphid Populations That Infest the Barley Field: the Discovery of Two Novel Groups of Nege/Kita-Like Viruses and Other Novel RNA Viruses. *Front. Microbiol.* 11:509. doi: 10.3389/fmicb.2020.00509

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol. Dir.* 1:29.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Ladner, J. T. (2021). Genomic signatures for predicting the zoonotic potential of novel viruses. *PLoS Biol.* 19:e3001403. doi: 10.1371/journal.pbio.3001403

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi: 10.1093/nar/gkw290

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265–D268. doi: 10.1093/nar/gkz991

Lühken, R., Jöst, H., Cadar, D., Thomas, S. M., Bosch, S., Tannich, E., et al. (2017). Distribution of Usutu Virus in Germany and Its Effect on Breeding Bird Populations. *Emerg. Infect. Dis.* 23, 1994–2001. doi: 10.3201/eid2312.171257

Luo, X.-L., Lu, S., Jin, D., Yang, J., Wu, S.-S., and Xu, J. (2018). Marmota himalayana in the Qinghai-Tibetan plateau as a special host for bi-segmented and unsegmented picobirnaviruses. *Emerg. Microbes Infect.* 7:20. doi: 10.1038/s41426-018-0020-6

Meerburg, B. G., Singleton, G. R., and Kijlstra, A. (2009). Rodent-borne diseases and their risks for public health. *Crit. Rev. Microbiol.* 35, 221–270. doi: 10.1080/10408410902989837

Mills, J. N. (2005). Regulation of rodent-borne viruses in the natural host: implications for human disease. *Arch. Virology. Suppl.* 2005, 45–57.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Mollentze, N., Babayan, S. A., and Streicker, D. G. (2021). Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS Biol.* 19:e3001390. doi: 10.1371/journal.pbio.3001390

Mu, F., Xie, J., Cheng, S., You, M. P., Barbetti, M. J., Jia, J., et al. (2017). Virome Characterization of a Collection of S. sclerotiorum from Australia. *Front. Microbiol.* 8:2540. doi: 10.3389/fmicb.2017.02540

Olival, K. J., Hosseini, P. R., Zambrana-Torrelio, C., Ross, N., Bogich, T. L., and Daszak, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature* 546, 646–650. doi: 10.1038/nature22975

Olsen, B., Munster, V. J., Wallensten, A., Waldenström, J., Osterhaus, A. D., and Fouchier, R. A. (2006). Global patterns of influenza a virus in wild birds. *Science* 312, 384–388. doi: 10.1126/science.1122438

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., et al. (2016). Uncovering Earth's virome. *Nature* 536, 425–430. doi: 10.1038/nature19094

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123–3124. doi: 10.1093/bioinformatics/btu494

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., et al. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526, 569–573. doi: 10.1038/nature 15697

Rabiee, M. H., Mahmoudi, A., Siahsarvie, R., Kryštufek, B., and Mostafavi, E. (2018). Rodent-borne diseases and their public health importance in Iran. *PLoS Negl. Trop. Dis.* 12:e0006256. doi: 10.1371/journal.pntd.0006256

Reed, K. D., Meece, J. K., Henkel, J. S., and Shukla, S. K. (2003). Birds, migration and emerging zoonoses: west nile virus, lyme disease, influenza A and enteropathogens. *Clin. Med. Res.* 1, 5–12.

Shi, M., Lin, X. D., Chen, X., Tian, J. H., Chen, L. J., Li, K., et al. (2018). The evolutionary history of vertebrate RNA viruses. *Nature* 556, 197–202. doi: 10. 1038/s41586-018-0012-7

Shi, M., Lin, X. D., Tian, J. H., Chen, L. J., Chen, X., Li, C. X., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature* 540, 539–543. doi: 10.1038/ nature20167

Smith, D. B., Simmonds, P., Izopet, J., Oliveira-Filho, E. F., Ulrich, R. G., Johne, R., et al. (2016). Proposed reference sequences for hepatitis E virus subtypes. *J. Gen. Virol.* 97, 537–542. doi: 10.1099/jgv.0.000393

Sorenson, M. D., Ast, J. C., Dimcheff, D. E., Yuri, T., and Mindell, D. P. (1999). Primers for a PCR-based approach to mitochondrial genome sequencing in birds and other vertebrates. *Mol. Phylogenet. Evol.* 12, 105–114.

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., et al. (2018). Origins and Evolution of the Global RNA Virome. *Mbio* 9, e02329–18. doi: 10.1128/mBio.02329-18

Woo, P. C., Lau, S. K., Lam, C. S., Lau, C. C., Tsang, A. K., Lau, J. H., et al. (2012). Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J. Virol.* 86, 3995–4008. doi: 10.1128/jvi.06540-11

Woolhouse, M. E. J., and Brierley, L. (2018). Epidemiological characteristics of human-infective RNA viruses. *Sci. Data* 5:180017. doi: 10.1038/sdata.2018.17

Wu, T., Perrings, C., Kinzig, A., Collins, J. P., Minteer, B. A., and Daszak, P. (2017). Economic growth, urbanization, globalization, and the risks of emerging infectious diseases in China: a review. *Ambio* 46, 18–29. doi: 10.1007/ s13280-016-0809-2

Wu, Z., Lu, L., Du, J., Yang, L., Ren, X., Liu, B., et al. (2018). Comparative analysis of rodent and small mammal viromes to better understand the wildlife origin of emerging infectious diseases. *Microbiome* 6:178. doi: 10.1186/s40168-018-0554-9

Yu, J. M., Li, L. L., Zhang, C. Y., Lu, S., Ao, Y. Y., Gao, H. C., et al. (2016). A novel hepatovirus identified in wild woodchuck Marmota himalayana. *Sci. Rep.* 6:22361. doi: 10.1038/srep22361

Zell, R., Delwart, E., Gorbalenya, A. E., Hovi, T., King, A. M. Q., Knowles, N. J., et al. (2017). ICTV Virus Taxonomy Profile: picornaviridae. *J. Gen. Virol.* 98, 2421–2422. doi: 10.1099/jgv.0.000911

Zhou, S., Wang, X., Wang, J., and Xu, L. (2006). A preliminary study on timing of the oldest Pleistocene glaciation in Qinghai-Xizang Plateau. *Quater. Int.* 15, 44–51. doi: 10.1016/j.quaint.2006.02.002

Zhou, Z. J., Qiu, Y., Pu, Y., Huang, X., and Ge, X. Y. (2020). BioAider: an efficient tool for viral genome analysis and its application in tracing SARS-CoV-2 transmission. *Sustain. Cities Soc.* 63:102466. doi: 10.1016/j.scs.2020.102466

Zhu, W., Yang, J., Lu, S., Lan, R., Jin, D., Luo, X. L., et al. (2021c). Beta- and Novel Delta-Coronaviruses Are Identified from Wild Animals in the Qinghai-Tibetan Plateau. China. *Virol. Sin.* 36, 402–411. doi: 10.1007/s12250-020-00325-z

Zhu, W., Yang, J., Lu, S., Jin, D., Wu, S., Pu, J., et al. (2021b). Discovery and Evolution of a Divergent Coronavirus in the Plateau Pika From China That Extends the Host Range of Alphacoronaviruses. *Front. Microbiol.* 12:755599. doi: 10.3389/fmicb.2021.755599

Zhu, W., Song, W., Fan, G., Yang, J., Lu, S., Jin, D., et al. (2021a). Genomic Characterization of a New Coronavirus from Migratory Birds in Jiangxi Province of China. *Virol. Sin.* 36, 1656–1659. doi: 10.1007/s12250-021-00402-x

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

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

Characterization of Two Novel Toti-Like Viruses Co-infecting the Atlantic Blue Crab, *Callinectes sapidus*, in Its Northern Range of the United States

Mingli Zhao[1], Lan Xu[2], Holly Bowers[3] and Eric J. Schott[4]*

[1]Institute of Marine and Environmental Technology, University of Maryland, Baltimore County, MD, United States, [2]Department of Marine Biotechnology, Institute of Marine and Environmental Technology, University of Maryland, Baltimore County, MD, United States, [3]Moss Landing Marine Laboratory, San Jose State University, San Jose, CA, United States, [4]Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science, Cambridge, MD, United States

The advancement of high throughput sequencing has greatly facilitated the exploration of viruses that infect marine hosts. For example, a number of putative virus genomes belonging to the *Totiviridae* family have been described in crustacean hosts. However, there has been no characterization of the most newly discovered putative viruses beyond description of their genomes. In this study, two novel double-stranded RNA (dsRNA) virus genomes were discovered in the Atlantic blue crab (*Callinectes sapidus*) and further investigated. Sequencing of both virus genomes revealed that they each encode RNA dependent RNA polymerase proteins (RdRps) with similarities to toti-like viruses. The viruses were tentatively named *Callinectes sapidus* toti-like virus 1 (CsTLV1) and *Callinectes sapidus* toti-like virus 2 (CsTLV2). Both genomes have typical elements required for −1 ribosomal frameshifting, which may induce the expression of an encoded ORF1–ORF2 (gag-pol) fusion protein. Phylogenetic analyses of CsTLV1 and CsTLV2 RdRp amino acid sequences suggested that they are members of two new genera in the family *Totiviridae*. The CsTLV1 and CsTLV2 genomes were detected in muscle, gill, and hepatopancreas of blue crabs by real-time reverse transcription quantitative PCR (RT-qPCR). The presence of ~40 nm totivirus-like viral particles in all three tissues was verified by transmission electron microscopy, and pathology associated with CsTLV1 and CsTLV2 infections were observed by histology. PCR assays showed the prevalence and geographic range of these viruses, to be restricted to the northeast United States sites sampled. The two virus genomes co-occurred in almost all cases, with the CsTLV2 genome being found on its own in 8.5% cases, and the CsTLV1 genome not yet found on its own. To our knowledge, this is the first report of toti-like viruses in *C. sapidus*. The information reported here provides the knowledge and tools to investigate transmission and potential pathogenicity of these viruses.

Keywords: dsRNA virus, totivirus, crustacean, disease ecology, host habitat expansion, next generation sequencing, taxonomy, climate

OPEN ACCESS

Edited by:
Pingfeng Yu,
Zhejiang University, China

Reviewed by:
Nicolas Bejerman,
CONICET, Argentina
Pedro Luis Ramos-González,
Biological Institute of São Paulo, Brazil

****Correspondence:***
Eric J. Schott
schott@umces.edu

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *15 January 2022*
Accepted: *14 February 2022*
Published: *03 March 2022*

Citation:
Zhao M, Xu L, Bowers H and Schott EJ (2022) Characterization of Two Novel Toti-Like Viruses Co-infecting the Atlantic Blue Crab, Callinectes sapidus, in Its Northern Range of the United States. Front. Microbiol. 13:855750. doi: 10.3389/fmicb.2022.855750

INTRODUCTION

With the wide application of next generation sequencing (NGS), a huge number of virus genomes have been described from studies of metagenomes and viromes. How to best use the massive amount of data produced by NGS remains a fundamental challenge. For instance, an increasing number of toti-like virus sequences have been revealed by metagenomic studies, but further characterization and investigation are lacking (Shi et al., 2016). Ideally, following NGS-based discovery, attention should also be paid to characterize the biological properties of putative viruses, especially their genetics, viral morphological characteristics, geographic range, and potential impacts on hosts.

Viruses of the *Totiviridae* family have a non-segmented, double-stranded RNA (dsRNA) genome, with two open reading frames (ORFs) encoding the putative capsid protein (Cp) and the RNA dependent RNA polymerase (RdRp). Most virions in this family are isometric with no projections, and are ∼40 nm in diameter (Wickner et al., 2011). At present, five genera are officially recognized by the International Committee on Taxonomy of Viruses (ICTV), including *Totivirus*, *Victorivirus*, *Giardiavirus*, *Leishmaniavirus*, and *Trichomonasvirus* (King et al., 2011; Wickner et al., 2011). Viruses belonging to the genera *Totivirus* and *Victorivirus* mainly infect fungi, whereas those in the genera *Giardiavirus*, *Leishmaniavirus*, and *Trichomonasvirus* are present in parasitic protozoa and do not appear to cause cytopathic effects (Ghabrial and Suzuki, 2009; Goodman et al., 2011). Recently, non-ICTV recognized totivirus species have been found in arthropod hosts, such as mosquitoes, ants, flies, as well as crustaceans (Poulos et al., 2006; Wu et al., 2010; Koyama et al., 2015, 2016). Novel toti-like viruses have also been found in fish and plant hosts (Haugland et al., 2011; Abreu et al., 2015; Chen et al., 2015). Two genera were proposed recently, including Artivirus which infect arthropod and fish hosts (Zhai et al., 2010), and Insevirus which infect insect hosts (Zhang et al., 2018).

Two totiviruses have been reported to cause crustacean disease: a Cherax *Giardiavirus*-like virus (CGV) in freshwater crayfish (*Cherax quadricarinatus*) and infectious myonecrosis (IMN) virus (IMNV) in the Pacific white shrimp (*Litopenaeus vannamei*). CGV was the first totivirus identified in crustaceans and caused high morbidity and mortality in infected juvenile crayfish (Edgerton et al., 1994). IMNV is the most well studied totivirus in crustaceans, which causes IMN in the Pacific white shrimp in Brazil and Indonesia (Lightner et al., 2004; Poulos et al., 2006; Senapin et al., 2007; Naim et al., 2014). Additionally, metagenomics studies have reported totivirus-like dsRNA genome sequences in sesarmid and charybdis crab (Refseq. NC_032566.1 and NC_032462.1) but these have not had further characterization or investigation (Shi et al., 2016).

The Atlantic blue crab, *Callinectes sapidus*, is an adaptable estuarine species that functions as both predator and prey in food webs and supports important fisheries from the United States mid-Atlantic coast to southern Brazil (Millikin, 1984; NOAA, 2020). *Callinectes sapidus* has greatly expanded its geographic habitat range since the Last Glacial Maximum when the seas became warmer (Macedo et al., 2019), and has been introduced to Asia and Europe waters as an invasive species since 1901 (Millikin, 1984; Mancinelli et al., 2021). Unique within the *Callinectes* genus, *C. sapidus* has the ability to inhabit high latitudes by becoming dormant in winter. As the climate and ocean temperatures have changed, the distribution of *C. sapidus* has shifted poleward and the abundance of *C. sapidus* has increased at high latitudes, as far north as Nova Scotia, Canada and as far south as Argentina (Piers, 1920; Gosner, 1978; Johnson, 2015).

Within the mid-Atlantic coast and Gulf of Mexico, a range of viruses have been described in blue crabs, in the families *Baculoviridae*, *Herpesviridae*, *Reoviridae*, *Picornaviridae*, *Roniviridae*, *Rhabdoviridae*, and *Bunyaviridae* (Johnson, 1978, 1983, 1984; Bowers et al., 2010; Shields et al., 2015; Zhao et al., 2021a,b). With one exception, the relationship of these viruses to the blue crab range, and climate factors are unknown. *Callinectes sapidus* reovirus 1 (CsRV1), which causes disease and mortality in *C. sapidus*, is more prevalent in blue crabs at higher latitudes (Zhao et al., 2020), which illustrates that host-pathogen interactions can be strongly affected by habitat and environmental changes. Therefore, investigations of the effects of climate-related range extension and variation on host-pathogen interactions of other viruses will advance the understanding of drivers for virus epizootiology and ecology. The feasibility of such studies has been dramatically accelerated by molecular technologies of qPCR and high throughput sequencing, enabling virus discovery and tracking (Maclot et al., 2020).

Here, we report the discovery and characterization of two novel toti-like virus genomes that co-infect *C. sapidus* along the northern Atlantic coast of the United States. Transmission electron microscope (TEM) revealed all virions are ~40 nm in diameter, suggesting that either the two viruses are in similar size or that only one of the viruses produces virions. Pathology caused by the viruses was revealed by histology. Additionally, probe-based real-time reverse transcription quantitative PCR (RT-qPCR) assays were developed to screen and quantify totivirus infections in large numbers of *C. sapidus* across a climatological gradient.

MATERIALS AND METHODS

Crab Sampling

Blue crabs were collected from coastal states of the United States, including Massachusetts (MA), Rhode Island (RI), New York (NY), Maryland (MD), Delaware (DE), North Carolina (NC), Texas (TX), and Louisiana (LA) between the years 2009 and 2021 (**Figure 1**). A portion of crabs collected prior to 2020 were also used in the analysis of CsRV1 prevalence (Zhao et al., 2020). Crab sex and carapace width (CW, measured laterally spine-to-spine), sampling date and locations were recorded during collection. Whole crab or two walking legs removed from each crab were chilled on ice at the time of harvest. For molecular analysis, frozen specimens were then shipped to the Institute of Marine and Environmental Technology (IMET) in Baltimore, MD and stored at −20°C until further analyses. Live crabs from

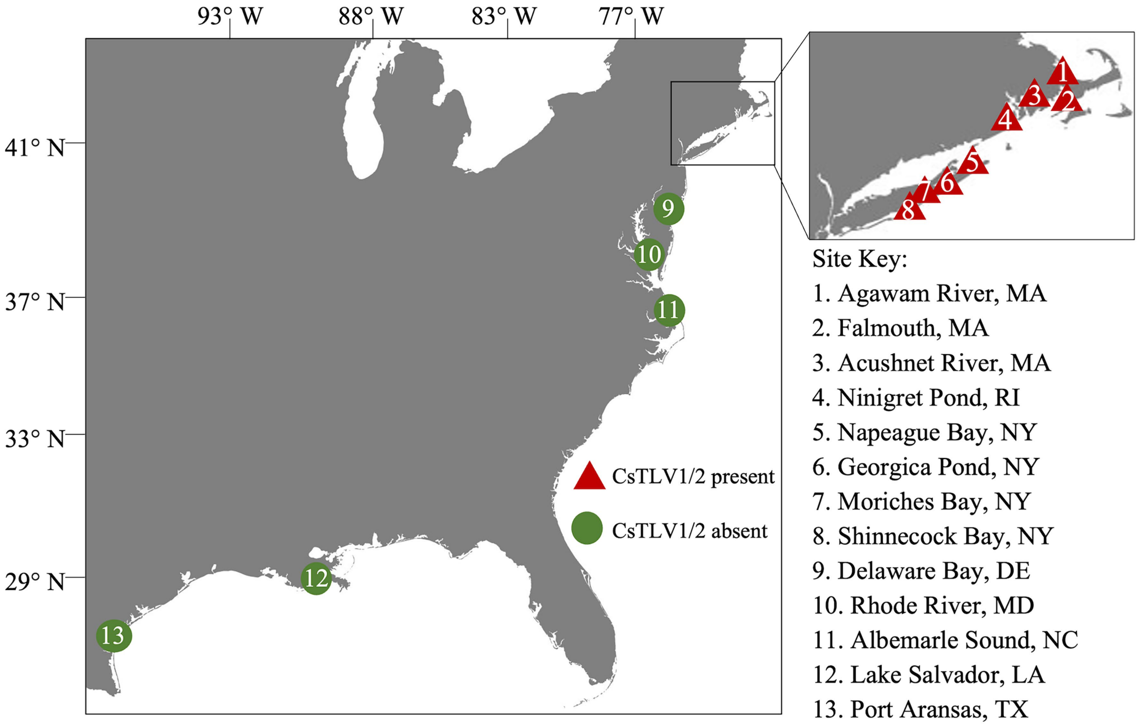


FIGURE 1 | Sampling sites in the northeastern Atlantic coast of the United States. Triangles 1–8 within the inset show sampling sites in MA, RI, and NY. Circles 9–13 indicate sampling sites in DE, MD, NC, LA, and TX. Red color indicates the presence of *Callinectes sapidus* toti-like virus 1 (CsTLV1) and *Callinectes sapidus* toti-like virus 2 (CsTLV2); and green indicates the absence of CsTLV1 and CsTLV2.

RI and Shinnecock Bay, NY, collected in the year 2021, were shipped chilled to IMET where tissues were collected for virus purification, histology, and electron microscope observations.

RNA Extraction

Crab dissections were performed with sterile wooden rods and single-use razor blades on a bench cleaned with ELIMINase™. After the external cuticle was cleaned with ELIMINase™, approximately 50 mg of muscle and hypodermis was dissected from a walking leg and homogenized in 1.0 ml of homemade Trizol (Rodriguez-Ezpeleta et al., 2009), with a Savant FastPrep™ FP120 homogenizer (MP Biomedicals, Santa Ana, CA, United States). RNA extraction followed protocols used by Spitznagel et al. (2019). After Trizol-chloroform separation of RNA and precipitation with isopropanol, two 12,000*g* centrifuge washes with 500 μl 75% ethanol were carried out. Resulting RNA pellets were dissolved in 50 μl 1 mM EDTA and stored at −80°C. Process control samples (muscle from frozen smelt) were extracted before and after sets of tested crab samples to monitor for cross contamination between each sample. RNA quality and concentration were determined by NanoDrop™ spectrophotometry (Thermo Scientific, Waltham, MA, United States). The dsRNA content of RNA extractions was revealed by electrophoresis on 1.0% agarose TBE gel and stained with ethidium bromide. The isolated dsRNA was agarose gel purified with the NucleoSpin® Gel and PCR Clean-Up Kit (Takara Bio, San Jose, CA, United States).

DNA Library Construction and High Throughput Sequencing

Purified dsRNA of a single crab infected with both *Callinectes sapidus* toti-like virus 1 (CsTLV1) and *Callinectes sapidus* toti-like virus 2 (CsTLV2), collected from Agawam River, MA, United States was used for cDNA synthesis with barcoded octamers (5’-GGCGGAGCTCTGCAGATATC-NNNNNNNN-3’) with the M-MLV Reverse Transcriptase 1st-Strand cDNA Synthesis Kit (Biosearch Technologies, Hoddesdon, United Kingdom). The resulting cDNA was amplified by PCR using the barcode primers (5’-GGCGGAGCTCTGCAGATATC-3’). Amplification was achieved through 40 cycles of 95°C for 5 s (denaturation), and 60°C for 30 s (annealing), followed by 72°C for 30 s (elongation). PCR products of 250–500 bp were obtained by agarose gel purification with a NucleoSpin® Gel and PCR Clean-Up Kit (Takara Bio, San Jose, CA, United States). The DNA library was constructed using the NEBNext R Ultra™ DNA Library Prep Kit for Illumina (NEB, Ipswich, MA, United States) following manufacturer’s instructions (NEB, Ipswich, MA, United States). The library was sequenced in a 2 × 250 paired-end configuration on the Illumina MiSeq platform with a MiSeq Reagent kit v3 (Illumina, San Diego, CA, United States).

Sequence Analyses

Sequencing barcodes were trimmed, and low quality and short reads were removed with CLC Genomics Workbench 9.5.2 (Qiagen, Hilden, Germany). The clean reads were collected and used for *de novo* assembly (Grabherr et al., 2011) with default settings (word-size = 45, Minimum contig length > =500). A preliminary set of contigs coding proteins of at least 150 amino acids were identified with ORF finder in CLC Genomics Workbench. ORFs of *de novo* derived contigs were used to search using the NCBI web server for non-redundant database using the BLASTp program. The conserved domains and motifs in the ORF were searched by NCBI Conserved Domain Database (CDD; http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Dotknot (Huang et al., 2005) was used to search the H-type pseudoknots with estimated free energy (EFE). Predicted RNA secondary structures were visualized by Pseudoviewer 2.5 (Byun and Han, 2006).

Rapid Amplification of cDNA Ends and Sequence Verification

Terminal sequences were determined using a SMARTer® Rapid Amplification of cDNA Ends (RACE) 5′/3′ Kit (Takara Bio, San Jose, CA, United States) with purified dsRNA as the initial template. The 3′ poly(A) tailing of RNA was performed at 37°C for 30 min using *E. coli* Poly(A) Polymerase (NEB, Ipswich, MA, United States). Poly (A)-tailed dsRNA was used for RACE First-strand cDNA synthesis of each terminus using 5′- or 3′- CDS primers as described by the manufacturer (Takara Bio, San Jose, CA, United States). Then, 5′-RACE and 3′-RACE PCR amplification was performed with viral gene specific primers (GSP) and universal primers (UMP), and then a nested PCR was performed with gene specific primers short (GSPS) and universal primer short (UPS; **Supplementary Table 1**) using Advantage 2 polymerase Mix (Takara Bio, San Jose, CA, United States). The conditions for amplification were 30 cycles at 94°C for 30 s, annealing at 68°C for 30 s, and elongation at 72°C for 2 min, with a final extension at 72°C for 10 min. Amplicons were then purified from the gel using NucleoSpin® Gel and PCR Clean-Up Kit (Takara Bio, San Jose, CA, United States), cloned into pGEM®-T Vector Systems (Promega Corporation, Madison, WI, United States), and sequenced. Sequence verification and filling of gaps between contigs were achieved with Sanger sequencing with primers in **Supplementary Table 1**. PCR conditions were 30 cycles at 94°C for 30 s, annealing at 56°C for 30 s, and elongation at 72°C for 90 s, with a final extension at 72°C for 10 min.

Sequence Alignment and Phylogenetic Analyses

Predicted RdRps from all reference sequences, and closest homologues from NCBI were aligned with RdRps of newly identified viruses in this study using MAFFT 7.0 (Katoh and Standley, 2013) with an accurate option (L-INS-i). The alignment was used for constructing the Maximum Likelihood (ML) phylogenetic tree with protein substitution model JTT in CLC Genomics Workbench. RdRp amino acid sequence of *Helminthosporium victoriae* virus 145S (HvV145S; Refseq. YP-052858) was used for the outgroup of the ML tree. Branch support values greater than 0.5 were shown in the tree. The accession numbers of the proteins and the corresponding virus names and acronyms are shown in **Supplementary Table 2**.

Reverse Transcription Quantitative PCR Development

To screen CsTLV1 and CsTLV2 infections in *C. sapidus*, a probe-based RT-qPCR assay was developed with primer pairs designed to detect a 193-bp region of CsTLV1 genome and a 183-bp region of CsTLV2 (**Table 1**) simultaneously. Probes designed for detecting CsTLV1 and CsTLV2 were also shown in **Table 1**. DsRNA standards were created by *in vitro* RNA synthesis. In brief, PCR products amplified by the primer pairs mentioned above were purified and cloned into pGEM®-T Vector Systems (Promega Corporation, Madison, WI, United States). Plasmids containing the targeted region were used as templates to synthesize each strand of the viral RNA standards by T7 or Sp6 RNA polymerase, respectively (Sigma-Aldrich, St. Louis, MO, United States). Viral RNAs were then quantified and annealed into dsRNA on ice, and serially diluted in 25 ng per μl yeast tRNA carrier. Standard curves were generated by RT-qPCR amplifications of a 10-fold dilution series of synthesized dsRNA containing 10–10e6 genome copies per μl. The qPCR cycling contained qScript® Virus 1-Step ToughMix® (Quantabio, Beverly, MA, United States) in 10 μl reactions with 0.25 μM each primer and 0.25 μM each probe for both genomes. To anneal PCR primers to dsRNA, primers and extracted RNA were combined, heated to 95°C for 5 min, and then cooled to 4°C prior to being added to the reverse transcriptase and Taq polymerase reaction mixture. Reverse transcription and amplification conditions were 50°C for 10 min (reverse transcription) followed by 1 min at 95°C (reverse transcriptase inactivation and template denaturation). Amplification was achieved through 40 cycles of 95°C for 10 s, and 61°C for 30 s. Gene target copies were then calculated as copies per mg of crab muscle, and samples with greater than 100 copies per mg were recorded as CsTLV1 and CsTLV2 positive, which was based on empirical observations of cross contamination in process control RNA extractions.

Statistical Analyses

All statistical tests were conducted using RStudio 1.1.456 (R Core Team, 2019). Significant correlations were defined as those where $p \leq 0.05$. To determine whether CsTLV1 and CsTLV2 infections were correlated with sex, crab size or latitude, binomial (infected vs. non-infected) generalized logistic regression models (GLM) were conducted (alpha = 0.05). Akaike's information

TABLE 1 | Primers and probes used in reverse transcription quantitative PCR (RT-qPCR).

Name		Sequences	Size (bp)
CsTLV1	forward	5’-GCAAAGGAGTGAAGGAGTGG-3’	193
	reverse	5’-GCAAGACGCATAGCACGATA-3’	
	Probe	5’6-FAM/TGCTTGCGG/ZEN/AGAAACTGAACGAGA/3’IABkFQ	
CsTLV2	forward	5’-ACGGCGACTTTGTTGAGT TT-3’	183
	reverse	5’-ACGGTAACCCAGACCATTGA-3’	
	probe	5’Cy5/AGTTGGGAG/TAO/GCAGAGATGTGTGTT/3’IAbRQSp	

criterion (AIC) was used to choose the best GLM to determine, which factors best correlate with CsTLV1 and CsTLV2 infection (Aho et al., 2014). The Pearson correlation was used to test the correlation between the variables (Kirch, 2008).

Histology and Electron Microscopy

Callinectes sapidus collected from RI and Long Island (NY) with CsTLV1 and CsTLV2 confirmed and quantified by RT-qPCR were dissected with muscle, gill and hepatopancreas tissues removed for further examination for virus presence. These crabs were also tested by RT-qPCR (methods in Zhao et al., 2020) to confirm they were not infected by CsRV1. For histological analyses, the tissues were fixed in Bouin's solution at 4°C overnight, and then placed into 75% ethanol for long-term storage. Preserved tissues were processed according to the standard operating procedures for embedding, sectioning and Hematoxylin and Eosin (H&E) staining (Luna, 1968). Slides were then observed with an Echo Revolve Microscope (San Diego, CA, United States).

For electron microscopy examination, crab tissues were immersion fixed in fixative buffer (2% paraformaldehyde, 2.5% glutaraldehyde, 2 mM CaCl2 in 0.1 M PIPES Buffer, and pH 7.35) at 4°C overnight. Tissue fragments were then trimmed into ~1 mm^3 cubes, post-fixed with 1% osmium tetroxide, washed in water and stained *en bloc* with 1% (w/v) uranyl acetate for 1 h. Specimens were then washed and dehydrated using 30, 50, 70, 90, and 100% ethanol in series. After dehydration, specimens were embedded in Araldite-Epoxy resin (Araldite, Embed 812, Electron Microscopy Sciences, Hatfield, PA, United States; Vorimore et al., 2013). Ultrathin sections (~ 70 nm) were cut and examined in a Tecnai T12 TEM (FEI, Hillsboro, OR, United States) operated at 80 KV. Digital images were acquired using an AMT bottom mount CCD camera and AMT600 software (Advanced Microscopy Techniques, Woburn, MA, United States). Crab samples infected with only CsTLV2 were not preserved well enough to be included in TEM and histology examinations.

RESULTS

Detection of Putative Viral dsRNA

In a search for viral dsRNA in blue crabs, RNA extracted from leg muscle analyzed on agarose gels. The RNA of 31% of sampled of crabs (9 of 29) harvested from the Agawam River, MA in the summer of 2008, showed two prominent dsRNA bands, termed dsRNA-S and dsRNA-L. The apparent molecular weights of the two dsRNA segments were ∼6.5 and ∼7.5 kbp, respectively (**Figure 2**).

Sequence Analyses of dsRNA-S and dsRNA-L

Total dsRNA, containing both dsRNA-S and dsRNA-L was sequenced with NGS. Assembly of trimmed and quality filtered reads (16,650 reads: 42.4% of the total reads) resulted in two contigs for each CsTLV1 and CsTLV2, at 383- and 120-fold average coverage, respectively. The longest contig was 4,016 nucleotides (nt) in length. The 5′ and 3′ untranslated regions (UTRs) of both genomes were obtained by RACE PCR and Sanger sequencing, to reveal two toti-like genome sequences of 6,444 and 7,421 nt, and designated as CsTLV1 and CsTLV2, respectively. Each genome contained two ORFs (ORF1 and ORF2) encoding Cp and RdRp proteins, respectively (**Figure 3**). Genomic sequences were not detected for any other virus in the NGS library.

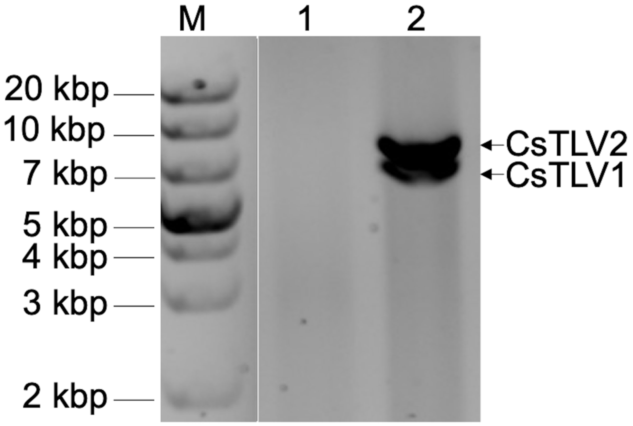


FIGURE 2 | Agarose gel electrophoresis of apparent viral double-stranded RNA (dsRNA) bands of CsTLV1 and CsTLV2. M, marker; Lane 1, control RNA of uninfected blue crab leg muscle; and Lane 2, RNA extracted from infected blue crab leg muscle showing two dsRNA bands of dsRNA-S and dsRNA-L, respectively.

The predicted Cp and RdRp proteins of CsTLV1 and CsTLV2 showed limited amino acid sequence identity with each other (21% for Cp and 27% for RdRp, respectively). CsTLV1 RdRp amino acid sequences showed >33% identity with the corresponding predicted RdRp of Beihai barnacle virus 15, Ahus virus, Parry's Creek toti-like virus 1, and Diatom colony associated dsRNA virus 17 genome type A (Shi et al., 2016; Urayama et al., 2016; Pettersson et al., 2019; Williams et al., 2020). CsTLV2 RdRp amino acid sequence showed >40% identity with the RdRp encoded in *Plasmopara viticola* lesion associated totivirus-like 5, Hubei toti-like virus 5, and Beihai sesarmid crab virus 7 (Shi et al., 2016; Chiapello et al., 2020; **Supplementary Table 3**). A search of the CDD and multiple protein alignment confirmed that the predicted RdRp domains of CsTLV1 and CsTLV2 contain eight conserved motifs (I–VIII), including the GDD motif, which are the typical characteristics of virus RdRps (**Figure 4**). Sequence analyses of CsLTV1 and CsTLV2 indicated that there is an overlap region between ORFs 1 and 2 (**Figure 3**), that allows ORF 2 to be translated as a fusion protein with ORF 1 through a-1 ribosomal frameshift motif "GGAUUUU" at 3,199–3,205 nt positions in CsTLV1, and "AAGAAAA" at positions of 2,685–2,691 nt in CsTLV2. An H-type pseudoknot structure was predicted in the downstream of each putative slippery site at positions 3,227–3,259 nt of CsTLV1 genome and 2,697–2,760 nt of CsTLV2 (**Figure 3**).

Phylogeny of CsTLV1 and CsTLV2

A maximum likelihood phylogenetic tree was used to show the relationships between CsTLV1 and CsTLV2 and other

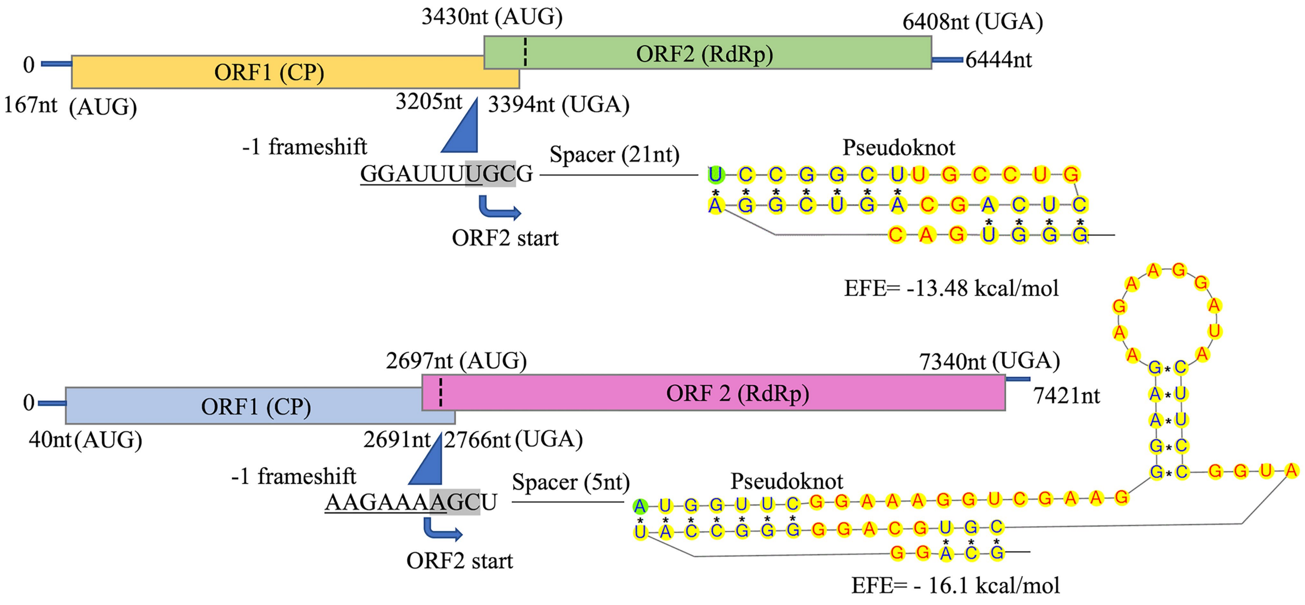


FIGURE 3 | Schematic representation of CsTLV1 and CsTLV2 genome. The two overlapping open reading frames (ORFs) and the untranslated regions (UTRs) are shown by boxes and a single line, respectively. Nucleotide positions of ORFs and the putative slippery site for −1 frameshifting, spacer and pseudoknot are indicated too. EFE (kcal/mol) indicates the minimal free energy.

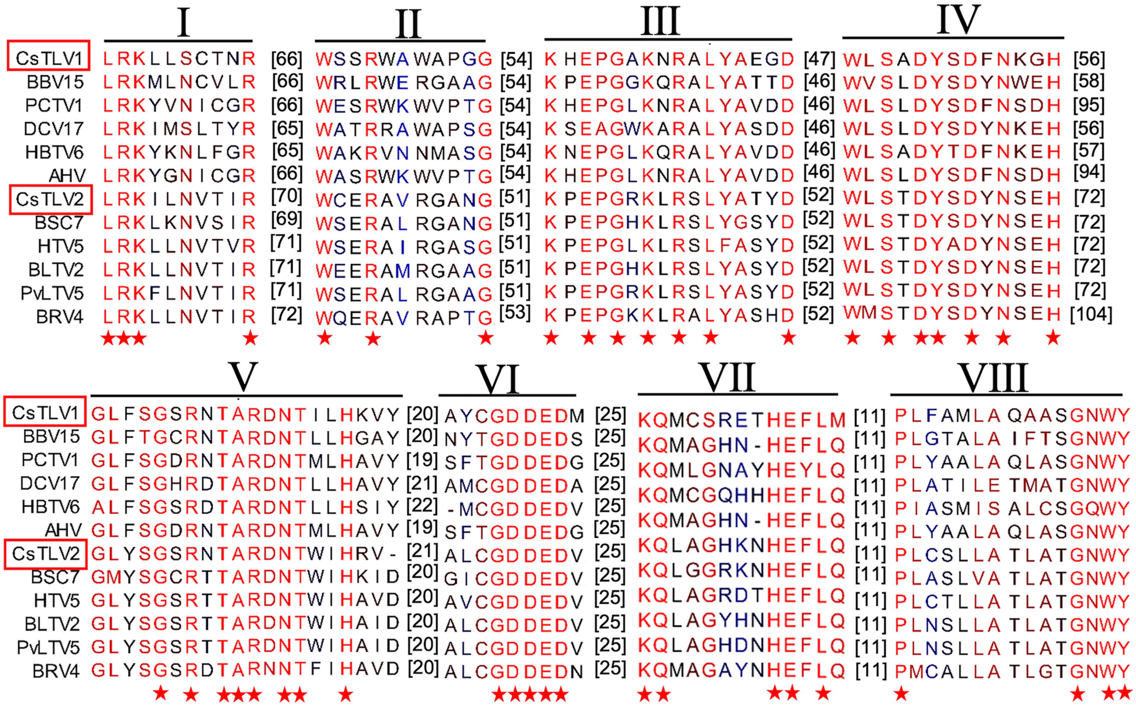

I
II
III
IV
CsTLV1 LRKLLSCTNR [66] WSSRWAWAPGG [54] KHEPGAKNRALYAEGD [47] WLSADYSDFNKGH [56]
BBV15 LRKMLNCVLR [66] WRLRWERGAAG [54] KPEPGGKQRALYATTD [46] WVSLDYSDYNWEH [58]
PCTV1 LRKYVNICGR [66] WESRWKWVPTG [54] KHEPGLKNRALYAVDD [46] WLSLDYSDFNSDH [95]
DCV17 LRKIMSLTYR [65] WATRRAWAPSG [54] KSEAGWKARALYASDD [46] WLSLDYSDYNKEH [56]
HBTV6 LRKYKNLFGR [65] WAKRVNNMASG [54] KNEPGLKQRALYAVDD [46] WLSADYTDFNKEH [57]
AHV LRKYGNICGR [66] WASRWKWVPTG [54] KHEPGLKNRALYAVDD [46] WLSLDYSDFNSDH [94]
CsTLV2 LRKILNVTIR [70] WCERAVRGANG [51] KPEPGRKLRSLYATYD [52] WLSTDYSDYNSEH [72]
BSC7 LRKLKNVSIR [69] WSERALRGANG [51] KPEPGHKLRSLYGSYD [52] WLSTDYSDYNSEH [72]
HTV5 LRKLLNVTVR [71] WSERAIRGASG [51] KPEPGLKLRSLFASYD [52] WLSTDYADYNSEH [72]
BLTV2 LRKLLNVTIR [71] WEERAMRGAAG [51] KPEPGHKLRSLYASYD [52] WLSTDYSDYNSEH [72]
PvLTV5 LRKFLNVTIR [71] WSERALRGAAG [51] KPEPGRKLRSLYASYD [52] WLSTDYSDYNSEH [72]
BRV4 LRKLLNVTIR [72] WQERAVRAPTG [53] KPEPGKKLRALYASHD [52] WMSTDYSDYNSEH [104]
V
VI
VII
VIII
CsTLV1 GLFSGSRNTARDNTILHKVY [20] AYCGDDEDM [25] KQMCSRETHEFLM [11] PLFAMLAQAASGNWY
BBV15 GLFTGCRNTARDNTLLHGAY [20] NYTGDDEDS [25] KQMAGHN-HEFLQ [11] PLGTALAIFTSGNWY
PCTV1 GLFSGDRNTARDNTMLHAVY [19] SFTGDDEDG [25] KQMLGNAYHEYLQ [11] PLYAALAQLASGNWY
DCV17 GLFSGHRDTARDNTLLHAVY [21] AMCGDDEDA [25] KQMCGQHHHEFLQ [11] PLATILETMATGNWY
HBTV6 ALFSGSRDTARDNTLLHSIY [22] -MCGDDEDV [25] [11] PIASMISALCSGQWY
AHV GLFSGDRNTARDNTMLHAVY [19] SFTGDDEDG [25] KQMAGHN-HEFLQ [11] PLYAALAQLASGNWY
CsTLV2 GLYSGSRNTARDNTWIHRV- [21] ALCGDDEDV [25] KQLAGHKNHEFLQ [11] PLCSLLATLATGNWY
BSC7 GMYSGCRTTARDNTWIHKID [20] GICGDDEDV [25] KQLGGRKNHEFLQ [11] PLASLVATLATGNWY
HTV5 GLYSGSRTTARDNTWIHAVD [20] AVCGDDEDV [25] KQLAGRDTHEFLQ [11] PLCTLLATLATGNWY
BLTV2 GLYSGSRTTARDNTWIHAID [20] ALCGDDEDV [25] KQLAGYHNHEFLQ [11] PLNSLLATLATGNWY
PvLTV5 GLYSGSRTTARDNTWIHAID [20] ALCGDDEDV [25] KQLAGHDNHEFLQ [11] PLNSLLATLATGNWY
BRV4 GLYSGSRDTARNNTFIHAVD [20] ALCGDDEDN [25] KQMAGAYNHEFLQ [11] PMCALLATLGTGNWY

FIGURE 4 | Conserved motifs in RNA dependent RNA polymerase protein (RdRp) of CsTLV1 and CsTLV2. Amino acid sequences alignment of CsTLV1 and CsTLV2 with closely related toti-like viruses from NCBI database. Horizontal lines above the alignment indicate the eight motifs, numbers in brackets indicate the amino acid sequence lengths between the motifs, asterisks indicate identical amino acid residues, and color gradients indicate the similarity level of amino acid residues. Virus notations are as in **Supplementary Table 2**.

selected totivirids. As shown in **Figure 5**, RdRp amino acid sequence multiple alignments of CsTLV1 and CsTLV2 and the corresponding toti and toti-like viral sequences revealed that CsTLVs is most closely related to, but distinct from, *Totivirus*, and Artivirus which is a proposed genus that includes IMNV and IMNV-like viruses (Zhai et al., 2010). CsTLV1 and the five toti-like viruses with the highest identity from GenBank formed a cluster in the tree (bootstrap value = 77%), adjacent to but different from the cluster of CsTLV2 and its close toti-like virus species (bootstrap value = 100%).

RT-qPCR Assay Performance

The probe-based RT-qPCR assay consistently detected as few as 10 copies of the target when tested on a dilution series of synthesized dsRNA standards (**Table 2**). Efficiency and sensitivity of the assay were evaluated by running10 RT-qPCR standard curves. The mean slope was 3.29 with a SD of 0.06 for CsTLV1, and the mean slope was 3.25 with a SD of 0.07 for CsTLV2. There is an average efficiency of 100.1 and 100.3% for CsTLV1 and CsTLV2 respectively, under typical use with the synthesized dsRNA standard.

Prevalence of CsTLV1 and CsTLV2 Infections and Co-infections

The prevalence of CsTLV1 and CsTLV2 was investigated by RT-qPCR in 875 crabs from the US Atlantic and Gulf of Mexico coasts. CsTLV1 and two infections were detected in the northern states (MA, RI, and NY, $n = 496$) but not the

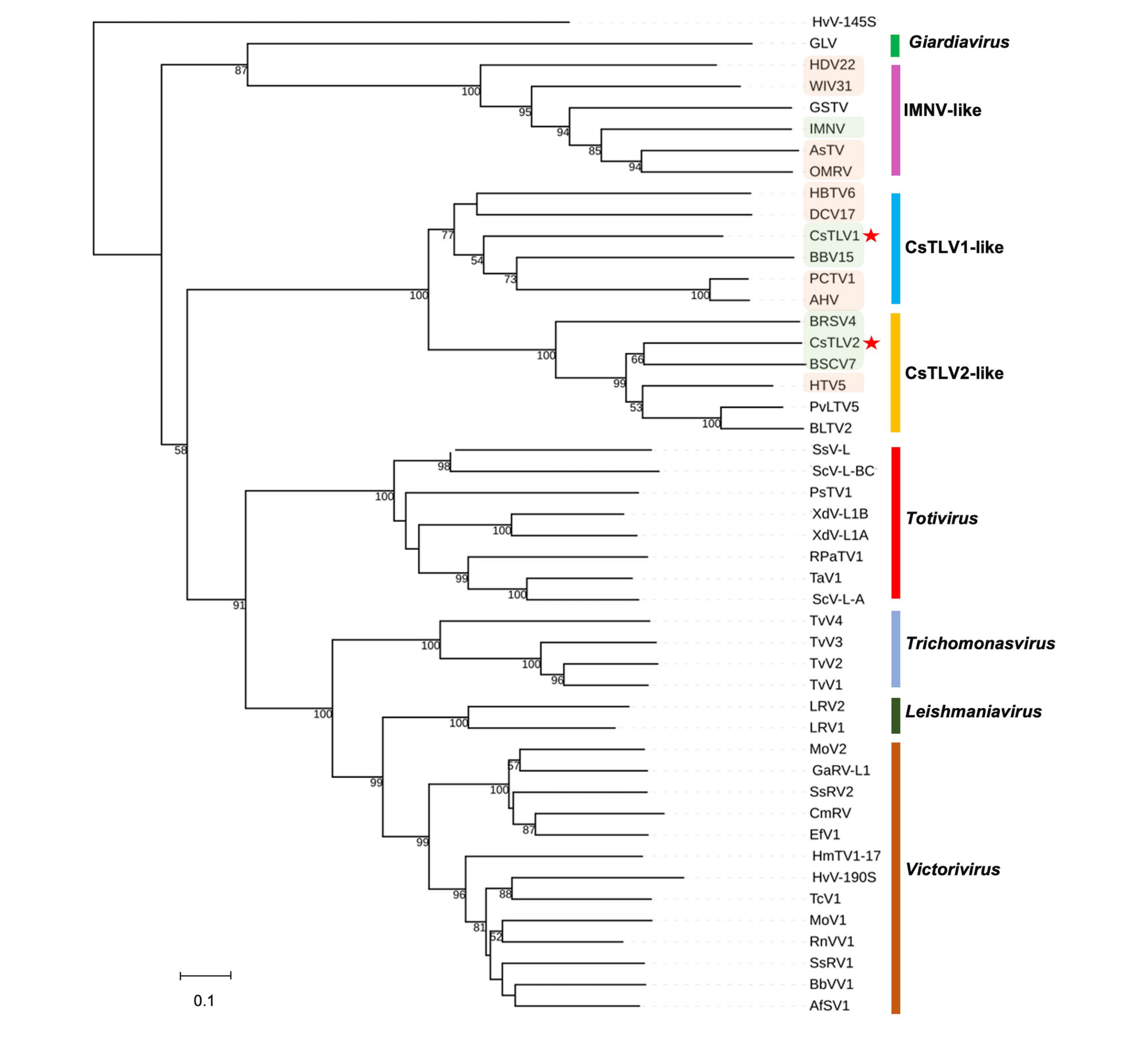


FIGURE 5 | Phylogenetic relationships between putative RdRp amino acids of CsTLV1 and CsTLV2 with other selected *Totiviridae* members. A phylogenetic tree was generated using the maximum likelihood (ML) method with 1,000 bootstrap replicates. CsTLV1 and CsTLV2 are highlighted with asterisks. Green shade indicates totiviruses identified in crustacean hosts and orange shade indicates totiviruses in other arthropod hosts. Virus notations are as in **Supplementary Table 2**.

TABLE 2 | RT-qPCR efficiency with dsRNA standards.

Genome	Slope	*Y*-intercept	R^2	Efficiency
CsTLV1	3.29	38.730	0.995	100.1
CsTLV2	3.25	37.624	0.995	100.3

The threshold cycles for a log10 dilution series are used to assess efficiency relative to 100% theoretical efficiency for a slope of 3.32 based on 10 replicates.

lower latitudinal Atlantic states (DE, MD, and NC, $n = 299$) and Gulf states (LA and TX, $n = 80$). CsTLV1 and CsTLV2 infections in *C. sapidus* were detected in all three estuaries sampled in MA: Agawam River, Acushnet River, and Falmouth (**Figure 1; Table 3**). CsTLV1 RNA prevalence in *C. sapidus* sampled from MA ($n = 198$) varied from 11.8% (Agawam River; $n = 127$) to 30.6% (Acushnet River; $n = 49$), and CsTLV2 RNA prevalence ranged from 12.6 to 36.7%. In crabs from RI, CsTLV1 prevalence was 27.6% and CsTLV2 prevalence was 32.8%. Viral RNA was detected in crab specimens collected from Moriches Bay, Shinnecock Bay, and Napeague Bay in Long Island, NY ($n = 133$) with an average prevalence of 42.1 and 43.6% for CsTLV1 and CsTLV2, respectively. CsTLV1 RNA was detected in Georgica Pond, NY, with low prevalence in 2012 (5.5%; $n = 18$), but not in 2013, 2020, or 2021, and CsTLV2 RNA was detected in 2012 (5.5%; $n = 18$) and 2013 (5.2%; $n = 19$), but not in 2020 ($n = 33$) or 2021 ($n = 37$).

Overall, CsTLV1 was never observed in the absence of CsTLV2, and co-infections of CsTLV1 were detected in 91.5% (107/117) of CsTLV2 positive specimens (**Table 3**). The dsRNA copy number per mg muscle ranged from 6.5×10e2 to 1.2×10e8 for CsTLV1, and 1.3×10e2 to 6.3×10e8 for CsTLV2.

Correlation of CsTLV1 and CsTLV2 Infection With Latitude and Crab Size

A binomial (infected vs. non-infected) generalized linear model (GLM) was used to test whether latitude, crab size, or sex could predict CsTLV1 and CsTLV2 infection status (**Table 4**). The 415 male and 140 female specimens, from 20 to 196 mm in carapace width (CW), were used for GLM analysis. Specimens that were PCR-positive for CsTLV1 and CsTLV2 ranged from 29 to 150 mm in CW. Prevalence for male and female crabs was 12.0 and 10.7%, respectively. Pearson correlation tests showed no significant correlation between latitude and crab size or sex. The full model analyzing the association between CsTLV1 and CsTLV2 infections and latitude, crab sex, and carapace width (CW), differed significantly from null models ($p < 0.01$), in which latitude and CW were significant fixed effects ($p < 0.01$). The reduced model, including only the association between CsTLV1 and CsTLV2 infection and latitude, sex, or CW, reinforced that latitude and CW were the significant factors correlated with CsTLV1 and CsTLV2 prevalence ($p < 0.01$; **Table 4**). CsTLV1 and CsTLV2 prevalence was positively related to latitude in the reduced model (slope is 2.33 for CsTLV1 and 2.38 for CsTLV2; $p < 0.01$), and CW showed a negative association with CsTLV1 and CsTLV2 prevalence (slope = −1.00 for CsTLV1 and slope = −1.13 for CsTLV2; $p < 0.01$). In both full and reduced models, the association between CsTLV1/CsTLV2 prevalence and crab sex was not significant ($p > 0.1$).

Electron Microscopy: Observation of Viral Particles

Crabs assessed to be infected with CsTLV1 and CsTLV2 by RT-qPCR were selected for TEMs observation (**Supplementary Figure 1**). TEM revealed the presence of isometric virus particles, with a diameter of ~40 nm in *C. sapidus* muscle, gill, and hepatopancreas tissues (**Figure 6**). Completed virions were present in the connective tissue and hemocytes of these tissues. We observed putative viroplasm in the gill of CsTLV1 and CsTLV2 infected crab and packed arrays of mature virions in the hepatopancreas of infected crab.

Histopathology of CsTLV1 and CsTLV2 Co-infected Blue Crab Tissues

Histological analysis of muscle, hepatopancreas and gills tissues of crabs naturally infected with CsTLV1 and CsTLV2 showed necrosis and hemocyte infiltration. Skeletal muscle in normal uninfected crabs is generally smooth, striated and with few circulating hemocytes (**Figure 7A**). Infected skeletal muscle had general necrosis and showed vacuolated areas with increased numbers of circulating hemocytes (**Figure 7B**). Hepatopancreas tubules in normal uninfected crabs have defined outer membranes and moderate numbers of circulating hemocytes circulating within hemal spaces between tubules (**Figure 7D**). Infected hepatopancreas often showed massive hemocytic infiltration (**Figure 7E**). Gills of normal uninfected crabs have moderate numbers of circulating hemocytes in hemal spaces (**Figure 7G**). Infected gills had considerably increased numbers of circulating hemocytes within necrotic areas (**Figure 7H**). At higher magnification, infected hemocytes in muscle, hepatopancreas, and gills often had pyknotic or karyorrhectic nuclei (magenta arrows) as well as opaque, slightly eosinophilic intracytoplasmic inclusion bodies (blue arrows; **Figures 7C,F,I**).

DISCUSSION

Molecular approaches for discovery of virus-like genomes have verified that viruses are an important and universal feature of the life history of marine organisms (Munn, 2006; Suttle, 2007). Beyond the discovery of new viruses, the characterization of these newly discovered viruses contributes to better understanding of their diversity, evolution, and ecology in marine environments. Partial toti-like virus sequences, reported in some crabs by metagenomics (e.g., in Shi et al., 2016), have only documented virus-like genome elements, but not the presence of virus particles. In this study, we sequenced and characterized the genomes of two new putative *C. sapidus* totiviruses—CsTLV1 and CsTLV2, and showed that viral particles are present in tissues of CsTLV1 and CsTLV2 co-infected crabs and are associated with pathology. This study is the first description of an endemic infection of totivirus in *C. sapidus*.

TABLE 3 | CsTLV1 and CsTLV2 prevalence in *C. sapidus*. Specimens were collected from locations along the US Atlantic coasts and Gulf coasts of the United States.

Location	Collection date (Month-Year)	Latitude	Longitude	Total *N*	CsTLV1		CsTLV2	
					Infected	Pre	Infected	Pre
					(*N*)	(%)	(*N*)	(%)
Agawam River, MA	Aug-2009	41.7619°N	71.6773°W	29	11	37.9	11	37.9
	Aug-2012			47	4	8.5	4	8.5
Falmouth, MA	Sep-2018	41.5388°N	70.6266°W	51	0	0	1	2
	Sep-2018			22	4	18	4	18
Acushnet River, MA	Aug-2012	41.6617°N	70.9182°W	49	15	30.6	18	36.7
Rhode Island, RI	Aug-2021	41.3697°N	71.6426°W	58	16	27.6	19	32.8
Napeague Bay, NY	Jul-2021	40.9987°N	72.0972°W	10	6	60	6	60
Georgica Pond, NY	Aug-2012	40.9361°N	72.2138°W	18	1	5.5	1	5.5
	Jul-2013			19	0	0	1	5.2
	Jul-2020			33	0	0	0	0
	Jul-2021			37	0	0	0	0
Moriches Bay, NY	Jul-2018	40.7738°N	72.8052°W	32	7	21.8	8	36.7
	Jul-2021			25	4	16	5	20
Shinnecock Bay, NY	Jul-2021	40.8426°N	72.4762°W	28	18	64.3	18	64.3
	Sep-2021			21	10	47.6	10	47.6
				17	11	64.7	11	64.7
Delaware Bay, DE	Apr-2019	38.9108°N	75.5277°W	51	0	0	0	0
	Aug-2021			38	0	0	0	0
Rhode River, MD	Mar-2015	38.8795°N	76.5216°W	33	0	0	0	0
	Jul-2018			52	0	0	0	0
	Aug-2020			30	0	0	0	0
Albemarle Sound, NC	Oct-2019	33.8772° N	76.1248° W	95	0	0	0	0
Port Aransas, TX	Jan-2021	27.8339° N	97.0611° W	40	0	0	0	0
Lake Salvador, LA	Jan-2021	29.7192° N	90.2432° W	40	0	0	0	0

Pre: Prevalence.

TABLE 4 | Generalized linear model (GLM) with potential effects on CsTLV1 and CsTLV2 infection.

Model	Predictor variable	Estimate (slope)	Standard Error	*p*-value
A. Full model:				
CsTLV1 Infection~ Latitude + Size + Sex (AIC = 339.98; df = 551)	Latitude	2.54	0.64	6.64e−05***
	Size	−1.00	0.25	9.47e−05***
	Sex	−0.31	0.35	0.37
CsTLV2 Infection1 ~ Latitude + Size + Sex (AIC = 340.89; df = 551)	Latitude	2.68	0.66	4.77e−05***
	Size	−1.12	0.26	1.12e−05***
	Sex	−0.30	0.35	0.39
B. Reduced model:				
CsTLV1 Infection ~ Latitude (AIC = 352.24; df = 554)	Latitude	2.33	0.54	1.43e−05***
CsTLV1 Infection ~Size (AIC = 382.85; df = 553)	Size	−1.13	0.24	2.24e−06***
CsTLV1 Infection ~Sex (AIC = 404.68; df = 553)	Sex	−0.13	0.31	0.67
CsTLV2 Infection ~ Latitude (AIC = 357.72; df = 554)	Latitude	2.38	0.54	9.78e−06***
CsTLV2 Infection ~Size (AIC = 385.34; df = 553)	Size	−1.25	0.24	1.42e−07***
CsTLV2 Infection ~Sex (AIC = 412.8; df = 553)	Sex	−0.08	0.30	0.78

*The model has the lowest Akaike's information criterion (AIC) of all combinations of predictor variables (see as full and reduced interactions). ***denotes significance ($p < 0.001$).*

Both CsTLV1 and CsTLV2 genomes contained two ORFs encoding the conserved domains of Cp and RdRp, respectively. Moreover, the two viruses contain a-1 ribosomal frameshifting in their genomes (**Figure 3**), which could facilitate the translation of ORF1 and ORF2 as a fusion polyprotein (Dinman et al., 1991). The predicted ORF2 coding strategy of CsTLV1 and CsTLV2 was consistent with other viruses in the family *Totiviridae*, such as *Saccharomyces cerevisiae* virus L-A (ScVL-A; Dinman et al., 1991) and IMN virus (IMNV; Nibert, 2007). CsTLV1 and CsTLV2 have all three elements that are required to accomplish −1 ribosomal frameshifting in RNA viruses: a slippery heptamer motif, an RNA pseudoknot shortly downstream of the site and a short spacer region between the slippery site and the pseudoknot (Rice et al., 1985; Dinman et al., 1991; Khalifa and MacDiarmid, 2019). The classical slippery site sequence is "XXXYYYZ" (where X is A/C/G/U, Y is A/U, and Z is A/C/U) within the overlapping region (Bekaert and Rousset, 2005). The slippery site of CsTLV1 (GGAUUUU) is

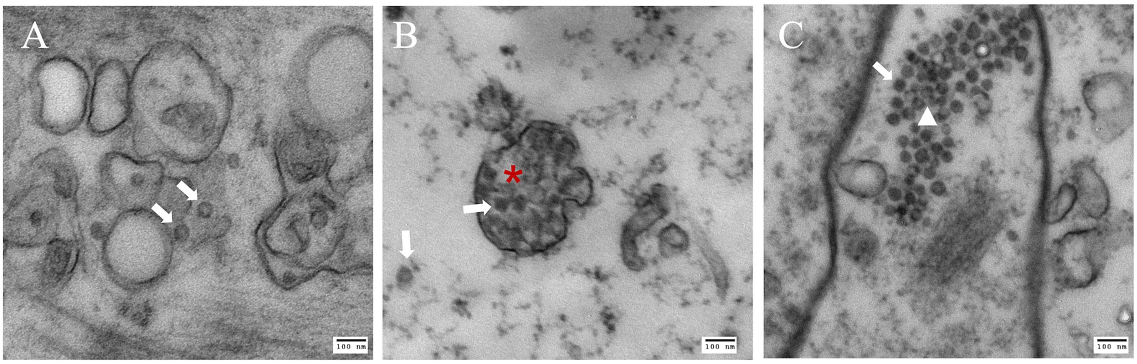


FIGURE 6 | Electron microscopy images of putative CsTLV1 and CsTLV2 viral particles in muscle, gill, and hepatopancreas tissues of *Callinectes sapidus*. **(A)** muscle; **(B)** gills; and **(C)** hepatopancreas. White arrow: virions; Red star: putative "viroplasm" in gill. White triangle: dense arrangement of virions in hepatopancreas. Scar bar, 100 nm.

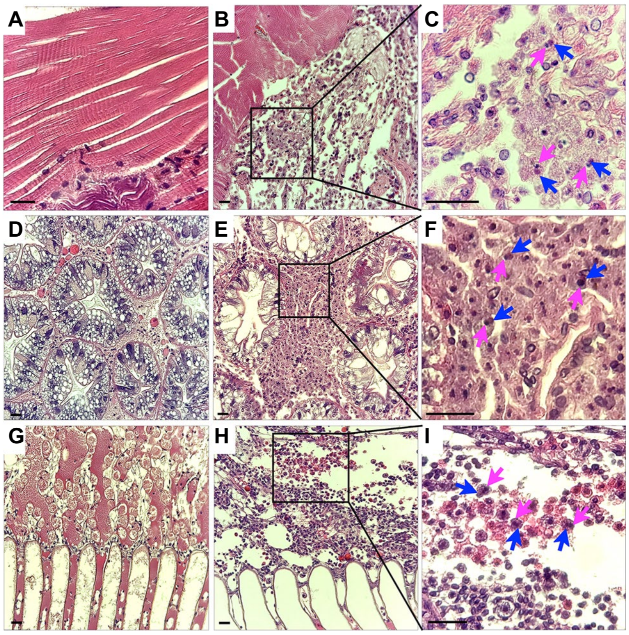


FIGURE 7 | Histology of CsTLV1 and CsTLV2 infections in muscle, gill, and hepatopancreas of *Callinectes sapidus*. **(A,D,G)** Muscle, hepatopancreas, and gill of uninfected crabs; **(B,E,H)** Muscle, hepatopancreas, and gill of CsTLV1 and CsTLV2 infected blue crabs; **(C,F,I)** Magnified view of boxed area in **(B,E,H)**, respectively. Infected hemocytes in muscle, hepatopancreases, and gill often had pyknotic or karyopyknotic nuclei (magenta arrows) as well as opaque, slightly eosinophilic intracytoplasmic inclusion bodies (blue arrows). Scale bar, 25 μm.

the same to the slippery heptamer nucleotides found in other totiviruses, such as *Xanthophyllomyces dendrorhous* viruses (GGAUUUU; Baeza et al., 2012), *Puccinia striiformis* totiviruses (PsTVs; GGG/AUUUU; Zheng et al., 2017) and red clover powdery mildew-associated totiviruses (RPaTVs; GGG/AUUUU; Kondo et al., 2016). Meanwhile, the slippery site is "AAGAAAA" in CsTLV2, which is the same as that used by plant associated astro-like virus (Lauber et al., 2019).

In the current ICTV scheme of totivirus taxonomy, 50% sequence identity of Cp/RdRp proteins is generally considered a threshold to define different species (Wickner et al., 2011). CsTLV1 and CsTLV2 share only 21% identity for Cp and 27% for RdRp, indicating they are distinct species in the family *Totiviridae*. Phylogenetic analyses of RdRp amino acid sequences showed that CsTLV1 and CsTLV2 formed a distinct branch from other genera in the family *Totiviridae* but clustered into two subgroups (**Figure 5**). CsTLV1, together with toti-like viruses identified from arthropod and crustacean hosts were classified into one group (Shi et al., 2016), and CsTLV2 formed another group with totiviruses sequenced from spirurian nematodes, sesarmid crab, and razor shell clam *Ensis magnus* (Shi et al., 2016). Compared to other genera of the family *Totiviridae*, members of CsTLV1-like and CsTLV2-like groups have the highest similarity between each other. Taken together with their genome structure and phylogenetic position, CsTLV1 and CsTLV2 may represent two new viral species within two novel genera of the family *Totiviridae*.

Co-infection by two distinct viruses has been reported in *C. sapidus* such as reovirus and RhVA (Johnson, 1978, 1983), and bunya-like virus (Zhang et al., 2004). Co-infection of distinct totiviruses has also been commonly reported, such as in *Sphaeropsis sapinea* and *Chalara elegans* (Preisig et al., 1998; Park et al., 2005). Recently, co-infection of three dsRNA viruses *Trichomonas vaginalis* virus (TVV1, TVV2, and TVV3) were revealed (Bokharaei-Salim et al., 2020). Co-infection of CsTLV1 was detected in more than 90% CsTLV2-positive specimens (**Table 3**), suggesting that there is a significant relationship between these two totiviruses in *C. sapidus*. Interestingly, although independent infection of CsTLV2 was identified, no crab was ever found that contained the CsTLV1 genome alone. One possible explanation for this observation may be that the CsTLV1 genome or virus cannot replicate or be encapsulated in the absence of CsTLV2. The relationship between CsTLV1 and 2 does not have the characteristics of defective virus genomes (Vignuzzi and López, 2019); the CsTLV1 genome does not have obvious deletions or frame shifts, although the CsTLV2 genome is over 1,000 nt longer than the CsTLV1 genome. A similar phenomenon has been revealed that *Helminthosporium* *victoriae* virus 190S (HvV190S; *Totiviridae*) and *Helminthosporium victoriae* virus 145S (HvV145S; *Chrysoviridae*) co-infect the pathogenic fungus *Helminthosporium victoriae*. HvV145S has never been found alone but is always associated with HvV190S virus. HvV145S was originally thought to be the cause of the diseases, however, a recent study suggested that HvV190S alone is the cause of diseases, and the co-infection is not required (Xie et al., 2016). In our study, TEM of co-infected blue crabs revealed all virions had a diameter of ~40 nm, suggesting that either CsTLV1 is indistinguishable in size or appearance from CsTLV2, or that only one of the viruses produces virions.

Most members of *Totiviridae* infecting fungi and protozoans lack extracellular transmission; instead, they are transmitted vertically during cell division, sporogenesis, and cell fusion (Ghabrial and Suzuki, 2009). However, some totiviruses with fiber-like protrusions on their surface, such as IMNV and Omono River virus (OmRV), are capable of extracellular transmission in their metazoan hosts (Poulos et al., 2006; Tang et al., 2008; Dantas et al., 2015; Shao et al., 2021). The transmission mechanism for CsTLV1 and CsTLV2 in the blue crab is yet unknown. Attempts to transmit the viruses by injection of previously frozen material (CsTLV1 and CsTLV2) into naïve crabs have been so far unsuccessful (Zhao and Schott, unpublished data). Necrosis and massive hemocyte infiltration in CsTLV1 and CsTLV2 infected muscle, gill, and hepatopancreas suggested that the viruses are detrimental to the health of blue crabs. CsTLV1 and CsTV2 infections were negatively correlated with crab size in GLM analyses, which suggested that juveniles may be more susceptible to infection, or that older animals infected with CsTLV1 and CsTLV2 either die or clear the virus as they mature or age. All these results provide the fundamental knowledge for future studies to investigate how these viruses are transmitted and how they affect the ecology of blue crabs.

The significant correlation between CsTLV1 and CsTLV2 infections and latitude has also been identified in another blue crab dsRNA virus-CsRV1, which also showed significantly higher prevalence at higher latitudinal locations compared to lower latitudes (Flowers et al., 2016, 2018; Zhao et al., 2020). However, compared to the wide geographic range of CsRV1 infections in blue crabs, infections of CsTLV1 and CsTLV2 were restricted to the most northeastern estuaries we sampled in MA, RI, and NY, but absent from the lower latitudinal estuaries of DE, MD, NC, LA, and TX. Although factors driving the emergence of viruses and the gradient of virus prevalence at different geographic locations could be complex, two likely covariates in our study are water temperature and length of the active period for blue crabs, which have strong correlations to latitudes (Zhao et al., 2020). It is notable that the virus is present in crabs at the northern edge of their geographic range. Microbiome community changes and emergence of novel pathogens have been widely reported during the dispersal of host invasion and extension range (Engering et al., 2013; Dragičević et al., 2021). The extensive poleward expansion of *C. sapidus* in its native range along the western Atlantic and its successful invasion to European waters (Johnson, 2015; Mancinelli et al., 2021), made *C. sapidus* a well-suited model to study virus evolution, diversity, and viral ecology of marine animals during host habitat expansion and invasion. In Rhode Island, state managers are beginning to survey blue crab abundance in anticipation of a growing commercial and recreational fishery (K. Rodigue, personal communication). Therefore, further systematic and comprehensive studies on the virome of *C. sapidus*, including CsTLV1 and CsTLV2, at different geographical locations are urgently needed for a better understanding of the virus ecology and epidemiology with the host habitat expansion.

In conclusion, two putative viral dsRNA sequences in *C. sapidus* were characterized with NGS, and shown to be associated with virus particles and histopathology. Based on their genomic organizations, phylogenetic relationships, and conserved motifs, the viruses are tentatively named CsTLV1 and CsTLV2, and proposed to be members of two new genera in the family *Totiviridae*.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/genbank/, OL456199, OL456200.

AUTHOR CONTRIBUTIONS

ES and MZ designed the experiments and analyzed the results and drafted the manuscript. MZ, LX, and HB performed the experiments. ES, MZ, LX, and HB revised the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by NSF Division of Ocean Sciences—Biological Oceanography awards 1658466 (ES), Maryland Sea Grant—Graduate student research support grant (ES, MZ). MZ and LX were supported by awards from the China Scholarship Council.

ACKNOWLEDGMENTS

We thank many managers, fishermen (John Rocchetta and Dominick DiMartino) and collaborators for assistance in collecting and processing samples, especially Bruce Estrella (MA Division of Marine Fisheries), Christopher Paparo (Stony Brook University, School of Marine and Atmospheric Sciences), and Katie Rodrigue and Patrick Barrett (Rhode Island Department of Environmental Management). We thank Gretchen Messick (Cooperative Oxford Laboratory, NOAA) for assistance in diagnoses and description of histology specimens. We thank USDA scientists Janet Warg (APHIS) and John Neil (ARS) who generated preliminary NGS sequences of the mysterious Agawam River dsRNA; Olivia Pares (UMCES-Institute of Marine Environmental Technology) and Jennifer Akanoh (UMBC) for their help with preliminary experiments. We are very grateful to Ten-Tsao Wong (UMBC-IMET) for guidance on RACE. We also thank the Electron Microscopy Core Imaging Facility and Histology core at University of Maryland Baltimore (UMB) for helping with TEM and Histology. We appreciate the thoughtful feedback of the the manuscript reviewers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.855750/full#supplementary-material

REFERENCES

Abreu, E. F., Daltro, C. B., Nogueira, E. O., Andrade, E. C., and Aragao, F. J. (2015). Sequence and genome organization of papaya meleira virus infecting papaya in Brazil. *Arch. Virol.* 160, 3143–3147. doi: 10.1007/s00705-015-2605-x

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95, 631–636. doi: 10.1890/13-1452.1

Baeza, M., Bravo, N., Sanhueza, M., Flores, O., Villarreal, P., and Cifuentes, V. (2012). Molecular characterization of totiviruses in *Xanthophyllomyces dendrorhous*. *Virol. J.* 9:140. doi: 10.1186/1743-422X-9-140

Bekaert, M., and Rousset, J. P. (2005). An extended signal involved in eukaryotic−1 frameshifting operates through modification of the E site tRNA. *Mol. Cell* 17, 61–68. doi: 10.1016/j.molcel.2004.12.009

Bokharaei-Salim, F., Esteghamati, A., Khanaliha, K., Esghaei, M., Donyavi, T., and Salemi, B. (2020). The first detection of co-infection of double-stranded RNA virus 1, 2 and 3 in Iranian isolates of *Trichomonas vaginalis*. *Iran. J. Parasitol.* 15, 357–363. doi: 10.18502/ijpa.v15i3.4200

Bowers, H. A., Messick, G. A., Hanif, A., Jagus, R., Carrion, L., Zmora, O., et al. (2010). Physicochemical properties of double-stranded RNA used to discover a reo-like virus from blue crab *Callinectes sapidus*. *Dis. Aquat. Org.* 93, 17–29. doi: 10.3354/dao02280

Byun, Y., and Han, K. (2006). PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res.* 34, W416–W422. doi: 10.1093/nar/gkl210

Chen, S., Huang, Q., Wu, L., and Qian, Y. (2015). Identification and characterization of a maize-associated *mastrevirus* in China by deep sequencing small RNA populations. *Virol. J.* 12, 156–159. doi: 10.1186/s12985-015-0384-3

Chiapello, M., Rodríguez-Romero, J., Ayllón, M. A., and Turina, M. (2020). Analysis of the virome associated to grapevine downy mildew lesions reveals new mycovirus lineages. *Virus Evol.* 6:veaa058. doi: 10.1093/ve/veaa058

Dantas, M. D. A., Chavante, S. F., Teixeira, D. I. A., Lima, J. P. M., and Lanza, D. C. (2015). Analysis of new isolates reveals new genome organization and a hypervariable region in infectious myonecrosis virus (IMNV). *Virus Res.* 203, 66–71. doi: 10.1016/j.virusres.2015.03.015

Dinman, J. D., Icho, T., and Wickner, R. B. (1991). A-1 ribosomal frameshift in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. *Proc. Natl. Acad. Sci.* 88, 174–178. doi: 10.1073/pnas.88.1.174

Dragičević, P., Grbin, D., Maguire, I., Blažević, S. A., Abramović, L., Tarandek, A., et al. (2021). Immune response in crayfish is species-specific and exhibits changes along invasion range of a successful invader. *Biology* 10:1102.

Edgerton, B., Owens, L., Giasson, B., and De Beer, S. (1994). Description of a small dsRNA virus from freshwater crayfish *Cherax quadricarinatus*. *Dis. Aquat. Org.* 18, 63–69. doi: 10.3354/dao018063

Engering, A., Hogerwerf, L., and Slingenbergh, J. (2013). Pathogen–host–environment interplay and disease emergence. *Emerg. Microbe, Infect.* 2, 1–7. doi: 10.1038/emi.2013.5

Flowers, E. M., Johnson, A. F., Aguilar, R., and Schott, E. J. (2018). Prevalence of the pathogenic crustacean virus *Callinectes sapidus* reovirus 1 near flow-through blue crab aquaculture in Chesapeake Bay, USA. *Dis. Aquat. Org.* 129, 135–144. doi: 10.3354/dao03232

Flowers, E. M., Simmonds, K., Messick, G. A., Sullivan, L., and Schott, E. J. (2016). PCR-based prevalence of a fatal reovirus of the blue crab, *Callinectes sapidus* (Rathbun) along the northern Atlantic coast of the USA. *J. Fish Dis.* 39, 705–714. doi: 10.1111/jfd.12403

Ghabrial, S. A., and Suzuki, N. (2009). Viruses of plant pathogenic fungi. *Annu. Rev. Phytopathol.* 47, 353–384. doi: 10.1146/annurev-phyto-080508-081932

Goodman, R. P., Ghabrial, S. A., Fichorova, R. N., and Nibert, M. L. (2011). Trichomonasvirus: a new genus of protozoan viruses in the family *Totiviridae*. *Arch. Virol.* 156, 171–179. doi: 10.1007/s00705-010-0832-838

Gosner, K.L., (1978). *A Field Guide to the Atlantic Seashore. The Peterson field guide series*. Boston, Massachusetts: Houghton Mifflin Co.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Haugland, Ø., Mikalsen, A. B., Nilsen, P., Lindmo, K., Thu, B. J., Eliassen, T. M., et al. (2011). Cardiomyopathy syndrome of Atlantic salmon (*Salmo salar* L.) is caused by a double-stranded RNA virus of the *Totiviridae* family. *J. Virol.* 85, 5275–5286. doi: 10.1128/JVI.02154-10

Huang, C. H., Lu, C. L., and Chiu, H. T. (2005). A heuristic approach for detecting RNA H-type pseudoknots. *Bioinformatics* 21, 3501–3508. doi: 10.1093/bioinformatics/bti568

Johnson, P. T. (1978). Viral diseases of the blue crab, *Callinectes sapidus*. *Mar. Fish. Rev.* 40, 13–15.

Johnson, P. T. (1983). "Diseases caused by viruses, rickettsiae, bacteria, and fungi," in *The Biology of Crustacea, 6, Pathology*. ed. A. J. Provenzano (Academic Press: New York), 1–78.

Johnson, P. T. (1984). Viral diseases of marine invertebrates. *Helgoländer Meeresun.* 37, 65–98. doi: 10.1007/BF01989296

Johnson, D. S. (2015). The savory swimmer swims north: a northern range extension of the blue crab *Callinectes sapidus*? *J. Crustac. Biol.* 35, 105–110. doi: 10.1163/1937240X-00002293

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Khalifa, M. E., and MacDiarmid, R. M. (2019). A novel totivirus naturally occurring in two different fungal genera. *Front. Microbiol.* 10:2318. doi: 10.3389/fmicb.2019.02318

King, A. M. Q., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J. (2011). *Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego, California: Elsevier Academic Press.

Kirch, W (ed.) (2008). *Pearson's correlation coefficient, 1090–1091*. Dordrecht: Springer

Kondo, H., Hisano, S., Chiba, S., Maruyama, K., Andika, I. B., Toyoda, K., et al. (2016). Sequence and phylogenetic analyses of novel totivirus-like double-stranded RNAs from field-collected powdery mildew fungi. *Virus Res.* 213, 353–364. doi: 10.1016/j.virusres.2015.11.015

Koyama, S., Sakai, C., Thomas, C. E., Nunoura, T., and Urayama, S. I. (2016). A new member of the family *Totiviridae* associated with arboreal ants (*Camponotus nipponicus*). *Arch. Virol.* 161, 2043–2045. doi: 10.1007/s00705-016-2876-x

Koyama, S., Urayama, S. I., Ohmatsu, T., Sassa, Y., Sakai, C., Takata, M., et al. (2015). Identification, characterization and full-length sequence analysis of a novel dsRNA virus isolated from the arboreal ant *Camponotus yamaokai*. *J. Gen. Virol.* 96, 1930–1937. doi: 10.1099/vir.0.000126

Lauber, C., Seifert, M., Bartenschlager, R., and Seitz, S. (2019). Discovery of highly divergent lineages of plant-associated astro-like viruses sheds light on the emergence of potyviruses. *Virus Res.* 260, 38–48. doi: 10.1016/j.virusres.2018.11.009

Lightner, D. V., Pantoja, C. R., Poulos, B. T., Tang, K. F. J., Redman, R. M., Andrade, T. P. D., et al. (2004). Infectious myonecrosis: new disease in Pacific white shrimp. *Glob. Aquac. Advocat.* 7:85.

Luna, L.G. (1968). *Manual of Histological Staining Methods of the Armed Forces Institute of Pathology*. 3rd Edn. New York: McGraw-Hill Book Company, 1–258

Macedo, D., Caballero, I., Mateos, M., Leblois, R., McCay, S., and Hurtado, L. A. (2019). Population genetics and historical demographic inferences of the blue crab *Callinectes sapidus* in the US based on microsatellites. *PeerJ* 7:e7780. doi: 10.7717/peerj.7780

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C. M., Roumagnac, P., van der Vlugt, R., et al. (2020). Illuminating an ecological blackbox: using high throughput sequencing to characterize the plant virome across scales. *Front. Microbiol.* 11:578064. doi: 10.3389/fmicb.2020.578064

Mancinelli, G., Bardelli, R., and Zenetos, A. (2021). A global occurrence database of the Atlantic blue crab *Callinectes sapidus*. *Sci. Data* 8, 111–110. doi: 10.1038/s41597-021-00888-w

Millikin, M.R. (1984). Synopsis of Biological Data on the Blue Crab, Callinectes *Sapidus* Rathbun (No. 138). National Oceanic and Atmospheric Administration, National Marine Fisheries Service.

Munn, C. B. (2006). Viruses as pathogens of marine organisms-from bacteria to whales. *J. Mar. Biol.* 86, 453–467. doi: 10.1017/S002531540601335X

Naim, S., Brown, J. K., and Nibert, M. L. (2014). Genetic diversification of penaeid shrimp infectious myonecrosis virus between Indonesia and Brazil. *Virus Res.* 189, 97–105. doi: 10.1016/j.virusres.2014.05.013

Nibert, M. L. (2007). '2A-like' and 'shifty heptamer' motifs in penaeid shrimp infectious myonecrosis virus, a monosegmented double-stranded RNA virus. *J. Gen. Virol.* 88, 1315–1318. doi: 10.1099/vir.0.82681-0

NOAA (2020) NOAA Landings. Available at: https://foss.nmfs.noaa.gov (Accessed November 24, 2021).

Park, Y., James, D., and Punja, Z. K. (2005). Co-infection by two distinct totivirus-like double-stranded RNA elements in *Chalara elegans* (*Thielaviopsis basicola*). *Virus Res.* 109, 71–85. doi: 10.1016/j.virusres.2004.10.011

Pettersson, J. H. O., Shi, M., Eden, J. S., Holmes, E. C., and Hesson, J. C. (2019). Meta-transcriptomic comparison of the RNA viromes of the mosquito vectors *Culex pipiens* and *Culex torrentium* in northern Europe. *Viruses* 11:1033. doi: 10.3390/v11111033

Piers, H. (1920). The blue crab (*Callinectes sapidus* Rathbun): extension of its range northward to near Halifax, Nova Scotia. *Proc. Nova Scot, Instit. Sci.* 15, 1918–1922.

Poulos, B., Tang, K., Pantoja, C., Bonami, J. R., and Lightner, D. (2006). Purification and characterization of infectious myonecrosis virus of penaeid shrimp. *J. Gen. Virol.* 87, 987–996. doi: 10.1099/vir.0.81127-0

Preisig, O., Wingfield, B. D., and Wingfield, M. J. (1998). Coinfection of a fungal pathogen by two distinct double-stranded RNA viruses. *Virology* 252, 399–406. doi: 10.1006/viro.1998.9480

R Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rice, N. R., Stephens, R. M., Burny, A., and Gilden, R. V. (1985). The gag and pol genes of bovine leukemia virus: nucleotide sequence and analysis. *Virology* 142, 357–377. doi: 10.1016/0042-6822(85)90344-90347

Rodriguez-Ezpeleta, N., Teijeiro, S., Forget, L., Burger, G., and Lang, B. F. (2009). Construction of cDNA libraries: focus on protists and fungi. *Methods Mol. Biol.* 533, 33–47. doi: 10.1007/978-1-60327-136-3_3

Senapin, S., Phewsaiya, K., Briggs, M., and Flegel, T. W. (2007). Outbreaks of infectious myonecrosis virus (IMNV) in Indonesia confirmed by genome sequencing and use of an alternative RT-PCR detection method. *Aquaculture* 266, 32–38. doi: 10.1016/j.aquaculture.2007.02.026

Shao, Q., Jia, X., Gao, Y., Liu, Z., Zhang, H., Tan, Q., et al. (2021). Cryo-EM reveals a previously unrecognized structural protein of a dsRNA virus implicated in its extracellular transmission. *PLoS Pathog.* 17:e1009396. doi: 10.1371/journal.ppat.1009396

Shi, M., Lin, X. D., Tian, J. H., Chen, L. J., Chen, X., Li, C. X., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature* 540, 539–543. doi: 10.1038/nature20167

Shields, J. D., Williams, J. D., and Boyko, C. B. (2015). "Parasites and diseases of Brachyura," in *Treatise on Zoology-Anatomy, Taxonomy, Biology. The Crustacea. Part C, Vol. 9*. (Leiden, Netherlands: Brill), 639–774.

Spitznagel, M. I., Small, H. J., Lively, J. A., Shields, J. D., and Schott, E. J. (2019). Investigating risk factors for mortality and reovirus infection in aquaculture production of soft-shell blue crabs (*Callinectes sapidus*). *Aquaculture* 502, 289–295. doi: 10.1016/j.aquaculture.2018.12.051

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812. doi: 10.1038/nrmicro1750

Tang, J., Ochoa, W. F., Sinkovits, R. S., Poulos, B. T., Ghabrial, S. A., Lightner, D. V., et al. (2008). Infectious myonecrosis virus has a totivirus-like, 120-subunit capsid, but with fiber complexes at the fivefold axes. *Proc. Natl. Acad. Sci.* 105, 17526–17531. doi: 10.1073/pnas.0806724105

Urayama, S. I., Takaki, Y., and Nunoura, T. (2016). FLDS: a comprehensive dsRNA sequencing method for intracellular RNA virus surveillance. *Microbes Environ.* 31, 33–40. doi: 10.1264/jsme2.ME15171

Vignuzzi, M., and López, C. B. (2019). Defective viral genomes are key drivers of the virus–host interaction. *Nat. Microbiol.* 4, 1075–1087. doi: 10.1038/s41564-019-0465-y

Vorimore, F., Hsia, R. C., Huot-Creasy, H., Bastian, S., Deruyter, L., Passet, A., et al. (2013). Isolation of a new chlamydia species from the feral sacred ibis (*Threskiornis aethiopicus*): *chlamydia ibidis*. *PLoS One* 8:e74823. doi: 10.1371/journal.pone.0074823

Wickner, R. B., Ghabrial, S. A., Nibert, M. L., Patterson, J. L., and Wang, C. C. (2011). "Family Totiviridae," in *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*. eds. A. M. Q. King, M. J. Adams, E. B. Carstens and E. J. Lefkowits (Tokyo: Elsevier Academic Press), 639–650.

Williams, S. H., Levy, A., Yates, R. A., Somaweera, N., Neville, P. J., Nicholson, J., et al. (2020). The diversity and distribution of viruses associated with *Culex annulirostris* mosquitoes from the Kimberley region of western Australia. *Viruses* 12:717. doi: 10.3390/v12070717

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E. C., Li, W. X., et al. (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci.* 107, 1606–1611. doi: 10.1073/pnas.0911353107

Xie, J., Havens, W. M., Lin, Y. H., Suzuki, N., and Ghabrial, S. A. (2016). The victorivirus *Helminthosporium victoriae* virus 190S is the primary cause of disease/hypovirulence in its natural host and a heterologous host. *Virus Res.* 213, 238–245. doi: 10.1016/j.virusres.2015.12.011

Zhai, Y., Attoui, H., Jaafar, F. M., Wang, H. Q., Cao, Y. X., Fan, S. P., et al. (2010). Isolation and full-length sequence analysis of *Armigeres subalbatus* totivirus, the first totivirus isolate from mosquitoes representing a proposed novel genus (*Artivirus*) of the family *Totiviridae*. *J. Gen. Virol.* 91, 2836–2845. doi: 10.1099/vir.0.024794-0

Zhang, P., Liu, W., Cao, M., Massart, S., and Wang, X. (2018). Two novel totiviruses in the white-backed planthopper, *Sogatella furcifera*. *J. Gen. Virol.* 99, 710–716. doi: 10.1099/jgv.0.001052

Zhang, S., Shi, Z., Zhang, J., and Bonami, J. R. (2004). Purification and characterization of a new reovirus from the Chinese mitten crab, *Eriocheir sinensis*. *J. Fish Dis.* 27, 687–692. doi: 10.1111/j.1365-2761.2004.00587.x

Zhao, M., Behringer, D. C., Bojko, J., Kough, A. S., Plough, L., dos Santos Tavares, C. P., et al. (2020). Climate and season are associated with prevalence and distribution of trans-hemispheric blue crab reovirus (*Callinectes sapidus* reovirus 1). *Mar. Ecol. Prog. Ser.* 647, 123–133. doi: 10.3354/meps13405

Zhao, M., dos Santos Tavares, C. P., and Schott, E. J. (2021a). Diversity and classification of reoviruses in crustaceans: a proposal. *J. Invertebr. Pathol.* 182:107568. doi: 10.1016/j.jip.2021.107568

Zhao, M., Flowers, E. M., and Schott, E. J. (2021b). Near-complete sequence of a highly divergent Reovirus genome recovered from *Callinectes sapidus*. *Microbiol. Resour. Announc.* 10, e01278–e01320. doi: 10.1128/MRA.01278-20

Zheng, L., Lu, X., Liang, X., Jiang, S., Zhao, J., Zhan, G., et al. (2017). Molecular characterization of novel totivirus-like double-stranded RNAs from *Puccinia striiformis* f. sp. tritici, the causal agent of wheat stripe rust. *Front. Microbiol.* 8:1960. doi: 10.3389/fmicb.2017.01960

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

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

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 31 March 2022
doi: 10.3389/fmicb.2022.866382

Enterococcus faecalis Bacteriophage vB_EfaS_efap05-1 Targets the Surface Polysaccharide and ComEA Protein as the Receptors

Lingqiong Huang[1,2†], Wenqiong Guo[3†], Jiahui Lu[4], Wuliang Pan[5], Fuqiang Song[6] and Peng Wang[1]**

[1] Yunnan Provincial Key Laboratory for Zoonosis Control and Prevention, Yunnan Institute of Endemic Diseases Control and Prevention, Dali, China, [2] School of Public Health, Dali University, Dali, China, [3] School of Nursing, Chengdu Medical College, Chengdu, China, [4] School of Clinical Medicine, Chengdu Medical College, Chengdu, China, [5] School of Pharmacy, Chengdu Medical College, Chengdu, China, [6] Department of Medical Laboratory, The General Hospital of Western Theater Command, Chengdu, China

Enterococcus faecalis is a Gram-positive opportunistic pathogen that causes nosocomial infections in humans. Due to the growing threat of antibiotic resistance of *E. faecalis,* bacteriophage therapy is a promising option for treating of *E. faecalis* infection. Here, we characterized a lytic *E. faecalis* bacteriophage vB_EfaS_efap05-1 with a dsDNA genome of 56,563 bp. Phage vB_EfaS_efap05-1 had a prolate head and a tail, and belongs to Saphexavirus which is a member of *Siphoviridae*. Efap05-1 uses either surface polysaccharide or membrane protein ComEA as the receptor because the mutation of both genes (*ComEA* and UDP-glucose 4-epimerase *galE*) prevents phage adsorption and leads to phage resistance, and complementation of *ComEA* or *galE* could recover its phage sensitivity. Our results provided a comprehensive analysis of a new *E. faecalis* phage and suggest efap05-1 as a potential antimicrobial agent.

Keywords: bacteriophage, phage receptor, *Enterococcus faecalis*, exopolysaccharide, ComEA protein

OPEN ACCESS

Edited by:
Mao Ye,
Institute of Soil Science (CAS), China

Reviewed by:
Mengzhe Li,
Beijing University of Chemical Technology, China
Xiaoxiao Liu,
South China Sea Institute of Oceanology (CAS), China

****Correspondence:***
Fuqiang Song
349657161@qq.com
Peng Wang
wp030801@126.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *31 January 2022*
Accepted: *08 March 2022*
Published: *31 March 2022*

Citation:
Huang L, Guo W, Lu J, Pan W, Song F and Wang P (2022) Enterococcus faecalis Bacteriophage vB_EfaS_efap05-1 Targets the Surface Polysaccharide and ComEA Protein as the Receptors. Front. Microbiol. 13:866382. doi: 10.3389/fmicb.2022.866382

INTRODUCTION

Enterococcus faecalis is a gram-positive bacterium that could cause intestinal dysbiosis or infections in humans, such as nosocomial sepsis, urinary tract, and surgical site infections (Hayakawa et al., 2013; Jahansepas et al., 2018). In addition, the cytolysin-positive *E. faecalis* strains are correlated with mortality in patients with alcoholic hepatitis (Duan et al., 2019). However, the emergence of antibiotic resistance, especially vancomycin and daptomycin resistance, is especially troubling (Hu et al., 2018). Thus, new therapeutic approaches are needed to treat *E. faecalis* infections.

Bacteriophages are viruses that infect bacteria and are promising agents for antimicrobial treatment (Kortright et al., 2019; Wu et al., 2021). Recently, phage therapy clinical trials are initiated in many countries, and the number of case reports describing patients being treated increased significantly word-wide. For example, COVID-19 patients with carbapenem-resistant *Acinetobacter baumannii* infection were treated with a pre-optimized 2-phage cocktail and the infection was significantly relieved (Wu et al., 2021). However, phage resistance is a potential barrier to successful phage therapy (Egido et al., 2021). To confront this issue, the molecular mechanisms of phage resistance, as well as the genomic and biological characteristics of a phage, should be studied to provide the foundation for rational selection of the phages for therapy (Pires et al., 2020).

In this study, we isolated a new *E. faecalis* bacteriophage vB_EfaS_efap05-1 with a dsDNA genome of 56,563 bp. It belongs to Saphexavirus which is a member of *Siphoviridae*, and efap05-1 could use either surface polysaccharide or membrane protein ComEA as the receptor to adsorb to the bacterial surface. Thus, phage-resistant mutants had both mutations of *ComEA* and *galE*. In summary, this study provided a detailed characterization of an *E. faecalis* bacteriophage and suggests efap05-1 as a potential antimicrobial agent.

EXPERIMENTAL PROCEDURES

Bacterial Strains and Phages

Enterococcus strains were collected from the Department of Clinical Laboratory Medicine and were grown aerobically on Brain-Heart Infusion Broth (BHI) broth at 37°C with shaking.

Bacteriophage was isolated from hospital sewage as previously described (Yang et al., 2016). Briefly, the sewage was pelleted, and the supernatant was filtered through a 0.22 μm-pore-size filter. Then, 500 μl of the sample was immediately mixed with 200 μl of bacterial culture, and 4 ml of molten BHI soft agar (0.4%) was added and poured onto BHI agar plates. After overnight culture, the formed plaque was picked, deposited in 1 ml of BHI, followed by a 10-fold dilution and double-layer agar assay to purify the phage. Then, one plaque from the third round of the purification process was picked for this study.

Transmission Electron Microscopy

Phage particles were dropped on carbon-coated copper grids for 10 min. Then, the grids were stained with phosphotungstic acid (pH 7.0) for 15 s. The sample was examined under a Philips EM 300 electron microscope. The sizes of the phage were measured using AxioVision LE based on five randomly selected images.

Phage Titering and MOI Experiment

The double-layer agar plate assay was used to calculate the phage titer. Briefly, 10-fold dilutions of phage solution were mixed with 200 μl of host bacteria, then mixed with 4 ml of molten BHI broth with 0.4% agar. Then, pour the mixture on a 1.5% agar plate. After overnight incubation at 37°C, the number of plaques was calculated as a plaque-forming unit (pfu). MOI experiments were performed by mixing log-phase bacteria (OD600 = 0.6) with a different number of phages, and the titer in the coculture was calculated using a double-layer agar plate assay after 5 h.

One-Step Growth

The one-step growth curve of efap05-1 was determined as described (Zhong et al., 2020). Briefly, 1 ml of log-phase bacteria and 1 ml of efap05-1 were mixed at an MOI of 10 and incubated at 37°C for 10 min. Then, the mixture was centrifuged for 1 min at a speed of 10,000 × *g*, and the pellet was resuspended in 6 ml of BHI medium. And samples were taken at the given time points, which are immediately pelleted and phage titer in the supernatant was measured immediately.

Adsorption Rate Experiments

Bacteriophage adsorption assay with various time points was performed as previously described (Al-Zubidi et al., 2019). Briefly, the log phase bacterial cultures were pelleted and resuspended in medium to a final concentration of 3×10^8 CFU/ml. Then, phage was added to a final titer of 3×10^6 pfu/ml. The samples were cultured at 37°C for 10 min, and the phage titer in the supernatant were measured using the double-agar plating assays. The adsorption rate was calculated as (the original phage titer—the remaining phage titer)/the original phage titer.

Determination of Host Range

Ten *E. faecalis* strains were selected to test the host range of efap05-1 through spot testing by dropping 1 μl of phage onto the double-layer soft agar premixed with the bacterial and cultured at 37°C for overnight. The formation of a clear plaque is considered as sensitive to phage efap05-1 infection.

EOP Assay

Two microliter of serial 10-fold dilutions of phage efap05-1 were spotted on double layer agar plates containing a bacterial host. The number of plaques observed after overnight incubation were compared to the number obtained on the strain efa05.

Genome Sequencing and Annotation

The phage DNA extraction is performed as previously described (Khan et al., 2021). Then, phage genomic DNA was sequenced using an Illumina Hiseq 2,500 platform (~1 Gbp/sample). Fastp (Chen et al., 2018) was used for adapter trimming and filtering the raw reads. The data were assembled using the *de novo* assembly algorithm Newbler Version2.9 with default parameters, and the assembled genome was annotated using RAST (Overbeek et al., 2014). The DNA and protein sequences were checked for homologs with BLAST manually. The genome map was drawn by SnapGene 4.1.8. The sequence data is available in the NCBI under accession number OL505085.

Stability Studies

The stability of phage under various conditions was tested by treating 10^9 pfu of phage under different pH (pH 2–13), temperature (0°C, 30°C, 40°C, 50°C, and 60°C), or chloroform concentration (10%, 25%, 50%, 75%, and 95%) for 60 min, the then the titer of the phage was determined by double-layer agar assay.

Selection of the Phage-Resistant Mutants

The phage resistant mutants were selected as previously described (Shen et al., 2018). The log phase bacteria were mixed with phage efap05-1 and cultured until the bacteria was lysed. Then, the lysate was inoculated onto the BHI agar. After overnight incubation, the single colonies were checked for its resistance against phage using the double-layer agar assay, which confirmed that all the colonies on the plates are resistant to phage infection.

Bacterial Genome Sequencing

The wide type strain efa05 and phage resistant mutant strain efa05R were selected for sequencing. Bacterial genomic DNA was extracted using UNlQ-10 Column Bacterial Genomic DNA Isolation Kit (sangon bitotec: SK1202), and then sent to Novogene Corporation for sequencing using the Illumina Hiseq 2,500 platform. Trimmomatic was used to remove adapter sequences and low-quality bases (Bolger et al., 2014). BWA was used to map clean reads to the reference genome sequence of efa05. Samtools (Li et al., 2009) was then used to prepare the data for use with the Integrative Genomics Viewer (IGV). DNA mutation locations were manually checked with IGV and SeqKit (Shen et al., 2016).

Complementation of *ComEA* and *galE*

The *ComEA* gene and plasmid pMGP23:mCherry were amplified by PCR (the primers for the amplification of *ComEA* gene and the plasmid pMGP23:mCherry are listed in **Table 1**), and the PCR products were purified. The *ComEA* and plasmid were ligated by Gibson assembly to generate pMG-ComEA, and the constructed plasmid was comfired by sanger sequencing. The efa05R complementation strain was generated by electroporation of pMG-ComEA into strain efa05R followed by selection on BHI agar erythromycin (20ug/ml). The *galE* gene was complemented with the same protocol.

Statistical Analysis

All the experiments were performed three times. The statistical analysis was performed using One-way ANOVA or *t*-test, and statistical significance was assumed if the value of *p* was <0.05.

RESULTS

The Biological Characterization of an *Enterococcus faecalis* Phage

An *E. faecalis* phage was isolated using plaque assay. It forms a clear plaque on the host strain efa05 in the double layer agar plates (**Figure 1A**). The phage particle was observed by transmission electron microscopy. It is a non-enveloped, head-tail structural particle. The prolate head is approximately 100 nm in length and 40 nm in width (**Figure 1B**). Thus, based on the morphology, phage vB_EfaS_efap05-1 belongs to Saphexavirus which is a member of *Siphoviridae*. The optimal multiplicity of infection (MOI) was 0.001, and the phage titer could reach approximately 8*10^{10} pfu/ml (**Figure 1C**). The one-step growth curve of efap05-1 indicates that this phage replicates quickly with a lysis period of about 20 min, and the phage titer approached plateau after 20 min (**Figure 1D**), and the burst size was estimated as about 20 pfu per bacterium.

TABLE 1 | Bacterial strains, phages, plasmids, and primers used in this study.

Names	Characteristics and descriptions	Source
Enterococcus faecalis		
efa01	Human blood isolate	This study
efa02	Human blood isolate	
efa03	Human blood isolate	
efa04	Human blood isolate	
efa05	Human urine isolate	
efa06	Human urine isolate	
efa07	Human urine isolate	
efa08	Human urine isolate	
efa09	Human urine isolate	
efa10	Human urine isolate	
efa05R	phage resistant mutant	
efa05R::*ComEA*	Complementation of *ComEA* strain	
efa05R::*galE*	Complementation of *galE* strain	
Phage		
efap05-1	*Siphoviridae*; isolated from sewage	This study
Plasmids		
pMGP23	Modified from plasmid pMG36e, which contains erythromycin and kanamycin resistance gene	This study
pMG-ComEA	pMGP23 expressing *ComEA* from the native promoter	This study
pMG-galE	pMGP24 expressing *galE* from the native promoter	This study
Primers		
ComEA-F	aaaatattcggaggaattttgaaatggattggttgaaacagttac	
ComEA-R	atatcgtagcgccggttaacggtaaattctaatagcattctctttaca	
galE-F	atatcgtagcgccggtcatctctgcttaccttccg	
galE-R	aaaatattcggaggaattttgaagtggaatcatttctaatcacagg	
pMGP23-F	ccggcgctacgatatt	
pMGP23-R	ttcaaaattcctccgaat	

TABLE 2 | The host range of phage efap05-1.

Strain	Origin	LG1 sensitivity
Enterococcus faecalis efa01	Blood	–
Enterococcus faecalis efa02	Blood	–
Enterococcus faecalis efa03	Blood	+
Enterococcus faecalis efa04	Blood	–
Enterococcus faecalis efa05	Urine	+
Enterococcus faecalis efa06	Urine	+
Enterococcus faecalis efa07	Urine	–
Enterococcus faecalis efa08	Urine	–
Enterococcus faecalis efa09	Urine	+
Enterococcus faecalis efa10	Urine	+

+ Indicates the strain is sensitive to phage efap05-1, and EOP ranges between 0.001 and 1.

The host range of efap05-1 was estimated by EOP (efficiency of plating) assays. Ten clinically isolated *E. faecalis* strains were tested and five strains could be lysed by efap05-1, indicating a modest host range (**Table 2**).

Stability of efap05-1

The stability of efap05-1 under various conditions was tested. It could maintain stability under pH 5–10, and other pH solutions could impair the viability of efap05-1 (**Figure 2A**).

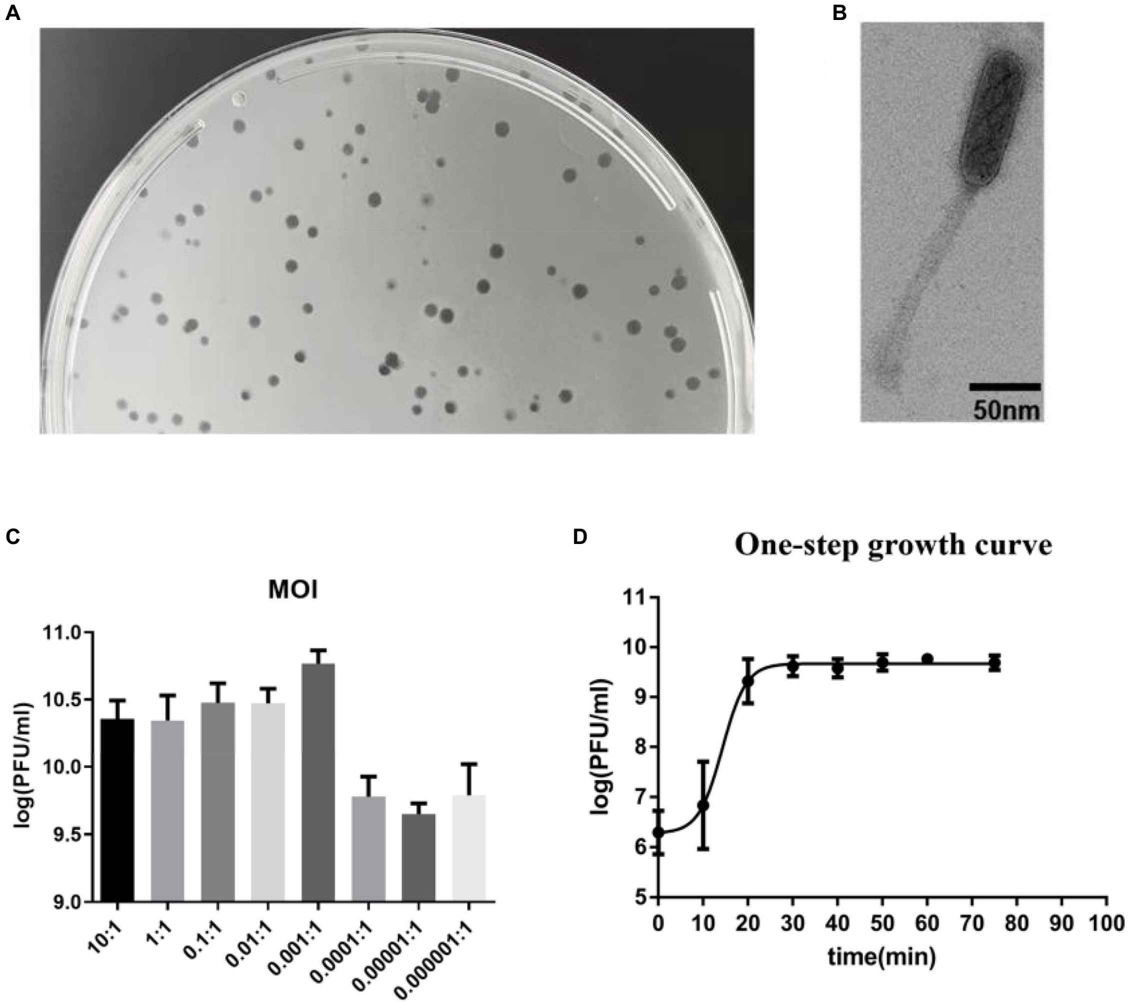


FIGURE 1 | Biological characterization of *E. faecalis* phage vB_EfaS_efap05-1. **(A)** Phage efap05-1 forms clear plaques on the agar plate. **(B)** The transmission electron micrograph reveals that efap05-1 is a Saphexavirus which is a member of Siphoviridae. **(C)** The optimal MOI of phage efap05-1 is 0.001. **(D)** The one-step growth curve of efap05-1.

And efap05-1 is stable under 50°C because its titer was not changed after 60 min incubation at 50°C (**Figure 2B**), and is completely inactivated over 70°C. Besides, chloroform treatment did not affect the viability of phage efap05-1, indicating that it is a non-enveloped phage (**Figure 2C**).

Genome Sequence Analysis of an *Enterococcus faecalis* Phage

Phage efap05-1 is a double-stranded (ds) DNA phage with a linear genome of 56,564 base pairs (bp). Its G+C content is 40% and encodes 99 ORFs and one tRNA (**Figure 3**), which are predicted by RAST (Overbeek et al., 2014) and visualized by SnapGene.

Most of the ORFs are functionally unknown, and 22 ORFs are functionally annotated, which can be categorized into several functional modules, including phage DNA replications, lysis, phage structural protein (**Figure 3**). However, efap05-1 did not encode any antibiotic-resistant gene or virulence gene, indicating that it is a safe candidate for phage therapy.

Phage Resistant Mutant Contains Two Mutations

To study the phage resistant mechanism of *E. faecalis* efa05 against phage efap05-1, we mixed phage with host and cultured until the bacteria are lysed. And then inoculate the lysate on BHI agar plate. One phage-resistant mutant efa05R was selected, and its resistance against efap05-1 was confirmed by EOP experiment (**Figure 4B**).

Then, efa05R and wild-type strain efa05 was sent for whole-genome sequencing, and the mutation sites of these two strains were detected as Late competence protein *ComEA* and UDP-glucose 4-epimerase *galE* (**Figure 4A**). ComEA is a cell membrane protein that binds to the double-stranded DNA and initiates the DNA uptake process (Burghard-Schrod et al., 2021). And UDP-glucose 4-epimerase is involved in the

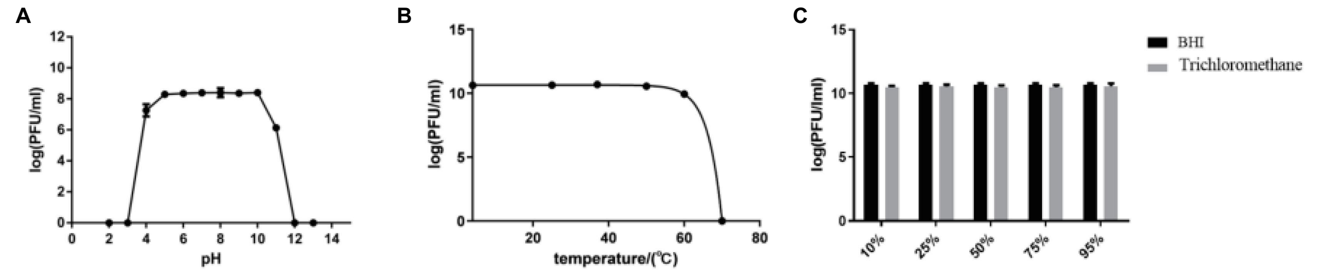


FIGURE 2 | Stability of efap05-1: **(A)** Phage efap05-1 is stable under pH5 ~ 10, and is completely inactived under pH3. **(B)** Phage efap05-1 is inactivated by 70°C treatment. **(C)** Phage efap05-1 is resistant to chloroform because the titer was stable after chloroform treatment.

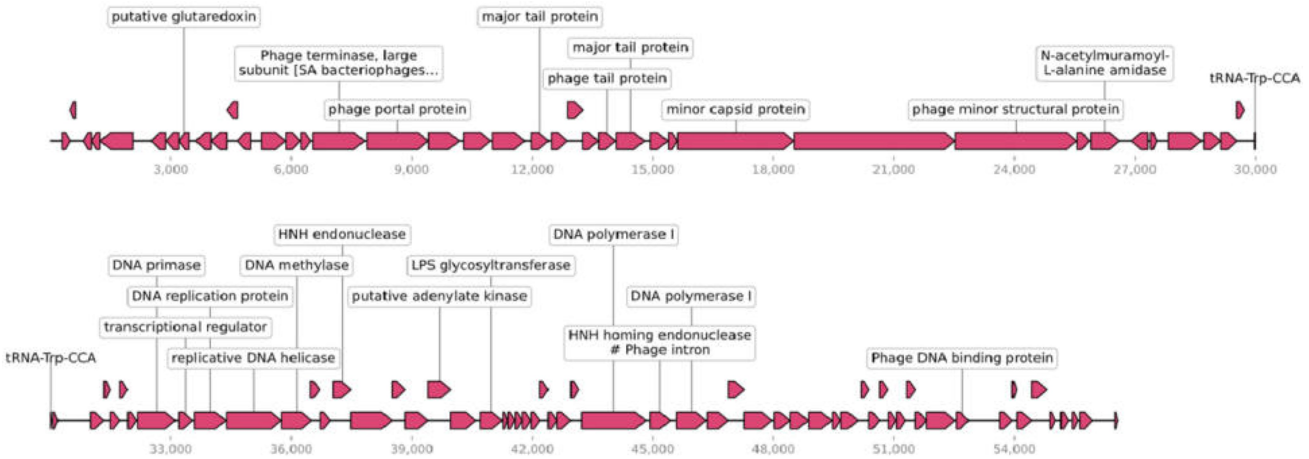


FIGURE 3 | Genomic characterization of efap05-1. Phage efap05-1 is a dsDNA phage that encodes 99 predicted proteins and one tRNA.

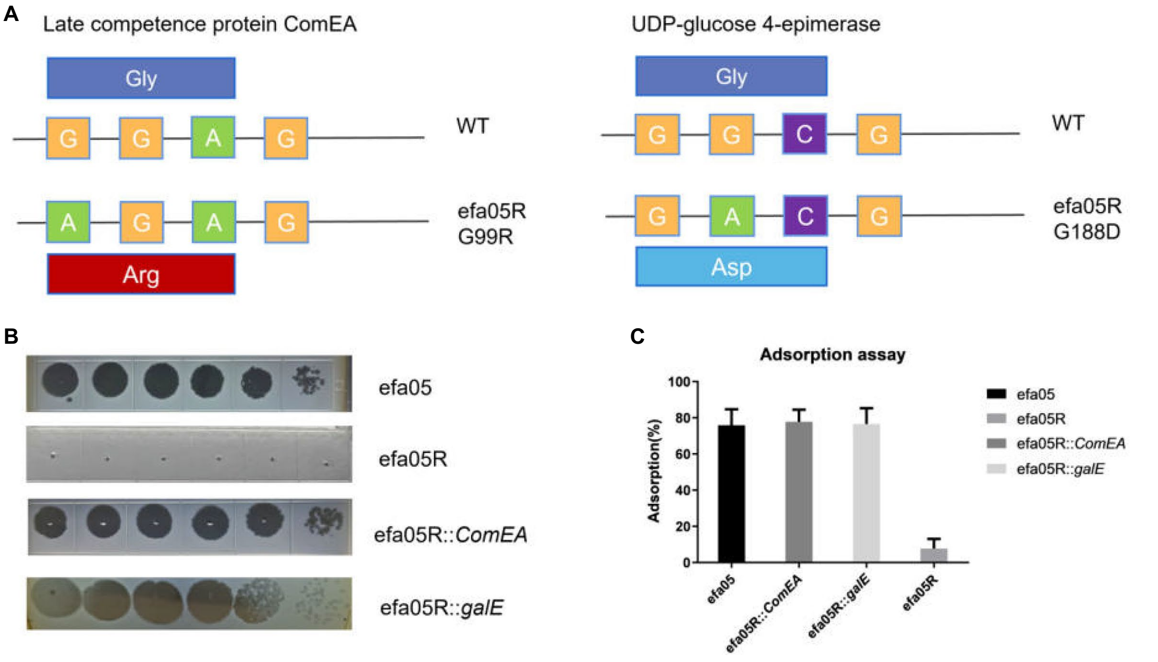


FIGURE 4 | Characterization of the phage-resistant mutant. **(A)** The mutation site in efa05R was detected as G99R in *ComEA* and G188D in *galE*. **(B)** The EOP experiment of phage against wild type strain, phage resistant efa05R, and the complemented strains. **(C)** Adsorption assay of phage onto each strain.

biosynthesis of cell wall polysaccharides (Boels et al., 2001; Lee et al., 2014).

Usually, phage resistance is selected with one key mutation site (Li et al., 2018), the mutation of both genes detected in phage resistant mutant efa05R indicates that both genes are required for phage infection. As expected, the complementation of either *ComEA* or *galE* in efa05R could recover the phage sensitivity through enabling phage adsorption (**Figures 4B,C**). Thus, these data indicate that phage could bind to either polysaccharides or protein ComEA to initiate the phage infection cycle.

DISCUSSION

Enterococcus. faecalis have developed resistance to antibiotics, including vancomycin and daptomycin (Hayakawa et al., 2013; Jahansepas et al., 2018). Thus, phage therapy is a renewed interest to treat multidrug-resistant *E. faecalis* infection. The biological and genomic characterization of a phage is essential before applications in phage therapy (Barbu et al., 2016; El Haddad et al., 2019; Pires et al., 2020). In this study, we isolated an *E. faecalis* bacteriophage vB_EfaS_efap05-1 with a prolate head. It is a completely lytic phage without the antibiotic-resistant genes or virulence genes, indicating it as a potential candidate for phage therapy.

The identification of phage resistance mechanisms is important for rational select phage or designing a phage cocktail for therapy (Labrie et al., 2010; Duerkop et al., 2016). Phage resistance is quite common and is important for phage therapy because it could lead to treatment failure. Selecting different phages that target different receptors is a rational approach in selecting phages for therapy (Yang et al., 2020). And *in vitro*, most phage resistance is selected through modifications of the receptors (Castillo et al., 2015; Duerkop et al., 2016). For example, *Pseudomonas aeruginosa* phage resistant mutants are O-antigen deficient to prevent phage adsorption (Shen et al., 2018). In this study, phage resistance mutants are selected with mutations in two genes. And the complementation of each gene could restore the phage sensitivity as well as the phage adsorption. Thus, it is reasonable to infer that phage efap05-1 uses either polysaccharides or protein ComEA as the receptors. Polysaccharides are common phage receptors for a lot of *E. faecalis* phages (Duerkop et al., 2016; Chatterjee et al., 2020).

ComEA is a membrane protein, and the loss of the ComEA decreases the binding of DNA to the competent cell surface and the internalization of DNA and impairs DNA transformability (Inamine and Dubnau, 1995). However, protein ComEA, to our knowledge, is the first report to serve as a phage receptor, which is an interesting biological phenomenon.

The limitation of this study is the lack of identification of the receptor binding proteins in the phages. Most phages use either polysaccharides or protein as receptors (Lim et al., 2021). However, it is not common that phage uses two different receptors of polysaccharides and membrane protein, because the phage tail fiber is very specifically targeting the receptors, and usually one phage tail fiber could not adsorb to two different structural receptors. And the *E.coli* phage phi92 could adsorb to both encapsulated and nonencapsulated bacteria due to the presence of four different types of tail fibers and tail spikes in the viral particles, which enable the phage to use attachment different sites on the host cell surface (Schwarzer et al., 2012). And staphylococcal Twort-like phage ΦSA012 possesses two receptor binding proteins to expand its host range (Takeuchi et al., 2016). In our study, the current data suggest that phage efap05-1 might also encode different receptor binding proteins that enable it to adsorb to both polysaccharides and membrane proteins. Three tail fiber proteins are annotated in the genome of efap05-1 (**Figure 3**), which needs further study to demonstrate the function of each tail fiber protein.

In conclusion, we isolated and characterized an *E. faecalis* phage efap05-1, which is a candidate for the development of phage cocktails or phage-antibiotic combinations treatment for *E. faecalis* infections. The characterization of the phage-resistant mutant bacterium could help to develop a cocktail to avoid phage resistance.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

PW and FS designed the research. LH, WG, JL, and WP performed the laboratory work and collected the data. LH, WG, PW, and FS wrote the first draft of the manuscript and prepared figures. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the National Natural Science Foundation of China (NSFC, 31660043 to WP) and Yunnan Medical leading talent project (L-2019001 to WP).

REFERENCES

Al-Zubidi, M., Widziolek, M., Court, E. K., Gains, A. F., Smith, R. E., Ansbro, K., et al. (2019). Identification of novel bacteriophages with therapeutic potential that target *Enterococcus faecalis*. *Infect. Immun.* 87, e00512–e00519. doi: 10.1128/IAI.00512-19

Barbu, E. M., Cady, K. C., and Hubby, B. (2016). Phage therapy in the era of synthetic biology. *Cold Spring Harb. Perspect. Biol.* 8:a023879. doi: 10.1101/cshperspect.a023879

Boels, I. C., Ramos, A., Kleerebezem, M., and de Vos, W. M. (2001). Functional analysis of the *Lactococcus lactis* galU and galE genes and their impact on sugar nucleotide and exopolysaccharide biosynthesis. *Appl. Environ. Microbiol.* 67, 3033–3040. doi: 10.1128/AEM.67.7.3033-3040.2001

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Burghard-Schrod, M., Kilb, A., Kramer, K., and Graumann, P. L. (2021). Single molecule dynamics of DNA receptor ComEA, membrane permease ComEC and taken up DNA in competent Bacillus subtilis cells. *J. Bacteriol.* doi: 10.1128/jb.00572-21 [Epub ahead of print].

Castillo, D., Christiansen, R. H., Dalsgaard, I., Madsen, L., and Middelboe, M. (2015). Bacteriophage resistance mechanisms in the fish pathogen *Flavobacterium psychrophilum*: linking genomic mutations to changes in bacterial virulence factors. *Appl. Environ. Microbiol.* 81, 1157–1167. doi: 10.1128/AEM.03699-14

Chatterjee, A., Willett, J. L. E., Nguyen, U. T., Monogue, B., Palmer, K. L., Dunny, G. M., et al. (2020). Parallel genomics uncover novel enterococcal-bacteriophage interactions. *mBio* 11, e03120–e031219. doi: 10.1128/mBio.03120-19

Chen, S., Zhou, Y., Chen, Y., and Jia, G. (2018). Fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Duan, Y., Llorente, C., Lang, S., Brandl, K., Chu, H., and Jiang, L., 2019. et al. Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. *Nature* 575, 505–511, doi: 10.1038/s41586-019-1742-x.

Duerkop, B. A., Huo, W., Bhardwaj, P., Palmer, K. L., and Hooper, L. V. (2016). Molecular basis for lytic bacteriophage resistance in Enterococci. *mBio* 7, e01304–e01316. doi: 10.1128/mBio.01304-16

Egido, J. E., Costa, A. R., Aparicio-Maldonado, C., Haas, P. J., and Brouns, S. J. J. (2021). Mechanisms and clinical importance of bacteriophage resistance. *FEMS Microbiol. Rev.* 46:fuab048. doi: 10.1093/femsre/fuab048

El Haddad, L., Harb, C. P., Gebara, M. A., Stibich, M. A., and Chemaly, R. F. (2019). A systematic and critical review of bacteriophage therapy against multidrug-resistant ESKAPE organisms in humans. *Clin. Infect. Dis.* 69, 167–178. doi: 10.1093/cid/ciy947

Hayakawa, K., Marchaim, D., Palla, M., Gudur, U. M., Pulluru, H., Bathina, P., et al. (2013). Epidemiology of vancomycin-resistant *Enterococcus faecalis*: a case-case-control study. *Antimicrob. Agents Chemother.* 57, 49–55. doi: 10.1128/AAC.01271-12

Hu, F., Zhu, D., Wang, F., and Wang, M. (2018). Current status and trends of antibacterial resistance in China. *Clin. Infect. Dis.* 67, S128–S134. doi: 10.1093/cid/ciy657

Inamine, G. S., and Dubnau, D. (1995). ComEA, a Bacillus subtilis integral membrane protein required for genetic transformation, is needed for both DNA binding and transport. *J. Bacteriol.* 177, 3045–3051. doi: 10.1128/jb.177.11.3045-3051.1995

Jahansepas, A., Ahangarzadeh Rezaee, M., Hasani, A., Sharifi, Y., Rahnamaye Farzami, M., Dolatyar, A., et al. (2018). Molecular epidemiology of Vancomycin-resistant enterococcus faecalis and enterococcus faecium isolated from clinical specimens in the northwest of Iran. *Microb. Drug Resist.* 24, 1165–1173. doi: 10.1089/mdr.2017.0380

Khan, F. M., Gondil, V. S., Li, C., Jiang, M., Li, J., Yu, J., et al. (2021). A novel *Acinetobacter baumannii* bacteriophage Endolysin LysAB54 With high antibacterial activity Against multiple gram-negative microbes. *Front. Cell. Infect. Microbiol.* 11:637313. doi: 10.3389/fcimb.2021.637313

Kortright, K. E., Chan, B. K., Koff, J. L., and Turner, P. E. (2019). Phage therapy: a renewed approach to combat antibiotic-resistant bacteria. *Cell Host Microbe* 25, 219–232. doi: 10.1016/j.chom.2019.01.014

Labrie, S. J., Samson, J. E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* 8, 317–327. doi: 10.1038/nrmicro2315

Lee, M. J., Gravelat, F. N., Cerone, R. P., Baptista, S. D., Campoli, P. V., Choe, S. I., et al. (2014). Overlapping and distinct roles of *Aspergillus fumigatus* UDP-glucose 4-epimerases in galactose metabolism and the synthesis of galactose-containing cell wall polysaccharides. *J. Biol. Chem.* 289, 1243–1256. doi: 10.1074/jbc.M113.522516

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, G., Shen, M., Yang, Y., Le, S., Li, M., Wang, J., et al. (2018). Adaptation of Pseudomonas aeruginosa to phage PaP1 predation via O-antigen polymerase mutation. *Front. Microbiol.* 9:1170. doi: 10.3389/fmicb.2018.01170

Lim, A. N. W., Yen, M., Seed, K. D., Lazinski, D. W., and Camilli, A. (2021). A tail fiber protein and a receptor-binding protein mediate ICP2 bacteriophage interactions with vibrio cholerae OmpU. *J. Bacteriol.* 203:e0014121. doi: 10.1128/JB.00141-21

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226

Pires, D. P., Costa, A. R., Pinto, G., Meneses, L., and Azeredo, J. (2020). Current challenges and future opportunities of phage therapy. *FEMS Microbiol. Rev.* 44, 684–700. doi: 10.1093/femsre/fuaa017

Schwarzer, D., Buettner, F. F., Browning, C., Nazarov, S., Rabsch, W., Bethe, A., et al. (2012). A multivalent adsorption apparatus explains the broad host range of phage phi92: a comprehensive genomic and structural analysis. *J. Virol.* 86, 10384–10398. doi: 10.1128/JVI.00801-12

Shen, W., Le, S., Li, Y., and Hu, F. Q. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 11:e0163962. doi: 10.1371/journal.pone.0163962

Shen, M., Zhang, H., Shen, W., Zou, Z., Lu, S., Li, G., et al. (2018). Pseudomonas aeruginosa MutL promotes large chromosomal deletions through non-homologous end joining to prevent bacteriophage predation. *Nucleic Acids Res.* 46, 4505–4514. doi: 10.1093/nar/gky160

Takeuchi, I., Osada, K., Azam, A. H., Asakawa, H., Miyanaga, K., and Tanji, Y. (2016). The presence of two receptor-binding proteins contributes to the wide host range of staphylococcal Twort-Like phages. *Appl. Environ. Microbiol.* 82, 5763–5774. doi: 10.1128/AEM.01385-16

Wu, N., Dai, J., Guo, M., Li, J., Zhou, X., and Li, F. (2021). Pre-optimized phage therapy on secondary *Acinetobacter baumannii* infection in four critical COVID-19 patients. *Emerg. Microb. Infect.* 10, 612–618. doi: 10.1080/22221751.2021.1902754

Yang, Y. H., Lu, S. G., Shen, W., Zhao, X., Shen, M. Y., et al. (2016). Characterization of the first double-stranded RNA bacteriophage infecting *Pseudomonas aeruginosa*. *Sci. Rep.* 6:38795. doi: 10.1038/srep38795

Yang, Y., Shen, W., Zhong, Q., Chen, Q., He, X., Baker, J. L., et al. (2020). Development of a bacteriophage cocktail to constrain the emergence of phage-resistant *Pseudomonas aeruginosa*. *Front. Microbiol.* 11:327. doi: 10.3389/fmicb.2020.00327

Zhong, Q., Yang, L., Li, L., Shen, W., Li, Y., Xu, H., et al. (2020). Transcriptomic analysis reveals the dependency of *Pseudomonas aeruginosa* genes for double-stranded RNA bacteriophage phiYY infection cycle. *iScience* 23:101437. doi: 10.1016/j.isci.2020.101437

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

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

ORIGINAL RESEARCH
published: 04 April 2022
doi: 10.3389/fmicb.2022.865287

Picorna-Like Viruses of the Havel River, Germany

Roland Zell[1], Marco Groth[2], Lukas Selinka[1] and Hans-Christoph Selinka[3]*

[1]Section of Experimental Virology, Institute for Medical Microbiology, Jena University Hospital, Friedrich Schiller University, Jena, Germany, [2]CF DNA Sequencing, Leibniz Institute on Aging, Fritz Lipmann Institute, Jena, Germany, [3]Section II 1.4 Microbiological Risks, Department of Environmental Hygiene, German Environment Agency, Berlin, Germany

To improve the understanding of the virome diversity of riverine ecosystems in metropolitan areas, a metagenome analysis was performed with water collected in June 2018 from the river Havel in Berlin, Germany. After enrichment of virus particles and RNA extraction, paired-end Illumina sequencing was conducted and assignment to virus groups and families was performed. This paper focuses on picorna-like viruses, the most diverse and abundant group of viruses with impact on human, animal, and environmental health. Here, we describe altogether 166 viral sequences ranging in size from 1 to 11.5 kb. The 71 almost complete genomes are comprised of one candidate iflavirus, one picornavirus, two polycipiviruses, 27 marnaviruses, 27 dicistro-like viruses, and 13 untypeable viruses. Many partial picorna-like virus sequences up to 10.2 kb were also investigated. The sequences of the Havel picorna-like viruses represent genomes of seven of eight so far known *Picornavirales* families. Detection of numerous distantly related dicistroviruses suggests the existence of additional, yet unexplored virus groups with dicistronic genomes, including few viruses with unusual genome layout. Of special interest is a clade of dicistronic viruses with capsid protein-encoding sequences at the 5′-end of the genome. Also, monocistronic viruses with similarity of their polymerase and capsid proteins to those of dicistroviruses are interesting. A second protein with NTP-binding site present in the polyprotein of solinviviruses and related viruses needs further attention. The results underline the importance to study the viromes of fluvial ecosystems. So far acknowledged marnaviruses have been isolated from marine organisms. However, the present study and available sequence data suggest that rivers and limnic habitats are relevant ecosystems with circulation of marnaviruses as well as a plethora of unknown picorna-like viruses.

Keywords: riverine ecosystems, viromes, *Picornavirales*, phylogenetic analysis, metagenomic, picorna-like viruses, RNA viruses

OPEN ACCESS

Edited by:
Pingfeng Yu,
Zhejiang University, China

Reviewed by:
Kayvan Etebari,
The University of Queensland, Australia
Marli Vlok,
University of British Columbia, Canada

****Correspondence:***
Roland Zell
roland.zell@med.uni-jena.de

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *29 January 2022*
Accepted: *14 March 2022*
Published: *04 April 2022*

Citation:
Zell R, Groth M, Selinka L and Selinka H-C (2022) Picorna-Like Viruses of the Havel River, Germany. Front. Microbiol. 13:865287. doi: 10.3389/fmicb.2022.865287

INTRODUCTION

The order *Picornavirales* is comprised of eight families, *Caliciviridae*, *Dicistroviridae*, *Iflaviridae*, *Marnaviridae*, *Picornaviridae*, *Polycipiviridae*, *Secoviridae*, and *Solinviviridae* with so far 100 genera and 323 species (as of 5 December 2021).[1] Viruses of this order have a single-stranded RNA genome with positive polarity and an icosahedral capsid with $T=3$ or $T=1$/pseudo T3 structures (Le Gall et al., 2008). The $T=1$/pseudo $T=3$ capsids of the *Picornavirales* members

[1] https://talk.ictvonline.org/taxonomy/

consist of 60 copies of three major capsid proteins (CP) with a characteristic jelly roll fold; an additional minor capsid protein (VP4) may be present (Rossmann and Johnson, 1989). As an exception, some viruses of the *Secoviridae* family with $T = 1$/pseudo T3 structure have either only one large CP with three jelly roll domains, or two CPs, a small CP with one jelly roll domain and a large CP with two (Chen et al., 1987; Chandrasekar and Johnson, 1998). Caliciviruses and solinviviruses have a $T = 3$ capsid which consists of 180 copies of a single CP (Prasad et al., 1999; Valles et al., 2014). Viral RNAs of all members of the *Picornavirales* are polyadenylated but exhibit a considerable variation of their genome organization, e.g., monopartite or dipartite genomes which encode one to five open reading frames (ORFs). The ORFs are expressed by various mechanisms including the usage of one or two internal ribosome entry sites (*Picornaviridae*, *Dicistroviridae, Iflaviridae, Marnaviridae, and Polycipiviridae*), ribosomal frame-shifting (Solenopsis invicta virus 3 of the *Solinviviridae*), transcription of subgenomic RNAs (*Caliciviridae, Labyrnavirus, and Solinviviridae*), or reinitiation (*Caliciviridae and Polycipiviridae*; Jan, 2006; Valles et al., 2014; Zinoviev et al., 2015; Olendraite et al., 2017). Despite all differences in genome layouts, all members of the *Picornavirales* and many yet unclassified picorna-like viruses (PLVs) share in common a set of phylogenetically related proteins, especially (i) a helicase (hel) with one or two conserved sequence motifs of P-loop ATPases (Walker A and B motifs), (ii) the chymotrypsin-like proteinase (pro) with a CxCG or CxSG active site sequence motif, (iii) the RNA-dependent RNA polymerase (RdRP or pol) with conserved SG, GDD, and LK sequence motifs, and (iv) one to three CPs with a jelly roll fold. A short genome-linked viral peptide (VPg) encoded by a gene region located between the hel and pro genes has been shown for six of the eight families. Polycipiviruses and solinviviruses are expected to possess a VPg as well, although an experimental proof is still lacking (Olendraite et al., 2017; Brown et al., 2019). The hel-pro-pol domains in this order are known as "hel-pro-pol core replicative module" or "hel-pro-pol replication block" (Le Gall et al., 2008; Sanfaçon et al., 2012). The CPs are encoded either (i) together with the nonstructural proteins (NSP) as part of a large polyprotein (e.g., *Iflaviridae*, most members of the *Picornaviridae*, and many members of the *Marnaviridae*), (ii) by a second ORF (e.g., *Dicistroviridae*, many viruses of the *Marnaviridae*, and dicipiviruses of the *Picornaviridae*), (iii) a subgenomic RNA (*Caliciviridae*, *Labyrnavirus* of the *Marnaviridae*, and *Solinviviridae*), (iv) a second RNA molecule (some members of the *Secoviridae*), or (v) overlapping ORFs expressed by ribosomal frame-shifting (*Solinviviridae*; for references, see Valles et al., 2007, 2014; Bonning and Miller, 2010; van der Vlugt et al., 2015; Thompson et al., 2017; Zell, 2018). In addition, genomes may show family-specific and genus-specific gene regions encoding conserved or unique proteins, like the leader protein of iflaviruses or many picornaviruses with various functions, the movement protein of some secoviruses, or the ovarian tumor domain of solinviviruses. The order of CP- and NSP-encoding gene regions may vary: the monocistronic iflaviruses, picornaviruses, and some secoviruses have large polyproteins with CP domains at the N-terminus and NSP domains at the C-terminus; polycipiviruses have CP-encoding ORFs at the 5′-end of their genomes, whereas caliciviruses, dicistroviruses, marnaviruses, and solinviviruses encode these proteins at the 3′-end of their genomes. Also, the position of the VP4, a minor CP of some *Picornavirales* families, may vary. Viruses of the *Picornavirales* infect a wide range of eukaryotic organisms. Hosts of six vertebrate classes are infected by picornaviruses, mammals, and fish by caliciviruses (Zell, 2018; Smertina et al., 2021), whereas dicistroviruses, iflaviruses, polycipiviruses, and solinviviruses are associated with arthropods (Valles et al., 2017a,b; Brown et al., 2019; Olendraite et al., 2019). Secoviruses infect (dicotyledonous) plants and marnaviruses diatoms, unicellular algae, and heterotrophic protists (Sanfaçon et al., 2009; Vlok et al., 2019).

It is long known that virioplankton, i.e., free-floating viruses in aquatic ecosystems, is quite ubiquitous, outnumber microbes and other organisms many times over and are important players in maintaining stable food webs and nutrient cycles (Fuhrman, 1999; Wilhelm and Suttle, 1999; Wommack and Colwell, 2000; Culley et al., 2003). Development of next-generation sequencing techniques and metagenomics pushed forward the description of a plethora of novel viruses infecting prokaryotes, protists, animals, and plants in marine, freshwater and terrestrial ecosystems which advanced our view on the virosphere (e.g., Culley and Steward, 2007; Rosario and Breitbart, 2011; Culley et al., 2014; Wommack et al., 2015; Fierer, 2017; Williamson et al., 2017; Coy et al., 2018; Lefeuvre et al., 2019). In particular, presence of PLVs in environmental samples indicates their abundant prevalence in marine ecosystems (Culley and Steward, 2007; Culley et al., 2014; Lang et al., 2018). However, culture-independent virus sequencing enabled also the identification of numerous PLVs in fecal and organ samples of many unexpected organisms indicating both wide distribution in organismic kingdoms but also suggesting unspecific virus uptake and accumulation without subsequent infection (e.g., Shi et al., 2016, 2018; Yinda et al., 2017; Dastjerdi et al., 2021).

In the present study, we demonstrate the presence of 166 mostly novel PLVs in the river Havel, designated Havel picorna-like virus (HPLV) 1 to −166. Seventy-one almost complete genomes (i.e., complete coding regions plus parts of the 5′- and 3′-untranslated regions) and many partial genomes of seven of eight families of the *Picornavirales* demonstrate a remarkable genetic diversity of this virus order in a riverine ecosystem whose overall complexity is still unexplored.

MATERIALS AND METHODS

Sample Collection and Virus Enrichment

A freshwater sample with a volume of 50 liters was collected from the river Havel in the metropolitan area of Berlin, Germany, on June 28th, 2018 (sampling site coordinates 52°30′46″N13°12′14″E). The sample was transported under cooled conditions to the laboratory and immediately processed. Five 10 L samples were homogenized by vortexing (20 min) and preconditioned for enhanced virus binding to negatively charged glass wool by adjusting the pH from 8.1 to 3.5. Thereafter, virus particles were concentrated by glass wool filtration as basically described by Wyn-Jones et al. (2011). Virus concentrates were slowly eluted from the column with a buffer (pH 9.5) containing 3% (w/v) beef extract in 0.05 M glycine and subsequently adjusted to pH 7.0 without further flocculation. The 180 ml eluate was filtered through a 0.45 μm filter to remove bacteria and detritus. Then, virus particles of the filtrate were sedimented by ultracentrifugation at 100,000 × g for 2.5 h at 4°C. Sediment was resolved in a total of 500 μl PBS and homogenized using a ball mill. RNA was extracted using the QIAamp Viral RNA mini kit (Qiagen, Hilden, Germany). One hundred nanogram RNA was employed for library preparation.

Virus Sequencing and Sequence Data Processing

The Illumina next-generation sequencing approach was used in our study (Bentley et al., 2008). The library was prepared from 100 ng of RNA using the TruSeq stranded mRNA library preparation kit (Illumina). In order to address all RNA molecules (not only polyadenylated RNA), the protocol was adapted as follows: RNA was precipitated using isopropanol and resolved in *Fragment, Prime, Finish Mix* (FPF). From this step on, the manufacturer's protocol was followed (p20, step 12, TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E, Illumina). The obtained library was quantified and quality-checked using the Agilent 2100 Bioanalyzer and the DNA 7500 kit (Agilent Technologies). Sequencing was done on a HiSeq 2500 running in rapid-mode, paired-end (2× 150 bp) by combining one 200 cycle and two 50 cycle SBS kits (Illumina, FC-402-4021 and FC-402-4022). Sequence data were extracted in FastQ format using Illumina's tool bcl2FastQ v2.19.1.403.

Reads were pre-processed before assembly. Adapter and quality trimming were done using cutadapt v1.8.3 (Martin, 2011); parameters: -q 10 -m 30 -a AGATCGGAAGAGCACACGT CTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAG GGAAAGAGTGT. Then, amplification duplicons were removed by comparing the sequence of all pairs against each other. In the next step, read pairs were assembled using two different software tools, metaSPAdes v3.15.3 (Nurk et al., 2017) with standard parameters (-k auto) and the clc_assembler v. 5.2.1 (part of the CLC Assembly Cell, Qiagen; parameters: -p fb ss 0850).

Sequence Data Analysis

The contigs, as result from both assembly tools, were used to search a protein database created with all NCBI GenBank entries for the Taxonomy ID 10239 (search term "viruses[organism]") utilizing DIAMOND (Buchfink et al., 2015) and BLAST+ version 2.6.0,[2] respectively. In parallel, contigs were compared with 12 reference data sets consisting of reference sequences of the families and sub-families of the *Picornavirales* order using tBLASTx. Then, translated candidate sequences with significant similarity to reference sequences were queried against the contig bank to identify contigs with similar sequences. Virus contigs were manually curated to generate the final full-length and partial genomes greater 1 kb. Protein domains were predicted using the Pfam conserved domain database (CDD) search tool of NCBI.[3]

For phylogenetic analyses, protein sequences of the present study and reference sequences of the GenBank were aligned with Mega version X (Kumar et al., 2018) and adjusted manually. Maximum likelihood trees were inferred with IQ-TREE 2.1.3 for Windows (Nguyen et al., 2015). Branch supports were assessed with UFBoot2 implemented in IQ-TREE software (Hoang et al., 2018). Usually, 50,000 ultrafast bootstrap replications (-B 50000) were conducted. Best-fitting nucleotide substitution models (-m MFP) were selected on basis of the Bayesian Information Criterion (BIC) with ModelFinder also implemented in IQ-TREE.

RESULTS

High-Throughput Sequencing, Assembly, and Identification of Viral Sequences

A 50-liter freshwater sample of the river Havel was used for virus enrichment, RNA extraction, and paired-end sequencing on an Illumina HiSeq 2,500 platform. Sequencing yielded a total of 144,395,487 read pairs. After removal of duplicons, 51,902,006 read pairs were obtained and utilized for assembly. The assembly process resulted in (i) metaSPAdes: 484,430 scaffolds (total length 132,507,247 nt; N50: 272 nt) and (ii) CLC assembler: 162,082 contigs (total length: 62,667,741 bp; N50: 382 bp). DIAMOND assigned 5,687 scaffolds (1.17%) and 3,902 CLC contigs (2.4%), respectively, to the *Picornavirales* order, but only 2,453 (41.6%) and 1,612 (41.3%), respectively, could be classified into one of the eight *Picornavirales* families. In a parallel approach, contigs with similarity to viruses of the *Picornavirales* (>100 identical amino acids) were identified by BLAST searches against 12 data sets of representative sequences of each of the *Picornavirales* families and sub-families. In this study, we further analyzed a total of 166 sequences with lengths greater 1 kb of which 71 sequences represented an almost complete virus genome (**Supplementary Table 1**).

Dicistroviridae and *Marnaviridae*

The dominant PLVs detected in the Havel river can be classified as dicistrovirus- and marnavirus-like as judged from preliminary phylogenetic analyses of the polymerase and helicase gene regions (data not shown). Eighty-five percent of the identified HPLV sequences (141/166) clustered with sequences of viruses of these families. For a further analysis, 109 HPLV polymerase sequences were aligned with 35 reference sequences representing all 10 genera of both families as well as 120 unclassified viruses with similarity

[2] https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.6.0

[3] https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

to our HPLVs (**Figure 1, Supplementary Figure 1**). The phylogenetic tree reveals two major branches with monophyletic clades corresponding to the genera *Bacillarnavirus*, *Kusarnavirus*, *Locarnavirus*, and *Triatovirus*. The remaining genera *Aparavirus*, *Cripavirus*, *Labyrnavirus*, *Marnavirus*, *Salisharnavirus*, and *Sogarnavirus* were not monophyletic. In addition, several clades of HPLVs and other unclassified viruses clustered on both major branches with various bootstrap values. Long maximum likelihood distances indicated substitution saturation of some sequences. In order to improve the robustness of the tree, a phylogenetic analysis was conducted with an alignment of the proteinase/polymerase sequences of these viruses. The topology of the resulting tree presented similar to that of the polymerase tree. Interestingly, in contrast to the polyphyletic sequences of sogarnaviruses and cripaviruses based on the polymerase gene alone, sequences of both genera were monophyletic by analyzing the proteinase/polymerase sequences together (**Figure 1, Supplementary Figure 1**). Altogether, the tree reveals one bacillarna-like, two marna-like, four kusarna-like, three sogarna-like, four salisharna-like, six labyrna-like, and 23 locarna-like HPLVs on the marnavirus branch. The newly identified HPLV-20 clustered with kusarnaviruses in both trees but exhibited an unusual genome layout: the CP-encoding gene region was located at the 5′-end of the genome, the NSP-encoding gene region at the 3′-end. Comparison of the polymerase and proteinase/polymerase trees revealed only few HPLV strains with inconsistent clustering (e.g., HPLV-10; HPLV-91).

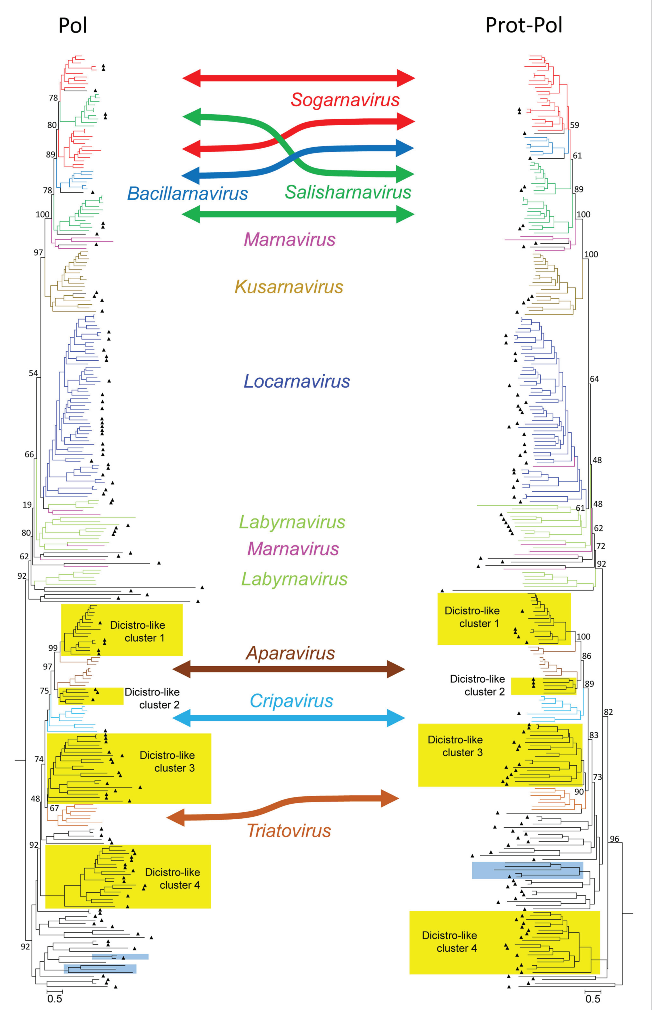


FIGURE 1 | Phylogenetic analysis of 267 polymerase- (left) and 264 proteinase/polymerase-encoding sequences (right) of dicistroviruses, marnaviruses, and unassigned viruses. Sequences were aligned with MEGA and manually adjusted. The trees were inferred with IQ-Tree 2, optimal substitution model: GTR+F+R9 for the pol tree and GTR+F+R10 for the prot/pol tree, respectively. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. The respective genera are indicated. Color codes: aparaviruses, dark brown; bacillarnaviruses, blue; cripaviruses, light blue; kusarnaviruses, ochre; labyrnaviruses, light green; locarnaviruses, dark blue; marnaviruses, magenta; salisharnaviruses, dark green; sogarnaviruses, red; triatoviruses, brown; and untypeable viruses, black. A triangle (▲) indicates viruses identified in the present study. Blue boxes indicate viruses with unusual genome layout (capsid protein-encoding gene region at the 5′-end; nonstructural polyprotein-encoding gene region at the 3′-end). Yellow boxes indicate four dicistrovirus-like sequence clusters. More details of this figure are presented in **Supplementary Figure 1**.

Moreover, as also shown in **Figure 1** (**Supplementary Figure 1**), the dicistrovirus branch of both trees was comprised of the three genera *Aparavirus*, *Cripavirus*, and *Triatovirus* plus several clades of unclassified viruses with both dicistronic and monocistronic genome layouts. Only HPLV-102 was identified as a dicistrovirus candidate on basis of the proteinase/polymerase sequence. However, many short contigs were identified with strong similarity to members of the acknowledged dicistroviruses (data not shown). The dicistrovirus-like clades 1 and 2 consistently clustered with high bootstrap support with aparaviruses and may represent new aparavirus species. Dicistrovirus-like clade 3 clusters with triatoviruses, whereas dicistrovirus-like clade 4 is less well supported. Both clades include altogether 28 sequences, but only three viruses were obtained from freshwater arthropods (KX883663, KX883640, and KX883650); the remaining viruses were detected in samples as diverse as Havel river water, grassland soil, fecal specimens/intestinal contents of mammals, cloacal swabs of birds, higher plants, freshwater snails, or marine oysters. Several viruses have been proposed as novel dicistroviruses previously (e.g., Reuter et al., 2014; Yinda et al., 2017; Duraisamy et al., 2018; Dastjerdi et al., 2021). Three clades contain viruses with monocistronic genomes. Also of interest are HPLV-32 and -150 with their unusual genome layout (CP-encoding gene region at the 5′-end; NSP-encoding gene region at the 3′-end).

As marnaviruses and dicistroviruses share a remarkable similarity of the structural proteins (Lang et al., 2018), we also conducted a phylogenetic analysis of the CPs and largely confirmed the major clades seen with the polymerase and proteinase/polymerase gene region (see **Supplementary Figure 2**). The CP tree revealed a second cripavirus candidate from the Havel river (HPLV-141), but also few viruses which displayed inconsistent clustering (e.g., HPLV-20, −38, −40, −52, −61, and −91).

Caliciviridae

Three new caliciviruses were identified. The partial genome of HPLV-93 corresponds to about 90% of a typical calicivirus genome and contains three partly overlapping ORFs. The available sequence comprises the almost complete NSP-encoding gene region of ORF1 [1821 amino acids (aa)], the complete ORF2 (533 aa) with the calicivirus coat protein (CCP) gene region, and a partial ORF3 region (48 aa). The NSP precursor comprises a helicase domain (modified Walker A motif: G_{118}xGKT), a second Walker A motif at aa position 378, a proteinase domain with a G_{949}xCG active site motif, and an RdRP which shows the conserved D_{1307}xxxxD, S_{1373}G, F_{1418}GDD, and L_{1470}KR sequence motifs (**Figure 2A**). Two other partial genomes of new caliciviruses, HPLV-90 (2,748 nt) and −126 (1,038 nt), contain polymerase to VP1 and VP1 sequences, respectively. Phylogenetic analysis of the CCP-encoding gene region (VP1) of these HPLVs suggests novel caliciviruses which show similarity to the classified caliciviruses but are more closely related to unclassified caliciviruses from bats, lizards, insects, annelids, and various bivalve shellfish (**Figure 2B**). This result is confirmed by analysis of the proteinase/polymerase gene region (**Figure 2C**). Genetic distances of the VP1 protein in pairwise comparisons with other caliciviruses reveal values greater 74%, suggesting that they belong to a new genus of the *Caliciviridae* family (**Supplementary Table 2**).

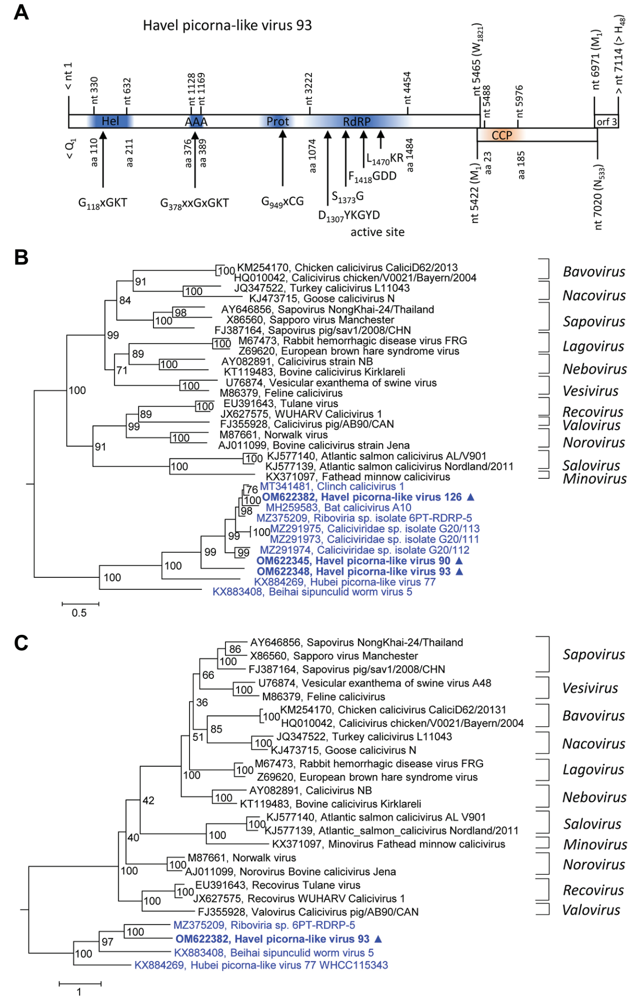

A
Havel picorna-like virus 93
< nt 1
nt 330
nt 632
nt 1128
nt 1169
nt 3222
nt 4454
nt 5465 (W_{1821})
nt 5488
nt 5976
nt 6971 (M_1)
> nt 7114 (> H_{48})
Hel
AAA
Prot
RdRP
CCP
orf 3
< Q_1
aa 110
aa 211
aa 376
aa 389
aa 1074
aa 1484
L_{1470}KR
F_{1418}GDD
S_{1373}G
D_{1307}YKGYD
active site
G_{118}xGKT
G_{378}xxGxGKT
G_{949}xCG
nt 5422 (M_1)
aa 23
aa 185
nt 7020 (N_{533})
B
KM254170, Chicken calicivirus CaliciD62/2013
HQ010042, Calicivirus chicken/V0021/Bayern/2004
JQ347522, Turkey calicivirus L11043
KJ473715, Goose calicivirus N
AY646856, Sapovirus NongKhai-24/Thailand
X86560, Sapporo virus Manchester
FJ387164, Sapovirus pig/sav1/2008/CHN
M67473, Rabbit hemorrhagic disease virus FRG
Z69620, European brown hare syndrome virus
AY082891, Calicivirus strain NB
KT119483, Bovine calicivirus Kirklareli
U76874, Vesicular exanthema of swine virus
M86379, Feline calicivirus
EU391643, Tulane virus
JX627575, WUHARV Calicivirus 1
FJ355928, Calicivirus pig/AB90/CAN
M87661, Norwalk virus
AJ011099, Bovine calicivirus strain Jena
KJ577140, Atlantic salmon calicivirus AL/V901
KJ577139, Atlantic salmon calicivirus Nordland/2011
KX371097, Fathead minnow calicivirus
MT341481, Clinch calicivirus 1
OM622382, Havel picorna-like virus 126 ▲
MH259583, Bat calicivirus A10
MZ375209, Riboviria sp. isolate 6PT-RDRP-5
MZ291975, Caliciviridae sp. isolate G20/113
MZ291973, Caliciviridae sp. isolate G20/111
MZ291974, Caliciviridae sp. isolate G20/112
OM622345, Havel picorna-like virus 90 ▲
OM622348, Havel picorna-like virus 93 ▲
KX884269, Hubei picorna-like virus 77
KX883408, Beihai sipunculid worm virus 5
Bavovirus
Nacovirus
Sapovirus
Lagovirus
Nebovirus
Vesivirus
Recovirus
Valovirus
Norovirus
Salovirus
Minovirus
0.5
C
AY646856, Sapovirus NongKhai-24/Thailand
X86560, Sapporo virus Manchester
FJ387164, Sapovirus pig/sav1/2008/CHN
U76874, Vesicular exanthema of swine virus A48
M86379, Feline calicivirus
KM254170, Chicken calicivirus CaliciD62/20131
HQ010042, Calicivirus chicken/V0021/Bayern/2004
JQ347522, Turkey calicivirus L11043
KJ473715, Goose calicivirus N
M67473, Rabbit hemorrhagic disease virus FRG
Z69620, European brown hare syndrome virus
AY082891, Calicivirus NB
KT119483, Bovine calicivirus Kirklareli
KJ577140, Atlantic salmon calicivirus AL V901
KJ577139, Atlantic_salmon_calicivirus Nordland/2011
KX371097, Minovirus Fathead minnow calicivirus
M87661, Norwalk virus
AJ011099, Norovirus Bovine calicivirus Jena
EU391643, Recovirus Tulane virus
JX627575, Recovirus WUHARV Calicivirus 1
FJ355928, Valovirus Calicivirus pig/AB90/CAN
MZ375209, Riboviria sp. 6PT-RDRP-5
OM622382, Havel picorna-like virus 93 ▲
KX883408, Beihai sipunculid worm virus 5
KX884269, Hubei picorna-like virus 77 WHCC115343
Sapovirus
Vesivirus
Bavovirus
Nacovirus
Lagovirus
Nebovirus
Salovirus
Minovirus
Norovirus
Recovirus
Valovirus
1

FIGURE 2 | Analysis of new caliciviruses. **(A)** The genome organization of Havel picorna-like virus 93 resembles the calicivirus genome. ORF1 shows conserved elements of helicase (Hel), proteinase (Prot), and polymerase (RdRP) domains, whereas ORF2 has similarity to the calicivirus capsid protein (CCP) family of caliciviruses. Conserved domains and sequence motifs were identified by searching the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Phylogenetic analyses of the VP1 (capsid protein) gene region **(B)** and the proteinase/polymerase gene region **(C)** of calicivirus prototype strains, unassigned caliciviruses, and Havel picorna-like viruses 93, −175, and −176 are shown. The trees were inferred with IQ-Tree 2, optimal substitution model: TVM + F + R4 **(B)** and TVMe+R4 **(C)**. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. Presented are GenBank acc. nos. and virus names. The respective genera are indicated to the right. Unassigned viruses are printed in blue. A triangle (▲) indicates the viruses of the present study. AAA, AAA protein with homology to ATPase family associated with various cellular functions, Hel, helicase, Prot, proteinase; RdRP, polymerase; and CCP, capsid protein with similarity to the calicivirus capsid protein.

Iflaviridae

Two contigs representing HPLVs with similarity to iflaviruses were investigated. Iflaviruses have a monocistronic genome encoding a large polyprotein with a leader protein and CP domains at the N-terminus and NSPs at the C-terminus. HPLV-14 (9,465 nt) has an almost complete genome with iflavirus gene layout, whereas the partial genome of HPLV-129 (3,028 nt) contains the helicase gene region only (**Figures 3A,B**). For the phylogenetic analysis shown in **Figure 3C**, aligned ORF sequences of 36 acknowledged and unclassified iflaviruses were investigated. The phylogenetic tree indicates closest similarity of HPLV-14 to solenopsis invicta virus 11. A second phylogenetic analysis based on partial sequences (VP1 to helicase gene region), which included both the HPLV-14 and HPLV-127 sequences, confirmed relationship of both HPLVs to the classified iflaviruses. Both viruses may belong to different novel iflavirus species (**Supplementary Figure 3**).

Picornaviridae

One almost complete picornavirus genome, HPLV-29 (8,847 nt), was detected in the Havel river sample. It shows a typical picornavirus genome organization (**Figure 4A**) and significant sequence homology (>75%) to the genomes of newt ampivirus (KP770140), Shahe picorna-like virus 13 from freshwater arthropods (KX883649; KX883657), and two virus sequences from cloacal swabs of birds (MT138174; MT138399). Divergences of the capsid protein VP1 of 41% in comparison with newt ampivirus and 35.3% in comparisons with Shahe picorna-like virus 13 and the viruses from avian swabs indicate a putative third genotype, ampivirus A3, within the species *Ampivirus A*. Important unique differences are a long insertion of 32 aa in VP1 and an N-terminal deletion of 41 aa of 3A (data not shown). Phylogenetic analyses of the P1- and 3CD-encoding gene regions confirm the close relation of HPLV-29 to ampiviruses (**Figures 4B,C**; **Supplementary Figures 4A,B**). In addition to

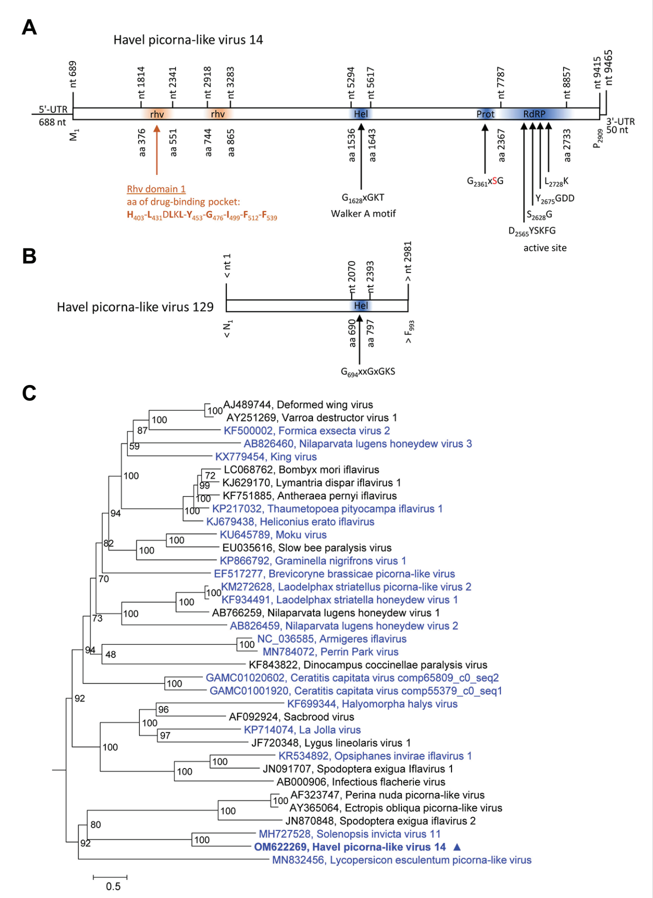

A
Havel picorna-like virus 14
5'-UTR 688 nt
nt 689
nt 1814
nt 2341
nt 2918
nt 3283
nt 5294
nt 5617
nt 7787
nt 8857
nt 9415
nt 9465
rhv
rhv
Hel
Prot
RdRP
3'-UTR 50 nt
M1
aa 376
aa 551
aa 744
aa 865
aa 1536
aa 1643
aa 2367
aa 2733
P2909
Rhv domain 1
aa of drug-binding pocket:
H403-L431DLKL-Y453-G476-I499-F512-F539
G1628xGKT
Walker A motif
G2361xSG
L2728K
Y2675GDD
S2628G
D2565YSKFG
active site
B
Havel picorna-like virus 129
< nt 1
nt 2070
nt 2393
> nt 2981
Hel
< N1
aa 690
aa 797
> F993
G694xxGxGKS
C
AJ489744, Deformed wing virus
AY251269, Varroa destructor virus 1
KF500002, Formica exsecta virus 2
AB826460, Nilaparvata lugens honeydew virus 3
KX779454, King virus
LC068762, Bombyx mori iflavirus
KJ629170, Lymantria dispar iflavirus 1
KF751885, Antheraea pernyi iflavirus
KP217032, Thaumetopoea pityocampa iflavirus 1
KJ679438, Heliconius erato iflavirus
KU645789, Moku virus
EU035616, Slow bee paralysis virus
KP866792, Graminella nigrifrons virus 1
EF517277, Brevicoryne brassicae picorna-like virus
KM272628, Laodelphax striatellus picorna-like virus 2
KF934491, Laodelphax striatella honeydew virus 1
AB766259, Nilaparvata lugens honeydew virus 1
AB826459, Nilaparvata lugens honeydew virus 2
NC_036585, Armigeres iflavirus
MN784072, Perrin Park virus
KF843822, Dinocampus coccinellae paralysis virus
GAMC01020602, Ceratitis capitata virus comp65809_c0_seq2
GAMC01001920, Ceratitis capitata virus comp55379_c0_seq1
KF699344, Halyomorpha halys virus
AF092924, Sacbrood virus
KP714074, La Jolla virus
JF720348, Lygus lineolaris virus 1
KR534892, Opsiphanes invirae iflavirus 1
JN091707, Spodoptera exigua Iflavirus 1
AB000906, Infectious flacherie virus
AF323747, Perina nuda picorna-like virus
AY365064, Ectropis obliqua picorna-like virus
JN870848, Spodoptera exigua iflavirus 2
MH727528, Solenopsis invicta virus 11
OM622269, Havel picorna-like virus 14 ▲
MN832456, Lycopersicon esculentum picorna-like virus
100
100
87
59
100
72
99
100
100
94
100
100
82
70
100
100
100
73
100
94
48
100
92
96
100
97
100
100
100
100
100
80
92
100
0.5

FIGURE 3 | Analysis of Havel picorna-like viruses 14 and 129. Their genome organization **(A,B)** corresponds to the iflavirus genome. Conserved domains and sequence motifs were identified by searching the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). **(C)** Phylogenetic analysis of the complete ORF of 15 acknowledged iflaviruses and 21 candidate strains including Havel picorna-like virus 14 and 129. The tree was inferred with IQ-Tree 2, optimal substitution model: GTR+F+R5. Numbers at nodes present bootstrap values obtained after 50,000 bootstrap replications. The scale indicates substitutions per site. Presented are GenBank acc. nos. and virus names. Unassigned viruses are printed in blue. A triangle (▲) indicates the virus of the present study. UTR, untranslated region; Hel, helicase; Prot, proteinase; RdRP, polymerase; and rhv, capsid protein with similarity to the rhinovirus capsid protein with jelly roll fold and drug-binding pocket.

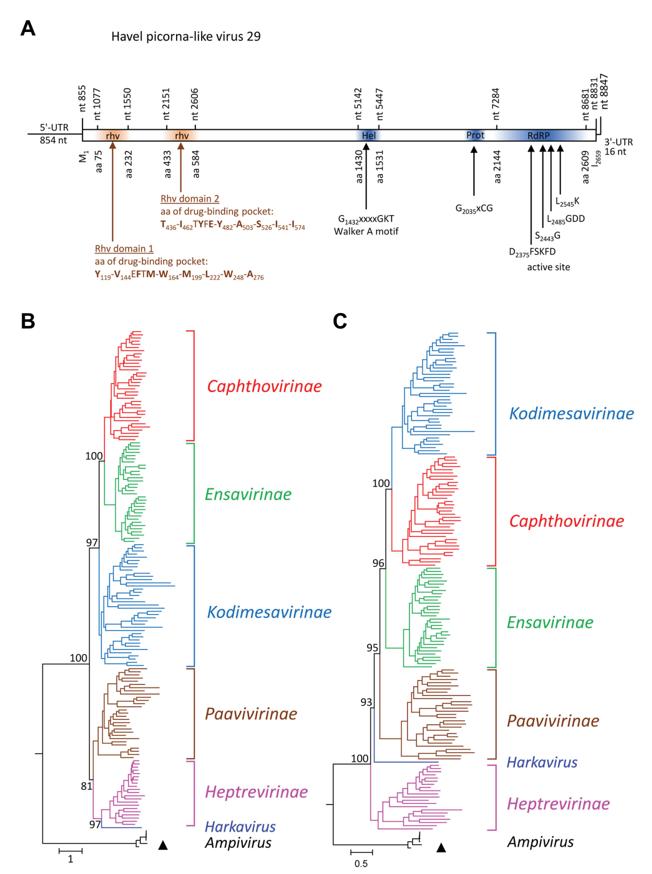


FIGURE 4 | Analysis of Havel picorna-like virus 29. **(A)** The genome organization corresponds to the picornavirus genome. Conserved domains and sequence motifs were identified by searching the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Phylogenetic analyses of the capsid protein-encoding gene region **(B)** and the proteinase/polymerase gene region **(C)** of Havel picorna-like virus 29 and 162 picornavirus strains. The tree was inferred with IQ-Tree 2, optimal substitution model: GTR+F+R9 **(B)** and GTR+F+R8 **(C)**. The respective sub-families and genera are given to the right. A triangle (▲) indicates the virus of the present study. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. More details are presented in **Supplementary Figures S4A,B**. UTR, untranslated region; Hel, helicase; Prot, proteinase; RdRP, polymerase; and rhv, capsid protein with similarity to the rhinovirus capsid protein with jelly roll fold and drug-binding pocket.

HPLV-29, one short contig corresponding to ampivirus A1 was obtained (data not shown). Further five contigs with similarity to the known ampiviruses suggest the existence of at least two other ampivirus genotypes (data not shown).

Polycipiviridae

HPLV-1 (11,517 nt) and −2 (11,670 nt) are candidate viruses of the genus *Chipolycivirus*, family *Polycipiviridae*. Their almost complete genomes have five ORFs with ORFs 1, 3, and 4 encoding capsid proteins with jelly roll domains and ORF5 a polyprotein with helicase, proteinase, and polymerase domains. Both viruses have GxSG active site sequences of their proteinase and AADD active site sequences of their RdRP, which is characteristic for chipoliciviruses (**Figure 5A**). Concatenated sequences of the three CP-encoding ORFs suggest affiliation to the chipolyciviruses (**Figure 5B**). This result was confirmed with an alignment of ORF5 sequences representing the helicase, proteinase, and polymerase gene regions (**Figure 5C**). At least 15 short contigs with high similarity to polycipiviruses indicated the occurrence of additional viruses of this family in the Havel river.

Secoviridae

No viruses with strong similarity to secoviruses were detected in our sample even though DIAMOND suggested several contigs.

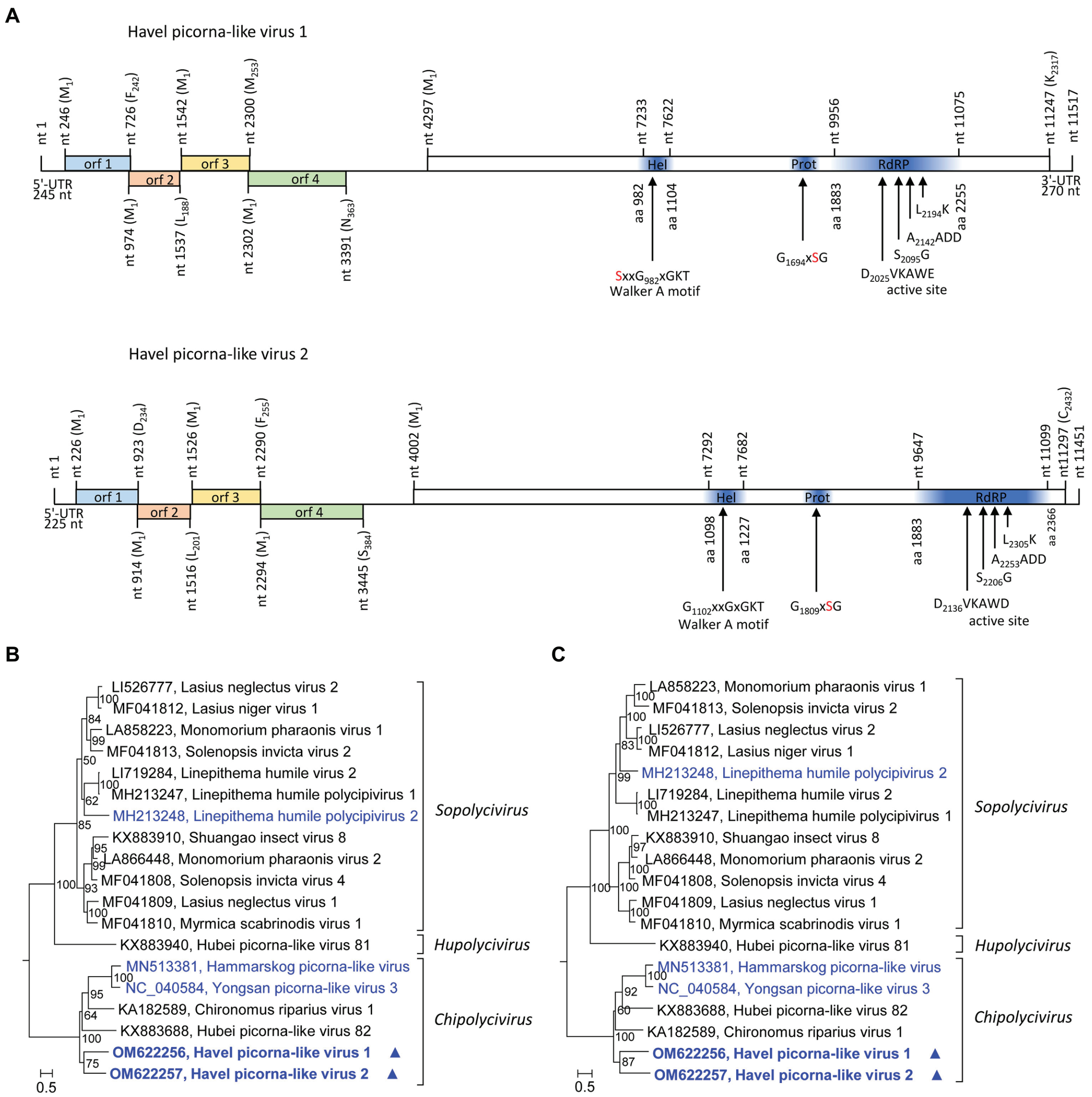

A
Havel picorna-like virus 1
nt 1
nt 246 (M1)
nt 726 (F242)
nt 1542 (M1)
nt 2300 (M253)
nt 4297 (M1)
nt 7233
nt 7622
nt 9956
nt 11075
nt 11247 (K2317)
nt 11517
orf 1
orf 2
orf 3
orf 4
Hel
Prot
RdRP
5'-UTR 245 nt
3'-UTR 270 nt
nt 974 (M1)
nt 1537 (L188)
nt 2302 (M1)
nt 3391 (N363)
aa 982
aa 1104
aa 1883
aa 2255
SxxG982xGKT Walker A motif
G1694xSG
D2025VKAWE active site
S2095G
A2142ADD
L2194K
Havel picorna-like virus 2
nt 1
nt 226 (M1)
nt 923 (D234)
nt 1526 (M1)
nt 2290 (F255)
nt 4002 (M1)
nt 7292
nt 7682
nt 9647
nt 11099
nt11297 (C2432)
nt 11451
orf 1
orf 2
orf 3
orf 4
Hel
Prot
RdRP
5'-UTR 225 nt
nt 914 (M1)
nt 1516 (L201)
nt 2294 (M1)
nt 3445 (S384)
aa 1098
aa 1227
aa 1883
aa 2366
G1102xxGxGKT Walker A motif
G1809xSG
D2136VKAWD active site
S2206G
A2253ADD
L2305K
B
LI526777, Lasius neglectus virus 2
MF041812, Lasius niger virus 1
LA858223, Monomorium pharaonis virus 1
MF041813, Solenopsis invicta virus 2
LI719284, Linepithema humile virus 2
MH213247, Linepithema humile polycipivirus 1
MH213248, Linepithema humile polycipivirus 2
KX883910, Shuangao insect virus 8
LA866448, Monomorium pharaonis virus 2
MF041808, Solenopsis invicta virus 4
MF041809, Lasius neglectus virus 1
MF041810, Myrmica scabrinodis virus 1
KX883940, Hubei picorna-like virus 81
MN513381, Hammarskog picorna-like virus
NC_040584, Yongsan picorna-like virus 3
KA182589, Chironomus riparius virus 1
KX883688, Hubei picorna-like virus 82
OM622256, Havel picorna-like virus 1 ▲
OM622257, Havel picorna-like virus 2 ▲
100
84
99
50
100
62
85
95
99
93
100
100
100
95
64
100
75
Sopolycivirus
Hupolycivirus
Chipolycivirus
0.5
C
LA858223, Monomorium pharaonis virus 1
MF041813, Solenopsis invicta virus 2
LI526777, Lasius neglectus virus 2
MF041812, Lasius niger virus 1
MH213248, Linepithema humile polycipivirus 2
LI719284, Linepithema humile virus 2
MH213247, Linepithema humile polycipivirus 1
KX883910, Shuangao insect virus 8
LA866448, Monomorium pharaonis virus 2
MF041808, Solenopsis invicta virus 4
MF041809, Lasius neglectus virus 1
MF041810, Myrmica scabrinodis virus 1
KX883940, Hubei picorna-like virus 81
MN513381, Hammarskog picorna-like virus
NC_040584, Yongsan picorna-like virus 3
KX883688, Hubei picorna-like virus 82
KA182589, Chironomus riparius virus 1
OM622256, Havel picorna-like virus 1 ▲
OM622257, Havel picorna-like virus 2 ▲
100
100
100
83
99
100
100
97
100
100
100
100
100
92
60
100
87
Sopolycivirus
Hupolycivirus
Chipolycivirus
0.5

FIGURE 5 | Analysis of Havel picorna-like viruses 1 and 2. **(A)** Genome organization of Havel picorna-like virus 1 (top) and 2 (below). Conserved domains and sequence motifs were identified by searching the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Phylogenetic analyses of the concatenated capsid proteins-encoding ORFs 1, 3, and 4 **(B)** and the helicase, proteinase, and polymerase gene regions of ORF5 **(C)**. Sequences of 17 polycipivirus strains plus Havel picorna-like virus 1 and 2 were included. The tree was inferred with IQ-Tree 2; optimal substitution model: GTR+F+R4 in both analyses. Presented are GenBank acc. nos. and virus names. The respective genera are presented to the right. Unassigned viruses are printed in blue. A triangle (▲) indicates the viruses of the present study. Numbers at nodes indicate bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. orf, open reading frame; UTR, untranslated region; Hel, helicase; Prot, proteinase; and RdRP, polymerase.

Solinviviridae

The partial genomes of HPLV-65 (6,945 nt) and HPLV-75 (6,038 nt) exhibit similarity to solinviviruses and encode nonstructural proteins. Most notably, both viruses exhibit a second domain with homology to P-loop ATPases (**Figure 6A**), a feature which is shared with the two acknowledged solinviviruses and a great number of related, unclassified candidate viruses. The proteinase/polymerase gene region of two solinviviruses (Solenopsis invicta virus 3; Nylanderia fulva virus 1), additional 47 candidate solinviviruses, HPLV-65, and −75 and 12 reference strains of other *Picornavirales* families was investigated in a phylogenetic analysis (**Figure 6B**). The data revealed a monophyletic branch with solinvivirus-like viruses plus several clades of viruses which despite some genetic diversity are characterized by a second helicase domain with a Walker A motif.

Viruses With Unusual Genome Organization

One contig of 10,207 nt exhibited similarity to satsuma dwarf virus (GenBank acc. no. BAA74537; e-value: 7.6e-26) by

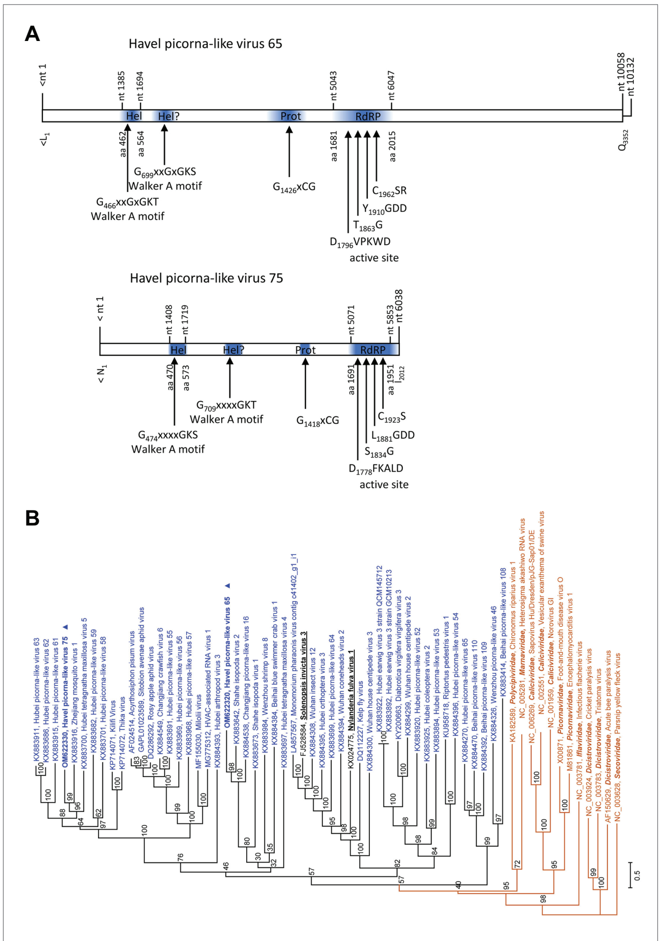


FIGURE 6 | Analysis of solinvivirus-like viruses. **(A)** Genome organization of Havel picorna-like virus 65 (top) and −75 (below). Conserved domains and sequence motifs were identified by searching the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). **(B)** Phylogenetic analyses of the proteinase/polymerase-encoding gene regions. Sequences of Havel picorna-like virus 65, −75, and two acknowledged solinviviruses (printed in bold and underlined), 47 unassigned solinvivirus candidates, and 12 reference viruses of the order *Picornavirales* (printed in brown) were included. The tree was inferred with IQ-Tree 2; optimal substitution model: TVM + F + R6. Presented are GenBank acc. nos. and virus names as well as genus names for the reference viruses. Unassigned viruses are printed in blue. A triangle (▲) indicates the viruses of the present study. Numbers at nodes indicate bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. Hel, helicase; Prot, proteinase; and RdRP, polymerase.

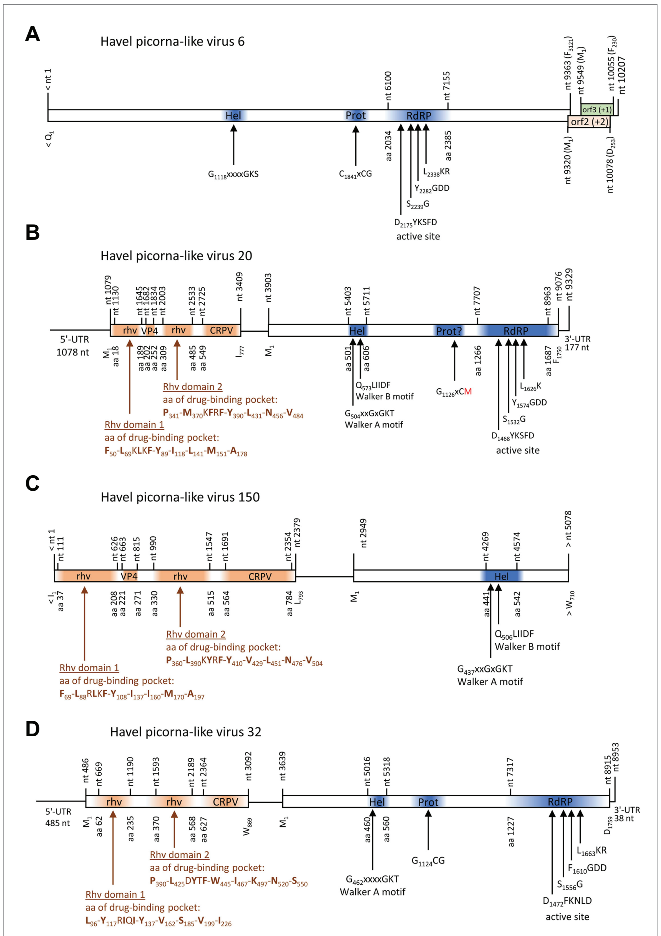


FIGURE 7 | Viruses with unusual genome layout. Genome organization of Havel picorna-like virus 6 **(A)**, −20 **(B)**, −150 **(C)**, and −32 **(D)**. Conserved domains and sequence motifs were identified by searching the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). UTR, untranslated region; Hel, helicase; Prot, proteinase; RdRP, polymerase; rhv, capsid protein with similarity to the rhinovirus capsid protein with jelly roll fold and drug-binding pocket; VP4, viral protein 4; and CRPV, capsid protein with similarity to the cricket paralysis virus capsid protein.

DIAMOND analysis and to maize chlorotic dwarf virus, potato U virus, and strawberry mottle virus of the *Secoviridae* by BLASTp search. This contig encodes a long ORF with helicase, proteinase, and RdRP domains plus two partly overlapping ORFs in the +1 and + 2 frames at the 3′-end (**Figure 7A**). No similarities of ORFs 2 and 3 to known proteins were found in additional BLAST and Pfam Conserved Domain Database searches. Phylogenetic analyses using the polymerase and proteinase/polymerase sequences (**Supplementary Figure 1**), as well as the helicase sequence (data not shown) yielded inconsistent results and failed to confirm a close relationship to secoviruses.

Three contigs, HPLV-32, −20, and −150, belonged to virus genomes with dicistronic layout but with the CP-encoding sequence located 5′ to the NSP-encoding gene region (**Figure 7**). Whereas almost complete genomes were obtained for HPLV-20 (9,252 nt) and -32 (8,785 nt), only a partial genome of HPLV-150 is available which is comprised of the CP region and the helicase domain of the NSP (5,078 nt). HPLV-32 clustered with other viruses with similar genome organization, namely, Trichosanthes kirilowii picorna-like virus pt111-pic-5, bat badiciviruses 1 and 2, the Aphis glycines virus 1, and soybean-associated bicistronic virus. Surprisingly, HPLV-20 grouped with the kusarnaviruses, indicating independent evolution. Phylogenetic analysis of the helicase indicates close relationship of HPLV-20 and -150 (data not shown).

DISCUSSION

Next,-generation sequencing techniques provide a powerful tool to analyze even unculturable viruses in various ecosystems. However, there is a disconcerting disparity in the small number of virus isolates with known hosts on the one hand and a plethora of sequences of uncultured viruses obtained from various sources on the other hand. For example, whereas the eight cultured viruses of the *Marnaviridae* infect protists like diatoms, raphidophyte, and thraustochytrids (Lang et al., 2021; Sadeghi et al., 2021),[4] at least a hundred times more candidate virus sequences are available in GenBank, obtained from very diverse sources such as environmental water and tissue/organ samples of marine invertebrates (e.g., Shi et al., 2016; Vlok et al., 2019; Wolf et al., 2020).

Previous studies of marine water samples have demonstrated a relative dominance of archaeal phages and bacteriophages, giant viruses, and single-stranded DNA and RNA viruses, but there is still a "vast viral unknown" (e.g., Hurwith and Sullivan, 2013; Labonté and Suttle, 2013; Schmidt et al., 2014; Roux et al., 2016; Liang et al., 2019; Callanan et al., 2020). At present, viromes of marine ecosystems are better studied than those of freshwater habitats and soils, even though interest in viruses of lotic and limnic systems, of the phytobiome or soils received increasing interest in recent years (Peduzzi, 2015; Williamson et al., 2017; Schoelz and Stewart, 2018; Roy et al., 2020). Meanwhile, available data indicate a considerably higher abundance of planktonic viruses in lakes than in sea water (Peduzzi, 2015). In the present study, we attempted a contribution to the understanding of virus diversity in a river. For this, we analyzed the enriched virus particles of a 50-liter water sample of the river Havel, taken within the metropolitan area. This river section is characterized by a near-natural river course with both recreational use and seasonal carefully controlled discharge of a wastewater treatment plant as well as occasionally drain water of the city of Berlin after heavy rainfalls. As characteristic for urban areas, the natural base discharge to this river is low, but wastewater effluent contributions under mean minimum discharge conditions vary from 30 to 50% (Karakurt et al., 2019). Previous analyses over a period of several years demonstrated little but significant pollution with human viruses like noroviruses, adenoviruses, hepatitis-E viruses, or cosaviruses, most notably in winter time, due to elaborate water safety management in the summer season (Beyer et al., 2020 and unpublished data and reports by the German Environment Agency). In our metagenomic river virome study, based on a large volume water sample from summer time, no PLVs similar to human, animal, or plant pathogens were detected. Likewise, posaviruses or related husa-, rasa-, basa-, or fisaviruses were also not identified. The latter ones are apparently apathogenic, enteric viruses indicating fecal pollution. This observation can also be explained by the assimilative capacity of the Havel river in summertime at elevated temperatures and high solar UV radiation. However, it should also be kept in mind that there is so far limited knowledge about the efficiency and putative inherent biases associated with virus enrichment protocols on metagenome analyses of viral richness, as demonstrated by Hjelmsø et al. (2017) for sewage metagenome analyses.

A total of 5,687 of 484,430 metaSPAdes scaffolds (1.17%) and 3,902 of 162,082 CLC contigs (2.4%) were identified by DIAMOND as PLV sequences (Taxonomy ID 464095). However, only 41.6 and 41.3%, respectively, of this fraction were assigned to a family, but many family and genus assignments are presumably incorrect as numerous randomly selected contigs were revealed to be misassigned (data not shown). Therefore, we adopted an alternative approach to identify contigs with similarity to one of the 12 *Picornavirales* families/sub-families in a tBLASTx search. Seventy-one almost complete HPLV genomes plus 93 partial genomes with lengths up to 10.95 kb were identified with this approach. At least 72 HPLVs are *Marnaviridae* candidates and about 60 were included in the pol, prot/pol, and CP trees (**Figure 1, Supplementary Figures 1, 2**). Whereas all so far acknowledged marnaviruses were detected in marine diatoms, unicellular algae, and heterotrophic protists, all recently proposed marnaviruses were from marine animals, marine algae, and coastal or estuarine water samples (Vlok et al., 2019; Lang et al., 2021). Presence of marnavirus candidates in the river Havel is compatible with the hypothesis that protists of freshwater habitats are also competent hosts. The present study, however, provides no data on possible host species. It is surprising that many marnavirus-like PLVs have been detected in plants, fecal samples, or cloacal swabs (e.g., sequences which were used in our phylogenetic analyses: GenBank acc. nos. KX644944, KY926885, MG995720, MN917672, MN917673, MN917674, MN823682, MN823683, MN823684, MN823685, MN823686, MN823687, MN823689, MN823691, MN823692, MT138127, MT138128, MT138129, MT138130, MT138131, MT138132, MT138133, and MT138336). Presence of PLVs in fecal samples or cloacal swabs is often explained by uptake of contaminated food or water. Occurrence of PLVs in plants of terrestrial habitats, however, requires either plant-protist contacts in the phytobiome or virus uptake *via* roots as has been shown for enteric viruses and bacteriophages (Murphy and Syverton, 1958; Ward and Mahler, 1982; Katzenelson and Mills, 1984; Urbanucci et al., 2009; Hirneisen et al., 2012).

Besides HPLV-102, −141, and − 159, no other dicistroviruses of the three genera *Aparavirus*, *Cripavirus*, and *Triatovirus* were identified among our HPLVs (**Figure 1, Supplementary Figure 1**). However, 50 sequences with significant similarity to dicistroviruses were detected. Three of four virus groups cluster close to sequences of the three dicistrovirus genera with high bootstrap values (>89% in the proteinase/polymerase tree). Other robust clades of the tree contain both viruses with dicistronic and monocistronic genomes. All clades are detectable in the CP tree indicating yet undefined virus groups which may represent new taxa. Of special interest is a clade comprising six dicistronic viruses with a CP polyprotein at the 5′-end of the genome.

HPLV-90, −93, and − 126 are calici-like viruses with similarity to viruses from lizards, bats, an unspecified insect, marine annelids, and bivalve shellfish. These viruses comprise a separate

[4] www.ictv.global/report/marnaviridae

clade in both the calicivirus VP1 tree and the proteinase/polymerase tree (**Figures 2B,C**). All acknowledged caliciviruses infect vertebrates, either mammals or fish. Detection in insects or bivalves may suggest invertebrate hosts but more likely contamination. Presence of such viruses in Havel river water and in filtrating shellfish is compatible with the assumption of a fish host and virus release into water. However, interspecies infection and ocean reservoirs have been described (Smith et al., 1980, 1998) and fish may be one of several possible hosts.

The *Picornaviridae* family is among the most divergent virus families in the virosphere. Sixty-eight genera, 158 species, and more than 650 types have been described (Zell et al., 2017).[5] The genome sequence of the first picornavirus infecting a fish was published in 2013 but descriptions of PLVs in fish, amphibia, and reptiles date back in the 1980s (Fichtner et al., 2013 and literature cited therein). Since then, numerous picornaviruses have been detected in lower vertebrates and meanwhile more than one hundred of such viruses are known (Shi et al., 2018). Therefore, it was surprising to detect only one picornavirus in Havel river water. HPLV-29 belongs to a third ampivirus type and additional sequences of the Havel river suggest the existence of further two ampivirus genotypes (**Supplementary Figures 4A,B**). The ampivirus was originally detected in fecal samples of smooth newts (Reuter et al., 2015). Later, similar viruses were found in unspecified freshwater arthropods (Shi et al., 2016) and (misclassified as totivirus) in cloacal swabs of flamingos and red-crowned cranes (see GenBank acc. nos. MT138174 and MT138399). As picornaviruses infect only vertebrates, freshwater arthropods are unlikely hosts but may have accumulated the virus in their gills. Experimental and natural accumulation of polioviruses and other enteroviruses in marine shellfish has been described (e.g., Hedström and Lycke, 1964; Metcalf and Stiles, 1965). Detection of ampiviruses in the Havel river water is compatible with the assumption of newts and other amphibia as hosts. However, a final answer whether amphibia or birds are true hosts of ampiviruses has to await virus isolation and clinical or experimental data.

Solinviviruses have only recently been described (Valles et al., 2014). We detected two viruses which belong to a monophyletic cluster of viruses comprised of the two acknowledged solinviviruses and a great number of related, unclassified candidate viruses (Brown et al., 2019).[6] One interesting feature of these viruses is the presence of a second protein domain with similarity to P-loop ATPases. All solinvivirus candidates have a second Walker A motif, either a complete version (GxxGxGK$^{S}/_{T}$) or a modified one.

Two viruses with unusual genome layouts, HPLV-20 and -150, exhibit marked similarity to members of the *Marnaviridae* family. Whereas the CP polyprotein shows similarity to the capsid of sogarnaviruses, proteinase and polymerase are kusarnavirus-like (**Supplementary Figures 1, 2**). Also the helicase is kusarnavirus-like (data not shown). Similar observations were described by Vlok et al. (2019). Several genera of the *Marnaviridae* exhibit a striking feature: mono- and dicistronic viruses are found in the same genus which is unusual in virus taxonomy. It has to be awaited additional sequence data to decide whether the genome layout of HPLV-20 is an exceptional feature or a third theme of marnavirus genome organization. Further, it is likely that some of our HPLVs represent members of novel virus genera and/or families. (i) Among such viruses are HPLV-9, a micalovirus-like dicistronic virus which clusters distinct from the known members of the *Picornavirales* but also distinct from other PLVs (see **Figure 1, Supplementary Figure S1**). Further candidates of novel families are the HPLVs with unusual genome layouts. (ii) The genome of HPLV-6 presents 3 ORFs with a missing 5′-end of ORF1 (**Figure 7**). Even so, ORF1 encodes a polyprotein of at least 3,121 aa which is rather long. No capsid proteins with jelly rolls motifs were detected by the Pfam conserved domain search tool. Also ORFs 2 and 3 lack similarity to other proteins of the NCBI protein database. However, presence of a Walker A motif, a CxCG proteinase active site sequence motif, and the characteristic RdRP active site motifs suggest a PLV. (iii) Another virus of interest is HPLV-32, a dicistronic virus with CP-encoding ORFs at the 5′-end of the genome (**Figure 7**). This virus clusters with Trichosanthes kirilowii picorna-like virus strain pt111-pic-5 and four other viruses with similar genome layout. Two of these viruses are from soybean aphids and two from feces of the straw-colored fruit bat (Yinda et al., 2017). Neither *Trichosanthes*, a plant of the cucumber family, nor aphids or bats are the likely hosts, but may be contaminated or may have ingested virus by food or water intake. As part of a complex phytobiome, aphids or fruit bats may play a role in virus distribution as well as the Havel river. (iii) Of note, viruses with monocistronic genome layouts but similarity of their proteinase/polymerase proteins to dicistroviruses are good candidates for novel virus taxa, as *per definitionem* they cannot be "dicistroviruses."

CONCLUSION

The Havel river in Berlin, Germany, is a near-natural river. It harbors highly divergent PLVs, most of which belong to the *Marnaviridae* family or are distant relatives of dicistroviruses likely representing new virus taxa. Whereas the majority of marnaviruses has been detected in seawater previously, an increasing number of marnavirus candidates are from freshwater habitats or terrestrial organisms. Dicistroviruses have been isolated from various arthropods, but there are many dicistrovirus-like sequences obtained from environmental water samples, plants, and vertebrate hosts. Both examples indicate more complex ecological connections of PLVs and their hosts. The role of rivers and lakes in the transmission and distribution of PLVs is so far insufficiently investigated and deserves more efforts to explore its relevance. However, as long as hosts are unknown, transmission cycles and the ecological importance of PLVs in the environment will remain obscure. Continuation of sequence data collection and progress in taxonomic classification are indispensable for an appropriate and advanced description of the virosphere diversity.

[5] www.ictv.global/report/picornaviridae

[6] www.ictv.global/report/solinviviridae

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: NCBI BioProject – PRJNA803428, BioSample SAMN25651207, GenBank (OM622256-OM622421).

AUTHOR CONTRIBUTIONS

RZ and H-CS: conception and study design and manuscript preparation. H-CS: responsibility for sampling, transport, and large scale virus enrichments. RZ: RNA preparation. MG: sequencing and sequence data processing. MG, RZ, and LS: data curation, bioinformatic analysis, and phylogenetic analyses. All authors have read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank M. Müller and I. Görlich for technical assistance and F. Kubek for help and expertise with sampling and virus enrichment techniques.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.865287/full#supplementary-material

Supplementary Figure 1 | Phylogenetic analysis of polymerase (left) and proteinase/polymerase-encoding sequences of dicistroviruses and marnaviruses. Two-hundred sixty-seven sequences of acknowledged dicistroviruses, marnaviruses and unassigned candidate viruses were aligned with MEGA for the polymerase tree, 264 sequences for the proteinase/polymerase tree. The trees were inferred with IQ-Tree 2, optimal substitution model: GTR + F + R9 for the pol tree and GTR + F + R10 for the prot/pol tree, respectively. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. Presented are GenBank acc. nos. and virus names. The respective genera are indicated. Colour code: aparaviruses, dark brown; bacillarnaviruses, blue; cripaviruses, light blue; kusarnaviruses, ochre; labyrnaviruses, light green; locarnaviruses, dark blue; marnaviruses, magenta; salisharnaviruses, dark green; sogarnaviruses, red; triatoviruses, brown, untypeable viruses, black. A triangle (▲) indicates viruses of the present study. Blue boxes indicate viruses with unusual genome layout (capsid protein-encoding gene region at 5'-end, nonstructural polyprotein-encoding gene region at the 3′-end). Yellow boxes indicate four dicistrovirus-like sequence clusters.

Supplementary Figure 2 | Phylogenetic analysis of the capsid protein-encoding sequences of dicistroviruses and marnaviruses. Two-hundred forty-two sequences of acknowledged dicistroviruses, marnaviruses and unassigned candidate viruses were aligned with MEGA. The tree was inferred with IQ-Tree 2, optimal substitution model: GTR + F + R9. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. Presented are GenBank acc. nos. and virus names. The respective genera are indicated. Colour code: aparaviruses, dark brown; bacillarnaviruses, blue; cripaviruses, light blue; kusarnaviruses, ochre; labyrnaviruses, light green; locarnaviruses, dark blue; marnaviruses, magenta; salisharnaviruses, dark green; sogarnaviruses, red; triatoviruses, brown, untypeable viruses, black. A triangle (▲) indicates viruses of the present study.

Supplementary Figure 3 | Phylogenetic analysis of the VP1-to-helicase gene region of 15 acknowleged iflavirus strains, 20 candidate strains and Havel picorna-like viruses 14 and -129. The tree was inferred with IQ-Tree 2, optimal substitution model: GTR + F + R5. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. Presented are GenBank acc. nos. and virus names. Unassigned viruses are printed in blue. A triangle (▲) indicates the viruses of the present study.

Supplementary Figure 4 | (A) Phylogenetic analysis of the capsid protein-encoding gene region (P1) of Havel picorna-like virus 29 and 162 picornaviruses. The tree was inferred with IQ-Tree 2, optimal substitution model: GTR+F+R9. Presented are GenBank acc. nos., genera (printed in italics and bold), virus names and strain designations (in square brackets). Sub-family names where available are given to the right. A triangle (▲) indicates the virus of the present study. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site. **(B)** Phylogenetic analysis of the proteinase/polymerase-encoding gene region (3CD) of Havel picorna-like virus 29 and 162 picornaviruses. The tree was inferred with IQ-Tree 2, optimal substitution model: GTR+F+R8. Presented are GenBank acc. nos., genera (printed in italics and bold), virus names and strain designations (in square brackets). Sub-family names where available are given to the right. A triangle (▲) indicates the virus of the present study. Numbers at nodes present bootstrap values obtained after 50,000 ultrafast bootstrap replications. The scale indicates substitutions per site.

REFERENCES

Bentley, D. R., Balasubramania, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Beyer, S., Szewzyk, R., Gnirss, R., Johne, R., and Selinka, H.-C. (2020). Detection and characterization of hepatitis-E virus genotype 3 in wastewater and urban surface waters in Germany. *Food Environ. Virol.* 12, 137–147. doi: 10.1007/s12560-020-09424-2

Bonning, B. C., and Miller, W. A. (2010). Dicistroviruses. *Annu. Rev. Entomol.* 55, 129–150. doi: 10.1146/annurev-ento-112408-085457

Brown, K., Olendraite, I., Valles, S. M., Firth, A. E., Chen, Y., Guérin, D. M. A., et al. (2019). ICTV virus taxonomy profile: Solinviviridae. *J. Gen. Virol.* 100, 736–737. doi: 10.1099/jgv.0.001242

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Callanan, J., Stockdale, S. R., Shkoporov, A., Draper, L. A., Ross, R. P., and Hill, C. (2020). Expansion of known ssRNA phage genomes: from tens to over a thousand. *Sci. Adv.* 6:eaay5981. doi: 10.1126/sciadv.aay5981

Chandrasekar, V., and Johnson, J. E. (1998). The structure of tobacco ringspot virus: a link in the evolution of icosahedral capsids in the picornavirus superfamily. *Structure* 6, 157–171. doi: 10.1016/S0969-2126(98)00018-5

Chen, Z., Stauffacher, C., Li, Y., Schmidt, T., Bomu, W., Kamer, G., et al. (1987). Protein-RNA interactions in an icosahedral virus at 3.0 Å resolution. *Science* 245, 154–159. doi: 10.1126/science.2749253

Coy, S. R., Gann, E. R., Pound, H. L., Short, S. M., and Wilhelm, S. W. (2018). Viruses of eukaryotic algae: diversity, methods for detection, and future directions. *Viruses* 10:487. doi: 10.3390/v10090487

Culley, A. I., Lang, A. S., and Suttle, C. A. (2003). High diversity of unknown picorna-like viruses in the sea. *Nature* 424, 1054–1057. doi: 10.1038/nature01886

Culley, A. I., Mueller, J. A., Belcaid, M., Wood-Charison, E. M., Poisson, G., and Steward, G. F. (2014). The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. *mBio* 5, e01210–e01214. doi: 10.1128/mBio.01210-14

Culley, A. I., and Steward, G. F. (2007). New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Appl. Environ. Microbiol.* 73, 5937–5944. doi: 10.1128/AEM.01065-07

Dastjerdi, A., Everest, D. J., Davies, H., Denk, D., and Zell, R. (2021). A novel dicistrovirus in a captive red squirrel (*Sciurus vulgaris*). *J. Gen. Virol.* 102. doi: 10.1099/jgv.0.001555

Duraisamy, R., Akiana, J., Davoust, B., Mdeiannikov, O., Michelle, C., Robert, C., et al. (2018). Detection of novel RNA viruses from free-living gorillas, republic of Congo: genetic diversity of picobirnaviruses. *Virus Genes* 54, 256–271. doi: 10.1007/s11262-018-1543-6

Fichtner, D., Philipps, A., Groth, M., Schmidt-Posthaus, H., Granzow, H., Dauber, M., et al. (2013). Characterization of a novel picornavirus isolate from a diseased European eel (*Anguilla anguilla*). *J. Virol.* 87, 10895–10899. doi: 10.1128/JVI.01094-13

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.* 15, 579–590. doi: 10.1038/nrmicro.2017.87

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548. doi: 10.1038/21119

Hedström, C. E., and Lycke, E. (1964). An experimental study on oysters as virus carriers. *Am. J. Hyg.* 79, 134–142.

Hirneisen, K. A., Sharma, M., and Kniel, K. E. (2012). Human enteric pathogen internalization by root uptake into food crops. *Foodborne Pathog. Dis.* 9, 396–405. doi: 10.1089/fpd.2011.1044

Hjelmsø, M. H., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., et al. (2017). Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. *PLoS One* 12:e0170199. doi: 10.1371/journal.pone.0170199

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hurwith, B. L., and Sullivan, M. B. (2013). The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* 8:e57355. doi: 10.1371/journal.pone.0057355

Jan, E. (2006). Divergent IRES elements in invertebrates. *Virus Res.* 119, 16–28. doi: 10.1016/j.virusres.2005.10.011

Karakurt, S., Schmid, L., Hübner, U., and Drewes, J. E. (2019). Dynamics of wastewater effluent contributions in streams and impacts on drinking water supply via riverbank filtration in Germany—A national reconnaissance. *Environ. Sci. Technol.* 53, 6154–6161. doi: 10.1021/acs.est.8b07216

Katzenelson, E., and Mills, D. (1984). Contamination of vegetables with animal viruses via the roots. *Monogr. Virol.* 15, 216–220. doi: 10.1159/000409140

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Labonté, J. M., and Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* 7, 2169–2177. doi: 10.1038/ismej.2013.110

Lang, A. S., Rise, M. L., Culley, A. I., and Steward, G. F. (2018). RNA viruses in the sea. *FEMS Microbiol. Rev.* 33, 295–323. doi: 10.1111/j.1574-6976.2008.00132.x

Lang, A. S., Vlok, M., Culley, A. I., Suttle, C. A., Takao, Y., Tomaru, Y., et al. (2021). ICTV virus taxonomy profile: *Marnaviridae* 2021. *J. Gen. Virol.* 102:001633. doi: 10.1099/jgv.0.001633

Le Gall, O., Christian, P., Fauquet, C. M., King, A. M. Q., Knowles, N. J., Nakashima, N., et al. (2008). *Picornavirales*, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture. *Arch. Virol.* 153, 715–727. doi: 10.1007/s00705-008-0041-x

Lefeuvre, P., Martin, D. P., Elena, S. F., Shepherd, D. N., Roumagnac, P., and Varsani, A. (2019). Evolution and ecology of plant viruses. *Nat. Rev. Microbiol.* 17, 632–644. doi: 10.1038/s41579-019-0232-3

Liang, Y., Wang, L., Wang, Z., Zhao, J., Yang, Q., Wang, M., et al. (2019). Metagenomic analysis of the diversity of DNA viruses in the surface and deep sea of the South China Sea. *Front. Microbiol.* 10:1951. doi: 10.3389/fmicb.2019.01951

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17:200. doi: 10.14806/ej.17.1.200

Metcalf, T. G., and Stiles, W. C. (1965). The accumulation of enteric viruses by the oyster *Crassostrea virginica*. *J. Inf. Dis.* 115, 68–76. doi: 10.1093/infdis/115.1.68

Murphy, W. H. Jr., and Syverton, J. T. (1958). Absorption and translocation of mammalian viruses by plants: II. Recovery and distribution of viruses in plants. *Virology* 6, 623–636. doi: 10.1016/0042-6822(58)90111-9

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi: 10.1101/gr.213959.116

Olendraite, I., Brown, K., Valles, S. M., Firth, A. E., Chen, Y., Guérin, D. M. A., et al. (2019). ICTV virus tasonomy profile: *Polycipiviridae*. *J. Gen. Virol.* 100, 554–555. doi: 10.1099/jgv.0.000672

Olendraite, I., Lukhovitskaya, N. I., Porter, S. D., Valles, S. M., and Firth, A. E. (2017). Polycipiviridae: a proposed new family of polycistronic picorna-like viruses. *J. Gen. Virol.* 98, 2368–2378. doi: 10.1099/jgv.0.000902

Peduzzi, P. (2015). Virus ecology of fluvial systems: a blank spot on the map? *Biol. Rev.* 91, 937–949. doi: 10.1111/brv.12202

Prasad, B. V., Hardy, M. E., Dokland, T., Bella, J., Rossmann, M. G., and Estes, M. K. (1999). X-ray crystallographic structure of the Norwalk virus capsid. *Science* 286, 287–290. doi: 10.1126/science.286.5438.287

Reuter, G., Boros, A., Toth, Z., Phan, T. G., Delwart, E., and Pankovics, P. (2015). A highly divergent picornavirus in an amphibian, the smooth newt (*Lissotritron vulgaris*). *J. Gen. Virol.* 96, 2607–2613. doi: 10.1099/vir.0.000198

Reuter, G., Pankovics, P., Gyöngyi, Z., Delwart, E., and Boros, A. (2014). Novel dicistrovirus brom bat guano. *Arch. Virol.* 159, 3453–3456. doi: 10.1007/s00705-014-2212-2

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297. doi: 10.1016/j.coviro.2011.06.004

Rossmann, M. G., and Johnson, J. E. (1989). Icosahedral RNA virus structure. *Annu. Rev. Biochem.* 58, 533–569. doi: 10.1146/annurev.bi.58.070189.002533

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537, 689–693. doi: 10.1038/nature19366

Roy, K., Ghosh, D., DeBruyn, J. M., Dasgupta, T., Wommack, K. E., Liang, X., et al. (2020). Temporal dynamics of soil virus and bacterial populations in agricultural and early plant successional soils. *Front. Microbiol.* 11:1494. doi: 10.3389/fmicb.2020.01494

Sadeghi, M., Toraru, Y., and Ahola, T. (2021). RNA viruses in aquatic unicellular eukaryotes. *Viruses* 13:362. doi: 10.3390/v13030362

Sanfaçon, H., Gorbalenya, A. E., Knowles, N. J., and Chen, Y. P. (2012). "Order *Picornavirales*," in *Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses*. eds. A. M. Q. King, M. J. Adams, E. B. Carstens and E. J. Lefkowitz (Amsterdam, Elsevier Academic Press), 835–839.

Sanfaçon, H., Wellink, J., Le Gall, O., Karasev, A., van der Vlugt, R., and Wetzel, T. (2009). Secoviridae: a proposed family of plant viruses within the order *Picornavirales* that combines the families *Sequiviridae* and *Comoviridae*, the unassigned genera *Cheravirus* and *Sadwavirus*, and the proposed genus Torradovirus. *Arch. Virol.* 154, 899–907. doi: 10.1007/s00705-009-0367-z

Schmidt, H. F., Sakowski, E. G., Williamson, S. J., Polson, S. W., and Wommack, K. E. (2014). Shotgun metagenomics indicates novel family A DNA polymerases predominate within marine virioplankton. *ISME J.* 8, 103–114. doi: 10.1038/ismej.2013.124

Schoelz, J. E., and Stewart, L. R. (2018). The role of viruses in the phytobiome. *Annu. Rev. Virol.* 5, 93–111. doi: 10.1146/annurev-virology-092917-043421

Shi, M., Lin, X. D., Chen, X., Tian, J. H., Chen, L. J., Li, K., et al. (2018). The evolutionary history of vertebrate RNA viruses. *Nature* 556, 197–202. doi: 10.1038/s41586-018-0012-7

Shi, M., Lin, X. D., Tian, J. H., Chen, L. J., Chen, X., Li, C. X., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature* 540, 539–543. doi: 10.1038/nature20167

Smertina, E., Hall, R. N., Urakova, N., Strive, T., and Frese, M. (2021). Calicivirus non-structural proteins: potential functions in replication and host cell manipulation. *Front. Microbiol.* 12:712710. doi: 10.3389/fmicb.2021.712710

Smith, A. W., Skilling, D. E., Cherry, N., Mead, J. H., and Matson, D. O. (1998). Calicivirus emergence from ocean reservoirs: zoonotic and interspecies movements. *Em. Inf. Dis.* 4, 13–20. doi: 10.3201/eid0401.980103

Smith, A. W., Skilling, D. E., Dardiri, A. H., and Latham, A. B. (1980). Calicivirus pathogenic for swine: a new serotype isolated from opaleye *Girella nigricans*, an ocean fish. *Science* 209, 940–941. doi: 10.1126/science.7403862

Thompson, J. R., Dasgupta, I., Fuchs, M., Iwanami, T., Karasev, A. V., Petrzik, K., et al. (2017). ICTV virus taxonomy profile: *Secoviridae*. *J. Gen. Virol.* 98, 529–531. doi: 10.1099/jgv.0.000779

Urbanucci, A., Myrmel, M., Berg, I., von Bonsdorff, C. H., and Maunula, L. (2009). Potential internalisation of caliciviruses in lettuce. *Int. J. Food Microbiol.* 135, 175–178. doi: 10.1016/j.ijfoodmicro.2009.07.036

Valles, S. M., Bell, S., and Firth, A. E. (2014). Solenopsis invicta virus 3: mapping of structural proteins, ribosomal frameshifting, and similarities to Acyrthosiphon pisum virus and kelp fly virus. *PLoS One* 9:e93497. doi: 10.1371/journal.pone.0093497

Valles, S. M., Chen, Y., Firth, A. E., Guérin, D. M. A., Hashimoto, Y., Herrero, S., et al. (2017a). ICTV virus taxonomy profile: *Dicistroviridae*. *J. Gen. Virol.* 98, 355–356. doi: 10.1099/jgv.0.000756

Valles, S. M., Chen, Y., Firth, A. E., Guérin, D. M. A., Hashimoto, Y., Herrero, S., et al. (2017b). ICTV virus taxonomy profile: *Iflaviridae*. *J. Gen. Virol.* 98, 527–528. doi: 10.1099/jgv.0.000757

Valles, S. M., Strong, S. A., and Hashimoto, Y. (2007). A new positive-strand RNA virus with unique genome characteristics from the red imported fire ant, *Solenopsis invicta*. *Virology* 365, 457–463. doi: 10.1016/j.virol.2007.03.043

van der Vlugt, R. A. A., Verbeek, M., Dullmans, A. M., Wintermantel, W. M., Cuellar, W. J., Fox, A., et al. (2015). Torradoviruses. *Annu. Rev. Phytopathol.* 53, 485–512. doi: 10.1146/annurev-phyto-080614-120021

Vlok, M., Lang, A. S., and Suttle, C. A. (2019). Application of a sequence-based taxonomic classification method to uncultivated and unclassified marine single-stranded RNA viruses in the order *Picornavirales*. *Virus Evol.* 5:vez056. doi: 10.1093/ve/vez056

Ward, R. L., and Mahler, R. J. (1982). Uptake of bacteriophage f2 through plant roots. *Appl. Environ. Microbiol.* 43, 1098–1103. doi: 10.1128/aem.43.5.1098-1103.1982

Wilhelm, S. W., and Suttle, C. A. (1999). Viruses and nutrient cycles in the sea. *Bioscience* 49, 781–788. doi: 10.2307/1313569

Williamson, K. E., Fuhrmann, J. J., Wommack, K. E., and Radosevich, M. (2017). Viruses in soil ecosystems: an unknown quantity within an unexplored territory. *Annu. Rev. Virol.* 4, 201–219. doi: 10.1146/annurev-virology-101416-041639

Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kaslauskas, D., et al. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat. Microbiol.* 5, 1262–1270. doi: 10.1038/s41564-020-0755-4

Wommack, K. E., and Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* 64, 69–114. doi: 10.1128/MMBR.64.1.69-114.2000

Wommack, K. E., Nasko, D. J., Chopyk, J., and Sakowski, E. G. (2015). Counts and sequences, observations that continue to change our understanding of viruses in nature. *J. Microbiol.* 53, 181–192. doi: 10.1007/s12275-015-5068-6

Wyn-Jones, A. P., Carducci, A., Cook, N., D'Agostino, M. D., Divizia, M., Fleischer, J., et al. (2011). Surveillance of adenoviruses and noroviruses in European recreational waters. *Water Res.* 45, 1025–1038. doi: 10.1016/j.watres.2010.10.015

Yinda, C. K., Zell, R., Deboutte, W., Zeller, M., Conceicao-Neto, N., Heylen, E., et al. (2017). Highly diverse population of *Picornaviridae* and other members of the *Picornavirales*, in Cameroonian fruit bats. *BMC Genomics* 18:249. doi: 10.1186/s12864-017-3632-7

Zell, R. (2018). *Picornaviridae*-the ever-growing virus family. *Arch. Virol.* 163, 299–317. doi: 10.1007/s00705-017-3614-8

Zell, R., Delwart, E., Gorbalenya, A. E., Hovi, T., King, A. M. Q., Knowles, N. J., et al. (2017). ICTV virus taxonomy profile: *Picornaviridae*. *J. Gen. Virol.* 98, 2421–2422. doi: 10.1099/jgv.0.000911

Zinoviev, A., Hellen, C. U. T., and Pestova, T. V. (2015). Multiple mechanisms of reinitiation on bicistronic calicivirus mRNAs. *Mol. Cell* 57, 1059–1073. doi: 10.1016/j.molcel.2015.01.039

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 07 April 2022
doi: 10.3389/fmicb.2022.848305

Distribution Characteristics of Soil Viruses Under Different Precipitation Gradients on the Qinghai-Tibet Plateau

Miao-Miao Cao[1,2], *Si-Yi Liu*[1,3], *Li Bi*[4], *Shu-Jun Chen*[5], *Hua-Yong Wu*[6], *Yuan Ge*[1], *Bing Han*[1,2], *Li-Mei Zhang*[1,2], *Ji-Zheng He*[4,7] *and Li-Li Han*[1]*

[1] *State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China,* [2] *University of Chinese Academy of Sciences, Beijing, China,* [3] *The Zhongke-Ji'an Institute for Eco-Environmental Sciences, Ji'an, China,* [4] *Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC, Australia,* [5] *Information Technology Center, Tsinghua University, Beijing, China,* [6] *State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, China,* [7] *Key Laboratory for Humid Subtropical Eco-Geographical Processes of the Ministry of Education, Fujian Normal University, Fuzhou, China*

OPEN ACCESS

Edited by:
Mao Ye,
Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science (CAS), China

Reviewed by:
Shuai Le,
Army Medical University, China
Jose Luis Balcazar,
Catalan Institute for Water Research, Spain

***Correspondence:**
Li-Li Han
llhan@rcees.ac.cn

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *04 January 2022*
Accepted: *07 February 2022*
Published: *07 April 2022*

Citation:
Cao M-M, Liu S-Y, Bi L, Chen S-J, Wu H-Y, Ge Y, Han B, Zhang L-M, He J-Z and Han L-L (2022) Distribution Characteristics of Soil Viruses Under Different Precipitation Gradients on the Qinghai-Tibet Plateau. Front. Microbiol. 13:848305. doi: 10.3389/fmicb.2022.848305

Viruses are extremely abundant in the soil environment and have potential roles in impacting on microbial population, evolution, and nutrient biogeochemical cycles. However, how environment and climate changes affect soil viruses is still poorly understood. Here, a metagenomic approach was used to investigate the distribution, diversity, and potential biogeochemical impacts of DNA viruses in 12 grassland soils under three precipitation gradients on the Qinghai-Tibet Plateau, which is one of the most sensitive areas to climate change. A total of 557 viral operational taxonomic units were obtained, spanning 152 viral families from the 30 metagenomes. Both virus-like particles (VLPs) and microbial abundance increased with average annual precipitation. A significant positive correlation of VLP counts was observed with soil water content, total carbon, total nitrogen, soil organic matter, and total phosphorus. Among these biological and abiotic factors, SWC mainly contributed to the variability in VLP abundance. The order *Caudovirales* (70.1% of the identified viral order) was the predominant viral type in soils from the Qinghai-Tibet Plateau, with the *Siphoviridae* family being the most abundant. Remarkably, abundant auxiliary carbohydrate-active enzyme (CAZyme) genes represented by glycoside hydrolases were identified, indicating that soil viruses may play a potential role in the carbon cycle on the Qinghai-Tibet Plateau. There were more diverse hosts and abundant CAZyme genes in soil with moderate precipitation. Our study provides a strong evidence that changes in precipitation impact not only viral abundance and virus–host interactions in soil but also the viral functional potential, especially carbon cycling.

Keywords: soil viruses, metagenome, precipitation, abundance, diversity, carbon cycle

INTRODUCTION

Viruses are the most abundant lifeform on Earth and highly encompass their biodiversity (Paez-Espino et al., 2016). Previous marine viral ecology studies demonstrated that viruses play a crucial part in the environment. Firstly, viruses can have an influence on microbial populations and evolution by modulating and controlling the abundance, diversity, and functional processes of a host through cell lysis (Thingstad, 2000) or lysogeny (Knowles et al., 2016). Two models were built to explain the viruses' role on microbial populations. For example, the "kill-the-winner" (KtW) model of lytic infection predicts that density- and frequency-dependent viral predation suppresses the blooms of rapidly growing hosts, maintaining and increasing the stability and diversity of host communities (Thingstad and Lignell, 1997). In contrast, the "piggyback-the-winner" (PtW) model of lysogenic infection predicts that increasing host density will enhance lysogenic incidence (i.e., "more microbes, fewer viruses") (Knowles et al., 2016).

Secondly, viruses also participate in biogeochemical cycles. For example, microorganisms are killed *via* viral infection and lysis, resulting in the release of nutrients and absorption by other microorganisms and plants, such as carbon, nitrogen, phosphorus, and sulfur (Kuzyakov and Mason-Jones, 2018). In marine systems, virus-driven carbon cycling is reported to account for 6%∼26% of the total carbon cycling (Weinbauer, 2004). Moreover, viruses indirectly impact biogeochemical cycles through abundant virus-encoded auxiliary metabolic genes (AMGs), including the genes related to the carbon (Hurwitz et al., 2013; Jin et al., 2019), nitrogen (Roux et al., 2016; Ahlgren et al., 2019; Gazitua et al., 2021), sulfur (Roux et al., 2016; Kieft et al., 2021), and phosphorus (Zeng and Chisholm, 2012; Goldsmith et al., 2015) cycles, which are discovered in marine and terrestrial ecosystems.

The previous researches about viruses were mainly focused on marine ecosystems. Compared with marine systems [ranging from 1 to 100; (Dion et al., 2020)], studies have shown a more highly variable range of virus-to-bacteria ratio (VBR) in soil [ranging from 0.001 to 8,200; (Williamson et al., 2017)]. In addition, viral communities in terrestrial ecosystems showed significant environmental specificity. Fierer et al. (2007) had demonstrated that the soil viral communities were significantly different from that in other environments (marine sediment, human fecal samples, and seawater environments) *via* metagenomic analysis. Therefore, facing the huge number of soil viruses and important ecological functions, we urgently need to pay more attention to the studies of soil viruses.

So far, studies on soil viral ecology have involved many soil types, such as desert soil (Adriaenssens et al., 2015), glacier soil (Han et al., 2017b), thawing permafrost soil (Trubl et al., 2016, 2018, 2019), forest soil (Jin et al., 2019; Liang et al., 2021), mud volcanic soil (Yu et al., 2019), agricultural soil (Bi et al., 2021), and wetlands (Jackson and Jackson, 2008). It has been confirmed that the abundance, diversity, and reproductive strategy of viruses in soil are affected by multiple factors. Soil pH was the main environmental driver of the viral community structure in agricultural soils (Bi et al., 2021) and also affected the viral attachment to soil particles (Gerba, 1984; Loveland et al., 1996). Soil type was significantly correlated with viral abundance (Williamson et al., 2017). In addition to the factors mentioned, the site altitude (Adriaenssens et al., 2017), temperature (Wen et al., 2004), soil organic matter (Chariou et al., 2019), soil water content (SWC) (Straub et al., 1992, 1993; Jin and Flury, 2002), etc., also impacted virus–host interaction. Wu et al. (2021) have verified that extreme changes in soil moisture have a great influence on the composition, activity, and potential functions of both DNA and RNA soil viruses. When the SWC decreased to < 5%, the infection ability of phage was significantly weakened, and the activity of phage was completely lost due to water volatilization (Straub et al., 1992, 1993). There was a distinct positive correlation between viral abundance and the SWC (Williamson et al., 2005).

Climate change is subtly changing the soil environment, affecting the species composition, community structure, and function of soil organisms, but how it affects soil viruses is still poorly understood (Kaisermann et al., 2017; Dong et al., 2020). For instance, the changes of precipitation lead to shift soil moisture, influencing soil biotic and abiotic properties (Cook et al., 2014). As mentioned, the SWC affects viral activity. However, it is currently unknown how changes in precipitation shape the soil virosphere and influence viral diversity, abundances, function, and replication strategies (lytic/lysogenic lifecycles) (Emerson, 2019; Van Goethem et al., 2019). The Qinghai-Tibet Plateau, known as the "the third pole" of the Earth, is considered to be one of the most sensitive areas to human activities and climate change and has become a research hotspot (Chen et al., 2013; Tao et al., 2018; Che et al., 2019). Despite the ecological importance of the Qinghai-Tibet Plateau, our knowledge on its biodiversity is notably limited. While some studies have focused on soil microbial community composition and diversity, including bacteria (Shen et al., 2019), fungi (Che et al., 2019), and archaea (Shi et al., 2019), there have been few reports of soil viruses. Based on the above views, this paper mainly focuses on (i) the abundance of soil viruses and microbes in the Qinghai-Tibet Plateau under different precipitation gradients and its potential environmental drivers; (ii) the differences of virus community composition and predicted hosts under different precipitation gradients; and (iii) the potential role in the biogeochemical cycling of soil viruses in the Qinghai-Tibet Plateau.

MATERIALS AND METHODS

Sample Collection and Soil Physicochemical Properties

A total of 36 soil samples, located in the Qinghai Province of China, were selected from 12 field sites under three precipitation gradients (**Figure 1**). Samples in low precipitation (LP, mean annual precipitation < 200 mm) include LP_32, LP_34, LP_35, and LP_36. Samples in moderate precipitation (MP, mean annual precipitation is between 200 and 400 mm) include MP_2, MP_7,

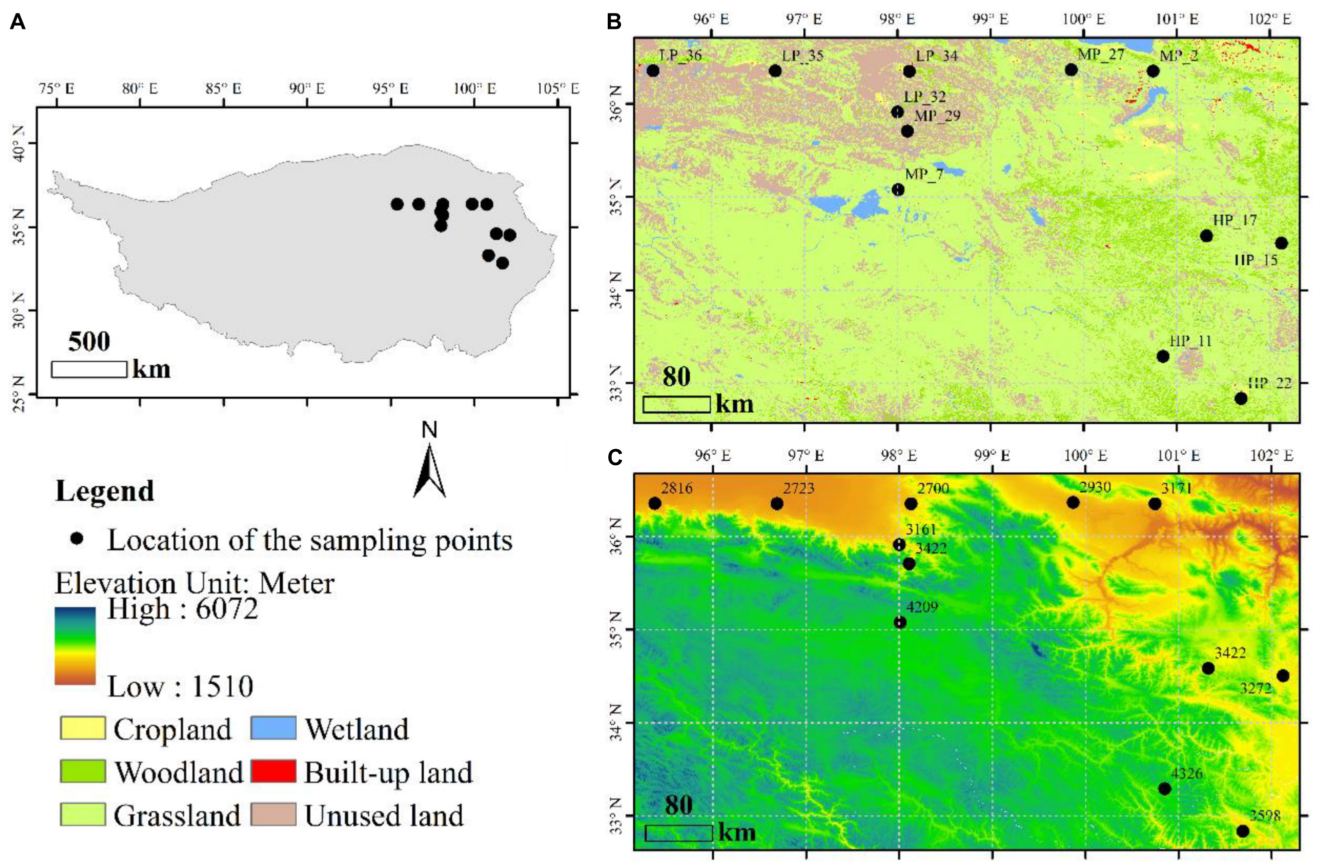


FIGURE 1 | Distribution and description of sample sites, including geographical location **(A)**, land use type **(B)** and altitude **(C)**.

MP_27, and MP_29. Samples in high precipitation (HP, mean annual precipitation > 400 mm) include HP_11, HP_15, HP_17, and HP_22. All samples were from grasslands (**Figure 1B**) with the detailed information in **Supplementary Table 1**. Three adjacent samples were randomly taken from each site by inserting the coring device 10 cm into the soil surface. The coring device was cleaned with ethanol (70%, v/v) between sites to avoid cross sample contamination. Each of the soil samples was sieved through a 2 mm sieve to remove rocks and roots, mixed evenly, and placed into a clean zip-lock bag for bioinformatics and physicochemical analysis.

The soil physicochemical characteristics were determined according to established protocols (Ge et al., 2021; Shi et al., 2021). Briefly, soil pH was determined with a soil- to-water ratio of 1:2.5 (w/w) suspension using a pH meter (DELTA-320, China). Soil electrical conductivity (EC) was measured with a soil-to- water ratio of 1:5. SWC was determined by the oven-drying method at 105°C to constant weight. Soil organic matter (SOM) was estimated by the $K_2Cr_2O_7$ oxidation-reduction colorimetric method. The TC and TN were determined by an elemental analyzer (Vario EL III-Elementar, Germany). $NH_4{}^+$-N and $NO_3{}^-$-N were extracted with 1 M KCl and measured by a continuous flow analyzer (SAN + +, Skalar, Holland). The soil's total phosphorus (TP) was extracted by NaOH solution and determined by Mo-Sb colorimetric method (Wang et al., 2014).

Epifluorescence Microscopy Enumeration

Virus-like particles (VLPs) and microbial abundances in each soil sample were estimated using epifluorescence microscopy (EFM) (Danovaro and Serresi, 2000; Thurber et al., 2009; Han et al., 2017a). 10 ml of 0.22 μm filtered amended 1% potassium citrate (AKC) buffer [1% potassium citrate resuspension buffer amended with 10% phosphate buffered-saline (PBS) and 150 mM magnesium sulfate ($MgSO_4$)] was added to 3 ± 0.5 g soil (Trubl et al., 2016). Viruses and microbes were physically dispersed *via* 15 min of shaking at 150 rpm at room temperature. The supernatant-contained viruses and microbes were collected to new tubes. The resuspension steps above were repeated two more times to obtain approximately 30 ml supernatants and then filtered through a 0.45 μm Millex filter to remove large particles. The moderate filtrate was disposed with 5 U/ml DNase I (Thermo Fisher Scientific, Lithuania, European Union) and incubated at 37°C for 1 h to remove free DNA, then filtered through a 0.02 μm Anodisc Al_2O_3 filter membrane (25 mm diameter, Whatman; GE Healthcare, Kent, United Kingdom) supported by a 0.45 μm filter. The dried Anodisc filter membrane was stained with SYBR Green I (Invitrogen, Eugene, OR, United States) working solution (1:400) for 20 min in darkness at room temperature. The stained filter membrane was then mounted on a glass slide with ab antifade solution [50% glycerol, 50% phosphate-buffered saline (0.05 M Na_2HPO_4, 0.85% NaCl; pH 7.5), 0.1% p-phenylenediamine]. Subsequently, ten fields of view were randomly selected for observation under EFM (Nikon, Melville, NY, United States), and the abundances of VLPs and microbial cells and VMR were calculated.

DNA Extraction, Library Construction, and Metagenomic Sequencing

The total DNA was extracted from 0.25 g of soil with the PowerLyser PowerSoil DNA isolation kit (Qiagen, Hilden, Germany), following the manufacturer's protocol. The DNA extract was fragmented to an average size of about 400 bp using Covaris M220 (Gene Company Limited, Shanghai, China) for paired-end library construction. A paired-end library was constructed using NEXTFLEX Rapid DNA-Seq (Bioo Scientific, Austin, TX, United States). Paired-end sequencing was performed on the Illumina NovaSeq platform (Illumina Inc., San Diego, CA, United States) at Majorbio Bio-Pharm Technology Co., Ltd. (Shanghai, China) using NovaSeq Reagent Kits according to the manufacturer's instructions.[1] The sequencing depth was 20 Gbp for each sample. The DNA content of LP_34 and LP_36 samples is too little to meet the requirements of library construction.

Analysis of Metagenomes

Quality Control, Assembly, and Identification of Viral Operational Taxonomic Units

The raw reads of the 30 samples obtained from the Illumina NovaSeq platform were cleaned using the Fastp tool to remove adapters and filter low-quality reads (length $<$ 50 bp or with a quality value $<$ 20 or having N bases) (Chen et al., 2018), followed by *de novo* assembly into contigs $\geq$ 500 bp in length under default conditions using MEGAHIT (Li et al., 2015) and clustered with PSI-CDHIT (Huang et al., 2010). Then, VirSorter (Roux et al., 2015), VIBRANT (Kieft et al., 2020), and DeepVirFinder (Ren et al., 2020) were used to detect dsDNA viral contigs from each assembly ($\geq$ 1,000 bp contigs). Only $\geq$ 10 kb contigs were retained according to Minimum Information about an Uncultivated Virus Genome (Roux et al., 2019). Based on the Discovery Environment 2.0,[2] only contigs from VirSorter categories 1, 2, 4, and 5 (high confidence) were retained, and the combined phages in VIBRANT were considered viral. DeepVirFinder runs according to its Python script,[3] and the contigs with scores $>$ 0.9 and $p < 0.05$ were considered viral. The identified viral contigs were then compiled and clustered at 98% nucleotide identity using cd-hit-est software (Li and Godzik, 2006), totally producing 557 viral operational taxonomic units (vOTUs).

Taxonomy and Functional Annotation and Host Prediction

Clean reads were classified using Kraken2 (Wood et al., 2019) against the reference sequences of the National Center for Biotechnology Information (NCBI) database (RefSeq, accessed December 2021) to identify bacterial, archaeal, and viral reads, and the relative abundance of viruses was directly obtained from the classification results of Kraken 2. The CAZymes identified from 557 vOTUs were automated on the dbCAN2 meta server based on CAZyme family specific HMMER (E-value $< 1e^{-15}$, coverage $>$ 0.35), DIAMOND (E-value $< 1e^{-102}$), and Hotpep (frequency $>$ 2.6, hits $>$ 6) together (Zhang et al., 2018). The information of integrase was selected from identified putative viral AMGs in 557 vOTUs by DRAM-v (Shaffer et al., 2020); only the results with viral_bitScore $>$ 60 and viral_E-value $< 1e^{-5}$ were retained. Similarly, putative hosts were predicted from 557 vOTUs as an input file using PHISDetector (default parameters), a web tool that detects diverse *in silico* phage–host interaction signals (Zhou et al., 2020).

Statistical Analyses of the Metagenomes

The difference analysis of VLPs and bacterial abundance under different precipitate-on gradients were completed by Statistical Product and Service Solutions (SPSS) Statistics 26. Random forest models (Breiman, 2001) used to evaluate the relative importance of various factors influencing VLP abundance were performed by the "randomForest" and "rfPermute" packages on the R platform. LEfSe (linear discriminant analysis effect size) analysis was performed based on $p < 0.05$ and an LDA score $>$ 2.0.[4] Virus–host prediction results were generated manually by the Adobe Illustrator CS6.

Data Availability

Metagenome read data are available in the NCBI Short Read Archive under BioProject ID PRJNA782356.

RESULTS

Virus-Like Particles and Microbial Abundance and Environmental Driving Factors

Viruses and microbes in soil were enumerated using EFM. The abundance of virus-like particles (VLPs) ranged from 2.0×10^7 (LP_36) to 1.0×10^{10} (HP_22) per gram of dry soil, and microbial abundance ranged from 1.0×10^8 (MP_29) to 8.2×10^8 (HP_15) per gram of dry soil. The virus-to-microbe ratio (VMR) also varied in different soils, ranging from 0.11 to 98.3 (**Table 1**). There were significant differences in VLP abundance among different precipitation gradients. The VLP abundance under HP was the highest, followed by MP and LP. Furthermore, the microbial abundance under HP was significantly higher than that under MP and LP. Interestingly, the VMR under HP was significantly higher than that under LP (**Table 1**).

[1] www.illumina.com

[2] https://de.cyverse.org

[3] https://github.com/jessieren/DeepVirFinder

[4] http://huttenhower.sph.harvard.edu/galaxy

Pearson correlations were used to evaluate the factors that might influence the soil VLP abundance, microbial abundance, and VMR. Across all samples, soil VLPs have a significant positive correlation with microbial abundance (r = 0.491, $p < 0.01$) (**Figure 2A**). The results showed significant positive correlations of VLPs and microbe counts with SWC (virus, r = 0.868, $p < 0.001$; microbe, r = 0.487, $p < 0.01$), total carbon (TC; virus, r = 0.701, $p < 0.001$; microbe, r = 0.399, $p < 0.05$), total nitrogen (TN; virus, r = 0.711, $p < 0.001$; microbe, r = 0.514, $p < 0.01$), soil organic matter (SOM; virus, r = 0.715, $p < 0.001$; microbe, r = 0.500, $p < 0.01$) and TP (virus, r = 0.652, $p < 0.001$; microbe, r = 0.500, $p > 0.05$). A significant negative correlation of VLPs was observed with C/N (r = −0.473, $p < 0.01$), soil pH (r = −0.677, $p < 0.001$), and EC (r = −0.350, $p < 0.05$). A significant negative correlation of microbial abundance was observed with C/N (r = −0.434, $p < 0.01$) and soil pH (r = −0.445, $p < 0.01$). Particularly, VMR was positively correlated with SWC (r = 0.371, $p < 0.05$) and VLP abundance (r = 0.484, $p < 0.01$). Furthermore, the random forest model was used to evaluate the relative importance of soil physical–chemical properties, biological factors, climate factors, and geographical location for predicting VLP abundance, which indicated that SWC was the main environmental factor affecting VLP abundance, followed by TP ($p < 0.05$) (**Figure 2B**).

Viral Community Composition Under Different Precipitation Gradients

A total of 54 viral orders, 152 families, and 362 genera were identified from the metagenomic data of 30 soil samples. At

TABLE 1 | Virus-like particles (VLPs) and microbial abundance under different precipitation gradients.

Annual precipitation	Site	VLP abundance × 10^9 gdw^{-1}		Microbial abundance × 10^8 gdw^{-1}		Virus-to-microbe ratio (VMR)	
< 200 mm	LP_32	1.66 ± 0.50 a	A	1.3 ± 0.32 a	A	12.09 ± 0.96 abc	A
	LP_34	0.83 ± 0.61 a		2.9 ± 0.61 abc		2.39 ± 1.43 ab	
	LP_35	0.2 ± 0.02 a		1.8 ± 0.19 ab		1.12 ± 0.18 a	
	LP_36	0.02 ± 0.00 a		1.4 ± 0.16 a		0.11 ± 0.02 a	
200–400 mm	MP_2	1.21 ± 0.37 a	B	3.3 ± 1.28 abc	A	7.29 ± 4.72 ab	AB
	MP_7	4.29 ± 0.21 a		4.4 ± 1.71 abc		13.61 ± 5.15 abc	
	MP_27	0.23 ± 0.03 a		3.7 ± 0.57 abc		0.63 ± 0.03 a	
	MP_29	9.21 ± 3.59 b		1.0 ± 0.49 a		98.3 ± 13.00 d	
> 400 mm	HP_11	9.82 ± 2.22 b	C	5.4 ± 1.01 bcd	B	18.14 ± 1.76 bc	B
	HP_15	9.76 ± 1.21 b		8.2 ± 2.77 d		13.57 ± 2.45 abc	
	HP_17	9.68 ± 0.83 b		5.7 ± 0.91 cd		18.52 ± 5.10 bc	
	HP_22	10.02 ± 2.61 b		3.7 ± 0.27 abc		26.42 ± 5.77 c	

Statistical differences within and between groups were determined using one-way ANOVA based on Duncan test, and all groups were labeled accordingly (i.e., a, b, c, A, B, and C; p < 0.05).

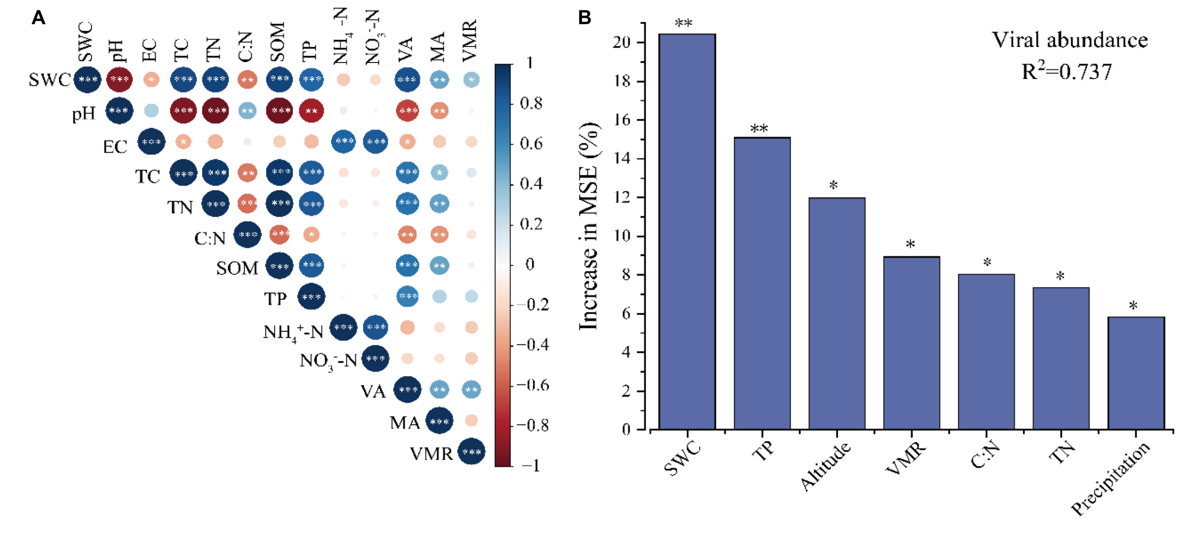


FIGURE 2 | Environmental factors influencing virus-like particles (VLPs) abundance. Correlation among VLPs abundance, microbial abundance, VMR and soil physical chemical properties **(A)**. Random forest model evaluating the relative importance of various factors influencing VLP abundance **(B)**. Only the factors with significant influence are shown in the figure (top 7), *$p < 0.05$, **$p < 0.01$.

the family level, the average relative abundance of *Siphoviridae* (40.1%–54.5%) belonging to the order *Caudovirales* was the highest in all samples. The family *Pandoraviridae* (10.1%–13.9%), *Myoviridae* (7.2%–13.1%), and *Herpesviridae* (5.0%–8.2%) also accounted for a large proportion (**Figure 3A**). The relative abundance of *Podoviridae* (22.6%) in LP_35 was 5–7 times higher than that in other samples (2.7%–3.6%). Finally, other viral families like *Poxviridae*, *Baculoviridae*, *Autographiviridae*, *Phycodnaviridae*, and *Adenoviridae* accounting for a small proportion were also detected in each soil. The distribution of dominant viral families was greatly similar among most samples under three precipitation gradients. LEfSe analysis identified five orders and eight families of soil viruses that showed significantly different abundance among three precipitation gradients (**Figure 3B**). A total of 70.1% belonged to *Caudovirales*, and most of these belonged to *Siphoviridae* (50.2%). Only *Imitervirales* and *Miniviridae* were significantly enriched in MP. Four orders (*Priklausovirales*, *Bunyavirales*, *Mononegavirales*, and *Geplafuvirales*) and seven families were significantly enriched in HP.

Differences Between Virus and Host Under Three Precipitation Gradients

The predicted hosts including bacteria (primarily) and archaea were identified from 8.44% of vOTUs (47 out of 557 vOTUs) *via* the PHISDetector tool (**Figure 4** ad **Supplementary Table 2**). Most of the individual links (47) occurred *via* CRISPRs (black solid line), with three *via* BLAST (blue dashed line). These hosts spanned 18 genera among four phyla (Bacteria: *Actinobacteria*, *Proteobacteria*, and *Tenericutes*; Archaea: *Euryarchaeota*). Most viruses were linked to only one host at the genus level; only three viruses were, respectively, linked to two hosts but always within the same order (vOTU_4, vOTU_346, and vOTU_349). Of all the hosts, *Rubellimicrobium* (involving 21 vOTUs) has the most association with viruses, followed by *Streptomyces* (involving 6 vOTUs). The overall number of virus–host linkages (pairs) under MP (38) was significantly higher than that under LP (6) and HP (6).

Abundance and Diversity of Putative Auxiliary Carbohydrate Metabolic Genes in Soil Viruses

To clarify the viral role in carbon cycling in soils, 557 vOTUs were further annotated for carbohydrate-active enzymes (CAZymes) by the dbCAN server based on the recognition of the CAZyme signature domain. According to results, 23.0% of vOTUs (128 of 557 vOTUs) carried 137 CAZyme genes (**Figure 5A**). These genes belonged to six CAZyme functional classes, most of which are affiliated to glycoside hydrolases (GHs, 42.3% of 137 viral CAZymes genes), followed by glycosyl-transferase (GTs, 28.5%) and carbohydrate-binding modules (CBMs, 20.4%) (**Figure 5A**). The most annotated CAZymes genes were found in soil viruses under MP, especially MP_29, followed by LP. However, the soil viruses under HP carried fewer CAZyme genes, only involving CE, GH, and GT (**Figure 5B**). CBM50 and GH23 were the most widely distributed, with different distributions in five samples (**Supplementary Figure 1**).

DISCUSSION

Soil Water Content Was the Main Driving Factor for Viral Abundance on the Qinghai-Tibet Plateau

It has been confirmed that the abundance, diversity, and reproductive strategy of viruses in soil are affected by multiple factors, such as soil types, physical and chemical properties, cover

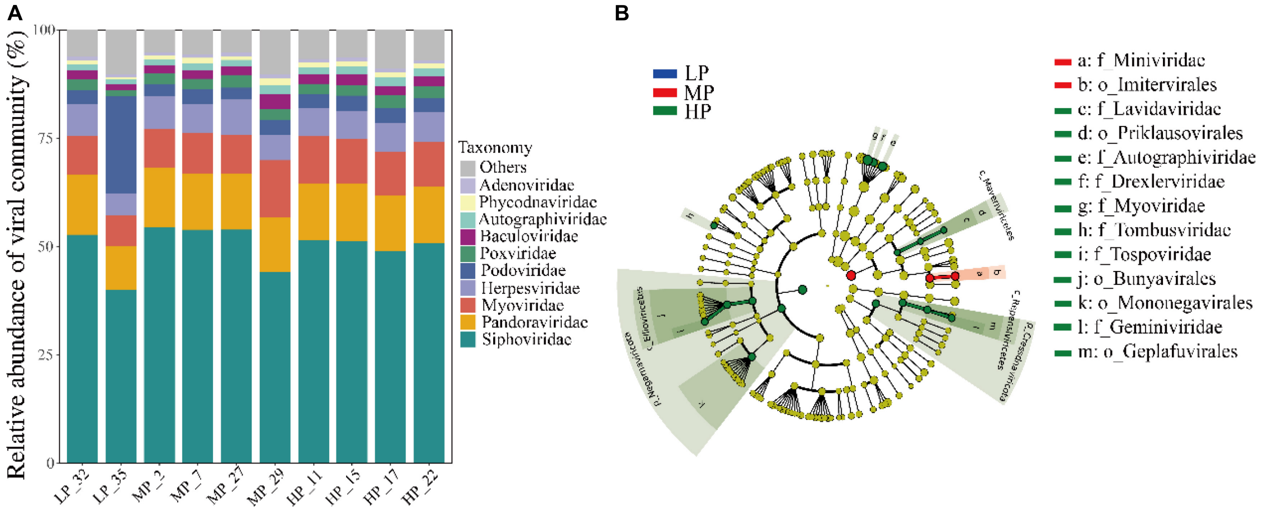


FIGURE 3 | Taxonomic composition of soil viruses in the Tibetan Plateau, assessed for all virus-associated reads at the family level **(A)**. Only the top ten viral families were shown. Viral biomarkers in different precipitation based on LEfSe analysis **(B)**. Different colors represent different precipitation gradients and the circles from inside to outside correspond to kingdom to family.

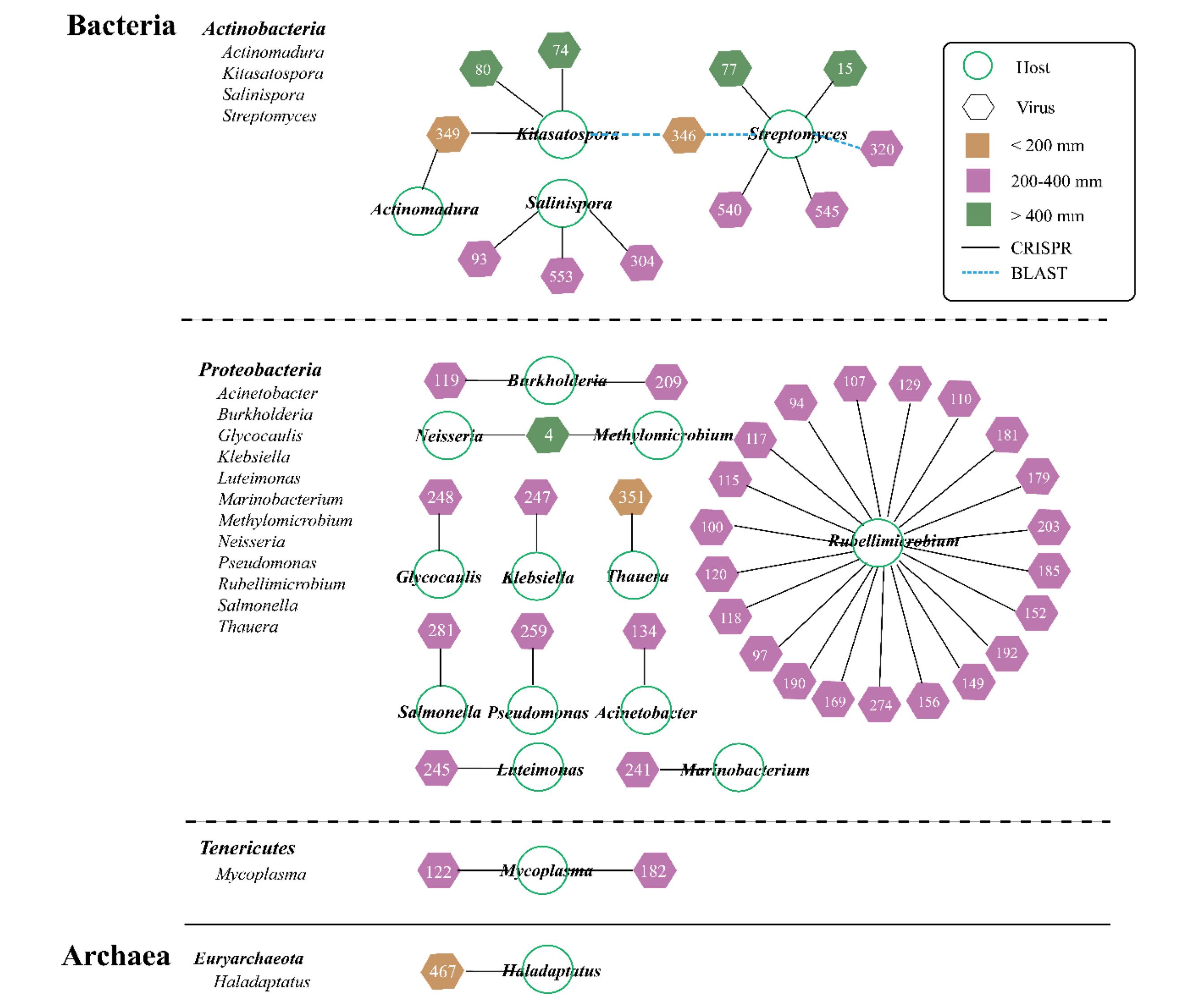


FIGURE 4 | Predicted viral-host linkages under three precipitation gradients. About 47 vOTUs are linked to 4 host lineages by multiple lines of evidence, and the three prediction methods are represented by the different color-coded lines. Node shape denotes organism (circle for microbe and hexagon for virus), and number represents vOTU (**Supplementary Table 2**). Viral shapes are color-coded by three precipitation gradients (yellow for LP, purple for moderate precipitation, and green for high precipitation).

plants, and so on (Williamson et al., 2017; Liang et al., 2021). In order to expand our understanding of the distribution of soil viruses across different precipitation gradients, the abundance of microorganisms and viruses in 36 soil samples across three precipitation gradients was quantified by fluorescence microscopy. This method only quantified free and adsorbed virus particles in soils, ignoring temperate viruses, so the true abundance of soil viruses was underestimated. Our data showed that there were significant differences in both VLPs and the microbial abundance among different precipitation gradients; both of them increased with precipitation and reached the highest at HP (**Table 1**), which was consistent with previous studies that soil tends to have more abundant viruses and bacteria under the condition of high moisture content and organic matter (Srinivasiah et al., 2008). In our data, high carbon and nitrogen nutrients in wet soil are beneficial to the growth of microorganisms (**Supplementary Table 3**), which may make it more conducive for the encounter of a virus and host, resulting in more viruses. Srinivasiah et al. (2008) found that bacterial abundance increased by 84-fold after the addition of C- and N-rich substrates. Also, VLP counts were highly correlated with multiple soil physical and chemical factors compared with microbial abundance, such as SWC, TN, TC, SOM, TP, pH, EC, and C/N (**Figure 2A**); the result of the random forest showed that SWC was the main environmental factor affecting VLP abundance ($p < 0.05$) (**Figure 2B**). Therefore, we supposed that viruses were more sensitive to environmental factors than microbes and SWC was a key environmental factor driving VLP

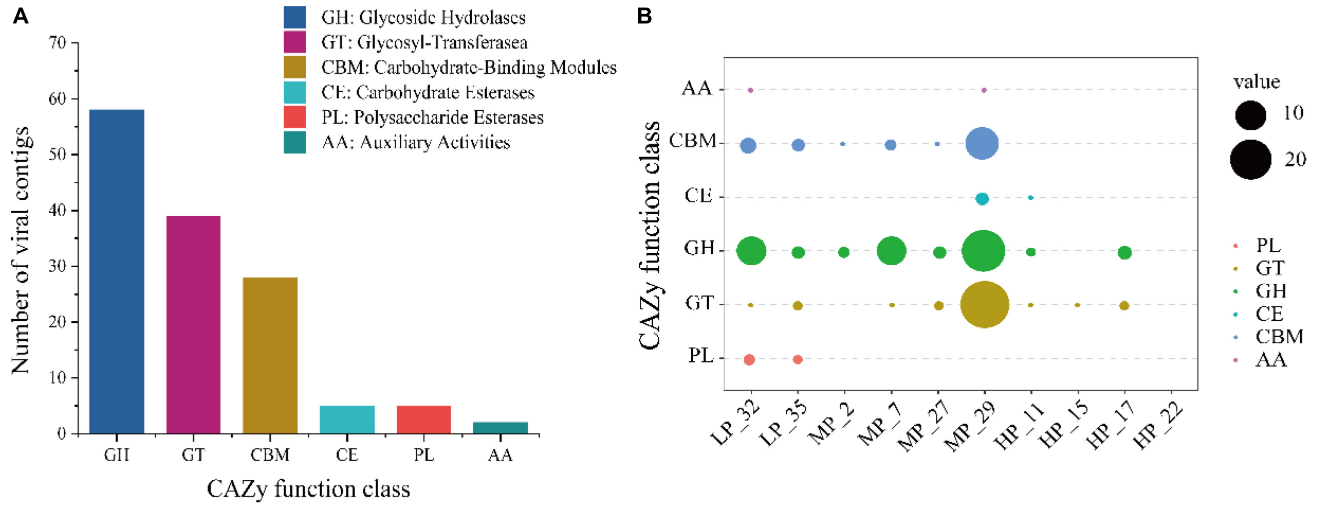


FIGURE 5 | Abundant auxiliary CAZyme from surface soil viruses in the Tibetan Plateau. Annotation of viral carbohydrate metabolism-related contigs (> 10 kb) in the CAZy database **(A)**. The distribution of viral auxiliary CAZyme in each sample **(B)**.

abundance in soil, which indirectly reflected the impact of climate change (average annual precipitation) on VLP abundance and provided a theoretical basis for the conjecture of Williamson et al. (2017) who strongly suspected that the SWC was a key environmental parameter driving both bacterial and viral abundance in soils.

In this study, the VMR, an index to measure viral activity and study the virus–host interaction in the environment (Ogunseitan et al., 1990; Wommack and Colwell, 2000; Parikka et al., 2017), fluctuated over four orders of magnitude (from 0.11 to 98.3) and was significantly higher in HP than that in LP (**Table 1**). Since VMR is the ratio of VLPs to microbial abundance, its value has to do with the factors controlling both VLPs and microbial abundance. As previous results showed, both VLPs and microbial abundance have a significant positive correlation with SWC, and SWC mainly contributed to the variability in VLP abundance (**Figure 2**). In addition, soil VLPs have a significantly positive correlation with microbial abundance (r = 0.491, p < 0.01) (**Figure 2A**),as has been observed in other soils (Williamson et al., 2005; Liang et al., 2019a, 2020). With regard to the correlations between viral abundance and replication strategy and microbial abundance, several hypothesized models have been previously reported. The KtW model proposed that viral predation was density and frequency dependent, while lysogeny is facilitated at low host density (Thingstad and Lignell, 1997). Moreover, a study on coral reef samples has found that both viral abundance and VMR were reduced at high host density and confirmed that lysogeny is more prevalent in an environment with high host densities. The "PtW" model was thus presented (Knowles et al., 2016). Based on our results, the viral lytic efficiency was higher in soils with higher water content, and the dynamic changes between viruses and hosts were more consistent with the KtW model (Thingstad and Lignell, 1997). While the sampling pool of this work was limited, this study indicates the interesting results of viral distribution and replication strategies likely being influenced by SWC and sheds light on the drivers of viral community dynamics in the complicated soil environment. Future effort is needed to reveal the relationship of soil properties and other factors with viral distribution and production and, in turn, the role of viruses in biogeochemical cycling in the soil of the Qinghai-Tibet Plateau.

Viral Diversity and Host Prediction Across Three Precipitation Gradients

Taxonomic diversity analysis revealed that *Caudovirales* (70.1%) was the major viral group in soil. The relative abundance of the *Siphoviridae* family was the highest in all samples (**Figure 3A**). This result is consistent with that in southeastern United States agricultural soil and Antarctic soil (Adriaenssens et al., 2017; Liang et al., 2019a). Many previous studies have shown that *Siphoviridae* was the dominant viral family in both soil and marine systems (Jin et al., 2019; Gao et al., 2020; Zheng et al., 2021). The family *Pandoraviridae*, *Myoviridae*, and *Herpesviridae* also accounted for a large proportion. Pandoravirus, the second largest giant virus after Mimivirus, can infect amoeba. Its DNA genomes can reach 2.5 Mb, much larger than that of other viruses (Philippe et al., 2013). Pandoraviruses were rarely detected in soil habitats; however, a most recent study by Legendre et al. (2018) found that *Pandoravirus quercus* can be isolated from ground soil in Marseille (France). In our results, *Pandoraviridae* family accounted for 10.1%–13.9% in each soil sample, which may provide convenient conditions for studying the diversity and evolution of *Pandoraviridae* family in soil. Importantly, the distribution of dominant viral families (top 10) was greatly similar among most samples under three precipitation gradients, indicating that the change of precipitation had little influence on the composition of soil viruses. While the viral species composition at a family level was similar among samples, different dominant virus populations were significantly enriched in soil with MP and HP as the LEfSe analysis result showed, respectively. For example, the families *Imitervirales* and *Miniviridae* were significantly enriched in soil under MP (**Figure 3B**). However, these results are highly dependent on the viral database that is still much incomplete, which will be further refined, as more environmental viruses are discovered and the database is improved.

In order to examine these viruses' impacts on the microbial communities and processes, we sought to link them to their hosts *via* the PHISDetector tool (Zhou et al., 2020). In this study, only 8.44% of vOTUs has been assigned to hosts including bacteria and archaea that spanned 18 genera among four phyla (Bacteria: *Actinobacteria*, *Proteobacteria*, and *Tenericutes*; Archaea: *Euryarchaeota*) (**Figure 4**). Consistent with previous reports, the majority of hosts were annotated as bacteria. It was reported that both *Proteobacteria* and *Actinobacteria* were the dominant soil bacteria taxa (Liang et al., 2021; Wu et al., 2021; Zheng et al., 2021). In addition, compared with grassland soils in Kansas (Wu et al., 2021) and samples from Lake Michigan (Malki et al., 2015), the host range was narrow because most viruses were linked to only one host at the genus level, which was similar to the findings of Trubl et al. (2018) and ter Horst et al. (2021). Particularly, the overall number of virus–host linkages (pairs) in MP (38) was significantly higher than that in LP (6) and HP (6). The site-specific vOTUs had different predicted hosts, thus illustrating the unique assemblages of soil viruses and hosts across sites with differences in precipitation. Interestingly, only vOTU_467 from LP_35 sample located in the Qaidam Basin was assigned to archaea, probably resulting from the highest salt concentration (**Supplementary Table 3**). The majority of archaea phages originated in thermophilic or extremely halophilic environments (Mochizuki et al., 2010). The Qaidam Basin was originally a huge lagoon. Due to the continuous elevation of the Qinghai-Tibet Plateau, reduction of rainfall, and evaporation of water, this huge natural salty lake basin was formed. Overall, further studies of viral communities and the virus–host interaction can fuel our knowledge of viral ecology in different soil matrices and how biotic/abiotic factors play a role in structuring viral communities.

Viruses May Play a Potential Role in Soil Carbon Cycling on the Qinghai-Tibet Plateau

Soil is the largest carbon pool in terrestrial ecosystems (Post et al., 1982), and the Tibetan Plateau stores the most abundant soil organic carbon (Stockmann et al., 2015; Liang et al., 2019b). In the past, many studies focused on the influence of microbes on the soil carbon cycle, including bacteria and fungi, but ignored the contribution of viruses. It was confirmed that viruses affect ecosystem carbon processing *via* the controls of top–down (lysing dominant microbial hosts) and bottom–up (carrying AMGs) (Brum and Sullivan, 2015; Trubl et al., 2018). Virus-encoded diverse AMGs could enhance or expand the host metabolic pathways, thus opening up new ecological niches and affecting biogeochemistry (Roux et al., 2016; Gazitua et al., 2021). For example, photosystem I and II genes were obtained by the marine cyanophages from cyanobacteria, and the expression of these genes during infection promoted the photosynthetic output of host cells (Thompson et al., 2011). Thus, to reveal the potential contribution of viruses to the carbon cycle on the Tibetan Plateau, potential CAZymes in the soil viral genomes were annotated. Finally, 22.2% of vOTUs (128 vOTUs) carrying 59 CAZyme genes (categories) that spanned 137 CAZyme genes were further annotated, with most of CAZymes affiliated to polysaccharide hydrolase activities, implying that viruses may play a potential role in the decomposition of organic carbon on the Qinghai-Tibet Plateau (**Figure 5A**). However, compared with farmlands (10 CAZymes genes) (Bi et al., 2021) and mangroves (27 CAZyme genes) (Jin et al., 2019), soil viruses in the Qinghai-Tibet Plateau carried more diverse CAZyme genes (species), such as GH5, GH8, CE11, CE14, and so on, indicating the higher occurrence of CAZymes and non-negligible roles of viruses in the soil carbon cycle. This difference might result from the environmental specificity of CAZymes (Bi et al., 2021). Our data suggested that the GHs responsible for the breakdown of complex organic matter were the most abundant, supporting previous findings (Emerson et al., 2018). Further, we found that CAZyme genes were the most abundance in samples under MP (**Figure 5B**), especially in MP_29, which corresponded to the result of host prediction. However, there were the highest VLPs abundance and the lowest CAZyme genes under HP, probably resulting from virus-carrying AMGs that mainly exist in the lysogenic state. The distribution of integrase genes also supported high abundant lysogenic viruses presented in MP (**Supplementary Figure 2** and **Supplementary Table 4**). In addition, we also identified more specific auxiliary carbohydrate metabolism genes, including GHs, glycosyl transferases, polysaccharide lyases, carbohydrate esterases, and carbohydrate-binding modules (**Supplementary Figure 1**), implying that more ecological functions of viruses are yet to be discovered. Together, these results suggest that viral infections contribute to soil ecosystem functioning and that further interrogation of soil viral communities will yield a more comprehensive understanding of complex functional networks and ecosystem processes in soil.

CONCLUSION

This study provides a new knowledge of the abundance, composition and host diversity, and potential biogeochemical impacts of soil viruses in the grassland soil of Qinghai-Tibet Plateau across three precipitation gradients. VLP abundance is positively correlated with microbial abundance; both of them reach the highest in wet soil, and the SWC mainly contributes to the variability in VLP abundance. In addition, based on LEfSe analysis, we find that different dominant virus populations were significantly enriched in soil with MP and HP. Interestingly, Pandoraviruses, the second largest giant virus after Mimivirus, are detected abundantly in soil. High host diversity and abundant CAZyme genes were shown in soils with moderate precipitation. Finally, abundant CAZymes represented by GHs were identified in our study, indicating that soil viruses may play a potential role in the carbon cycle on the Qinghai-Tibet Plateau, and some novel auxiliary carbohydrate metabolism genes were also identified. Overall, the results of this study indicate major differences in soil viruses along the precipitation gradient and provide a theoretical basis for the influence of climate change on the soil virosphere.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

L-MZ, J-ZH, and L-LH designed the research. L-MZ, S-YL, H-YW, and YG collected the soil samples. M-MC, S-YL, and L-LH lead the laboratory work. M-MC, S-JC, and BH performed the data analysis. M-MC, LB, and L-LH interpreted the results and wrote the manuscript in close consultation from all authors. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP), Grant Nos. 2019QZKK0306 and 2019QZKK0308, the National Science Foundation of China (Grant No. 41771289), and Foundation of President of the Zhongke-Ji'an Institute for Eco-Environmental Sciences.

ACKNOWLEDGMENTS

We thank Ganlin Zhang's group (Institute of Soil Science, Chinese Academy of Sciences) for their assistance in soil sampling, Baoming Yao, Donghui Zhao, and Yabo Zhang (University of Chinese Academy of Sciences) for their assistance in soil sampling and laboratory work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.848305/full#supplementary-material

REFERENCES

Adriaenssens, E. M., Kramer, R., Van Goethem, M. W., Makhalanyane, T. P., Hogg, I., and Cowan, D. A. (2017). Environmental drivers of viral community composition in Antarctic soils identified by viromics. *Microbiome* 5, 83. doi: 10.1186/s40168-017-0301-7

Adriaenssens, E. M., Van Zyl, L., De Maayer, P., Rubagotti, E., Rybicki, E., Tuffin, M., et al. (2015). Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* 17, 480–495. doi: 10.1111/1462-2920.12528

Ahlgren, N. A., Fuchsman, C. A., Rocap, G., and Fuhrman, J. A. (2019). Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *ISME J.* 13, 618–631. doi: 10.1038/s41396-018-0289-4

Bi, L., Yu, D. T., Du, S., Zhang, L. M., Zhang, L. Y., Wu, C. F., et al. (2021). Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils. *Environ. Microbiol.* 23, 588–599. doi: 10.1111/1462-2920.15010

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brum, J. R., and Sullivan, M. B. (2015). Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* 13, 147–159. doi: 10.1038/nrmicro3404

Chariou, P. L., Dogan, A. B., Welsh, A. G., Saidel, G. M., Baskaran, H., and Steinmetz, N. F. (2019). Soil mobility of synthetic and virus-based model nanopesticides. *Nat. Nanotechnol.* 14, 712–718. doi: 10.1038/s41565-019-0453-7

Che, R., Wang, S., Wang, Y., Xu, Z., Wang, W., Rui, Y., et al. (2019). Total and active soil fungal community profiles were significantly altered by six years of warming but not by grazing. *Soil Biol. Biochem.* 139:107611. doi: 10.1016/j.soilbio.2019.107611

Chen, H., Zhu, Q. A., Peng, C. H., Wu, N., Wang, Y. F., Fang, X. Q., et al. (2013). The impacts of climate change and human activities on biogeochemical cycles on the Qinghai-Tibetan Plateau. *Global. Change. Biol.* 19, 2940–2955. doi: 10.1111/gcb.12277

Chen, S. F., Zhou, Y. Q., Chen, Y. R., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, 884–890. doi: 10.1093/bioinformatics/bty560

Cook, B. I., Smerdon, J. E., Seager, R., and Coats, S. (2014). Global warming and 21st century drying. *Clim. Dynam.* 43, 2607–2627.

Danovaro, R., and Serresi, M. (2000). Viral density and virus-to-bacterium ratio in deep-sea sediments of the Eastern Mediterranean. *Appl. Environ. Microb.* 66, 1857–1861. doi: 10.1128/AEM.66.5.1857-1861.2000

Dion, M. B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* 18, 125–138. doi: 10.1038/s41579-019-0311-5

Dong, S. K., Shang, Z. H., Gao, J. X., and Boone, R. B. (2020). Enhancing sustainability of grassland ecosystems through ecological restoration and grazing management in an era of climate change on Qinghai-Tibetan Plateau. *Agr. Ecosyst. Environ.* 287:106684. doi: 10.1016/j.agee.2019.106684

Emerson, J. B. (2019). Soil Viruses: A New Hope. *Msystems* 4, e00120–19. doi: 10.1128/mSystems.00120-19

Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. B., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* 3, 870–880. doi: 10.1038/s41564-018-0190-y

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., et al. (2007). Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria. Archaea, Fungi, and Viruses in Soil. *Appl. Environ. Microbiol.* 73, 7059–7066. doi: 10.1128/aem.00358-07

Gao, S. M., Schippers, A., Chen, N., Yuan, Y., Zhang, M. M., Li, Q., et al. (2020). Depth-related variability in viral communities in highly stratified sulfidic mine tailings. *Microbiome* 8:89. doi: 10.1186/s40168-020-00848-3

Gazitua, M. C., Vik, D. R., Roux, S., Gregory, A. C., Bolduc, B., Widner, B., et al. (2021). Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. *ISME J.* 15, 981–998. doi: 10.1038/s41396-020-00825-6

Ge, A. H., Liang, Z. H., Xiao, J. L., Zhang, Y., Zeng, Q., Xiong, C., et al. (2021). Microbial assembly and association network in watermelon rhizosphere after soil fumigation for Fusarium wilt control. *Agric. Ecosyst. Environ.* [preprint]. doi: 10.1016/j.agee.2021.107336

Gerba, C. P. (1984). Applied and Theoretical Aspects of Virus Adsorption To Surfaces. *Adv. Appl. Microbiol.* 30, 133–168. doi: 10.1016/S0065-2164(08)70054-6

Goldsmith, D. B., Parsons, R. J., Beyene, D., Salamon, P., and Breitbart, M. (2015). Deep sequencing of the viral phoH gene reveals temporal variation, depth-specific composition, and persistent dominance of the same viral phoH genes in the Sargasso Sea. *PeerJ.* 3:e997. doi: 10.7717/peerj.997

Han, L. L., Yu, D. T., Zhang, L. M., Shen, J. P., and He, J. Z. (2017a). Genetic and functional diversity of ubiquitous DNA viruses in selected Chinese agricultural soils. *Sci Rep*. 7:45142. doi: 10.1038/srep45142

Han, L. L., Yu, D. T., Zhang, L. M., Wang, J. T., and He, J. Z. (2017b). Unique community structure of viruses in a glacier soil of the Tianshan Mountains. *China. J. Soils Sediments* 17, 852–860. doi: 10.1007/s11368-016-1583-2

Huang, Y., Niu, B. F., Gao, Y., Fu, L. M., and Li, W. Z. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Hurwitz, B. L., Hallam, S. J., and Sullivan, M. B. (2013). Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol*. 14:R123. doi: 10.1186/gb-2013-14-11-r123

Jackson, E. F., and Jackson, C. R. (2008). Viruses in wetland ecosystems. *Freshw. Biol*. 53, 1214–1227. doi: 10.1111/j.1365-2427.2007.01929.x

Jin, M., Guo, X., Zhang, R., Qu, W., Gao, B., and Zeng, R. (2019). Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* 7:58. doi: 10.1186/s40168-019-0675-9

Jin, Y., and Flury, M. (2002). Fate and transport of viruses in porous media. *Adv. Agron*. 77, 39–102. doi: 10.1016/S0065-2113(02)77013-2

Kaisermann, A., de Vries, F. T., Griffiths, R. I., and Bardgett, R. D. (2017). Legacy effects of drought on plant-soil feedbacks and plant-plant interactions. *New Phytol*. 215, 1413–1424. doi: 10.1111/nph.14661

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8:90. doi: 10.1186/s40168-020-00867-0

Kieft, K., Zhou, Z., Anderson, R. E., Buchan, A., Campbell, B. J., Hallam, S. J., et al. (2021). Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat. Commun*. 12:3503. doi: 10.1038/s41467-021-23698-5

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobian-Guemes, A. G., et al. (2016). Lytic to temperate switching of viral communities. *Nature* 531, 466–470. doi: 10.1038/nature17193

Kuzyakov, Y., and Mason-Jones, K. (2018). Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil. Biol. Biochem*. 127, 305–317. doi: 10.1016/j.soilbio.2018.09.032

Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J. M., et al. (2018). Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun*.[preprint]. doi: 10.1038/s41467-018-04698-4

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly *via* succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, W. Z., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liang, X. L., Wagner, R. E., Zhuang, J., DeBruyn, J. M., Wilhelm, S. W., Liu, F., et al. (2019a). Viral abundance and diversity vary with depth in a southeastern United States agricultural ultisol. *Soil Biol. Biochem*. [preprint]. doi: 10.1016/j.soilbio.2019.107546

Liang, X. L., Wang, Y. S., Zhang, Y., Zhuang, J., and Radosevich, M. (2021). Viral abundance, community structure and correlation with bacterial community in soils of different cover plants. *Appl. Soil Ecol*. [preprint]. doi: 10.1016/j.apsoil.2021.104138

Liang, X. L., Zhang, Y. Y., Wommack, K. E., Wilhelm, S. W., DeBruyn, J. M., Sherfy, A. C., et al. (2020). Lysogenic reproductive strategies of viral communities vary with soil depth and are correlated with bacterial diversity. *Soil Biol. Biochem*. 144:107767. doi: 10.1016/j.soilbio.2020.107767

Liang, Z. Z., Chen, S. C., Yang, Y. Y., Zhao, R. Y., Shi, Z., and Rossel, R. A. V. (2019b). National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. *Geoderma* 335, 47–56. doi: 10.1016/j.geoderma.2018.08.011

Loveland, J. P., Ryan, J. N., Amy, G. L., and Harvey, R. W. (1996). The reversibility of virus attachment to mineral surfaces. *Colloid Surf. A Physicochem. Eng. Asp*. 107, 205–221. doi: 10.1016/0927-7757(95)03373-4

Malki, K., Kula, A., Bruder, K., Sible, E., Hatzopoulos, T., Steidel, S., et al. (2015). Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virol. J*. 12:164. doi: 10.1186/s12985-015-0395-0

Mochizuki, T., Yoshida, T., Tanaka, R., Forterre, P., Sako, Y., and Prangishvili, D. (2010). Diversity of viruses of the hyperthermophilic archaeal genus Aeropyrum, and isolation of the Aeropyrum pernix bacilliform virus 1, APBV1, the first representative of the family Clavaviridae. *Virology* 402, 347–354. doi: 10.1016/j.virol.2010.03.046

Ogunseitan, O. A., Sayler, G. S., and Miller, R. V. (1990). Dynamic Interactions of *Pseudomonas*-Aeruginosa and Bacteriophages in Lake Water. *Microb. Ecol*. 19, 171–185. doi: 10.1007/BF02012098

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., et al. (2016). Uncovering Earth's virome. *Nature* 536, 425–430. doi: 10.1038/nature19094

Parikka, K. J., Le Romancer, M., Wauters, N., and Jacquet, S. (2017). Deciphering the virus-to-prokaryote ratio (VPR): insights into virus-host relationships in a variety of ecosystems. *Biol. Rev*. 92, 1081–1100. doi: 10.1111/brv.12271

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., et al. (2013). Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* 341, 281–286.

Post, W. M., Emanuel, W. R., Zinke, P. J., and Stangenberger, A. G. (1982). Soil Carbon Pools and World Life Zones. *Nature* 298, 156–159. doi: 10.1038/298156a0

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., et al. (2020). Identifying viruses from metagenomic data using deep learning. *Quant biol*. 8, 64–77. doi: 10.1007/s40484-019-0187-4

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., et al. (2019). Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol*. 37, 29–37. doi: 10.1038/nbt.4306

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537, 689–693. doi: 10.1038/nature19366

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ*. 3:e985. doi: 10.7717/peerj.985

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., et al. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res*. 48, 8883–8900. doi: 10.1093/nar/gkaa621

Shen, C. C., Shi, Y., Fan, K. K., He, J. S., Adams, J. M., Ge, Y., et al. (2019). Soil pH dominates elevational diversity pattern for bacteria in high elevation alkaline soils on the Tibetan Plateau. *FEMS Microbiol. Ecol*. 95:fiz003. doi: 10.1093/femsec/fiz003

Shi, W., Zhao, H. Y., Chen, Y., Wang, J. S., Han, B., Li, C. P., et al. (2021). Organic manure rather than phosphorus fertilization primarily determined asymbiotic nitrogen fixation rate and the stability of diazotrophic community in an upland red soil. *Agric. Ecosyst. Environ*. 319. doi: 10.1016/j.agee.2021.107535

Shi, Y., Fan, K. K., Li, Y. T., Yang, T., He, J. S., and Chu, H. Y. (2019). Archaea Enhance the Robustness of Microbial Co-occurrence Networks in Tibetan Plateau Soils. *Soil Sci. Soc. Am. J*. 83, 1093–1099. doi: 10.2136/sssaj2018.11.0426

Srinivasiah, S., Bhavsar, J., Thapar, K., Liles, M., Schoenfeld, T., and Wommack, K. E. (2008). Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol*. 159, 349–357. doi: 10.1016/j.resmic.2008.04.010

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., et al. (2015). Global soil organic carbon assessment. *Glob. Food Secur.Agric.Policy*. 6, 9–16. doi: 10.1016/j.gfs.2015.07.001

Straub, T. M., Pepper, I. L., and Gerba, C. P. (1992). Persistence of Viruses in Desert Soils Amended with Anaerobically Digested Sewage-Sludge. *Appl. Environ. Microbiol*. 58, 636–641. doi: 10.1128/AEM.58.2.636-641.1992

Straub, T. M., Pepper, I. L., and Gerba, C. P. (1993). Virus Survival in Sewage-Sludge Amended Desert Soil. *Water Sci. Technol*. 27, 421–424. doi: 10.2166/wst.1993.0384

Tao, J., He, D. K., Kennard, M. J., Ding, C. Z., Bunn, S. E., Liu, C. L., et al. (2018). Strong evidence for changing fish reproductive phenology under climate warming on the Tibetan Plateau. *Glob. Change Biol*. 24, 2093–2104. doi: 10.1111/gcb.14050

ter Horst, A. M., Santos-Medellin, C., Sorensen, J. W., Zinke, L. A., Wilson, R. M., Johnston, E. R., et al. (2021). Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* 9:242. doi: 10.1186/s40168-021-01210-x

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr*. 45, 1320–1328. doi: 10.4319/lo.2000.45.6.1320

Thingstad, T. F., and Lignell, R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat. Microb. Ecol*. 13, 19–27. doi: 10.3354/ame013019

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., et al. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl. Acad. Sci.U.S.A*. 108, E757–E764. doi: 10.1073/pnas.1102164108

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nat. Protoc*. 4, 470–483. doi: 10.1038/nprot.2009.10

Trubl, G., Jang, H. B., Roux, S., Emerson, J. B., Solonenko, N., Vik, D. R., et al. (2018). Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *Msystems* 3, e00076–18. doi: 10.1128/mSystems.00076-18

Trubl, G., Roux, S., Solonenko, N., Li, Y. F., Bolduc, B., Rodriguez-Ramos, J., et al. (2019). Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ*. 7:e7265. doi: 10.7717/peerj.7265

Trubl, G., Solonenko, N., Chittick, L., Solonenko, S. A., Rich, V. I., and Sullivan, M. B. (2016). Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. *PeerJ*. 4:e1999. doi: 10.7717/peerj.1999

Van Goethem, M. W., Swenson, T. L., Trubl, G., Roux, S., and Northen, T. R. (2019). Characteristics of Wetting-Induced Bacteriophage Blooms in Biological Soil Crust. *mBio* 10, e02287–19. doi: 10.1128/mBio.02287-19

Wang, W., Wang, H., and Zu, Y. (2014). Temporal changes in SOM, N, P, K, and their stoichiometric ratios during reforestation in China and interactions with soil depths: Importance of deep-layer soil and management implications. *For. Ecol. Manage*. 325, 8–17. doi: 10.1016/j.foreco.2014.03.023

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *Fems Microbiol. Rev*. 28, 127–181. doi: 10.1016/j.femsre.2003.08.001

Wen, K., Ortmann, A. C., and Suttle, C. A. (2004). Accurate estimation of viral abundance by epifluorescence microscopy. *Appl. Environ. Microbiol*. 70, 3862–3867. doi: 10.1128/AEM.70.7.3862-3867.2004

Williamson, K. E., Fuhrmann, J. J., Wommack, K. E., and Radosevich, M. (2017). Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. *Annu.Rev. Virol*. 4, 201–219. doi: 10.1146/annurev-virology-101416-041639

Williamson, K. E., Radosevich, M., and Wommack, K. E. (2005). Abundance and diversity of viruses in six Delaware soils. *Appl. Environ. Microbiol*. 71, 3119–3125. doi: 10.1128/AEM.71.6.3119-3125.2005

Wommack, K. E., and Colwell, R. R. (2000). Virioplankton: Viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev*. 64, 69–114. doi: 10.1128/MMBR.64.1.69-114.2000

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol*. 20:257. doi: 10.1186/s13059-019-1891-0

Wu, R. N., Davison, M. R., Gao, Y. Q., Nicora, C. D., Mcdermott, J. E., Burnum-Johnson, K. E., et al. (2021). Moisture modulates soil reservoirs of active DNA and RNA viruses. *Commun. Biol*. 4:992 . doi: 10.1038/s42003-021-02514-2

Yu, D. T., He, J. Z., Zhang, L. M., and Han, L. L. (2019). Viral metagenomics analysis and eight novel viral genomes identified from the Dushanzi mud volcanic soil in Xinjiang. *China. J. Soil Sediments*. 19, 81–90. doi: 10.1007/s11368-018-2045-9

Zeng, Q., and Chisholm, S. W. (2012). Marine viruses exploit their host's two-component regulatory system in response to resource limitation. *Curr. Biol*. 22, 124–128. doi: 10.1016/j.cub.2011.11.055

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., et al. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic. Acids Res*. 46, W95–W101. doi: 10.1093/nar/gky418

Zheng, X. W., Liu, W., Dai, X., Zhu, Y. X., Wang, J. F., Zhu, Y. Q., et al. (2021). Extraordinary diversity of viruses in deep-sea sediments as revealed by metagenomics without prior virion separation. *Environ. Microbiol*. 23, 728–743. doi: 10.1111/1462-2920.15154

Zhou, F., Gan, R., Zhang, F., Ren, C., and Huang, Z. (2020). PHISDetector: a tool to detect diverse in silico phage-host interaction signals for virome studies. *bioRxiv*. [Preprint].doi: 10.1101/661074

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

The handling editor declared a shared affiliation with one of the authors H-YW at the time of review.

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

ORIGINAL RESEARCH
published: 08 April 2022
doi: 10.3389/fmicb.2022.797463

A Comprehensive Analysis of Citrus Tristeza Variants of Bhutan and Across the World

Dilip Kumar Ghosh[1], Amol Kokane[1], Sunil Kokane[1], Krishanu Mukherjee[2], Jigme Tenzin[3], Datta Surwase[1], Dhanshree Deshmukh[1], Mrugendra Gubyad[1] and Kajal Kumar Biswas[4]*

[1] Plant Virology Laboratory, ICAR-Central Citrus Research Institute, Nagpur, India, [2] Whitney Laboratory for Marine Biosciences, University of Florida, St. Augustine, FL, United States, [3] National Citrus Program, Department of Agriculture, Royal Government of Bhutan, Thimpu, Bhutan, [4] Department of Plant Pathology, Indian Agricultural Research Institute, New Delhi, India

Mandarin orange is economically one of the most important fruit crops in Bhutan. However, in recent years, orange productivity has dropped due to severe infection of citrus tristeza virus (CTV) associated with the gradual decline of citrus orchards. Although the disease incidence has been reported, very limited information is available on genetic variability among the Bhutanese CTV variants. This study used reverse transcription PCR (RT-PCR) to detect CTV in collected field samples and recorded disease incidence up to 71.11% in Bhutan's prominent citrus-growing regions. To elucidate the extent of genetic variabilities among the Bhutanese CTV variants, we targeted four independent genomic regions (5′ORF1a, p25, p23, and p18) and analyzed a total of 64 collected isolates. These genomic regions were amplified and sequenced for further comparative bioinformatics analysis. Comprehensive phylogenetic reconstructions of the GenBank deposited sequences, including the corresponding genomic locations from 53 whole-genome sequences, revealed unexpected and rich diversity among Bhutanese CTV variants. A resistant-breaking (RB) variant was also identified for the first time from the Asian subcontinent. Our analyses unambiguously identified five (T36, T3, T68, VT, and HA16-5) major, well-recognized CTV strains. Bhutanese CTV variants form two additional newly identified distinct clades with higher confidence, B1 and B2, named after Bhutan. The origin of each of these nine clades can be traced back to their root in the north-eastern region of India and Bhutan. Together, our study established a definitive framework for categorizing global CTV variants into their distinctive clades and provided novel insights into multiple genomic region-based genetic diversity assessments, including their pathogenicity status.

Keywords: citrus tristeza virus, genomic diversity, sequencing and phylogenetic analysis, RT-PCR, genomic regions

Edited by:
Tao Jin,
Guangdong Magigene Biotechnology Co., Ltd., China

Reviewed by:
Islam Hamim,
Bangladesh Agricultural University, Bangladesh
Susheel Kumar,
National Botanical Research Institute (CSIR), India

****Correspondence:***
Dilip Kumar Ghosh
ghoshdk@hotmail.com

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *18 October 2021*
Accepted: *19 January 2022*
Published: *08 April 2022*

Citation:
Ghosh DK, Kokane A, Kokane S, Mukherjee K, Tenzin J, Surwase D, Deshmukh D, Gubyad M and Biswas KK (2022) A Comprehensive Analysis of Citrus Tristeza Variants of Bhutan and Across the World. Front. Microbiol. 13:797463. doi: 10.3389/fmicb.2022.797463

INTRODUCTION

Tristeza, caused by the citrus tristeza virus (CTV), is a destructive disease affecting citrus plants. Over the past few decades, tristeza has damaged millions of productive citrus trees worldwide (Kokane et al., 2021c; Moreno et al., 2008; Ghosh et al., 2021). Bhutan is a small landlocked Himalayan country located between India and China, and is the likely place of origin of citrus

(Wu et al., 2018). Diverse agro-climatic conditions are prevalent in the country, favoring the production of a wide range of horticultural crops, among which the citrus is the most important fruit crop (Joshi and Gurung, 2009). Mandarin (*Citrus reticulata* Blanco) is a widely grown citrus cultivar in 17 out of the 20 districts, constituting more than 95% of the total citrus grown in Bhutan (Joshi and Gurung, 2009; Dorji et al., 2016). The prominent mandarin-growing geographical regions are Tsirang, Dagana, Zhemgang, and Sarpang (Tipu and Fantazy, 2014; Ghosh et al., 2021). However, the citrus productivity in these regions is very low because of several factors, including infection of virus and virus-like pathogens. Among these pathogens, CTV is considered a major pathogen responsible for reducing the citrus yield and quality and decline of fruit-bearing citrus groves in Bhutan (Tipu and Fantazy, 2014).

Local transmission of the virus within a citrus groove is by aphid species, such as *Aphis gossypii* Glover, *Aphis* (Toxoptera) *citricidus* Kirkaldy, and *Aphis spiraecola* Patch, in a semi-persistent manner, whereas transmission into a new geographical area or country occurs through the movement of infected budwood during nursery propagation (Bar-Joseph et al., 1989; Marroquín et al., 2004; Herron et al., 2006). CTV is a phloem-limited virus having long flexuous filamentous particles of 2,000 × 11 nm in size and belongs to the genus *Closterovirus* under the family *Closteroviridae*. The single-stranded positive-sense RNA genome of ~19.3 kb organized into 12 open reading frames (ORFs), which potentially encode 19 different proteins, and two untranslated regions (UTRs) located at the 5′ and the 3′ terminal (Manjunath et al., 1993; Pappu et al., 1993; Karasev et al., 1995; Biswas et al., 2018). The 5′ proximal ORF1a encodes a 349-kDa polyprotein that includes two cysteine papain proteins-like (PRO) domains, a methyltransferase-like (MT) and helicase-like (HEL) domains, and ORF1b encodes an RNA-dependent RNA polymerase (RdRp)-like domain. The 3′ genomic region of CTV consists of 10 ORFs (ORFs 2–11) that encode different proteins with diverse functions, namely, major (CP) and minor (CPm) coat proteins, p65 (a homolog of cellular HSP 70 proteins), and p61 that is required for virion assembly and movement along with the hydro-phobic p6 protein (Hilf et al., 1995; Satyanarayana et al., 2000; Dolja et al., 2006; Tatineni et al., 2008). Additionally, the p20 and p23 proteins function as suppressors of the host RNA silencing along with CP (Lu et al., 2004), and three genes (p33, p18, and p13) are needed for systemic infection and play a role in extending the virus–host range (Tatineni et al., 2008, 2011). There are numerous biological strains of CTV (Brlansky et al., 2003; Harper, 2013) that infect almost all commercial citrus species and induce a wide variety of symptoms including stem pitting, vein clearing, stunting, veins corking, chlorosis, leaf cupping, and slow or quick decline (Ghosh et al., 2008, 2009; Warghane et al., 2017a,b; Kokane A. et al., 2020). The expression of symptoms in field-infected plants depends on the type of host, virus strain, rootstock–scion combination, age of the citrus tree, and environmental conditions (Ghosh et al., 2018; Kokane A. et al., 2020). Biological indexing has been used as a classical method of detecting CTV for years, but it has certain limitations. Other techniques that have been used for CTV detection are enzyme-linked immunosorbent assay (ELISA) (Liu et al., 2016; Kokane S. B. et al., 2020), dot immunobinding assay (DIBA), and immunoelectron microscopy (IEM) with polyclonal or monoclonal antibodies (Ahlawat, 2012). However, the serological methods are not used extensively because they require a supply of good quality antisera.

Electron microscopy is another powerful technique, but it is very expensive, requires highly skilled personnel, and cannot distinguish viruses of similar size. However, the most sensitive virus detection methods that are being routinely used at present in different laboratories include reverse transcription polymerase chain reaction (RT-PCR) (Mehta et al., 1997), quantitative PCR (qPCR) (Ruiz-Ruiz et al., 2007; Kokane et al., 2021b), multiplex RT-PCR (Meena and Baranwal, 2016), and RT-LAMP (Warghane et al., 2017a; Kokane et al., 2021a). Recently, a rapid, sensitive, robust, reliable, and highly specific reverse transcription recombinase polymerase amplification technique coupled with a lateral flow immunochromatographic assay (CTV-RT-RPA-LFICA) has been developed for early detection of CTV (Ghosh et al., 2020). Apart from complete genome sequencing (Harper, 2013), the genetic diversity of CTV has also been determined based on different genomic regions by several researchers (Rubio et al., 2001; Martín et al., 2009). The phylogenetic analysis of CTV isolates using 5′ORF1a genomic region (Roy et al., 2005a; Atallah et al., 2016), coat protein region (p25) (Warghane et al., 2020), RNA binding protein gene (p23) (Flores et al., 2013; Kokane S. B. et al., 2020), and p18 gene has also been reported (Dawson et al., 2013). The major focus of our study was molecular detection, characterization, and determination of the genetic diversity of 64 CTV isolates based on sequence variations of four (5′ORF1a, p25, p23, and p18 gene) different genomic regions. The targeted regions span most of the CTV genome, and each region plays a specific role; for example, the highly variable region ORF1a encodes a polyprotein of MT, and HEL domains, p25 covers 95% coat protein, p23 plays a role as a major suppressor, and p18 is involved in systemic infection for extending the virus–host range.

Thus, we selected these four regions for a comprehensive analysis of CTV variants by comparing them with whole-genome sequences across the globe. By comparing the 64 CTV isolates along with the published and unpublished sequences deposited in the GenBank, we have shown that the Bhutanese isolates could be unambiguously classified under six (except for T30) of seven (RB, T36, T30, T3, T68, VT-B, and HA16-5) internationally recognized and two additional clades (B1 and B2) identified in this analysis. Our analysis provides a comprehensive framework and a thorough picture of the global categorization of the CTV isolates and their origin.

MATERIALS AND METHODS

Plant Acquisition and Virus Maintenance

Leaves and twigs from 90 citrus plants suspected of being infected by CTV were collected from different geographical regions of eight (Tsirang, Wangdue Phodrang, Punakha, Trashiyangste, Zhemgang, Dagana, Sarpang, and Chukha) districts of Bhutan (**Figures 1**, **2A** and **Table 1**). These samples were assayed for CTV using conventional RT-PCR, as reported earlier by Ghosh et al. (2021). We also used the biological indexing technique to test representative samples of each district. This was done by side and wedge grafting in 10–12 months old seedlings of acid lime (*Citrus aurantifolia*) that were maintained in an insect-proof screen house at the Indian Council of Agricultural Research-Central Citrus Research Institute (ICAR-CCRI) Nagpur, India.

Sample Processing and RNA Extraction

Symptomatic leaves from all collected samples were thoroughly washed with double-distilled water, wiped with 70% ethanol to avoid surface contamination, and blot dried. Midrib portions of the leaves were excised and ground in liquid nitrogen. Approximately 100 mg of ground sample was used for total RNA extraction using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) as per the manufacturer's protocol. The extracted RNA was dissolved into the Tris-EDTA (TE) buffer and stored at −80°C for further analysis. The concentration of total genomic RNA was assessed by a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Delaware, United States), and quality was determined by 2% agarose gel, stained with 0.5 μg/ml of ethidium bromide, and visualized in a Gel documentation system (G: Box, Syngene, Frederick, United States).

Primer Designing

Primers were designed against the CTV-targeted genomic locations of 5′ORF1a, p25, p23, and p18 using primer 3v.0.4.0 tool.[1] Primer specificity was then checked using primer BLAST software at National Center for Biotechnology Information (NCBI)[2] to avoid cross-reaction with other pathogens or targets. The primers were finally custom synthesized from IDT (Integrated DNA Technologies, Coralville, IA, United States) (**Table 2**).

Establishing Citrus Tristeza Virus Culture Confirmation by RT-qPCR and Electron Microscopy

Total RNA extracted from graft-inoculated samples were used for RT-qPCR assay using CTV-specific primer–probe combination (P25-F/p25-R) and corresponding CTV-FAM probe [labeled with 6-carboxy-fluorescein (FAM) reporter dye at the 5′ terminus and the Black Hole Quencher (BHQ)-1 dye at the 3′ terminus]. The TaqMan-qPCR assay for CTV was performed using a StepOne Real-Time PCR System (Applied Biosystems) in two steps as described by Ghosh et al. (2020) with the following conditions: 95°C for 2 min (initial denaturation), followed by 40 cycles at 95°C for 15 s, annealing, and primer extension simultaneously for 1 min at 60°C. All experimental reactions were conducted in triplicate along with non-template controls (NTC), and the data were analyzed using StepOne Software v2.1. Furthermore, the graft-inoculated samples were also tested by electron microscopy in leaf dip preparation as reported by Ghosh et al. (2009).

Detection of Citrus Tristeza Virus in Field Samples Using Conventional RT-PCR

The total genomic RNA extracted from the leaves of CTV-suspected samples were used to perform RT-PCR in two steps with CTV 5′ORF1a gene-specific primer set, 488F/491R (Roy et al., 2005a; Atallah et al., 2016). In the first step, cDNA was synthesized in a 15 μl reaction volume. The reaction contained 1× first-strand buffer, 0.5 mM deoxyribonucleotide triphosphates (dNTPs) (Promega, Madison, United States), 15.6 U of RNAsin (Promega, Madison, United States), 0.4 μM reverse primer (491R), 6 μl of total RNA, and 120 units of M-MLV reverse transcriptase (Promega, Madison, United States). The reaction was carried out in a thermal cycler (Bio-Rad 100 Thermal Cycler, California, United States) with extension at 42°C for 50 min and denaturation at 72°C for 10 min. In the second step, a 1.75 μl aliquot of cDNA was used as a template in a 25 μl reaction mixture containing 1× PCR buffer, 0.2 μM of each primer (488F/491R), 0.2 mM of the dNTPs mix, 1.5 mM $MgCl_2$, and 1.25 U of GoTaq DNA polymerase (Promega, Madison, United States). The amplification was with one cycle of 3 min at 94°C followed by 35 cycles of 0.30 min at 94°C, 0.45 min at 58°C, 1 min at 72°C, and final extension for 10 min at 72°C. The amplified RT-PCR products of the 5′ORF 1a fragment were analyzed on 1.2% agarose gel. Three more genomic regions of CTV, *viz.*, p25, p23, and p18 genes were also used for detection and molecular characterization of CTV isolates. The primer pairs specific for p25 (CN150/CN151) (Cevik et al., 1996), p23 (RBP-23F/RBP-23R) (Kokane S. B. et al., 2020), and p18 (AR18F/AR18R) (Roy et al., 2005b) were used to perform the RT-PCR (**Table 2**). The amplification program for the p25, p23, and p18 genes were the same as described above with modifications in annealing time and temperature, *i.e.,* 0.45 min at 61°C for p25, 0.40 min at 52°C for p23, and 0.35 min at 62°C for p18 gene. The amplified RT-PCR products were analyzed on 1.2% agarose gel.

Nucleotide Sequence Analysis of the 5′ and 3′-Terminal Regions of Citrus Tristeza Virus Genome

The amplified products of four genomic regions were excised and eluted from the agarose gel using the GenElute Gel Extraction Kit (Sigma-Aldrich, Bengaluru, India) and sequenced from both ends DNA sequencing facility (Eurofins Genomics, Bengaluru, India). The forward and reverse sequences were assembled into one complete contig of the target gene and eliminated the repeated sequences. To assess the sequence similarity, the prepared contigs were analyzed by the basic local alignment search tool (BLAST) of the NCBI. The confirmed nucleotide sequences were translated using the online software EXPASY translate tool.[3] Sequence similarities of proteins were identified using the BLASTp. Assembled sequences of each gene were then deposited into GenBank using BankIt-NCBI-NIH software.[4] Assembled nucleotide

[1] http://bioinfo.ut.ee/primer3-0.4.0/

[2] https://www.ncbi.nlm.nih.gov/tools/primer-blast/

[3] http://web.expasy.org/translate/

[4] https://www.ncbi.nlm.nih.gov/WebSub/

sequences of the four genomic regions were further used to analyze genetic variations biostatistically and pair-wise identity using GeneDoc software among the CTV isolates (Nicholas et al., 1997).

Sequence Retrieval, Extraction of Genomic Location, Sequence Alignment, and Phylogeny Reconstruction

Depending on the amino acids or nucleotide sequences, Blastp, TBlastn, or Blastn searches were performed with the default parameters using 5′ORF1a, p25, p23, and p18 as a query against all GenBank deposited sequences, including the whole-genome tristeza sequences available at the NCBI. Along with the GenBank deposited CTV variants, a total of 53 whole-genome CTV nucleotide sequences were retrieved and downloaded, and local standalone BLAST (Camacho et al., 2009) searches were performed against the retrieved genomic sequences. In local BLAST searches, the amino acid sequences of 5′ORF1a, p23, p18, and p25 were again used as a query against the whole-genomes CTV sequences in Tblastn searches to identify the corresponding amino acid sequences. The whole-genome nucleotide sequence was then translated in three frames at https://www.bioinformatics.org/sms2/trans_map.html, and the nucleotide sequence of the corresponding genomic region encoded by these proteins was extracted manually. Individual DNA and the protein sequences against these four genomic regions extracted for a particular isolate from the whole-genome sequences along with all other retrieved sequences from NCBI are presented in **Supplementary Excel File 1** and freely available for download. A novel phylogenetic reconstruction approach was then used in this analysis using the concatenated nucleotides and protein sequences of the four genomic regions of the Bhutanese variants and the GenBank deposited sequences. Due to high sequence similarities at the protein level, the phylogeny was performed using the corresponding DNA sequences to determine the greater sequence variations due to the presence of both synonymous (mutations in the codon that do not change the amino acids) and non-synonymous (mutations that alter the amino acids) changes. Both the amino acid and nucleotide sequences were aligned in MUSCLE (Edgar, 2004), and maximum likelihood (ML) trees were inferred using PhyML v3.0 (Guindon and Gascuel, 2003; Guindon et al., 2010), with the best-fit evolutionary model identified using the Akaike information criterion (AIC) criterion estimated by ProtTest (Abascal et al., 2005). The JTT substitution matrix was used for the amino acid sequences and the GTR substitution model for the nucleotide sequences while estimating the tree topology, branch lengths, amino acid equilibrium frequencies, fraction of invariable sites, and discrete-gamma distributed substitution rates. Clade support was calculated using the SH-like approximate likelihood ratio test (Anisimova et al., 2011). The resulting phylogenetic trees were viewed online and edited with iTol version 2.0 (Letunic and Bork, 2007). The vector graphics file was then imported onto Adobe Illustrator version CS6 for editing and final exporting of the high-resolution picture for publication.

RESULTS

Symptomatology, Bioassay, RT-qPCR, and Electron Microscopy

During field surveys in the different districts of Bhutan (**Figure 2A**), citrus trees showed the typical characteristic of tristeza symptoms, specifically chlorosis, yellow leaves, leaf cupping, vein clearing, vein flecking, declined condition, poor growth, and vigor. Stunting in diverse species was also observed, for instance, in mandarin (*Citrus reticulata*), pomelo (*Citrus grandis*), lime (*Citrus aurantifolia*), citron (*Citrus medica*), and other citrus cultivars or hybrids. However, few citrus trees found seemed to be healthy (**Table 1**). We performed Koch's postulate successfully for CTV in acid lime indicator plants. The virus-inoculated plants developed vein clearing, leaf cupping, temporary yellowing, and stunting of young seedlings (**Supplementary Figures 1A,B**). The titers of CTV in graft-inoculated plants were confirmed by RT-qPCR. However, the virus titer was varied from plant to plant, and Ct (cycle threshold) values were found ranging from 19.25 to 29.12 per 500 ng/μl of RNA extracted (**Supplementary Figure 1E**). Furthermore, under electron microscopy, CTV particles having the size of 2,000 × 11 nm were also observed (**Supplementary Figure 1D**).

RT-PCR Detection and Disease Incidence

RT-PCR detected the CTV variants in all the collected samples by separate targeted gene-specific primer sets (**Figure 2B** and **Table 2**). For example, the 5′ORF1a-specific primers pair 488F/491R targeting the genomic region between 1,082 and 1,484 nucleotides on the CTV genome resulted in an intense band of ∼404 bp. Of the 90 samples collected, 64 were found positive for CTV. These samples also tested positive against p25, p23, and p18 gene-specific primers and showed the expected amplicons of ∼672, ∼627, and ∼511 bp, respectively. Amplicons of 10 representative isolates for each gene were separated on a 1.2% agarose gel (**Figures 3A–D**). No amplification was observed with either the healthy citrus plant or the non-template control (NTC). The extent of the disease incidence was at a higher level; surprisingly, very few citrus trees were observed to be healthy (**Table 1**). The average CTV disease incidence was nearly 71.11% (**Table 1**). The percent of tree infection varied based on citrus cultivars and locations of the orchards. The highest CTV incidence was recorded in the Zhemgang district (83.33%), followed by Tsirang (78%), Dagana (70%), Chukha (66.66%), Sarpang (62.5%), Wangdue Phodrang (50%), and Trashiyangtse (33.33%).

Sequence Variations and Pair-Wise Identity Among the Citrus Tristeza Virus Isolates

The assembled sequences of four genomic regions of CTV were deposited into GenBank databases under the accession numbers listed in **Table 3**. Furthermore, these sequences were used to analyze genetic variations and pair-wise identity among

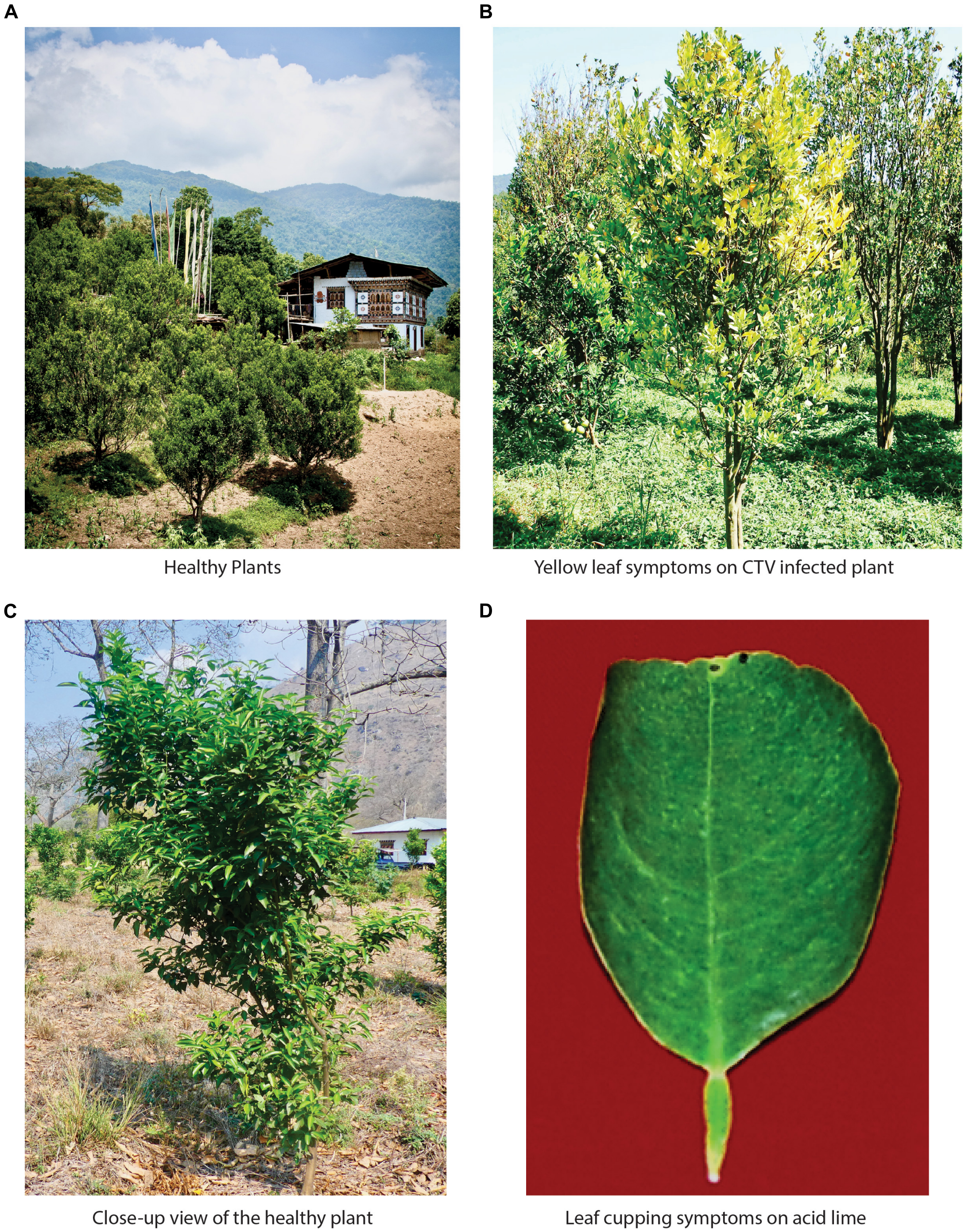
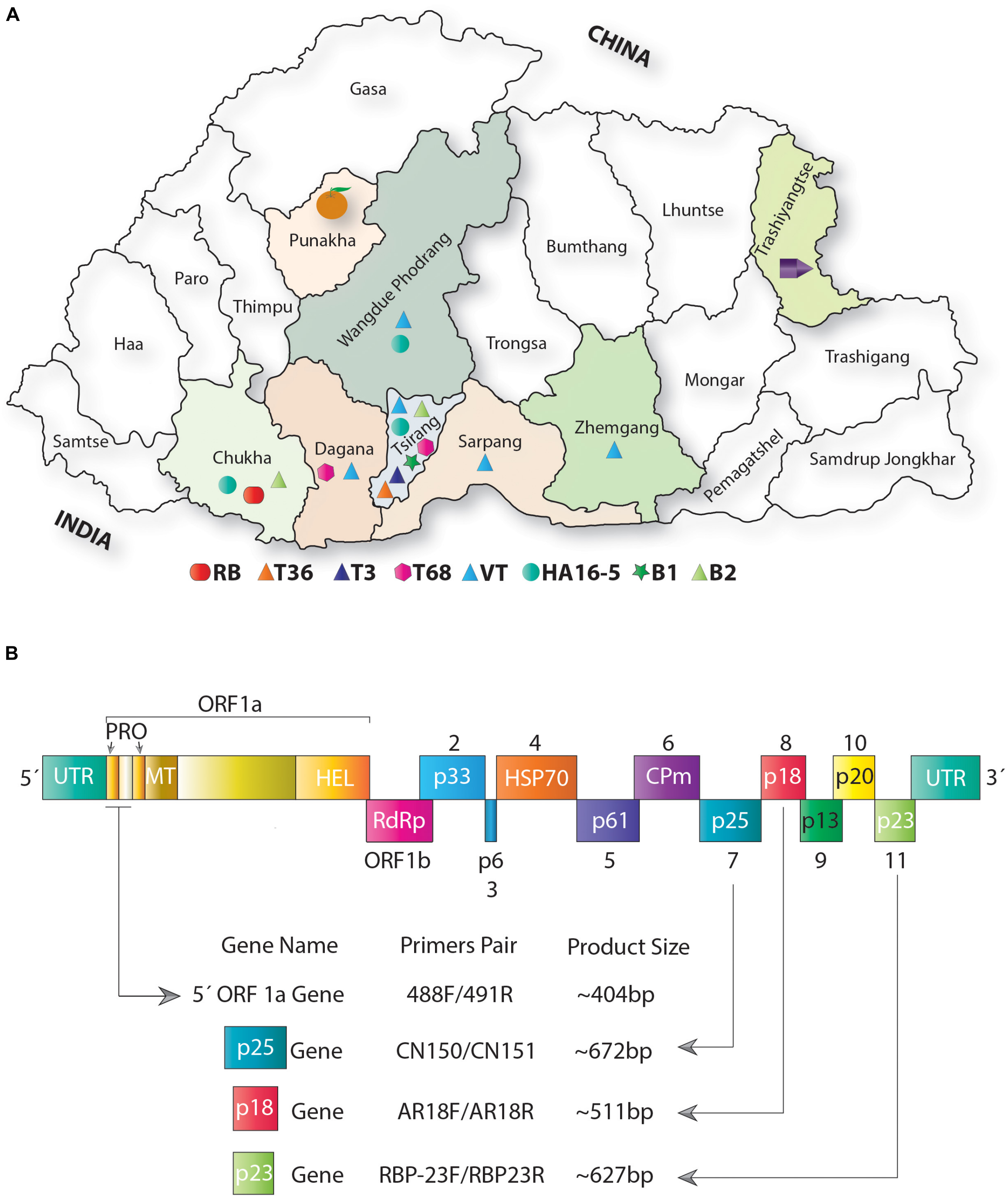

A
B
Healthy Plants
Yellow leaf symptoms on CTV infected plant
C
D
Close-up view of the healthy plant
Leaf cupping symptoms on acid lime

FIGURE 1 | Bhutan represents the richest diversity of tristeza variants found across the world. **(A)** Citrus cultivation scenario in Bhutan. Almost every household grows oranges in their backyard and surround their houses. **(B)** Typical CTV-infected plant in the field showing the yellowing of leaves. **(C)** A closer view of the healthy plant in the field. **(D)** One of the typical symptoms of a severely tristeza-infected plant is leaf cupping on acid lime.

all Bhutanese CTV variants. The nucleotide variation in the 5′ORF1a region ranged from 0.0 to 0.18 with an average of 0.06, and the pair-wise nucleotide identities were found to be 84–100% across all CTV variants. Genetic variations in the p25 gene varied from 0.02 to 0.10 with an average of 0.059 and 89–100% sequence identity within isolates. The p23 gene showed 88–100% nucleotide identity, and sequence variations ranged from 0.0 to 0.13 with an average of 0.48, whereas the p18 gene showed 91–100% identity, and nucleotide variation ranged from 0.0 to 0.11 with an average of 0.05.



Gene Name	Primers Pair	Product Size
5´ ORF 1a Gene	488F/491R	~404bp
p25 Gene	CN150/CN151	~672bp
p18 Gene	AR18F/AR18R	~511bp
p23 Gene	RBP-23F/RBP23R	~627bp

FIGURE 2 | Distribution of citrus tristeza virus (CTV) variants across major citrus growing regions of Bhutan. **(A)** Map showing the geographic locations of Bhutan. Districts filled with color from where the samples were collected. Phylogenetically distinct variants identified among Bhutanese CTV isolates are shown with different colored shapes and placed at the bottom of the map (please refer to **Figures 4**, **5** for detailed phylogenetic classification of CTV variants). The district "Tsirang" was identified as the hotspot of the CTV variants showing the presence of seven major CTV variants out of nine variants identified in this analysis. The district "Chukha" shows the unique presence of a novel variant RB found absent in other districts of Bhutan. Most districts where samples were collected show a high incidence of CTV. The district where CTV was detected with unknown variants is shown with a purple pentagon cartoon. Districts where no CTV and its variants were detected are indicated by an orange cartoon. **(B)** Schematic representation of the genome organization of CTV: colored boxes depict the complete open reading frame (ORF) of the protein-coding and the untranslated region (UTR). Protein coding genes are labeled from 1 to 11. Four CTV genomic locations were targeted for PCR amplification shown with the help of arrows. Primers pair was used to PCR amplify these four genomic locations, and the size of the PCR-amplified product obtained is indicated at the bottom of the image. PRO, papain-like protease domain; MT, methyltransferase domain; HEL, helicase domain; RdRp, RNA-dependent RNA polymerase protein; HSP, heat shock protein; and CPm, minor capsid protein.

TABLE 1 | Details of samples collected from different geographical regions of Bhutan and presence or absence of citrus tristeza virus (CTV) tested by RT-PCR.

Sr. no	Sample code	Citrus cultivar	Botanical name	Location	Symptoms	Target genomic region of CTV			
						5′ORF 1a	p25	p23	p18
1	Bhu-Ts-1	Local mandarin	*Citrus reticulata*	Tsirang	YL	+	+	+	+
2	Bhu-Ts-2	Local mandarin	*Citrus reticulata*	Tsirang	YL, Chl	+	+	+	+
3	Bhu-Ts-3	Local mandarin	*Citrus reticulata*	Tsirang	D, YL, PG	+	+	+	+
4	Bhu-Ts-4	Local mandarin	*Citrus reticulata*	Tsirang	D	+	+	+	+
5	Bhu-Ts-5	Local mandarin	*Citrus reticulata*	Tsirang	YL	+	+	+	+
6	Bhu-Ts-6	Local mandarin	*Citrus reticulata*	Tsirang	VC, Chl, St	+	+	+	+
7	Bhu-Ts-7	Local mandarin	*Citrus reticulata*	Tsirang	AH	+	+	+	+
8	Bhu-Ts-8	Local mandarin	*Citrus reticulata*	Tsirang	D	+	+	+	+
9	Bhu-Ts-9	Local mandarin	*Citrus reticulata*	Tsirang	D	+	+	+	+
10	Bhu-Ts-10	Local mandarin	*Citrus reticulata*	Tsirang	D	+	+	+	+
11	Bhu-Ts-11	Shemjong lime	*Citrus aurantifolia*	Tsirang	PG, Chl	+	+	+	+
12	Bhu-Ts-12	Teishuponkan	*Citrus poonensis*	Tsirang	VCc, Chl	+	+	+	+
13	Bhu-Ts-13	Tarku	*Citrus reticulata*	Tsirang	VC, VF, Chl	+	+	+	+
14	Bhu-Ts-14	Fortunella	*Citrus japonica*	Tsirang	D, St	+	+	+	+
15	Bhu-Ts-15	Dorokhamandarian	*Citrus reticulata*	Tsirang	D, VC. VF	+	+	+	+
16	Bhu-Ts-16	27/28	*Citrus reticulata*	Tsirang	Chl, St	+	+	+	+
17	Bhu-Ts-17	Okitsuwase	*Citrus sinensis*	Tsirang	Chl, D	+	+	+	+
18	Bhu-Ts-18	Othaponkan	*Citrus poonensis*	Tsirang	VC, VF, Chl, St	+	+	+	+
19	Bhu-Ts-19	Yushidaponkan	*Citrus poonensis*	Tsirang	D, Chl	+	+	+	+
20	Bhu-Ts-20	Clementine	*Citrus clementina*	Tsirang	D, VC	+	+	+	+
21	Bhu-Ts-21	Local mandarin	*Citrus reticulata*	Tsirang	AH	–	–	–	–
22	Bhu-Ts-22	Local mandarin	*Citrus reticulata*	Tsirang	Chl, D, PG	+	+	+	+
23	Bhu-Ts-23	Caracara	*Citrus sinensis*	Tsirang	Chl, VC	+	+	+	+
24	Bhu-Ts-24	Citron	*Citrus medica*	Tsirang	PG, D, VC	+	+	+	+
25	Bhu-Ts-25	Local mandarin	*Citrus reticulata*	Tsirang	AH	–	–	–	–
26	Bhu-Ts-26	Local mandarin	*Citrus reticulata*	Tsirang	Chl, D, VC	+	+	+	+
27	Bhu-Ts-27	Local mandarin	*Citrus reticulata*	Tsirang	PG, Chl	+	+	+	+
28	Bhu-Ts-28	Local mandarin	*Citrus reticulata*	Tsirang	Chl, PG, VC	+	+	+	+
29	Bhu-Ts-29	Local mandarin	*Citrus reticulata*	Tsirang	PG, Chl	+	+	+	+
30	Bhu-Ts-30	Local mandarin	*Citrus reticulata*	Tsirang	VC, VF, Chl, St	+	+	+	+
31	Bhu-Ts-31	Pomelo	*Citrus grandis*	Tsirang	YL	–	–	–	–
32	Bhu-Da-32	Local mandarin	*Citrus reticulata*	Dagana	D, YL	+	+	+	+
33	Bhu-Da-33	Rangpur lime	*Citrus limonia*	Dagana	AH	–	–	–	–
34	Bhu-Da-34	Local mandarin	*Citrus reticulata*	Dagana	D, YL	+	+	+	+
35	Bhu-Da-35	Rangpur lime	*Citrus limonia*	Dagana	AH	–	–	–	–
36	Bhu-Da-36	Local mandarin	*Citrus reticulata*	Dagana	YL	+	+	+	+
37	Bhu-Da-37	Rangpur lime	*Citrus limonia*	Dagana	Chl, YL	+	+	+	+
38	Bhu-Da-38	Local mandarin	*Citrus reticulata*	Dagana	YL, Chl	+	+	+	+
39	Bhu-Ts-39	Local mandarin	*Citrus reticulata*	Tsirang	AH	–	–	–	–
40	Bhu-Ts-40	Hayaka,	*Citrus reticulata*	Tsirang	AH, Chl	+	+	+	+
41	Bhu-Ts-41	Berti pomelo	*Citrus grandis*	Tsirang	AH	–	–	–	–
42	Bhu-Ts-42	Hayaka	*Citrus reticulata*	Tsirang	D	+	+	+	+
43	Bhu-Ts-43	Local T-13	*Citrus reticulata*	Tsirang	AH	–	–	–	–
44	Bhu-Ts-44	Clementine	*Citrus reticulata*	Tsirang	YL, D	+	+	+	+
45	Bhu-Ts-45	Salustiana	*Citrus sinensis*	Tsirang	VL, St	+	+	+	+
46	Bhu-Ts-46	Local mandarin	*Citrus reticulata*	Tsirang	YL	–	–	–	–
47	Bhu-Ts-47	Otsu-4	*Citrus reticulata*	Tsirang	YL, VC	+	+	+	+
48	Bhu-Ts-48	Ichang papeda	*Citrus ichangensis*	Tsirang	YL, VC, Chl	+	+	+	+
49	Bhu-Ts-49	Ryan	*Citrus sinensis*	Tsirang	AH	–	–	–	–
50	Bhu-Ts-50	Narng mandarin	*Citrus reticulata*	Tsirang	AH	–	–	–	–
51	Bhu-Ts-51	Dagana mandarin	*Citrus reticulata*	Tsirang	VC, Chl	+	+	+	+

(Continued)

TABLE 1 | (Continued)

Sr. no	Sample code	Citrus cultivar	Botanical name	Location	Symptoms	Target genomic region of CTV			
						5′ORF 1a	p25	p23	p18
52	Bhu-Ts-52	Samtse mandarin	*Citrus reticulata*	Tsirang	AH	–	–	–	–
53	Bhu-Ts-53	Khengkhar mandarin	*Citrus reticulata*	Tsirang	YL, St, VF	+	+	+	+
54	Bhu-Ts-54	Tsirang mandarin	*Citrus reticulata*	Tsirang	VC, VF,	+	+	+	+
55	Bhu-Ts-55	Jongkhar mandarin	*Citrus reticulata*	Tsirang	VF, Chl,	+	+	+	+
56	Bhu-Ts-56	Shumar mandarin	*Citrus reticulata*	Tsirang	Chl, St	+	+	+	+
57	Bhu-Ts-57	Chukha mandarin	*Citrus reticulata*	Tsirang	AH	–	–	–	–
58	Bhu-Wa-58	Local mandarin	*Citrus reticulata*	Wangdue Phodrang	AH	–	–	–	–
59	Bhu-Wa-59	Pomelo	*Citrus grandis*	Wangdue Phodrang	YL	–	–	–	–
60	Bhu-Wa-60	Euraka	*Citrus limon*	Wangdue Phodrang	VC, VF, Chl	+	+	+	+
61	Bhu-Wa-61	Grapefruit	*Citrus paradisi*	Wangdue Phodrang	AH	–	–	–	–
62	Bhu-Wa-62	Mandarin	*Citrus reticulata*	Wangdue Phodrang	VC, VF,	+	+	+	+
63	Bhu-Wa-63	Mandarin	*Citrus reticulata*	Wangdue Phodrang	Chl, PG	+	+	+	+
64	Bhu-Pu-64	Local mandarin	*Citrus reticulata*	Punakha	AH	–	–	–	–
65	Bhu-Tr-65	Local mandarin	*Citrus reticulata*	Trashiyangtse	AH	–	–	–	–
66	Bhu-Tr-66	Local mandarin	*Citrus reticulata*	Trashiyangtse	VC, VF	+	+	+	+
67	Bhu-Tr-67	Local mandarin	*Citrus reticulata*	Trashiyangtse	AH	–	–	–	–
68	Bhu-Zh-68	Local mandarin	*Citrus reticulata*	Zhemgang	VC, VF, Chl, PG	+	+	+	+
69	Bhu-Zh-69	Local mandarin	*Citrus reticulata*	Zhemgang	YL, VC	+	+	+	+
70	Bhu-Zh-70	Local mandarin	*Citrus reticulata*	Zhemgang	D, PG	+	+	+	+
71	Bhu-Zh-71	Local mandarin	*Citrus reticulata*	Zhemgang	D, St	+	+	+	+
72	Bhu-Zh-72	Local mandarin	*Citrus reticulate*	Zhemgang	VC, VF	+	+	+	+
73	Bhu-Zh-73	Local mandarin	*Citrus reticulata*	Zhemgang	AH	–	–	–	–
74	Bhu-Da-74	Local mandarin	*Citrus reticulata*	Dagana	AH	–	–	–	–
75	Bhu-Da-75	Local mandarin	*Citrus reticulata*	Dagana	D, PG	+	+	+	+
76	Bhu-Da-76	Local mandarin	*Citrus reticulata*	Dagana	PG, Chl, VC	+	+	+	+
77	Bhu-Sa-77	Local mandarin	*Citrus reticulata*	Sarpang	AH	–	–	–	–
78	Bhu-Sa-39	Local mandarin	*Citrus reticulata*	Sarpang	D, Chl	+	+	+	+
79	Bhu-Sa-79	Local mandarin	*Citrus reticulata*	Sarpang	D, PG	+	+	+	+
80	Bhu-Sa-80	Local mandarin	*Citrus reticulata*	Sarpang	VC, VF	+	+	+	+
81	Bhu-Sa-81	Local mandarin	*Citrus reticulata*	Sarpang	PG, YL	+	+	+	+
82	Bhu-Sa-82	Local mandarin	*Citrus reticulata*	Sarpang	YL, VC	+	+	+	+
83	Bhu-Sa-83	Local mandarin	*Citrus reticulata*	Sarpang	AH	–	–	–	–
84	Bhu-Sa-84	Local mandarin	*Citrus reticulata*	Sarpang	AH	–	–	–	–
85	Bhu-Ch-85	Dorokha mandarin	*Citrus reticulata*	Chukha	Chl, YL	+	+	+	+
86	Bhu-Ch-86	Wangkhartshalv-I	*Citrus reticulata*	Chukha	AH	–	–	–	–
87	Bhu-Ch-87	Wangkhartshalvngam	*Citrus reticulata*	Chukha	VC, VF, Chl, St	+	+	+	+
88	Bhu-Ch-88	Wangkhartshalvngam	*Citrus reticulata*	Chukha	AH	–	–	–	–
89	Bhu-Ch-89	Satsuma manadarin	*Citrus reticulata*	Chukha	PG, Chl, D	+	+	+	+
90	Bhu-Ch-90	Lemon euraka	*Citrus limon*	Chukha	VC, Chl	+	+	+	+
Total no of positive samples (disease incidence)							**64 (71.11%)**		

Chl, Chlorosis; AH, Apparently healthy; D, Declined; YL, Yellow leaves; PG, Poor growth; VC, Vein clearing; VF, Vein flecking; St, Stunting; +, CTV positive sample; –, CTV negative sample.

Molecular Characterization of Citrus Tristeza Virus Variants

Both the concatenated nucleotide and amino-acid-based maximum likelihood (ML) trees are presented in **Figures 4**, **5**. Worldwide CTV isolates have been classified under seven internationally recognized strains (Harper, 2013). However, the ML tree based on both the nucleotides and protein sequences in our analysis identified two (B1 and B2) additional isolates or variants (**Figures 4**, **5**). These trees show remarkable unity in their branching pattern and relationship with neighboring clades. Based on our analysis, Bhutanese and the worldwide CTV isolates could be robustly classified under the following variants described below:

Resistance-Breaking Isolate

RB isolate was named after discovering the founding member, *Poncirus trifoliate* resistance-breaking (RB) strain from New Zealand and shown to have 90% nucleotide sequence

TABLE 2 | Details of the primers used in the present study.

Sr. no	Primer code	Sequence	Annealing temp.	Amplicons size	Target genomic regions	References
1	488F 491R	5′ TGTTCCGTCCTGSGCGGAAYAATT 3′ 5′ GTGTARGTCCCRCGCATMGGAACC 3′	58°C	404 bp	5′ ORF 1a	Roy et al., 2005a
2	CN150 CN151	5′ATATATTTACTCTAGATCTACCATGGACGACGAAACAAA 3′ 5′ GAATCGGAACGCGAATTCTCAACGTGTTAAATTTCC 3′	61°C	672 bp	p25	Cevik et al., 1996
3	RBP-23F RBP-23R	5′ ATGAACGATACTAGCGGAC 3′ 5′ GATGAAGTGGTGTTCACGG 3′	52°C	627 bp	p23	Kokane S. B. et al., 2020
4	AR18F AR18R	5′ ATGTCAGGCAGCTTGGGAAATT 3′ 5′ TTCGTGTCTAAGTCRCGCTAAACA 3′	62°C	511 bp	P18	Roy et al., 2005b

identity against the neighboring clade T36 (Harper et al., 2010). The presence of an RB strain among the Puerto Rican isolates was also reported (Roy et al., 2013). The RB isolate has not been reported elsewhere from the world, including the Asian subcontinent. Besides establishing for the first time the RB strain from South Africa (B389-4) and Australia (PB61), based on our comparative bioinformatic analysis, here, we report for the first time the presence of RB strain among Bhutanese isolates (**Figure 4**). Validation in the glasshouse using the host assay system is subject to future further analysis. Our analysis also suggests that a small invariable pentapeptide signature motif "RVENV" is present at the amino terminus of the p23 protein sequence, which separates RB strains from the rest of the CTV isolates worldwide. The Bhutanese variant Bhu-Ch-90 carry these invariable amino acids in its p23 gene and confirms its classification under the RB clade. An Indian isolate (GFA-MH) with only the coat-protein (p25) gene sequence from Maharashtra falls under the RB clade in the nucleotide-based tree (**Figure 4**) and is segregated under the T3 clade in the amino acid-based tree (**Figure 5**). Together, these results would suggest the widespread presence of the RB variant than earlier realized poignantly located at the center of origin of citrus.

T36 Isolate

T36 isolate was named after the founding member of the Florida decline isolate for which the whole-genome sequence was published as early as 1995 (Karasev et al., 1995), and the whole-genome sequence of T36 from Turkey was revealed recently (Cevik et al., 2013). After analyzing the GenBank deposited sequences, we observed that the T36 strain is widespread in Taiwan, Mexico, Tunisia, and Italy (**Figures 4**, **5**). However, the presence of the T36 isolate has not been reported elsewhere, particularly from the Asian subcontinent, including north-eastern India, which has been considered the likely place of origin of citrus (Wu et al., 2018). For the first time in this analysis, we report and establish the presence of a T36 strain (Bhu-Ts-14) among Bhutanese CTV isolates, which form a strong clade along with other T36 strains (**Figures 4**, **5**).

T30 Isolate

This strain was named after Florida isolate T30 (Albiach-Martí et al., 2000). The whole-genome sequence, including the two other isolates from Florida and China, is available in the GenBank (Albiach-Martí et al., 2000). These T30 isolates form a distinct clade in nucleotide and protein-based ML trees (**Figures 4**, **5**). We were unable to identify any Bhutanese isolates that belonged to this clade. However, Bhutanese isolates (Bhu-Ts-12, 13, 18, and 20) form a distinct clade with solid statistical support within the T30 and T3 clade (**Figure 4**), and it could be the transitional state from which T30 or the T3 sequences might have evolved. However, these Bhutanese isolates come under the T3 clade when the amino-acids-based ML tree was generated (**Figure 5**). However, our finding of a Chinese isolate that belonged to the T30 group would, for the first time, establish its Asian origin (**Figures 4**, **5**). Besides, the robust sister clade relationship among RB, T30, and T36 with solid statistical support indicated that these isolates might have originated from their ancestral sequences that subsequently gave rise to these three clade members.

T3 Isolate

The T3 isolate was named after the founding member recovered from a lime tree in Florida, for which the whole-genome sequence is available in the GenBank. A strain from New Zealand (NZ-M16) has been classified as T3-like, for which the whole-genome sequence is available (Harper et al., 2009). This strain, in our analysis, forms its clade next to the T68 clade with robust statistical support together with other sequences from India that suggested its Asian origin (**Figures 4**, **5**). We retain the name of this clade as NZ-M16 to distinguish it from other CTV isolates (**Figures 4**, **5**). In addition, the GenBank deposited sequences analysis identified several T3 isolates from South Africa: Maxi, T3-KB, N10-64 (**Figures 4**, **5** and **Supplementary Excel File 1**). We also identified a T3 isolate from Brazil (GenBank Acc. DQ363400), for which only 5′ORF1a sequence data are available in the GenBank. Several Indian sequences (An-9, Kat1, Kat3, and K38) also fall within this clade with strong statistical support value. Bhutanese sequences, specifically Bhu-Ts-12, 13, 18, and 20, form a strong cluster within the T3 clade based on their amino acid sequences. These results suggest that the T3 isolate is being distributed worldwide, and its origin could be traced to the center of origin of citrus.

T68 Isolate

T68 belongs to the Florida isolates for which the whole-genome sequence is available in the GenBank. In addition, the full-genome sequences of two T68 strains that differ in their stem-pitting severity reported from South Africa are also published

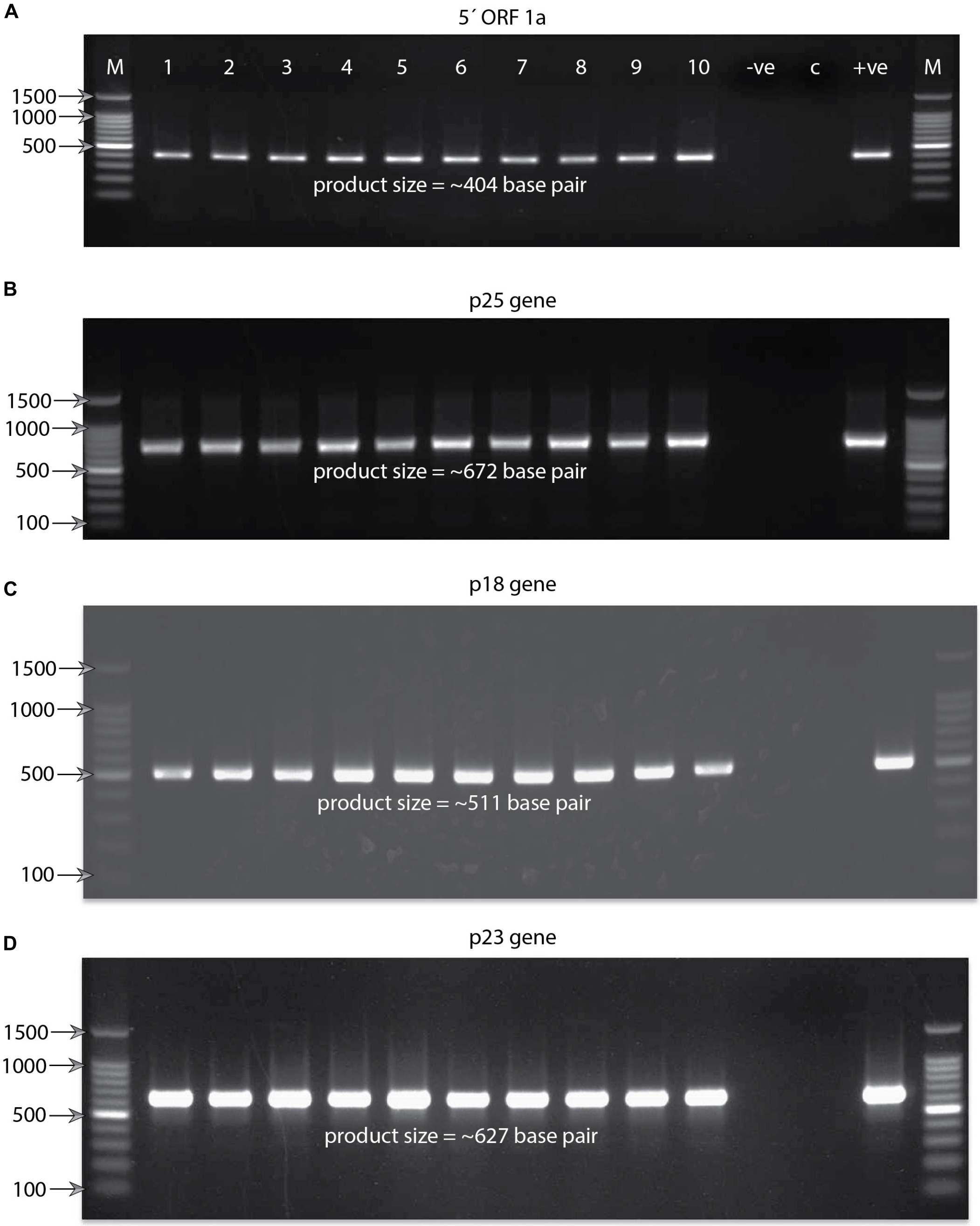

A
5´ ORF 1a
M
1
2
3
4
5
6
7
8
9
10
-ve
c
+ve
M
1500
1000
500
product size = ~404 base pair
B
p25 gene
1500
1000
500
100
product size = ~672 base pair
C
p18 gene
1500
1000
500
100
product size = ~511 base pair
D
p23 gene
1500
1000
500
100
product size = ~627 base pair

FIGURE 3 | Agarose gel electrophoresis pictures showing RT-PCR-amplified product. **(A)** Product size of (∼404) base pair obtained targeting the 5′ORF1a gene against the Bhutanese CTV variants. At the top of the gel lanes from right to left are labeled as follows: lane M, 100-base-pair DNA ladder; lanes 1–10, representatives of the Bhutanese CTV variants; lane −ve, reaction control; lane C, healthy plant control; and lane + ve, positive control. For clarity, labeling is not shown for the other three gel pictures. The same labeling of the first gel applies to all four gel pictures. Some of the significant DNA ladder sizes are labeled and shown with the help of arrows. **(B)** Gel showing the RT-PCR product size of ∼672 base pairs obtained targeting the p25 coat protein gene of the Bhutanese CTV variants. **(C)** Gel showing the RT-PCR product size of ∼511 base pairs obtained targeted against the p18 gene of the Bhutanese CTV variants. **(D)** Gel showing the RT-PCR product size of ∼627 base pairs targeted against the RNA binding p23 gene of the Bhutanese CTV variants. Primer pairs used to amplify these products are mentioned in **Figure 2B** and section "Materials and Methods."

and available in the GenBank (Cook et al., 2020, 2021). The complete genome sequence of an orange stem-pitting isolate (B165) from India (Roy and Brlansky, 2010) was classified under VT, but our analysis suggested reconsidering it to be classified under T68 isolate (**Figures 4**, **5**). Several isolates, specifically TG2, 3, and 5 from the north-eastern region of India from

Assam, form a strong cluster within the T68 clade along with K10, K14, and Kpg6 isolates from Darjeeling, West Bengal, India. The Bhutanese isolate Bhu-Ts-27 based on sequences of all four independent genomic locations formed a strong association within the T68 clade in both nucleotide and protein trees (**Figures 4**, **5**). However, Bhu-Ts-28 based only on the p18 genomic region was found to switch clades between T68 and VT-B. Therefore, we conclude that the Bhu-Ts-28 strain remained unclassified, and the additional genomic sequence would be required for confident classification. Together, our analysis suggests that T68 is distributed worldwide, possibly originating in the north-eastern region of India and Bhutan. We also observed that NZ-M16 formed a sister clade with T68 clade with strong statistical support, which indicated its Indian origin.

VT Isolate

The VT strain was named after discovering its founding member from Israel, for which the full-genome sequence is available. In addition to this VT isolate, two other VT strains have been reported from Florida, and all these three VT isolates form an independent clade in the nucleotide tree adjacent to the T68, NZ-M16 clade (**Figure 4**). Bhutan's unidentified VT-like sequences get segregated within the same clade (**Figure 4**). We have to refer to these Bhutanese VT-like sequences as an independent VT-B clade where the B is derived from Bhutan. In the protein tree, some of the sequences from India fall in the same clade as the Florida VT strains without significant statistical support (**Figure 5**). These Indian sequences, however, form a strong clade together with HA16-5 (Hawaii isolates of CTV), indicating that Florida VT strains are related and originated from any of these four ancestral clades, namely, T68, NZ-M16, VT-B, and HA16-5. In addition, our analyses also have suggested that the ancestry of these four clade members can be traced back to their roots in the north-eastern region of India and Bhutan (Wu et al., 2018).

HA16-5 Isolate

This isolate was named after the founding member from Hawaii and was classified as a new genotype for which the whole-genome sequence has been published (Melzer et al., 2010). The CTV isolate LMS6-6 from South Africa was recently classified under HA16-5 clade (Cook et al., 2016). Both LMS6-6 and HA16-5 in our analysis fall in the HA16-5 clade in both nucleotide and protein trees with firm statistical support (**Figures 4**, **5**). Four sequences are supposedly *Poncirus*-resistant breaking strains, namely, CA-RB-AT3 from California, United States (Yokomi et al., 2017), L13 from China (Wang et al., 2019), DSST17 from Uruguay (Benitez-Galeano et al., 2018), and B390-5 from South Africa (Cook et al., 2016) and have been classified under RB clade, which, in our analysis with both nucleotide and protein tree, form a strong clade together with HA16-5 and LMS6-6 (**Figures 4**, **5**). The clade HA16-5 also accommodates several isolates from the southern part of India and the north-eastern region, including Assam and Darjeeling, together with at least three isolates from Bhutan (Bhu-Ts-15, Bhu-Wa-62, and Bhu-Ch-87) (**Figures 4**, **5**). This result suggests that the HA16-5 strains have been distributed worldwide, with the root traced back to Bhutan and Northeast India.

B1 Isolate

A severe stem-pitting (SP) isolate from California (SY568) reported to have sequence similarities with Florida and Israel VT strain has been published (Yang et al., 1999). In addition, the whole-genome sequences of two Italian isolates Mac39 and SG29 sequences are also available in the GenBank, of which SG29 was shown to have clustered within the VT-Asian subtypes (Licciardello et al., 2015). Similarly, the whole-genome sequences of severe stem-pitting (SP) isolates from Spain (Ruiz-Ruiz et al., 2006) and Brazilian isolate CSL02 have been classified under the VT group (Matsumura et al., 2017). The full-length genome of NUagA isolate from Japan, which causes seedling yellows, and sequences of four unclassified isolates, namely, HU-PSTS, FN08, CT11A, and AT-1, from China are available in the GenBank. Together with the isolates from India and Bhutanese variants, all these isolates are mentioned here. Specifically, Bhu-Ts-17, Bhu-Ts-22, Bhu-Ts-23, Bhu-Ts-26, Bhu-Ch-89 form an unrelated cluster distinctly different from VT isolates in both nucleotide and protein trees (**Figures 4**, **5**). This result suggests that these isolates should be classified under distinct clade, which we prefer to name B1 clade after Bhutan. Thus, the B1 clade harboring the severe stem-pitting isolates has a worldwide distribution, and its root could be traced back to the north-eastern region of India and Bhutan (Wu et al., 2018).

B2 Isolate

B2 isolates form a distinctly different clade in the ML tree with solid statistical support next to the HA16-5 clade in nucleotide and amino acid tree (**Figures 4**, **5**). Isolates of this clade have so far been found only among Bhutanese variants (Bhu-Ts-11, Bhu-Ts-16, Bhu-Ts-24, Bhu-Ts-29) and variants from the north-eastern region of Southeast Asia, including Assam and Darjeeling.

Distribution of Citrus Tristeza Virus Variants in Bhutan

Phylogenetic analysis using the four genomic locations has allowed us to classify the CTV isolates into nine major groups (RB, T36, T30, T3, T68, VT or VT-B, HA16-5, B1, and B2). Except for the T30 group, Bhutanese isolates have representations in all eight groups indicating greater diversity. Except for the resistant-breaking strain RB, all other seven variants were observed mainly in the Tsirang district. The devastating VT-B strain occurs throughout Bhutan's major citrus growing districts (**Figure 2A**). The second most widely distributed variant was HA16-5, which, besides Tsirang, was found in two other regions, Wangdue Phodrang and Chukha districts, and the resistant-breaking RB strain was reported exclusively from the Chukha region (**Figure 2A**).

DISCUSSION

CTV, the largest and most complex member of the family *Closteroviridae*, is a phloem-limited virus that infects citrus and closely related species and produces a wide range of characteristic symptoms. Viruses having RNA as their genome have the potential for genetic variations due to their error-prone

TABLE 3 | Citrus tristeza virus (CTV) isolates collected from different geographic regions of Bhutan. Four different specific genomic regions are sequenced and their accession numbers are presented.

Sr. no	Sample code	Concatenated study based CTV groups	CTV accession			
			5′ ORF 1a	p25	p23	p18
1	Bhu-Ts-1	VT-B	SND	SND	MN104226	SND
2	Bhu-Ts-2	VT-B	SND	SND	MN104227	SND
3	Bhu-Ts-3	VT-B	SND	SND	MN104228	MN117985
4	Bhu-Ts-4	VT-B	SND	MN104221*	MN104229	MN117986
5	Bhu-Ts-5	VT-B	SND	SND	MN104230	MN117987
6	Bhu-Ts-6	VT-B	SND	MN104222*	MN117969	SND
7	Bhu-Sa-7	VT-B	MN384882	SND	SND	SND
8	Bhu-Ts-8	VT-B	SND	MN104223*	MN117970	MN117988
9	Bhu-Ts-9	VT-B	SND	MN104224*	MN117971	MN117989
10	Bhu-Ts-10	VT-B	SND	MN104225*	MN117972	MN117990
11	Bhu-Ts-11	B2	MN384885	SND	MN549939	MN580428
12	Bhu-Ts-12	T3	SND	MN366299	MN549947	SND
13	Bhu-Ts-13	T3	MN384883	MN366307	MN549941	MN580427
14	Bhu-Ts-14	T36	SND	MN366297	SND	MN580434
15	Bhu-Ts-15	HA16-5	MN384884	MN366302	MN549942	MN580429
16	Bhu-Ts-16	B2	SND	SND	MN549940	MN580430
17	Bhu-Ts-17	B1	MN651084	SND	MN549944	MN580432
18	Bhu-Ts-18	T3	MN651083	MN366301*	MN549948	MN580435
19	Bhu-Ts-19	VT-B	MN651088	MN366303*	MN549945	MN580436
20	Bhu-Ts-20	T3	MN651089	MN366300*	MN549951	MN580422
21	Bhu-Ts-22	B1	MN651087	MN366306	MN549954	MN580426
22	Bhu-Ts-23	B1	SND	SND	MN549949	MN580433
23	Bhu-Ts-24	B2	MN651085	SND	MN549953	MN580424
24	Bhu-Ts-26	B1	SND	SND	MN549952	MN580423
25	Bhu-Ts-27	T68	MN651086	MN366298	MN549950	MN580437
26	Bhu-Ts-28	VT-B	SND	SND	SND	MN580425
27	Bhu-Ts-29	B2	SND	MN366305	MN549943	MN580431
28	Bhu-Ts-30	VT-B	MN384880	MN366304	MN549946	MN580438
29	Bhu-Da-36	T68	SND	MN101752*	MN137882	MN137877
30	Bhu-Da-38	VT-B	SND	SND	MN137883	MN137878
31	Bhu-Wa-60	VT-B	MN651093	MN366309	MN398270	SND
32	BhuWa-62	HA16-5	MN651091	MN366310	SND	SND
33	Bhu-Zh-68	VT-B	MN651096	MN366311*	MN398271	SND
34	Bhu-Zh-71	VT-B	SND	MN101753*	MN137884	MN137879
35	Bhu-Zh-72	VT-B	MN651094	MN366316	MN398272	SND
36	Bhu-Da-76	VT-B	MN651097	MN366312*	MN398273	MN580439
37	Bhu-Sa-39	VT-B	SND	SND	MN137885	SND
38	Bhu-Sa-79	VT-B	SND	SND	MN137887	MN137881
39	Bhu-Sa-80	VT-B	MN651092	MN366313	MN398274	MN580440
40	Bhu-Sa-81	VT-B	MN651095	SND	SND	MN580441
41	Bhu-Sa-82	VT-B	SND	SND	MN137886	SND
42	Bhu-Ch-87	HA16-5	MN651090	MN366314	MN398277	MN580442
43	Bhu-Ch-89	B1	MN651098	MN366315	MN398278	SND
44	Bhu-Ch-90	RB	SND	MN366317	MN398279	SND
Major CTV strains						
45	T36	T 36	U16304	U16304	U16304	U16304
46	T30	T 30	AF260651	AF260651	AF260651	AF260651
47	VT	VT	EU937519	EU937519	EU937519	EU937519
48	T3	T 3	KC525952	KC525952	KC525952	KC525952
49	T68	T 68	JQ965169	JQ965169	JQ965169	JQ965169
50	RB	RB	FJ525434	FJ525434	FJ525434	FJ525434
51	HA16-5	HA16-5	GQ454870	GQ454870	GQ454870	GQ454870

*CTV, Citrus tristeza virus; *, Our earlier studied samples; SND, Sequencing not done.*

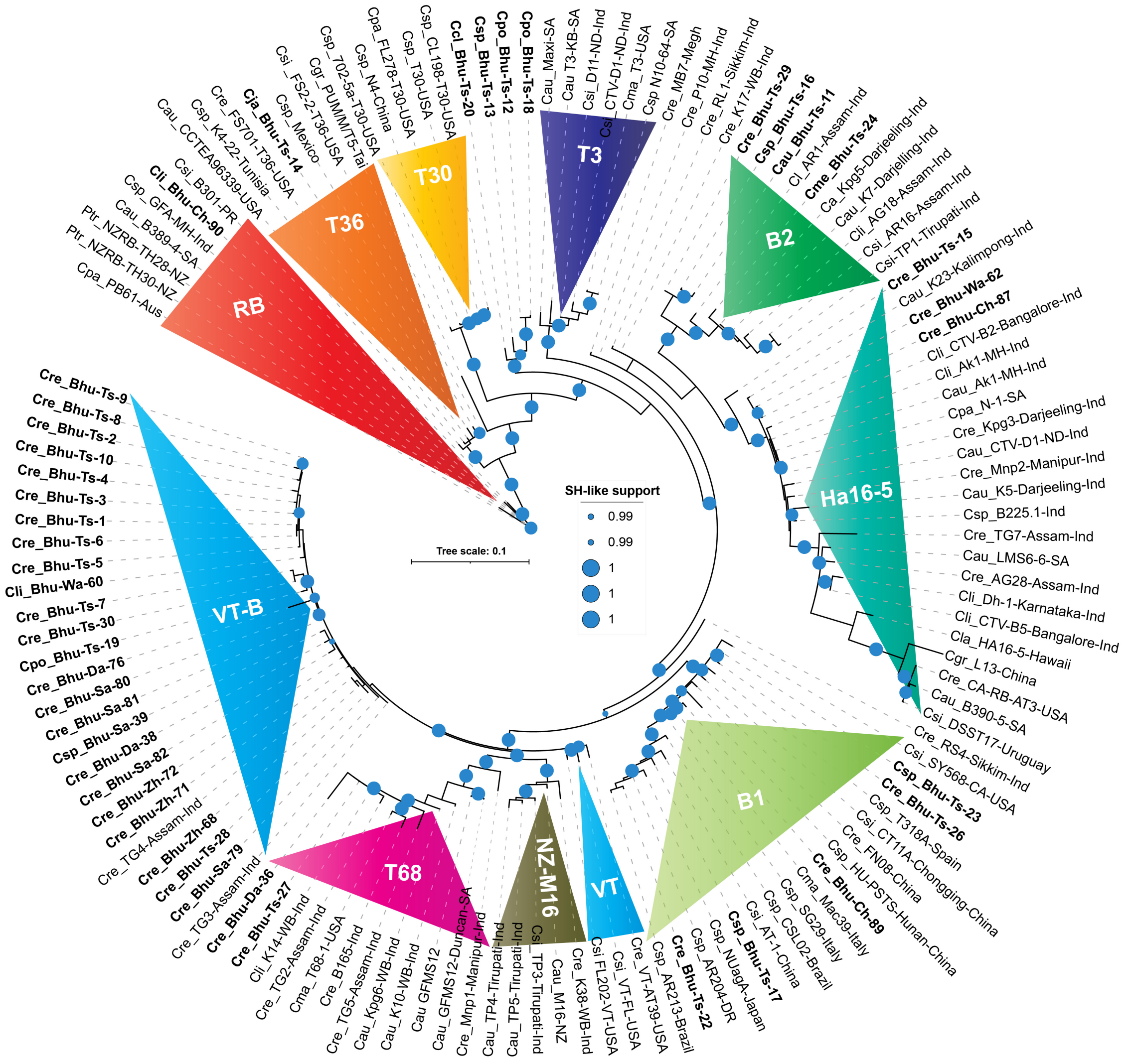


FIGURE 4 | Phylogenetic trees reconstructed using concatenated DNA correspond to four genomic locations showing unprecedented diversities among Bhutanese CTV variants. Maximum likelihood (ML) tree was reconstructed using PHYML3.2.2 (Guindon et al., 2010) with default parameters, except for the model parameter that was set to GTR, and both NNI (nearest-neighbor interchange) and SPR (sub-tree pruning and re-grafting) moves were adopted for their accuracy of branch placement along the tree (Hordijk and Gascuel, 2005). SH-like (Anisimova et al., 2011) support, which is known to outperform bootstrap support (Simmons and Norton, 2014), was used for branch stability. SH-like support of 99% or higher is shown with a cyan color circle. CTV strains were earlier classified under seven (T30, T36, VT, T3, T68, RB, and HA16-5) subtypes or variants (Harper, 2013). In contrast, our concatenated nucleotides-based ML tree showing an additional two (B1 and B2) clades could be confidently allotted with a high SH-like support value. Clades B1 and B2 are named after the Bhutanese variants. Labeling the ML tree was made clearer by using the first letter of the genus and two letters of the citrus species, followed by the isolate or genotype name, place, and the country where it was reported. For clarity, Bhutanese isolates are shown in bold font. For a complete list of the GenBank accessions, species name, isolate or variants name, country name, protein-coding region, and their respective DNA used to build this tree refer to **Supplementary Excel File 1**, and for the complete list of Bhutanese isolates, refer to **Tables 1**, **3**.

replication mechanism (Rubio et al., 2001). Similarly, CTV can evolve and exhibit variable pathogenesis on the citrus hosts. Several genomic regions have been targeted and characterized to determine the genetic diversity among the CTV isolates (Ghosh et al., 2008, 2021; Martín et al., 2009; Warghane et al., 2020). It was reported that the 3′ half of the CTV genome, among various isolates, is more conserved (90% identity), while most genetic diversity (44–88% identity) is found at the 5′ terminal half (Atallah et al., 2016). The genetic diversity based on the 5′ terminal region and based on the two combined regions (5′ORF1a and p25 genes) has been reported to discriminate CTV isolates from the northeastern and southern parts of India (Roy et al., 2005a; Tarafdar et al., 2013). In the present study, efforts have been carried out to determine the genetic diversity of CTV isolates from Bhutan using 5′ end ORF1a and three other potential genes of 3′ end, *viz.*, p25, p23, and p18.

In the present investigation, a novel approach of concatenating-independent genomic locations utilizing both

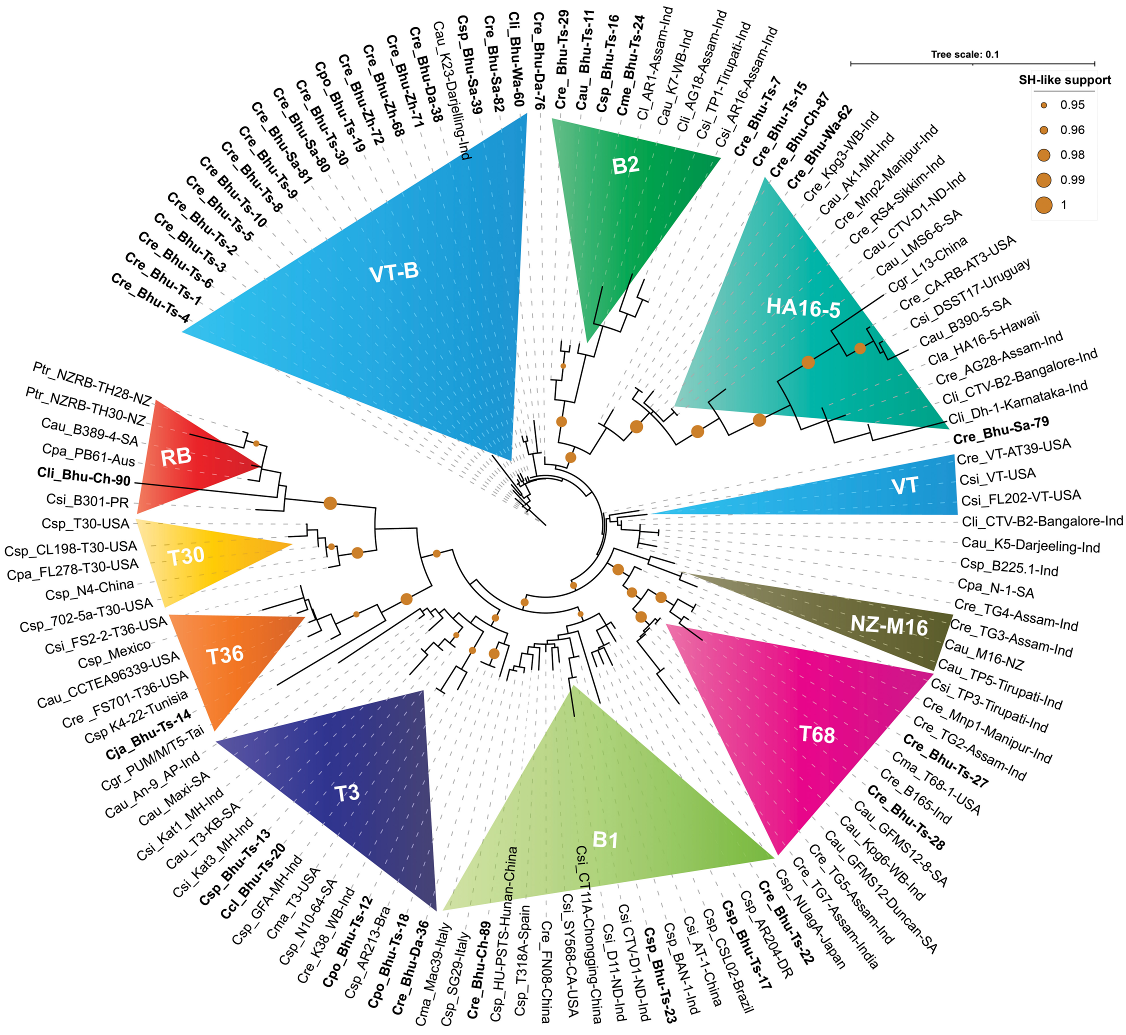


FIGURE 5 | Phylogenetic tree reconstructed using concatenated protein sequences congruent with the nucleotide tree (**Figure 4**). The maximum likelihood (ML) tree was reconstructed using PHYML3.2.2 (Guindon et al., 2010) with default parameters except for the model parameter that was set to JTT (Jones et al., 1992), and both NNI (nearest-neighbor interchange) and SPR (sub-tree pruning and re-grafting) moves were adopted for better accuracy (Hordijk and Gascuel, 2005). SH-like (Anisimova et al., 2011) support, which is known to outperform bootstrap support (Simmons and Norton, 2014), was used for branch stability. SH-like support of 95% or higher is shown with a mustard color circle. CTV strains were earlier classified under seven (T30, T36, VT, T3, T68, RB and HA16-5) subtypes or variants (Harper, 2013). In contrast, our concatenated protein-based ML tree shows an additional three (B1, B2, and NZ-M16) clades. Clades B1 and B2 are named after the Bhutanese variants. Labeling in the ML tree was made clearer by using the first letter of the genus and two letters of the citrus species, followed by the isolate or genotype name, place, and the country where it was reported. For clarity, Bhutanese isolates are shown in bold font. For a complete list of the GenBank accessions, species name, isolate or variants name, country name, and protein-coding amino acids used to build this tree, refer to **Supplementary Excel File 1**, and for the complete list of Bhutanese isolates, refer to **Tables 1**, **3**.

the nucleotide and their corresponding amino acid sequences for differentiation of tristeza variants from Bhutan and across the World has been used. This work provides a new framework for revisiting and re-classifying the existing tristeza variants in future studies. The four genomic locations used in this study were in the viral homologous recombination-free regions in the tristeza genome (Vives et al., 2005) and, thus, allowed us to build a robust phylogenetic relationship among global CTV isolates. Although consensus trees generated are robust, common sources of phylogenetic error such as long-branch attraction might skulk; we remain cautious of our interpretation while classifying the new clade such as B2, which is based solely on the 5′ORF1a sequence available at this moment for phylogenetic reconstruction. Despite limited sequence information available in some instances (used for phylogenetic reconstruction), an astounding congruence was noticed in the placement of the isolates within the clade and their sister–clade relationship in the concatenated nucleotide and the protein tree, which further validate the robustness of our classification.

The most striking feature of the results is the sequence diversity assessment among all the Bhutanese CTV variants and establishing their one-to-one relationship with existing worldwide-recognized isolates. Sequences were extracted from the whole-genome sequences and partially sequenced CTV isolates that were previously reported (Hilf et al., 2005; Roy et al., 2005a; Harper, 2013; Ghosh et al., 2021). An exhaustive search was done both for the whole-genome sequence information and the partial sequences available in the GenBank. It was beyond the scope to incorporate all sequences from the GenBank. However, we incorporated all worldwide-recognized CTV isolates that belonged to the seven well-known clades, namely, RB, T36, T30, T3, T68, VT, and HA16-5, and other full-genome and partial CTV isolate sequences supporting the clades. Concatenated nucleotide and protein-based trees are novel approaches that never have been used for plant virus isolates differentiation. We see some strains, such as Bhu-Ts-7, Bhu-Ts-28, and GFA-MH-Ind, switching the places between the clades, mainly because of very limited sequence information available for them. For example, out of four genomic locations targeted in this analysis, sequence information for only one genomic location was available for these isolates for phylogenetic reconstruction. More genomic information for these strains would be required to establish a stable phylogenetic relationship among these Bhutanese isolates. Another most striking result that appeared from our result is placing the roots of these nine clades identified in this analysis to the center of origin of citrus to the north-eastern region of India and Bhutan, which was not defined so far in earlier studies.

Recently, the association of CTV and *Candidatus* Liberibacter asiaticus with citrus decline has been recorded in Bhutan with higher incidence up to 70.58 and 27.45%, respectively (Ghosh et al., 2021). The present investigation also suggests an average incidence of 71.11% (64 out of 90 samples tested positive) of CTV occurring in Bhutan based on targeted genes (5′ORF1a, p25, p23, and p18) by RT-PCR test from eight different districts of Bhutan. The highest CTV incidence was recorded in Zhemgang and Tsirang followed by other districts. We also observed that most of the citrus orchards were neglected, and infestation of aphids was common in most of the surveyed orchards. The tristeza disease is a major threat to the citrus in the northeast region of India, including Bhutan (Borah et al., 2014; Warghane et al., 2020), and aphids may be the major source of virus spread. The evidence generated in the present study will be helpful in quarantine applications in Bhutan. Furthermore, sanitation and the use of virus-free propagation material will be the most powerful method to put the citrus industry of Bhutan on sound scientific footings for increased citrus productivity.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

DG, AK, and SK designed the study and developed the methods. DG, KM, SK, and AK prepared the data. All authors analyzed the results and wrote the manuscript.

FUNDING

This research was funded by the ICAR-Consortia Research Platform (CRP) on Vaccines and Diagnostics, Government of India, F.No. 16-11/pp/ICAR-CRP/17-18/06.

ACKNOWLEDGMENTS

We would like to thank Siddarame Gowda, University of Florida, Citrus Research and Education Centre, Lake Alfred, FL, United States, for reviewing the manuscript and providing valuable suggestions for improving the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.797463/full#supplementary-material

REFERENCES

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105. doi: 10.1093/bioinformatics/bti263

Ahlawat, Y. S. (2012). *Virus Diseases of Citrus & Management*. New Delhi: Studium Press (India).

Albiach-Martí, M., Mawassi, M., Gowda, S., Satyanarayana, T., Hilf, M. E., Shanker, S., et al. (2000). Sequences of *Citrus tristeza* virus separated in time and space are essentially identical. *J. Virol.* 74, 6856–6865. doi: 10.1128/jvi.74.15.6856-6865.2000

Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C., and Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* 60, 685–699. doi: 10.1093/sysbio/syr041

Atallah, O. O., Kang, S. H., El-Mohtar, C. A., Shilts, T., Bergua, M., and Folimonova, S. Y. (2016). A 5′-proximal region of the *Citrus Tristeza* virus genome encoding two leader proteases is involved in virus super infection exclusion. *Virology* 489, 108–115. doi: 10.1016/j.virol.2015.12.008

Bar-Joseph, M., Marcus, R., and Lee, R. F. (1989). The continuous challenge of *citrus tristeza* virus control. *Annu. Rev. Phytopathol.* 27, 291–316.

Benitez-Galeano, M. J., Vallet, T., Carrau, L., Hernandez-Rodriguez, L., Bertalmio, A., Rivas, F., et al. (2018). Complete genome sequence of a novel recombinant *Citrus tristeza* virus, a resistance-breaking isolate from Uruguay. *Genome Announc.* 6:e00442-18. doi: 10.1128/genomeA.00442-18

Biswas, K. K., Palchoudhury, S., Sharma, S. K., Saha, B., Godara, S., Ghosh, D. K., et al. (2018). Analyses of the 3′ half genome of *citrus tristeza* virus reveal the existence of distinct virus genotypes in citrus growing regions of India. *Virus Dis.* 29, 308–315. doi: 10.1007/s13337-018-0456-2

Borah, M., Nath, P. D., and Saikia, A. K. (2014). Biological and serological technique for detection of *citrus tristeza* virus affecting citrus species of Assam. *India Afr. J. Agric. Res.* 9, 3804–3810.

Brlansky, R. H., Damsteegt, V. D., Howd, D. S., and Roy, A. (2003). Molecular analyses of *Citrus tristeza* virus sub isolates separated by aphid transmission. *Plant Dis.* 87, 397–401. doi: 10.1094/PDIS.2003.87.4.397

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Cevik, B., Pappu, S. S., Pappu, H. R., Tight, D., Benscher, D., Futch, S. H., et al. (1996). Molecular cloning and sequencing of coat protein genes of *citrus tristeza* virus isolated from meyer lemon and homely tangor trees in Florida. *Int. Organ. Citrus Virol. Conf. Proc.* 13, 1957–2010.

Cevik, B., Yardimci, N., and Korkmaz, S. (2013). The first identified *Citrus tristeza* virus Isolate of Turkey contains a mixture of mild and severe strains. *Plant Pathol. J.* 29, 31–41. doi: 10.5423/PPJ.OA.09.2012.0141

Cook, G., Breytenbach, J. H., Steyn, C., de Bruyn, R., van Vuuren, S. P., Burger, J. T., et al. (2021). Grapefruit field trial evaluation of *citrus tristeza* virus T68-strain sources. *Plant Dis.* 105, 361–367. doi: 10.1094/PDIS-06-20-1259-RE

Cook, G., Coetzee, B., Bester, R., Breytenbach, J. H., Steyn, C., de Bruyn, R., et al. (2020). *Citrus tristeza* virus isolates of the same genotype differ in stem pitting severity in grapefruit. *Plant Dis*. 104, 2362–2368. doi: 10.1094/PDIS-12-19-2586-RE

Cook, G., van Vuuren, S. P., Breytenbach, J. H., Steyn, C., Burger, J. T., and Maree, H. J. (2016). Characterization of *citrus tristeza* virus single-variant sources in grapefruit in greenhouse and field trials. *Plant Dis.* 100, 2251–2256. doi: 10.1094/PDIS-03-16-0391-RE

Dawson, W. O., Garnsey, S. M., Tatineni, S., Folimonova, S. Y., Harper, S. J., and Gowda, S. (2013). *Citrus tristeza* virus-host interactions. *Front. Microbiol.* 4:88. doi: 10.3389/fmicb.2013.00088

Dolja, V. V., Kreuze, J. F., and Valkonen, J. P. (2006). Comparative and functional genomics of *closteroviruses*. *Virus Res.* 117, 38–51. doi: 10.1016/j.virusres.2006.02.002

Dorji, K., Lakey, L., Chophel, S., Dorji, S. D., and Tamang, B. (2016). Adoption of improved citrus orchard management practices: a micro study from Drujegang growers. Dagana, Bhutan. *Agric. Food Secur*. 5, 1–8.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113

Flores, R., Ruiz-Ruiz, S., and Soler, N. (2013). *Citrus tristeza* virus p23: a unique protein mediating key virus–host interactions. *Front. Microbiol* 4:98. doi: 10.3389/fmicb.2013.00098

Ghosh, D. K., Aglave, B., and Baranwal, V. K. (2008). Simultaneous detection of one RNA and one DNA virus from naturally infected citrus plants using duplex PCR technique. *Curr. Sci.* 25, 1314–1318.

Ghosh, D. K., Aglave, B., Roy, A., and Ahlawat, Y. S. (2009). Molecular cloning, sequencing and phylogenetic analysis of coat protein gene of a biologically distinct *Citrus tristeza* virus isolate occurring in central India. *J. Plant Biochem. Biotechnol.* 18, 105–108.

Ghosh, D. K., Kokane, A. D., Kokane, S. B., Tenzin, J., Gubyad, M. G., Wangdi, P., et al. (2021). Detection and molecular characterization of '*Candidatus* Liberibacter asiaticus' and *Citrus tristeza* virus associated with citrus decline in Bhutan. *Phytopathology* 111, 870–881. doi: 10.1094/PHYTO-07-20-0266-R

Ghosh, D. K., Kokane, S. B., and Gowda, S. (2020). Development of a reverse transcription recombinase polymerase based isothermal amplification coupled with lateral flow immunochromatographic assay (CTV-RT-RPA-LFICA) for rapid detection of *Citrus tristeza* virus. *Sci Rep.* 10:20593. doi: 10.1038/s41598-020-77692-w

Ghosh, D. K., Kokane, S. B., Kokane, A. D., Warghane, A. J., Motghare, M. R., Bhose, S., et al. (2018). Development of a recombinase polymerase based isothermal amplification combined with lateral flow assay (HLB-RPA-LFA) for rapid detection of '*Candidatus* Liberibacter asiaticus'. *PLoS One* 13:e0208530. doi: 10.1371/journal.pone.0208530

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Harper, S. J. (2013). *Citrus tristeza* virus: evolution of complex and varied genotypic groups. *Front. Microbiol.* 4:93. doi: 10.3389/fmicb.2013.00093

Harper, S. J., Dawson, T. E., and Pearson, M. N. (2009). Complete genome sequences of two distinct and diverse *Citrus tristeza* virus isolates from New Zealand. *Arch. Virol* 154, 1505–1510.

Harper, S. J., Dawson, T. E., and Pearson, M. N. (2010). Isolates of *Citrus tristeza* virus that overcome *Poncirus trifoliata* resistance comprise a novel strain. *Arch. Virol.* 155, 471–480. doi: 10.1007/s00705-010-0604-5

Herron, C. M., Mirkov, T. E., da Graça, J. V., and Lee, R. F. (2006). *Citrus tristeza* virus transmission by the *Toxoptera citricida* vector: in vitro acquisition and transmission and infectivity immuno neutralization experiments. *J. Virol. Methods* 134, 205–211.

Hilf, M. E., Karasev, A. V., Pappu, H. R., Gumpf, D. J., Niblett, C. L., and Garnsey, S. M. (1995). Characterization of *citrus tristeza* virus subgenomic RNAs in infected tissue. *Virology* 208, 576–582. doi: 10.1006/viro.1995.1188

Hilf, M. E., Mavrodieva, V. A., and Garnsey, S. M. (2005). Genetic marker analysis of a global collection of isolates of *citrus tristeza* virus: Characterization and distribution of CTV genotypes and association with symptoms. *Phytopathology* 95, 909–917. doi: 10.1094/PHYTO-95-0909

Hordijk, W., and Gascuel, O. (2005). Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* 21, 4338–4347. doi: 10.1093/bioinformatics/bti713

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Joshi, S. R., and Gurung, B. R. (2009). *Citrus in Bhutan: Value Chain Analysis. Department of Agricultural Marketing and Cooperatives*. Thimphu: Ministry of Agriculture and Forests, Royal Government of Bhutan.

Karasev, A. V., Boyko, V. P., Gowda, S., Nikolaeva, O. V., Hilf, M. E., Koonin, E. V., et al. (1995). Complete sequence of the *citrus tristeza* virus RNA genome. *Virology* 208, 511–520.

Kokane, A. D., Kokane, S. B., Warghane, A. J., Gubyad, M. G., Sharma, A. K., Reddy, M. K., et al. (2021a). A rapid and sensitive reverse transcription-loop-mediated isothermal amplification (RT-LAMP) assay for the detection of Indian citrus ringspot virus. *Plant Dis.* 105, 1346–1355. doi: 10.1094/PDIS-06-20-1349-RE

Kokane, A. D., Lawrence, K., Kokane, S. B., Gubyad, M. G., Misra, P., Reddy, M. K., et al. (2021b). Development of a SYBR Green-based RT-qPCR assay for the detection of Indian citrus ringspot virus. *3 Biotech* 11:359. doi: 10.1007/s13205-021-02903-8

Kokane, S. B., Misra, P., Kokane, A. D., Gubyad, M., Warghane, A. J., Surwase, D., et al. (2021c). Development of a real-time RT-PCR method for the detection of *Citrus tristeza* virus (CTV) and its implication in studying virus distribution in plant. *3 Biotech* 11:431. doi: 10.1007/s13205-021-02976-5

Kokane, A., Lawrence, K., Surwase, D., Misra, P., Warghane, A., and Ghosh, D. K. (2020). Development of reverse transcription duplex PCR (RT-d-PCR) for simultaneous detection of the *citrus tristeza* virus and Indian citrus ringspot virus. *Int. J. Innov. Hortic.* 9, 124–131.

Kokane, S. B., Kokane, A. D., Misra, P., Warghane, A. J., Kumar, P., Gubyad, M. G., et al. (2020). In-silico characterization and RNA-binding protein based polyclonal antibodies production for detection of *citrus tristeza* virus. *Mol. Cell Probes* 54:101654. doi: 10.1016/j.mcp.2020.101654

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127–128.

Licciardello, G., Scuderi, G., Ferraro, R., Giampetruzzi, A., Russo, M., Lombardo, A., et al. (2015). Deep sequencing and analysis of small RNAs in sweet orange grafted on sour orange infected with two *citrus tristeza* virus isolates prevalent in Sicily. *Arch. Virol.* 160, 2583–2589. doi: 10.1007/s00705-015-2516-x

Liu, Z., Chen, Z., Hong, J., Wang, X., Zhou, C., Zhou, X., et al. (2016). Monoclonal antibody-based serological methods for detecting *Citrus tristeza* virus in citrus groves. *Virol. Sin.* 31, 324–330. doi: 10.1007/s12250-016-3718-4

Lu, R., Folimonov, A., Shintaku, M., Li, W. X., Falk, B. W., Dawson, W. O., et al. (2004). Three distinct suppressors of RNA silencing encoded by a 20-kb viral RNA genome. *Proc. Natl. Acad. Sci. U.S.A,* 101, 15742–15747. doi: 10.1073/pnas.0404940101

Manjunath, K. L., Pappu, H. R., Lee, R. F., Niblett, C. L., and Civerolo, E. L. (1993). Studies on the coat protein genes of four isolates of *citrus tristeza closterovirus* from India: cloning, sequencing and expression. *Int. Organ. Citrus Virol. Conf. Proc.* 12, 1957–2010.

Marroquín, C., Olmos, A., Gorris, M. T., Bertolini, E., Martınez, M. C., Carbonell, E. A., et al. (2004). Estimation of the number of aphids carrying *citrus tristeza* virus that visit adult citrus trees. *Virus Res.* 100, 101–108. doi: 10.1016/j.virusres.2003.12.018

Martín, S., Sambade, A., Rubio, L., Vives, M. C., Moya, P., Guerri, J., et al. (2009). Contribution of recombination and selection to molecular evolution of *citrus tristeza* virus. *J. Gen. Virol.* 90, 1527–1538. doi: 10.1099/vir.0.008193-0

Matsumura, E. E., Coletta-Filho, H. D., Nouri, S., Falk, B. W., Nerva, L., Oliveira, T. S., et al. (2017). Deep sequencing analysis of RNA from citrus plants grown in a citrus sudden death-affected area reveals diverse known and putative novel viruses. *Viruses* 9:92. doi: 10.3390/v9040092

Meena, R. P., and Baranwal, V. K. (2016). Development of multiplex polymerase chain reaction assay for simultaneous detection of clostero-, badna-and mandari-viruses along with huanglongbing bacterium in citrus trees. *J. Virol. Methods* 235, 58–64. doi: 10.1016/j.jviromet.2016.05.012

Mehta, P., Brlansky, R. H., Gowda, S., and Yokomi, R. K. (1997). Reverse-transcription polymerase chain reaction detection of *Citrus tristeza* virus in aphids. *Plant Dis.* 81, 1066–1069. doi: 10.1094/PDIS.1997.81.9.1066

Melzer, M. J., Borth, W. B., Sether, D. M., Ferreira, S., Gonsalves, D., and Hu, J. S. (2010). Genetic diversity and evidence for recent modular recombination in Hawaii an *Citrus tristeza* virus. *Virus Genes* 40, 111–118. doi: 10.1007/s11262-009-0409-3

Moreno, P., Ambros, S., Albiach-Marti, M. R., Guerri, J., and Pena, L. (2008). *Citrus tristeza* virus: a pathogen that changed the course of the citrus industry. *Mol. Plant Pathol.* 9, 251–268. doi: 10.1111/j.1364-3703.2007.00455.x

Nicholas, K. B., Nicholas, Jr, H. B, and Deerfield, D. W. (1997). *Embnet News GeneDoc: Analysis and Visualization of Genetic Variation*, Vol. 4. Nijmegen: EMBnet Administration, 14.

Pappu, H., Pappu, S., Niblett, C., Lee, R., and Civerolo, E. (1993). Comparative sequence analysis of coat protein of biologically distinct *citrus tristeza* clostero virus isolate. *Virus Genes* 7, 255–264. doi: 10.1007/BF01702586

Roy, A., and Brlansky, R. H. (2010). Genome analysis of an orange stem pitting *citrus tristeza* virus isolate reveals a novel recombinant genotype. *Virus Res.* 151, 118–130. doi: 10.1016/j.virusres.2010.03.017

Roy, A., Choudhary, N., Hartung, J. S., and Brlansky, R. H. (2013). The prevalence of the *citrus tristeza* virus trifoliate resistance breaking genotype among puerto rican isolates. *Plant Dis*. 97, 1227–1234. doi: 10.1094/PDIS-01-12-0012-RE

Roy, A., Manjunath, K. L., and Brlansky, R. H. (2005a). Assessment of sequence diversity in the 5′-terminal region of *Citrus tristeza* virus from India. *Virus Res.* 11, 132–142. doi: 10.1016/j.virusres.2005.04.023

Roy, A., Fayad, A., Barthe, G., and Brlansky, R. H. (2005b). A multiplex polymerase chain reaction method for reliable, sensitive and simultaneous detection of multiple viruses in citrus trees. *J. Virol. Methods* 129, 47–55. doi: 10.1016/j.jviromet.2005.05.008

Rubio, L., Ayllón, M. A., Kong, P., Fernández, A., Polek, M., Guerri, J., et al. (2001). Genetic variation of *Citrus tristeza* virus isolates from California and Spain: evidence for mixed infections and recombination. *J. Virol.* 75, 8054–8062. doi: 10.1128/jvi.75.17.8054-8062.2001

Ruiz-Ruiz, S., Moreno, P., Guerri, J., and Ambros, S. (2006). The complete nucleotide sequence of a severe stem pitting isolate of *Citrus tristeza* virus from Spain: comparison with isolates from different origins. *Arch. Virol.* 151, 387–398. doi: 10.1007/s00705-005-0618-6

Ruiz-Ruiz, S., Moreno, P., Guerri, J., and Ambrós, S. (2007). A real-time RT-PCR assay for detection and absolute quantitation of *Citrus tristeza* virus in different plant tissues. *J. Virol. Methods* 145, 96–105. doi: 10.1016/j.jviromet.2007.05.011

Satyanarayana, T., Gowda, S., Mawassi, M., Albiach-Martí, M. R., Ayllón, M. A., Robertson, C., et al. (2000). Closterovirus encoded HSP70 homolog and p61 in addition to both coat proteins function in efficient virion assembly. *Virology* 278, 253–265. doi: 10.1006/viro.2000.0638

Simmons, M. P., and Norton, A. P. (2014). Divergent maximum-likelihood-branch-support values for polytomies. *Mol. Phylogenet Evol.* 73, 87–96. doi: 10.1016/j.ympev.2014.01.018

Tarafdar, A., Godra, S., Dwivedi, S., Jayakumar, B. K., and Biswas, K. K. (2013). Characterization of *Citrus tristeza* virus and determination of genetic variability in North-east and South India. *Indian Phytopathol.* 66, 302–307.

Tatineni, S., Robertson, C. J., Garnsey, S. M., Bar-Joseph, M., Gowda, S., and Dawson, W. O. (2008). Three genes of *Citrus tristeza* virus are dispensable for infection and movement throughout some varieties of citrus trees. *Virology* 376, 297–307. doi: 10.1016/j.virol.2007.12.038

Tatineni, S., Robertson, C. J., Garnsey, S. M., and Dawson, W. O. (2011). A plant virus evolved by acquiring multiple nonconserved genes to extend its host range. *Proc. Natl. Acad. Sci.* 108, 17366–17371.

Tipu, S. A., and Fantazy, K. A. (2014). Supply chain strategy, flexibility, and performance: a comparative study of SMEs in Pakistan and Canada. *Int J Logist Manag.* 25, 399–416.

Vives, M. C., Rubio, L., Sambade, A., Mirkov, T. E., Moreno, P., and Guerri, J. (2005). Evidence of multiple recombination events between two RNA sequence variants within a *Citrus tristeza* virus isolate. *Virology* 331, 232–237. doi: 10.1016/j.virol.2004.10.037

Wang, J., Zhou, T., Cao, M., Zhou, Y., and Li, Z. (2019). First report of *citrus tristeza* virus trifoliate resistance-breaking (RB) genotype in *Citrus grandis* in China. *J. Plant Pathol.* 101:451.

Warghane, A., Kokane, A., Kokane, S., Motghare, M., Surwase, D., Palchoudhury, S., et al. (2020). Molecular detection and coat protein gene based characterization of *citrus tristeza* virus prevalent in Sikkim state of India. *Indian Phytopathol*. 73, 135–143.

Warghane, A., Misra, P., Bhose, S., Biswas, K. K., Sharma, A. K., Reddy, M. K., et al. (2017a). Development of reverse transcription-loop mediated isothermal amplification (RT-LAMP) assay for rapid detection of *Citrus tristeza* virus. *J. Virol. Methods* 250, 6–10.

Warghane, A., Misra, P., Ghosh, D. K., Shukla, P. K., and Ghosh, D. K. (2017b). Diversity and characterization of *citrus tristeza* virus and '*Candidatus* Liberibacter asiaticus' associated with citrus decline in India. *Indian Phytopathol* 70, 359–367.

Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C., et al. (2018). Genomics of the origin and evolution of Citrus. *Nature* 554, 311–316. doi: 10.1038/nature25447

Yang, Z. N., Mathews, D. M., Dodds, J. A., and Mirkov, T. E. (1999). Molecular characterization of an isolate of *citrus tristeza* virus that causes severe symptoms in sweet orange. *Virus Genes* 19, 131–142. doi: 10.1023/a:1008127224147

Yokomi, R. K., Selvaraj, V., Maheshwari, Y., Saponari, M., Giampetruzzi, A., Chiumenti, M., et al. (2017). Identification and characterization of *citrus tristeza* virus isolates breaking resistance in trifoliate orange in California. *Phytopathol* 107, 901–908. doi: 10.1094/PHYTO-01-17-0007-R

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 14 April 2022
doi: 10.3389/fmicb.2022.811318

Updating the Phylodynamics of Yellow Fever Virus 2016–2019 Brazilian Outbreak With New 2018 and 2019 São Paulo Genomes

Ana Paula Moreira Salles[1,2], Ana Catharina de Seixas Santos Nastri[3], Yeh-Li Ho[3], Luciana Vilas Boas Casadio[3], Deyvid Emanuel Amgarten[2,4], Santiago Justo Arévalo[2,4,5], Michele Soares Gomes-Gouvea[1], Flair Jose Carrilho[1], Fernanda de Mello Malta[1,2] and João Renato Rebello Pinho[1,2,6]*

[1] Department of Gastroenterology (LIM07), Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil, [2] Clinical Laboratory of Hospital Israelita Albert Einstein, São Paulo, Brazil, [3] Department of Infectious and Parasitic Diseases, Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil, [4] Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil, [5] Facultad de Ciencias Biológicas, Universidad Ricardo Palma, Lima, Peru, [6] Division of Clinical Laboratories (LIM 03), Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil

OPEN ACCESS

Edited by:
Pingfeng Yu,
Zhejiang University, China

Reviewed by:
Filipe Vieira Santos De Abreu,
Instituto Federal de Educação, Ciência e Tecnologia do Norte de Minas Gerais (IFNMG), Brazil
Diogo Antonio Tschoeke,
Federal University of Rio de Janeiro, Brazil

****Correspondence:***
Fernanda de Mello Malta
femalta@yahoo.com

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

***Received:** 08 November 2021*
***Accepted:** 16 February 2022*
***Published:** 14 April 2022*

Citation:
Moreira Salles AP, de Seixas Santos Nastri AC, Ho Y-L, Vilas Boas Casadio L, Emanuel Amgarten D, Justo Arévalo S, Soares Gomes-Gouvea M, Jose Carrilho F, de Mello Malta F and Rebello Pinho JR (2022) Updating the Phylodynamics of Yellow Fever Virus 2016–2019 Brazilian Outbreak With New 2018 and 2019 São Paulo Genomes.
Front. Microbiol. 13:811318.
doi: 10.3389/fmicb.2022.811318

The recent outbreak of yellow fever (YF) in São Paulo during 2016–2019 has been one of the most severe in the last decades, spreading to areas with low vaccine coverage. The aim of this study was to assess the genetic diversity of the yellow fever virus (YFV) from São Paulo 2016–2019 outbreak, integrating the available genomic data with new genomes from patients from the Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP). Using phylodynamics, we proposed the existence of new IE subclades, described their sequence signatures, and determined their locations and time of origin. Plasma or urine samples from acute severe YF cases (*n* = 56) with polymerase chain reaction (PCR) positive to YFV were submitted to viral genome amplification using 12 sets of primers. Thirty-nine amplified genomes were subsequently sequenced using next-generation sequencing (NGS). These 39 sequences, together with all the complete genomes publicly available, were aligned and used to determine nucleotide/amino acids substitutions and perform phylogenetic and phylodynamic analysis. All YFV genomes generated in this study belonged to the genotype South American I subgroup E. Twenty-one non-synonymous substitutions were identified among the new generated genomes. We analyzed two major clades of the genotypes IE, IE1, and IE2 and proposed the existence of subclades based on their sequence signatures. Also, we described the location and time of origin of these subclades. Overall, our findings provide an overview of YFV genomic characterization and phylodynamics of the 2016–2019 outbreak contributing to future virological and epidemiological studies.

Keywords: yellow fever virus, next generation sequencing, outbreak, São Paulo, vaccine coverage

Abbreviations: YF, yellow fever; YFV, yellow fever virus; PCR, polymerase chain reaction; NGS, next-generation sequencing; NHPs, non-human primates; UTR, untranslated region; ORF, open reading frame; IGV, Integrative Genomics Viewer; SNP, single nucleotide polymorphism; NTPase, nucleoside triphosphatase.

INTRODUCTION

Yellow fever (YF) is a tropical short-term disease transmitted by the bite of infected female mosquitoes. It has a large spectrum of symptoms, from an asymptomatic form to a severe and deadly hemorrhagic fever in humans and non-human primates (NHPs) (Beasley et al., 2015; Tabachnick, 2016). It is estimated to cause approximately 30,000 deaths out of 200,000 infections worldwide, mostly in Africa (Monath and Vasconcelos, 2015).

Yellow fever virus (YFV) belonging to the *Flaviviridae* family is the etiological agent of the YF. It is a positive-sense single-stranded RNA virus with a genome of approximately 11 kb that consists of a 5' untranslated region (UTR), followed by a single open reading frame (ORF), and a 3'UTR (Chambers et al., 1990; Gómez et al., 2018). The ORF is further divided into three structural proteins (C, prM/M, and E) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5).

Yellow fever is found mostly in Africa and the Americas with recurrent epidemics from the seventeenth century until the beginning of the twentieth century in Europe and North America. Currently, almost all the infections are from the endemic areas of Africa, and Central and South America.

Seven lineages of YFV have been identified so far: five in Africa (West Africa I and II, East Africa, East/Central Africa, and Angola) with an estimated genetic variance at the nucleotide level ranging between 10 and 23% (Mutebi et al., 2001; von Lindern et al., 2006) and two in Central and South America (South America I and II) with an estimated genetic diversity at the nucleotide level of 7% (Mutebi et al., 2001; Mir et al., 2017).

The South American I is the most prevalent genotype in Brazil, and it is divided into subgroups IA to IE (de Souza et al., 2010; Nunes et al., 2012). Only subgroups ID and IE have been detected circulating in Brazil, but since 2008, only subgroup IE has been detected. Recent studies speculated that the YFV strain, associated with the recent outbreaks (genotype IE), would have originated in the central–west region and then probably reached the southeast region (Cunha et al., 2019a; Delatorre et al., 2019).

YFV has been sporadically detected in non-human primates (NHPs) and human populations from enzootic and endemic areas from northern and central–western regions of Brazil before the twentieth century. However, during the last two decades, YFV has spread to the southeast region, reaching the Atlantic rainforest. After the end of 2016, an important increase in human cases have been reported in the south and southeast of Brazil (Silva et al., 2020).

According to the last Brazilian epidemiological report from the Ministry of Health, the metropolitan area of São Paulo City had 538 confirmed human cases and 184 deaths (34.2%) in 2018 and 66 confirmed human cases and 12 deaths (18.2%) during 2019 (Governo do Estado de São Paulo, 2019).

The aim of this study was to assess the genetic diversity and phylodynamics of YFV from the 2016–2019 outbreak, integrating the available genomic data with new genomes from patients from the Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP). Using phylodynamics, we proposed the existence of new subclades, described their sequence signatures, and determined their locations and time of origin. This study may help in the surveillance, epidemiology studies, and to increase our understanding of the genetic diversity and spread of YFV.

MATERIALS AND METHODS

Patients and Samples

Among 192 suspected YF cases followed at HCFMUSP during the 2018 and 2019 outbreaks, 56 patients that had their YF infection confirmed by qPCR were enrolled in this study. Blood and urine samples were collected at hospital admission. The methodology applied for detection and quantification of YFV-RNA in serum and urine samples was the same as that described by Casadio et al. (2019), noticing that all of them were tested for both wild and vaccine YFV strains, using specific primers and probes for each one of them. After that, we chose the earliest positive available sample of each patient to perform the viral genome sequencing.

We also performed a geopositioning analysis using the previously collected information and available data on patient residence and year of infection and mapped them using Google Maps® (Google, 2021) tools available at the Google platform.

This study was conducted in compliance with the institutional guidelines, approved by the Ethical Committee from the Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (CEP/HCFMUSP; CAAE: 74535417.3,1001.0068), and all individuals signed written informed consent forms.

Amplification and Sequencing of Yellow Fever Virus Genome

Viral RNA was isolated from 140 μl of serum or urine using QIAamp® Viral RNA Mini Kit (Qiagen^{TM}, Hilden, Germany), according to the manufacturer protocol. After extraction, the RNA was reverse transcribed to cDNA using random primers and M-MLV Reverse Transcriptase 200 U/μl (Invitrogen^{TM}, Thermo Fisher Scientific Brand, Carlsbad, CA, United States), according to the manufacturer's instructions. The cDNA was amplified by PCR using 12 different pairs of primers generating 11 overlapping PCR fragments covering the YFV genome (each fragment was amplified separately, 11 PCR reactions by sample; **Supplementary Table 1**).

The singleplex PCR reactions contained 35 μl of RNase-free H_2O, 5 μl 10 × buffer, 1 μl of dNTP mix (10 nM), 1.5 μl of $MgCl_2$ (50 nM), 1 μl of a set of primers (20 nM), 5 U of Platinum Taq DNA Polymerase (Invitrogen^{TM}, Thermo Fisher Scientific Brand, Carlsbad, CA, United States), and 5 μl of cDNA. The cycling protocol was initial denaturation at 94°C for 5 min, then 45 cycles of denaturation at 94°C for 30 s, annealing at 65°C for 30 s, and extension at 72°C for 90 s, followed by 72°C for 10 min, and 10°C up to the next step.

Among the 56 samples, we were able to amplify all overlapping PCR fragments from 40 YFV-positive samples (mean of Ct value = 26.7), at least one PCR fragment from the other 12 samples (mean of Ct value = 29.2), and we were not able to amplify any fragment in 4 samples (mean of Ct value = 30.8). The most difficult YFV genome region to amplify was the 3'UTR. As an alternative, we used another pair of primers (F11D) that does not cover all the genomes (approximately 10.336 bp).

PCR products were quantified using the fluorimetric method (Qubit® 4 Fluorometer^{TM}; Thermo Fisher Scientific, Waltham, MA, United States), and the DNA concentration from each amplicon was adjusted before amplicon pooling (each sample has one pool with all 11 fragments). The DNA concentration of the amplicon pool was adjusted to 0.8 ng/μl to perform the Nextera® XT DNA Sample Library Preparation protocol (Illumina, Inc., San Diego, CA, United States).

The library was purified using AMPure XP® beads (Beckman Coulter^{TM}; Life Sciences Division Headquarters; Indianapolis, IN, United States) once quantified and diluted to 2 nM, and denatured according to the manufacturer's protocol (Preparing DNA Libraries for Sequencing, Miseq Guide). Denatured libraries were loaded in MiSeq Reagent Cartridge v2 (300-cycle) and paired-end sequenced on the MiSeq platform (Illumina, Inc., San Diego, CA, United States).

Nearly complete virus genomes from 40 samples were sequenced with a mean average coverage depth of 5.149× and a breadth coverage ranging from 92.75 to 100% (**Supplementary Table 2**). Only one sample did not reach quality metrics, and therefore. it was excluded. Phylogenetic analysis was performed using the 39 samples with good quality metrics (average coverage, Q30, and number of reads).

Sequence Analysis of Yellow Fever Virus Genome

All sequences were trimmed and filtered. Short unpaired reads and low-quality bases and reads were removed using Cutadapt 2.10 (Martin, 2011). Human genome reference was downloaded (GCF_000001405.12), and all unmapped reads were filtered using Samtools (Li et al., 2009), then all FASTQ data were extracted.

FASTQ data were analyzed using SPAdes (Nurk et al., 2013) (trusted contigs using MF538786.2/RJ104 as the sequence reference and *de novo* assembly) and IVA software (Hunt et al., 2015) combining the results to create the most reliable consensus sequence for all samples (Pipeline available at: https://github.com/deyvidamgarten/YFV/wiki/Montagem_genoma_YFV).

The sequence files were downloaded from BaseSpace and checked for quality score (Q30) and trimmed using Cutadapt (Martin, 2011). The next step was mapping and indexing the sequences using BWA (Li and Durbin, 2009) to align our YFV sequences back to the reference. Samtools view (Li et al., 2009) was used to remove the reads with secondary alignment or with low quality of mapping (<30) and/or no mapping at all when compared with the reference in the BAM archive that we had previously generated. Then the sequences were sorted and indexed generating the clean and sorted BAM archive that can be visualized using Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013).

Sequence variations in the library were detected using single-nucleotide polymorphism (SNP) and short indel detection function using Freebayes (Garrison and Marth, 2012) and GATK haplotype caller (McKenna et al., 2010) software for each sample, generating a VCF file that then was merged for analysis using BCF tools (Li, 2011).

Phylogenetic Analysis of Yellow Fever Virus Genomes

Phylogenetic analysis was performed using all the YFV sequences available as complete genomes in NCBI plus those described in Nunes et al. (2012), Gómez et al. (2018), Abreu et al. (2019), Cunha et al. (2019a,b), Delatorre et al. (2019), and Hill et al. (2020). All those complete genomes were included to reconstruct a first phylogeny (complete list of NCBI IDs of YFV genomes used herein are available in the **Supplementary Material**). A total of 314 genomes plus 39 genomes generated in this study were first analyzed. Sequences with less than 7,000 bp and with more than 1% Ns were removed. The 342 genomes that remained were aligned using MAFFT (Katoh and Standley, 2013), and the region corresponding to positions 143 to 10,309 with respect to NC_002031.1 was used for maximum-likelihood analysis. IQ-TREE2 (Minh et al., 2020) was used for phylogenetic inference. ModelFinder (Kalyaanamoorthy et al., 2017) was used to select the substitution model GTR+F+I+G4 according to BIC. A total of 1,000 replicates of UF-Boot (Hoang et al., 2018) and SH-aLRT (Guindon et al., 2010) was also used to measure consistency and support of nodes. The consensus tree generated in the previous inference (**Supplementary Figure 1**) was used to determine the available sequences most related to the sequences generated in this study. Thus, the two largest clades near the clade containing the sequences generated in this study (clades A, B, and C in **Supplementary Figure 1**) were selected for further analysis.

From the 237 sequences belonging to the previously mentioned clades, we removed seven sequences from the Netherlands that were isolated from travelers from Brazil (MK760660, MK760661, MK760662, MK760663, MK760664, MK760665, and MK760666) and other two sequences from Brazil without a collection place information (MF465805 and MH560359). The region corresponding to positions 143 to 10,309 with respect to NC_002031.1 of the 228 remaining sequences were aligned using MAFFT (Katoh and Standley, 2013). The alignment was used for a new maximum-likelihood inference. Again, IQ-TREE2 (Minh et al., 2020) was used for phylogenetic inference. ModelFinder (Kalyaanamoorthy et al., 2017) was used to select the substitution model TIM2+F+I+G4 according to BIC. A total of 1,000 replicates of UF-Boot (Hoang et al., 2018) and SH-aLRT (Guindon et al., 2010) was also used to measure consistency and support of nodes. The NCBI codes and all the available metadata used in this analysis are available in **Supplementary Table 3**.

Phylodynamic Analysis of Genomes

The rate of nucleotide substitution, the time to the most recent common ancestors, and the ancestral state reconstruction were estimated using the Markov chain Monte Carlo (MCMC) algorithms implemented in BEAST2 (Bouckaert et al., 2019) with BEAGLE library (Suchard and Rambaut, 2009) to speed up the run time. The same alignment of the 228 sequences previously described was used for this analysis. The evolutionary process was estimated from the sampling year of the sequences (considering the mid of the year of collection) using the GTR substitution model, a strict molecular clock model (Ferreira and Suchard, 2008) or an uncorrelated lognormal molecular clock model (Drummond et al., 2006), and a Bayesian Skyline coalescent tree prior (Drummond et al., 2005). Comparisons among the two clock models were performed using nested sampling with four independent runs for each model (Russel et al., 2019). Migration events throughout the phylogeny were reconstructed using a reversible discrete phylogeographic model (Lemey et al., 2009). A discrete state was assigned for each sequence corresponding to the state (Brazilian sequences) of infection (sequences from our study) or the state reported by the authors (sequences not generated in this study). For the sequences generated in this study, the city/state of YFV infection was collected and mapped using the geolocation tool available at Google Maps® (**Supplementary Figure 2**) (Google, 2021). MCMC was run sufficiently long to ensure stationarity and convergence. Uncertainty of parameter estimates were assessed after excluding the initial 10% of the run by calculating the effective sample size (ESS) and the 95% highest probability density (HPD) values using TRACER (Rambaut et al., 2018). Tree annotator (Drummond et al., 2012) was used to summarize the posterior tree distribution, and the R package GGTREE (Yu, 2020) was used to visualize and generate the final tree figures.

Analysis of Synonymous and Non-synonymous Substitutions

For the analysis of synonymous and non-synonymous substitutions, consensus nucleotide sequences were aligned using CLUSTAL W (Larkin et al., 2007). To analyze the presence of these substitutions, the alignment was translated using MEGA7 program (Kumar et al., 2016), Freebayes, and Haplotype Caller software (McKenna et al., 2010; Garrison and Marth, 2012).

TABLE 1 | Demographic data of enrolled samples.

Variable	2018 (*n* = 24)	2019 (*n* = 32)	*p*-Value
Sex			
Male, *n* (%)	19 (79.2)	29 (90.6)	0.268^F
Female, *n* (%)	5 (20.8)	3 (9.4)	
Age (years)			
Mean (min–max)	43.7 (19–74)	45.7 (19–88)	0.596^U
Days after onset***			
Mean (±sd)	5.6 (±2.2)	5.2 (±2.2)	0.536^U
Viral load (log10)			
Mean (±sd)	6.3 (±1.32)	6.4 (±1.4)	0.308^U
Ct			
Mean (±sd)	27.7 (±4.9)	27.4 (±3.9)	0.328^U

*N, number of samples; min, minimum; max, maximum; sd, standard deviation; Ct, cycle threshold; ^{F}Fisher exact test, ^{U}Mann–Whitney test, $p < 0.005$; ***before first day of onset.*

RESULTS

No Demographic Differences Were Found Between Patients of Different Years

Demographic data analysis from 56 patients enrolled in this study indicates that 85.7% are male aged between 19 and 88 years, old and all of them were RT-qPCR positive for YFV besides its viral load quantified using a standard curve (Casadio et al., 2019). In addition, no significant statistical differences were found after patients were divided into groups considering their respective year of infection (**Table 1**).

Phylogenetic Analysis of Yellow Fever Virus Complete Genomes

The maximum-likelihood tree obtained from 228 YFV complete genomes showed representatives belonging to five genotypes: South American II, South American IB, IC, ID, and IE (**Supplementary Figure 3**). The monophyletic clade that contains the representatives of ID and IE is well supported (SH-aLRT = 98) (**Supplementary Figure 3**). A closer look on this monophyletic clade showed well-supported ID and IE sister clades (SH-aLRT = 100 for both ID and IE clades) (**Supplementary Figure 4**), with all the sequences obtained in this study grouped inside genotype IE (**Supplementary Figure 4**).

Going deeper in our analysis, we further explore the monophyletic clade IE. This clade can be subdivided into three groups: (i) basal IE, a paraphyletic group with the oldest IE sequences (one sequence from 2002 and one from 2008), (ii) IE1, and (iii) IE2 (all the sequences obtained in the present study belongs to the subclade IE2) (**Supplementary Figure 5**).

To determine differences between major clades IE1 and IE2, we integrated to the phylogenetic tree the year and state of collection, and an alignment with the genomic positions that allow us to distinguish between clade IE1 and IE2 (**Figure 1**). Most of the IE1 sequences were isolated from Espirito Santo (ES), Minas Gerais (MG), or Rio de Janeiro (RJ), just one sequence from this clade was isolated from São Paulo (SP) (**Figure 1**). On the other hand, almost all IE2 sequences come from SP, and only three sequences were isolated from other states [two from Goias (GO) and one from MG] (**Figure 1**) (the two sequences from GO in the IE2 clade have an uncertain position, see below).

Although the geographical structure of IE1 and IE2 major clades can be well differentiated, we cannot observe a clear temporal distribution (**Figure 1**). Both clades contain sequences spanning from 2015 to 2019. Most of the sequences isolated in 2019 belong to the subclade IE2; however, just one 2019 sequence not from this study is available. Because all the sequences from this study were isolated in SP, we cannot see the distribution of genotypes in other states in 2019.

Three genomic positions unambiguously allow us to distinguish between subclades IE1 and IE2 (3,097, 5,533, and 7,240 with respect to the sequence NC_002031.1). The other four positions (3,502, 4,948, 5,452, and 6,199) are distinguishable except for one sequence from one of the subclades (**Figure 1**). Additionally, the other 11 positions provide differences between

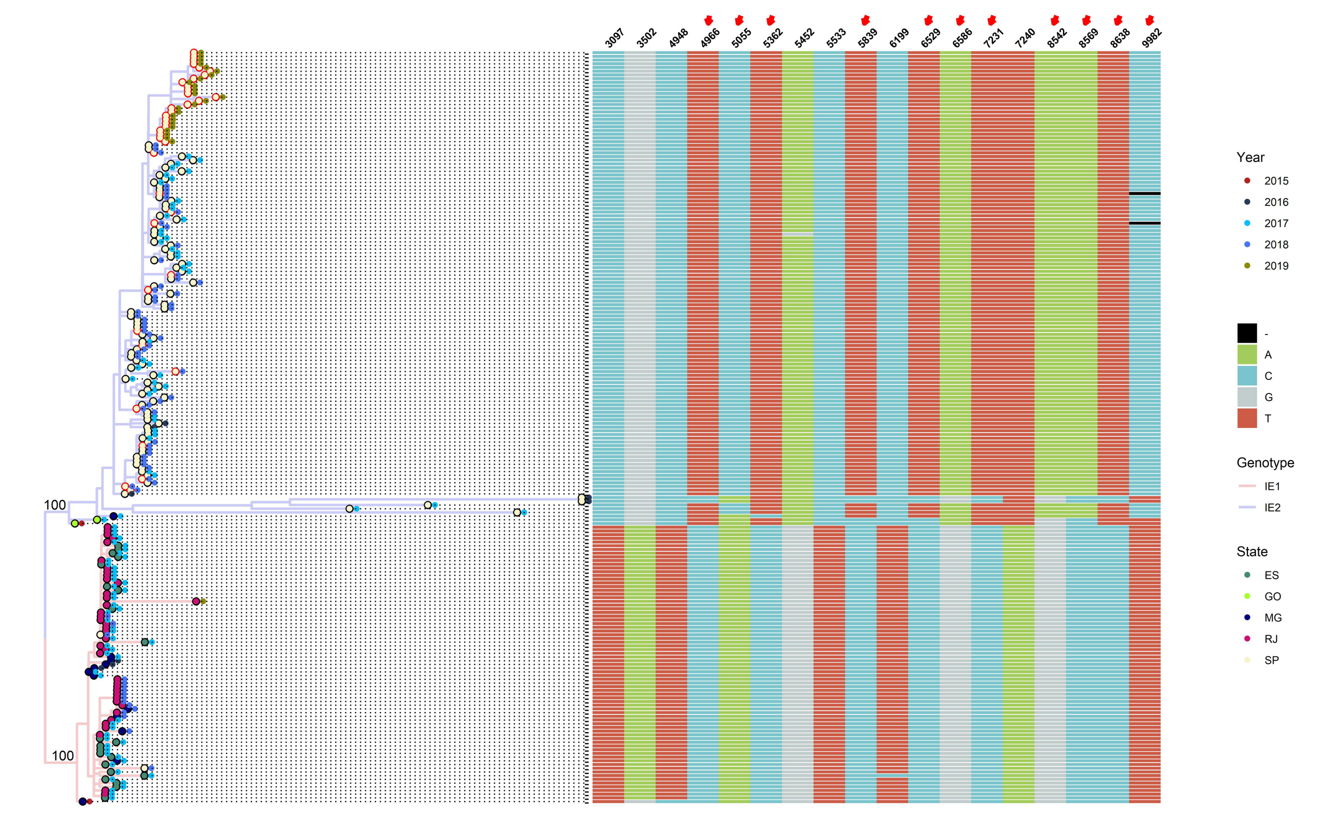


FIGURE 1 | Genotype IE can be subdivided into two major subclades: IE1 and IE2. Maximum-likelihood tree of sequences belonging to genotype IE, showing subdivision in two well-supported clades (SH-aLRT = 100, label of parent nodes), IE1 (pink branches), and IE2 (blue branches). External nodes are colored according to state of collection (internal circle) and year of collection (external circle). Red border in the first circle indicates that the sequence was generated in this study. At the right of the phylogenetic tree, an alignment of relevant positions to discriminate between genotypes is shown.

subclades IE1 and IE2 but are intermingled in the basal sequences of clade IE2 (**Figure 1**).

Further subclassification of clade IE1 into IE1_basal, IE1_1, IE1_trans, and IE1_2 based on specific genomic positions are available in **Supplementary Figure 6**. Subclassification of subclade IE2 into IE2_Basal, IE2_1, IE2_2, IE2_3, and IE2_4 also based on specific genomic position is available in **Supplementary Figure 7**.

Phylodynamic Analysis of Major Clades IE1 and IE2

To gain insights in the phylodynamics of the major clades (IE1 and IE2), we performed Bayesian inferences to estimate the time of MRCA and the most probable state of divergence of clades IE1 and IE2. An alignment of 228 complete genomes (see Materials and Methods section) was used to perform this inference. With a strict clock model, the 95% HPD interval estimated for the substitution rate was 2.63E-4–3.42E-4. On the other hand, an uncorrelated lognormal relaxed clock estimated the 95% HPD interval of the mean between 4.21E-4 and 7.90E-4, and the variance between 1.51E-7 and 1.83E-6. To determine which of these models is better adjusted with the data, we used nested sampling (Russel et al., 2019) to estimate the log Bayes factor. Log Bayes factor was 115.31 in favor of the uncorrelated lognormal relaxed clock.

Based on the inference with the uncorrelated lognormal relaxed clock, the divergence time of South America I and South America II genotypes has high margin of uncertainty (1801–1955) (**Supplementary Figure 8**) with dates in concordance with other studies (Bryant et al., 2007; Auguste et al., 2010; de Souza et al., 2010). This analysis also allows us to estimate the divergence time of genotypes IC (1940–1971), IB (1957–1979), and ID from IE (1978–1992) (**Supplementary Figure 8**). It was not possible to determine the Brazilian states where those divergences took place (several states appeared with similar probabilities) (**Supplementary Figure 8**).

Our phylodynamic analysis showed that the major clades IE1 and IE2 diverge between mid-2011 and the last months of 2014 (**Figure 2**). Again, the state where this divergence took place could not be estimated with certainty (**Figure 2**). Additionally, this analysis permits us to determine that the MRCA for clade IE1 exists between the last months of 2014 and mid-2015 in MG (**Figure 2**). The MRCA of the subclade IE1_1 was estimated to exist between mid-2016 and the first months of 2017 in ES (**Figure 2**). In contrast, the subclade IE1_2 have its origin in MG from where it was introduced to ES and RJ between the last months of 2016 and the first months of 2017 (**Figure 2**).

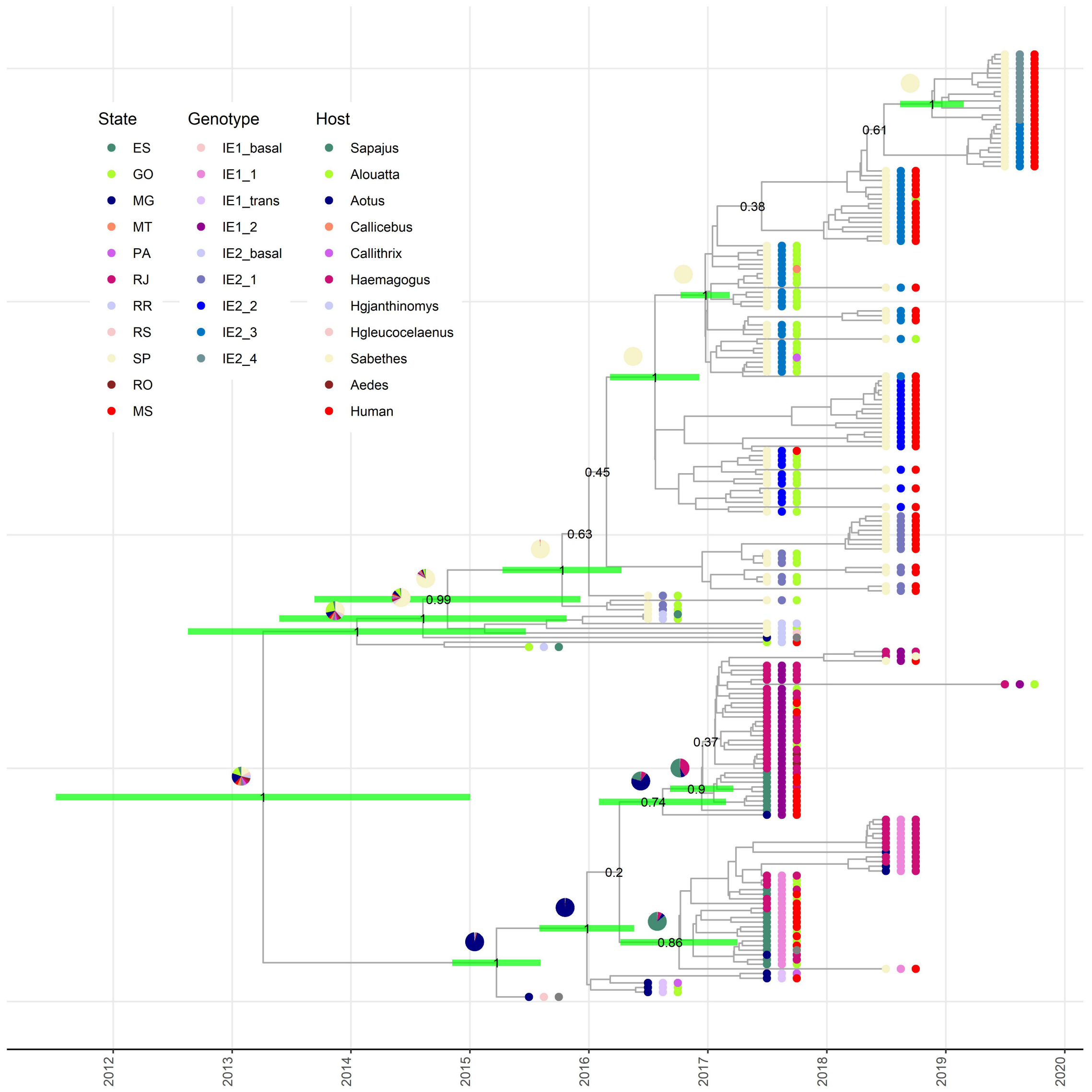


FIGURE 2 | Origin of subclades IE1 and IE2 are revealed by phylodynamic analysis. Time-scaled Bayesian maximum clade credibility tree showing the nodes of divergence between subclades IE1 and IE2 in their respective subdivisions. Green bars in the selected internal nodes show 95% HPD intervals of divergence times. Pie graphics on the internal nodes represent the probability of the state where this node existed. Numbers in the selected internal nodes represent the posterior value. External node points are colored according the state of collection (internal circle), subgenotype (middle circle), and host (external circle).

The MRCA of the major clade IE2 existed between the last months of 2012 and mid-2015, but the state of origin is uncertain, with highest probabilities for SP and GO (**Figure 2**). After its appearance, it is clear that it continues diverging in subclades in SP (**Figure 2**). The four IE2 subclades described above also appear well supported in our phylodynamic analysis. Thus, our analysis estimated that all these subclades have their origin in SP. The subclade IE2_1 diverged between the first months of 2015 and the first months of 2016, IE2_2 during 2016, IE2_3 between the last months of 2016 and the first months of 2017, and the most recent subclade, IE2_4, between the last months of 2018 and the beginning of 2019.

These results support the hypothesis that a single lineage introduced to SP gave rise to the establishment of the IE2 clade in SP (**Figure 2**). On the other hand, just one sequence of IE1 subclade has been described in SP in 2018, and this observation confirms an independent introduction of subclade IE1 to SP, but apparently IE1 clade has not dispersed in SP as effectively as IE2. At the moment, any of the IE2 subclades has been found in a state different from SP.

All the IE1 and IE2 subclades (except IE2_4) described here have at least one representant isolated from NHPs or mosquitoes. Thus, these subclades could arise in humans, NHPs, or mosquitoes. Anyway, this is a clear evidence of frequent interchange between human and NHPs YFVs.

Synonymous and Non-synonymous Substitutions Analysis

From the substitution analysis, we found 46 nucleotide substitutions, including 20 non-synonymous substitutions in the amino acid level: three in the capsid protein (K26R, I43V, and K60R), four in the envelope (H301Y, A341V, N555D, and D597G), three in the NS1 (Y953H, L994P, and G1067R), one in the NS2A (A1209V), three in the NS3 (N1646T, T1826M, and P1953H), two in the NS4A (V2136G and L2137P), and four in the NS5 (R2535W, M2620V, A3149V, and T3229I). Only N1646T in the NS3 are present in all the 39 samples, as an SNP signature for these outbreaks. Furthermore, the substitution T3329I in the NS5 is present in almost all sequences (30/39), and K26R in capsid protein is present only in 15/29 sequences, all of them from the 2019 strain (**Supplementary Table 4**).

DISCUSSION

In this study, we analyzed 56 yellow fever virus patients and generated 39 new YFV nearly complete genomic sequences from samples from humans, collected in HCFMUSP during the 2018 and 2019 outbreaks. First, we conducted a demographic data analysis indicating a homogeneity between the 2018 and 2019 sample groups. In accordance with the disease monitoring carried out by Brazil's Ministry of Health, we report a higher percentage of YF in men (82.1%) since it is considered a reflection of the work activities performed by them in or near forest areas (Vasconcelos, 2003).

Phylogenetic analysis was made corroborating the fact that all 39 sequences belong to the South American IE. In order to determine the existence of possible subclades, we analyzed a phylogenetic inference of 228 YFV complete genomes. Similar to that describe by Delatorre et al. (2019), we proposed that the IE subclade can be further divided into two major clades: IE1 and IE2 [named YFV$_{MG/ES/RJ}$ and YFV$_{MG/SP}$, respectively, by Delatorre et al. (2019)]. We renamed those clades as IE1 and IE2 to allow easy naming of new subclades as IE1_1, IE1_2, IE2_1, IE2_2, IE2_3, and IE2_4 described here. Based on this analysis, we classified the genomes generated in this study as belonging to IE2 with representants of all four subclades (**Supplementary Table 3**).

Phylodynamic analyses showed a strong geographical structure of the major clades IE1 and IE2. However, our analysis was not able to determine the state of divergence of these major clades. Delatorre et al. (2019) mentioned GO as the most likely state (0.57 probability) where this divergence took place. However, no sequence of 2016 from SP was included in that study. The inclusion of several sequences from SP in our study increases the uncertainty of the state of divergence of the major clades IE1 and IE2 (**Figure 2**). On the other hand, our estimations of the date of divergence of these clades are similar to those described by Delatorre et al. (2019) (2011–2015) and inside the 95% HPD interval (2014–2016) mentioned by Rezende et al. (2018) inferred from 1,038-nt sequences.

Our deeper analysis of clade IE1 showed that it originated in MG from where it was introduced to ES to form the subclade IE1_1. From here, subclade IE1_1 was dispersed to RJ and returned to MG (**Figure 2**). These results are in accordance with one of the subclades of IE1 (YFV$_{MG/ES/RJ}$) as shown by Delatorre et al. (2019). In the case of subclade IE1_2, Delatorre et al. (2019) indicated its origin in ES from where it was introduced to RJ. However, Delatorre et al. (2019) did not include the genome from MG with NCBI code MF370533 that in our analysis appeared basal to subclade IE1_2 (**Figure 2**). The inclusion of this genome modifies the origin of this subclade and established the origin in MG from where it moves to ES and RJ (**Figure 2**). Dates of the MRCA of the new proposed subclades IE1_1 and IE1_2 match with the report of Delatorre et al. (2019) (**Figure 2**).

Cunha et al. (2019a) hypothesized that the major clade IE2 (or YFV$_{MG/SP}$) originated in MG. However, their inference was done without any genome of neither 2016–2017 SP nor GO. In contrast, the study of Hill et al. (2020) that included earlier SP genomes and GO genomes were not able to accurately determine the state of the MRCA of IE2. Our analysis, which also includes GO and early SP genomes, cannot accurately determine the state of origin of IE2, but showed GO and SP as the most likely states with similar probabilities (**Figure 2**). Importantly, we were able to accurately determine the state of origin of the new proposed subclades IE2_1, IE2_2, IE2_3, and IE2_4 and their respective times of divergence (**Figure 2**).

Subclades IE2_3 and IE2_4 were only observed during 2019. If this is a sample bias or if this is the process of lineage replacement in SP is an open question that has to be answered in the next studies.

We analyzed the 39 YFV genomes generated in this study, searching for synonymous and nonsynonymous mutations among them. Nine of the non-synonymous substitutions involve changes in the amino acid functional classes. Interestingly, two of these nine amino acid changes are located in two important proteins of the viral replicase complex: NS3 protein with RNA helicase, serine protease (Chen et al., 2017), and nucleoside triphosphatase (NTPase) (Brand et al., 2017) domains and NS5 protein, the largest and highly conserved protein in flavivirus considered a key for viral replication (Baleotti et al., 2003). Amino acid changes in these conserved proteins may have an impact on viral infectivity, both in humans and NHPs, as well as in mosquitoes (Gómez et al., 2018).

Comparing the 2018 and 2019 sequences, it was possible to observe a non-synonymous mutation at nucleotide 195 (K26R) in the capsid protein, which occurs in 15/25 samples in 2019, but found in just one YFV genome from 2018 (not from this study). This mutation, together with mutations in positions 2,545, 2,623, and 9,406 are the fingerprint of IE2_4. Studies associate the capsid protein with the packaging of the viral genome and the formation of the nucleocapsid (Patkar et al., 2007).

Aiming to explore more about the genomes that were sequenced in this study, we analyzed if patients who were also attended in HCFMUSP and evolved to death had similar variants when compared with those who survived. For this goal, we explored the studies made by Cunha et al. (2019a) that sequenced 36 YFV whole genomes from patients who evolved to death. Phylogenetic analysis did not find specific clades with higher percentages of patients that evolved to death.

These findings reinforce the idea that continued genomic surveillance strategies are needed to assist in the monitoring and understanding of YFV epidemics aiming to help public health actions and the management of infections. As shown here, inclusion of new genomes can update our hypotheses and confirm those of others helping to better understand the epidemiology of YFV.

Monitoring of the new subclades described here could help in determining interconnections between southern states. It is intriguing why clades IE1 and IE2 have a strong geographical structure despite the high human transit between the southern states, especially RJ and SP. YFV strains from NHPs from Brazil's southeast are necessary to determine if the subclade IE2_4 here described has been maintained in cycles in humans and/or NHPs.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI GenBank MZ604838–MZ604876.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by this study was conducted in compliance with institutional guidelines, approved by the ethical committee from the Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (CEP/HCFMUSP; CAAE: 74535417.3,1001.0068) and all individuals signed written informed consent forms. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AM, FM, and DE conceived and designed the experiments. FM and MS performed the experiments. Y-LH and LV were essential to the data collection. SJ performed the phylogenetic inferences. DE, FM, and SJ analyzed the data. FM and SJ drafted the work. DE, JR, FJ, SJ, and Y-LH revised the manuscript critically for important intellectual content. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank the HCFMUSP Infectious Diseases Department healthcare workers for their dedication in assisting patients.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.811318/full#supplementary-material

REFERENCES

Abreu, F. V. S., Ribeiro, I. P., Ferreira-de-Brito, A., dos Santos, A. A. C., de Miranda, R. M., Bonelly, I. S., et al. (2019). *Haemagogus leucocelaenus* and *Haemagogus janthinomys* are the primary vectors in the major yellow fever outbreak in Brazil, 2016–2018. *Emerg. Microbes Infect.* 8, 218–231. doi: 10.1080/22221751.2019.1568180

Auguste, A. J., Lemey, P., Pybus, O. G., Suchard, M. A., Salas, R. A., Adesiyun, A. A., et al. (2010). Yellow fever virus maintenance in Trinidad and its dispersal throughout the Americas. *J. Virol.* 84, 9967–9977. doi: 10.1128/JVI.00588-10

Baleotti, F. G., Moreli, M. L., and Figueiredo, L. T. (2003). Brazilian *Flavivirus phylogeny* based on NS5. *Mem. Inst. Oswaldo Cruz* 98, 379–382. doi: 10.1590/s0074-02762003000300015

Beasley, D. W., Mcauley, A. J., and Bente, D. A. (2015). Yellow fever virus: genetic and phenotypic diversity and implications for detection, prevention and therapy. *Antiviral Res.* 115, 48–70. doi: 10.1016/j.antiviral.2014.12.010

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Brand, C., Bisaillon, M., and Geiss, B. J. (2017). Organization of the Flavivirus RNA replicase complex. *Wiley Interdiscip. Rev. RNA* 8:10.1002/wrna.1437. doi: 10.1002/wrna.1437

Bryant, J. E., Holmes, E. C., and Barrett, A. D. T. (2007). Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. *PLoS Pathog.* 3:e75. doi: 10.1371/journal.ppat.0030075

Casadio, L. V. B., Salles, A. P. M., Malta, F. M., Leite, G. F., Ho, Y. L., Gomes-Gouvêa, M. S., et al. (2019). Lipase and factor V (but not viral load) are prognostic factors for the evolution of severe yellow fever cases. *Mem. Inst. Oswaldo Cruz* 114:e190033. doi: 10.1590/0074-0276019 0033

Chambers, T. J., Hahn, C. S., Galler, R., and Rice, C. M. (1990). Flavivirus genome organization, expression, and replication. *Annu. Rev. Microbiol.* 44, 649–688. doi: 10.1146/annurev.mi.44.100190.003245

Chen, S., Wu, Z., Wang, M., and Cheng, A. (2017). Innate immune evasion mediated by Flaviviridae non-structural proteins. *Viruses* 9:291. doi: 10.3390/v9100291

Cunha, M. D. P., Duarte-Neto, A. N., Pour, S. Z., Ortiz-Baez, A. S., Černý, J., Pereira, B. B. D. S., et al. (2019a). Origin of the São Paulo yellow fever epidemic of 2017–2018 revealed through molecular epidemiological analysis of fatal cases. *Sci. Rep.* 9:20418.

Cunha, M. S., da Costa, A. C., de Azevedo Fernandes, N. C. C., Guerra, J. M., dos Santos, F. C. P., Nogueira, J. S., et al. (2019b). Epizootics due to yellow fever virus in São Paulo State, Brazil: viral dissemination to new areas (2016-2017). *Sci. Rep.* 9:5474. doi: 10.1038/s41598-019-41950-3

de Souza, R. P., Foster, P. G., Sallum, M. A., Coimbra, T. L., Maeda, A. Y., Silveira, V. R., et al. (2010). Detection of a new yellow fever virus lineage within the South American genotype I in Brazil. *J. Med. Virol.* 82, 175–185. doi: 10.1002/jmv.21606

Delatorre, E., De Abreu, F. V. S., Ribeiro, I. P., Gómez, M. M., Dos Santos, A. A. C., Ferreira-De-Brito, A., et al. (2019). Distinct YFV lineages co-circulated in the central-western and southeastern Brazilian regions from 2015 to 2018. *Front. Microbiol.* 10:1079. doi: 10.3389/fmicb.2019.01079

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88. doi: 10.1371/journal.pbio.0040088

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192. doi: 10.1093/molbev/msi103

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29, 1969–1973. doi: 10.1093/molbev/mss075

Ferreira, M. A. R., and Suchard, M. A. (2008). Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.* 36, 355–368. doi: 10.1002/cjs.5550360302

Garrison, E. P., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv* [Preprint] arXiv: 1207.3907,

Gómez, M. M., Abreu, F. V. S., Santos, A., Mello, I. S., Santos, M. P., Ribeiro, I. P., et al. (2018). Genomic and structural features of the yellow fever virus from the 2016-2017 Brazilian outbreak. *J. Gen. Virol.* 99, 536–548. doi: 10.1099/jgv.0.001033

Google (2021). *Google Maps [Online]*. Available online at: https://cloud.google.com/maps-platform/ (accessed October 30, 2021).

Governo do Estado de São Paulo (2019). *Boletim Epidemiológico da Febre Amarela*. Available online at: https://www.saude.sp.gov.br/resources/cve-centro-de-vigilancia-epidemiologica/areas-de-vigilancia/doencas-de-transmissao-por-vetores-e-zoonoses/doc/famarela/2019/fa19_boletim_epid_1811.pdf (accessed October 30, 2020).

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hill, S. C., de Souza, R., Thézé, J., Claro, I., Aguiar, R. S., Abade, L., et al. (2020). Genomic surveillance of yellow fever virus epizootic in São Paulo, Brazil, 2016 – 2018. *PLoS Pathog.* 16:E1008699. doi: 10.1371/journal.ppat.1008699

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hunt, M., Gall, A., Ong, S. H., Brener, J., Ferns, B., Goulder, P., et al. (2015). IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* 31, 2374–2376. doi: 10.1093/bioinformatics/btv120

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5:e1000520. doi: 10.1371/journal.pcbi.1000520

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:3. doi: 10.1089/cmb.2017.0096

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534.

Mir, D., Delatorre, E., Bonaldo, M., Lourenço-De-Oliveira, R., Vicente, A. C., and Bello, G. (2017). Phylodynamics of yellow fever virus in the Americas: new insights into the origin of the 2017 Brazilian outbreak. *Sci. Rep.* 7:7385. doi: 10.1038/s41598-017-07873-7

Monath, T. P., and Vasconcelos, P. F. (2015). Yellow fever. *J. Clin. Virol.* 64, 160–173.

Mutebi, J. P., Wang, H., Li, L., Bryant, J. E., and Barrett, A. D. (2001). Phylogenetic and evolutionary relationships among yellow fever virus isolates in Africa. *J. Virol.* 75, 6999–7008. doi: 10.1128/JVI.75.15.6999-7008.2001

Nunes, M. R., Palacios, G., Cardoso, J. F., Martins, L. C., Sousa, E. C. Jr., De Lima, C. P., et al. (2012). Genomic and phylogenetic characterization of Brazilian yellow fever virus strains. *J. Virol.* 86, 13263–13271. doi: 10.1128/JVI.00565-12

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., et al. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* 20, 714–737. doi: 10.1089/cmb.2013.0084

Patkar, C. G., Jones, C. T., Chang, Y. H., Warrier, R., and Kuhn, R. J. (2007). Functional requirements of the yellow fever virus capsid protein. *J. Virol.* 81, 6471–6481. doi: 10.1128/JVI.02120-06

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rezende, I. M., Sacchetto, L., de Mello, E. M., Alves, P. A., Iani, F. C. M., Adelino, T. E. R., et al. (2018). Persistence of Yellow Fever virus outside the Amazon Basin, causing epidemics in Southeast Brazil, from 2016 to 2018. *PLoS Negl. Trop. Dis.* 12:e0006538. doi: 10.1371/journal.pntd.0006538

Russel, P. M., Brewer, B. J., Klaere, S., and Bouckaert, R. R. (2019). Model selection and parameter inference in phylogenetics using nested sampling. *Syst. Biol.* 68, 219–233. doi: 10.1093/sysbio/syy050

Silva, N. I. O., Sacchetto, L., De Rezende, I. M., Trindade, G. D. S., Labeaud, A. D., De Thoisy, B., et al. (2020). Recent sylvatic yellow fever virus transmission in Brazil: the news from an old disease. *Virol. J.* 17:9. doi: 10.1186/s12985-019-1277-7

Suchard, M. A., and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25, 1370–1376. doi: 10.1093/bioinformatics/btp244

Tabachnick, W. J. (2016). Climate change and the arboviruses: lessons from the evolution of the Dengue and yellow fever viruses. *Annu. Rev. Virol.* 3, 125–145. doi: 10.1146/annurev-virology-110615-035630

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Vasconcelos, P. F. (2003). [Yellow Fever]. *Rev. Soc. Bras. Med. Trop.* 36, 275–293.

von Lindern, J. J., Aroner, S., Barrett, N. D., Wicker, J. A., Davis, C. T., and Barrett, A. D. T. (2006). Genome analysis and phylogenetic relationships between east, central and west African isolates of yellow fever virus. *J. Gen. Virol.* 87, 895–907. doi: 10.1099/vir.0.81236-0

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* 69:e96. doi: 10.1002/cpbi.96

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

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

ORIGINAL RESEARCH
published: 09 June 2022
doi: 10.3389/fmicb.2022.913023

Bacteriophage-Resistant Mutant of *Enterococcus faecalis* Is Impaired in Biofilm Formation

Jiazhen Liu[1], *Yanpeng Zhu*[2,3], *Yang Li*[4], *Yuwen Lu*[4], *Kun Xiong*[5]*, *Qiu Zhong*[1]* and *Jing Wang*[2]*

[1]Department of Clinical Laboratory Medicine, Daping Hospital, Army Medical University, Chongqing, China, [2]Department of Microbiology, Army Medical University, Chongqing, China, [3]Department of Oral and Maxillofacial Surgery, Southwest Hospital, Army Medical University, Chongqing, China, [4]Medical Center of Trauma and War Injury, Daping Hospital, Army Medical University, Chongqing, China, [5]Department of Frigidzone Medicine, College of High Altitude Military Medicine, Army Medical University, Chongqing, China

Edited by:
Mao Ye,
Institute of Soil Science (CAS), China

Reviewed by:
Mengzhe Li,
Beijing University of Chemical Technology, China
Jingmin Gu,
Jilin University, China

***Correspondence:**
Kun Xiong
xiongkun515@sina.com
Qiu Zhong
410835830@qq.com
Jing Wang
wangjing2008@aliyun.com

Specialty section:
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

Received: 05 April 2022
Accepted: 19 May 2022
Published: 09 June 2022

Citation:
Liu J, Zhu Y, Li Y, Lu Y, Xiong K, Zhong Q and Wang J (2022) Bacteriophage-Resistant Mutant of Enterococcus faecalis Is Impaired in Biofilm Formation. Front. Microbiol. 13:913023. doi: 10.3389/fmicb.2022.913023

Enterococcus faecalis is a common gram-positive non-spore-forming bacterium in nature and is found in the upper respiratory tract, intestine, and mouth of healthy people. *E. faecalis* is also one of the common pathogens causing nosocomial infections and is resistant to several antibiotics commonly used in practice. Thus, treating drug-resistant *E. faecalis* with antibiotics is challenging, and new approaches are needed. In this study, we isolated a bacteriophage named EFap02 that targets *E. faecalis* strain EFa02 from sewage at Southwest Hospital. Phage EFap02 belongs to the *Siphoviridae* family with a long tail of approximately 210 nm, and EFap02 can tolerate a strong acid and alkali environment and high temperature. Its receptor was identified as the capsular polysaccharide. Phage-resistant mutants had loss-of-function mutations in glycosyltransferase (*gtr2*), which is responsible for capsular polysaccharide biosynthesis, and this caused the loss of capsular polysaccharide and interruption of phage adsorption. Although phage-resistant mutants against EFap02 can be selected, such mutants are impaired in biofilm formation due to the loss of capsular polysaccharide, which compromises its virulence. Therefore, this study provided a detailed description of the *E. faecalis* EFap02 phage with the potential for treating *E. faecalis* infection.

Keywords: bacteriophage, *Enterococcus faecalis*, phage resistance, phage receptor, capsular polysaccharide, biofilm

INTRODUCTION

Enterococcus faecalis is a common opportunistic pathogen that causes blood and urinary tract infections. It is one of the primary pathogens that infect root canals (Khalifa et al., 2016). *E. faecalis* is intrinsically resistant to several commonly used antibiotics, such as cephalosporin and aminoglycoside (García-Solache and Rice, 2019). *Vancomycin-resistant Enterococcus* (VRE) is resistant to vancomycin (Miller et al., 2016). With the emergence of multidrug-resistant strains of *E. faecalis* and its ability to form biofilms (Ch'ng et al., 2019), *E. faecalis* infection has caused great concern in practice (Fiore et al., 2019), and there is an urgent need to find new treatments.

In recent years, phage therapy has renewed interest as a promising alternative to antibiotics (Khalifa et al., 2016). Phage therapy has successfully treated patients with multidrug-resistant bacterial infections. In Belgium and France, a randomized, controlled, double-blind phase I/II trial (PhagoBurn) used a cocktail of 12 natural lytic phages to treat burn wounds infected with *Pseudomonas aeruginosa*. Compared to standard care, the bacterial load was reduced in the phage cocktail treatment group, indicating that phage therapy is promising, although the quality of the phage product needs improvement (Jault et al., 2019). Furthermore, in China, a 63-year-old woman who developed a recurrent urinary tract infection with extensively drug-resistant *Klebsiella pneumoniae* was cured by antibiotic and phage synergism (Bao et al., 2020). Only one human study described the treatment of chronic prostatitis associated with *E. faecalis* by phage therapy in 2007. Three patients who suffered chronic *E. faecalis* infection were not cured with antibiotics, autovaccines, and laser biostimulation. They were then treated with phage therapy twice daily for 1 month. The pathogen was eradicated, and clinical symptoms were relieved without recurrence (Letkiewicz et al., 2009).

In this study, we successfully isolated a lytic phage, EFap02, from the sewage of Southwest Hospital in China. Phage EFap02 belongs to the *Siphoviridae* family and has a potent lytic effect against *E. faecalis*. The capsular polysaccharide was identified as the EFap02 receptor. The phage-resistant mutant had loss-of-function mutations in glycosyltransferase Group 2 (*gtr2*), which is responsible for capsular polysaccharide biosynthesis. The mutations caused the loss of capsular polysaccharide and interruption of phage adsorption. The loss of capsular polysaccharide significantly decreased the biofilm formation capability compromising the EPap02 virulence. Overall, this study suggested that EFap02 could be a promising candidate for *E. faecalis* phage therapy.

MATERIALS AND METHODS

Strains and Cultural Conditions

The bacterial strains and phages are listed in **Table 1**. The EFa02 strain was isolated from the Daping Hospital Department of Clinical Laboratory Medicine. *E. faecalis* was cultured in brain heart infusion (BHI) medium at 37°C. When necessary, erythromycin (20 µg/ml) was added to the medium. The stock cultures were stored in medium supplemented with 20% glycerol at −80°C.

TABLE 1 | Bacterial strains and phages used in this study.

Strain or phage	Description	Source
EFa02	Wild type *Enterococcus faecalis* strain	This study
EFa02R	Phage-resistant mutant	This study
EFa02R*::gtr2*	EFa02R complemented with *gtr2*	This study
EFap02	Phage targets EFa02	This study

Isolation of the Phage That Infects *Enterococcus faecalis*

Isolation of bacteriophages was performed as previously described (Shen et al., 2018). Sewage (10 ml) from the Southwest Hospital was centrifuged at 10,000 × g for 10 min. The supernatant was filtered with a 0.45 µm aseptic filter, and 2 ml of the filtered supernatant was added to a 50 ml clean centrifuge tube. The logarithmic phase host bacteria EFa02 was added and then cultured in a 37°C shaking incubator overnight. The mixture was centrifuged at 21,000 × g for 1 min, and the supernatant was filtered using a 0.22 µm aseptic filter. Then, 10 µl supernatant was mixed with 100 µl host bacteria EFa02 in a 15 ml centrifuge tube. BHI soft agar (5 ml) was added, and the content was poured onto the surface of agar plates. The plates were incubated at 37°C overnight, and the plaques were observed on the top agar.

Electronic Microscope

Phage morphology was observed by transmission electron microscopy (TEM; HT7700, Hitachi, Japan; Lee et al., 2019). The filtered phage lysate was dropped onto the prepared Formvar/carbon-coated copper grid and incubated for 1 min. The grid was negatively stained with 2% uranium acetate for 10 min and then observed using TEM at an acceleration voltage of 80 kV. Phages are classified according to the guidelines of the International Committee on the Taxonomy of Viruses (ICTV; Lefkowitz et al., 2018).

Determination of the Optimal Multiplicity of Infection

The optimal multiplicity of infection (MOI) is the ratio of bacteriophages to bacteria at the time of infection. The optimal MOI is the number of infections when the phage can achieve the best growth state (Lee et al., 2019). According to different MOIs, the phage (2 ml) and host bacteria (2 ml) were mixed. BHI medium was added to the mixture to a final volume of 10 ml. The mixture was incubated at 37°C with shaking at 220 rpm for 6 h. The 1 ml mixture was centrifuged at 12,000 × g for 1 min and filtered with a 0.45 µm filter. The phage titer was determined by the double-layer agar (DLA) method (Lee et al., 2019). The MOI with the highest phage titer was the optimal MOI. Three biological repeats were performed.

One-Step Growth Curve

As previously described, one-step phage growth was performed (Lu et al., 2013). EFa02 was cultivated to the early logarithmic stage ($OD_{600} = 0.5$). The phage was mixed with EFa02 (10^8 CFU/ml) at an MOI of 0.01 and incubated at 37°C with shaking at 220 rpm. Samples were taken at time points 0, 10, 20, 30, 40, 50, 60, and 90 min. The phage titer was measured using the DLA method. Three biological replicates were performed.

pH and Thermal Stability of EFap02

The BHI medium was adjusted with HCl or NaOH to pH 2.0–14.0. Then, 990 µl of BHI medium was mixed with 10 µl phage stock solution (3 × 10^8 PFU/ml). After incubation at 37°C for 60 min, the phage titer was determined by the DLA method.

BHI medium (900 μl) was mixed with 100 μl phage stock solution (9 × 10^9 PFU/ml) and treated at 4, 25, 37, 50, 60, and 70°C. After 60 min, the samples were removed and cooled to room temperature. The phage titer was determined by the DLA method (Ding et al., 2020).

Sensitivity of EFap02 to Chloroform

Phage EFap02 (10^{10} PFU/ml) was mixed with chloroform at ratios of 0, 10, 25, 50, 75, and 95%. The mixtures were incubated at 37°C with shaking at 220 rpm for 60 min. And the phage titer was determined using the DLA method (Oduor et al., 2020).

Selection of Phage-Resistant Mutants

Isolating phage-resistant mutants was performed as described previously (Yang et al., 2020). Briefly, 10 μl of *E. faecalis* culture (OD_{600} = 0.2) was added to EFap02 (10^9 PFU), and the mixture was placed on BHI agar and incubated at 37°C for 24 h. Then, the phage resistance of the colonies was validated by the DLA assay.

Phage Genome Extraction, Sequencing, and Analysis

The extraction of the phage genome was performed as previously described with a slight modification (Yuan et al., 2020). To remove contaminated DNA and RNA from the phage stocks, DNase I and RNase A were added to a final concentration of 1 μg/ml, and the phage stocks were incubated at 37°C for 60 min. EDTA (0.5 mol/l) was added to a final concentration of 0.02 mol/l. Proteinase K (20 g/l) and 10% SDS were added (final concentration 0.5%). The mixture was treated at 56°C for 60 min. An equal volume of balanced phenol (pH = 8.0) was added to extract nucleic acids, and then, the mixture was centrifuged at 10,000 × g for 5 min. The upper aqueous phase was transferred to a new 1.5 ml centrifuge tube. An equal volume of chloroform was added. The tube was gently mixed and centrifuged at 10,000 × g for 10 min. The supernatant was transferred to a new centrifuge tube, mixed with a 0.6 volume of isoamyl alcohol, and placed at −20°C for 1 h. The mixture was centrifuged at 12,000 × g for 20 min to collect the DNA pellet. The DNA pellet was washed with 75% ice-cold ethanol and centrifuged at 12,000 × g for 20 min. Finally, the pellet was dried at room temperature, dissolved in 50 μl of water, and stored at −20°C for future use.

The Bacterial Genomic DNA Kit (Tiangen-DP302) was used for DNA extraction. The phage genome was sequenced using the Illumina HiSeq 2,500 platform (~1 Gbp/sample) and assembled with Newbler (Version 2.9) under default parameters. The characteristics of the phage genes were predicted by RAST[1] (Aziz et al., 2008) and FgeneSV.[2] The homologous DNA sequences and proteins were searched and analyzed using BlastN and BlastP in the BLAST program[3] (Altschul et al., 1997). Visualization of the phage circular genome was performed using the CGView server database (Stothard and Wishart, 2005). Phage virulence factors were analyzed by searching a virulence database (Underwood et al., 2005). A comparative analysis of the EFa02 genomes and phage-resistant mutants was performed by Breseq (Barrick et al., 2014). Analysis of phage DNA termini was performed using PhageTerm[4] (Garneau et al., 2017), and the phylogenetic tree was constructed and displayed by MEGA 7.0[5] with the neighbor-joining method (Tamura et al., 2013).

Bacterial Growth Curve and Construction of *Enterococcus faecalis Strains*

EFa02 and EFap02 were cocultured for 72 h to explore the effect of phage on bacterial growth. The OD_{600} of the culture medium was measured at given time points. Three biological replicates were performed. The *gtr2* gene was amplified by PCR using the primers listed in **Table 2** to complement *gtr2* in EFa02R. The PCR product was ligated into the plasmid pGM23 by Gibson assembly to generate pGM-*gtr2,* and pGM-*gtr2* was electroporated into EFa02R, and the strain was named EFa02R::*gtr*2.

Phage Adsorption Assay and Efficiency of Plating Assay

Phage adsorption tests were performed on different *E. faecalis* strains according to a previously reported protocol (Kęsik-Szeloch et al., 2013; Yang et al., 2020). The following steps were performed to verify the difference in the adsorption capacity of phage EFap02 to *E. faecalis* EFa02, EFa02R, and EFa02R::*gtr2*. The phage filtrate was diluted to 10^9 PFU/mL in sterile water to form the initial titer. The host bacteria (1 ml) and phage (10^5 PFU) were mixed and left in the EP tube for 5 min and centrifuged at 21,000 × g for 2 min. The phage titer was calculated by the DPA method (residual titer). The phage adsorption rate was calculated as [(initial titer − residual titer)/initial titer] × 100%. Three biological replicates were performed.

The phage was diluted in tenfold increments. Then, 100 μl of *E. faecalis* and 4 ml of BHI soft agar were mixed and poured onto a BHI agar plate. Then, 1 μl of the diluted phage was pipetted onto agar plates and incubated at 37°C overnight (Forti et al., 2018).

Capsule Staining

Bacterial was stained with crystal violet solution for 5 min and then stained with 20% copper sulfate solution. Then, the bacteria were observed under the microscope immediately.

Biofilm Experiment

Crystal violet staining was used to monitor biofilms (Forti et al., 2018). EFa02, EFa02R, and EFa02R::*gtr2* were inoculated

[1] https://rast.nmpdr.org/
[2] http://linux1.softberry.com/
[3] https://blast.ncbi.nlm.nih.gov/Blast.cgi
[4] https://galaxy.pasteur.fr
[5] http://www.megasoftware.net

TABLE 2 | Primers used in this study.

Primer use or name	Sequence (5′–3′)
Complementing with *gtr2*	
pMG44-F	TTCAAAATTCCTCCGAAT
pMG44-R	CCGGCGCTACGATATT
pMG44-02R-F	AAAATATTCGGAGGAATTTTGAAATGCCCAAAATTAGTATTATTGTTCC
pMG44-02R-R	ATATCGTAGCGCCGGTTAACTATTCTTTTTATTATTTGCTTCTTGTAATTTAGG

in 1 ml of BHI medium and cultured at 37°C overnight. The OD_{600} was measured, and the culture was diluted to an OD_{600} of 0.02. A diluted bacterial solution (1 ml) was added to a 24-well plate and incubated at 37°C for 24 h, 48 h, or 72 h. The medium was then gently removed, and the wells were washed three times with 1 ml of phosphate-buffered saline (PBS). Finally, the plate was stained with 1 ml of 0.1% crystal violet at room temperature for 20 min. The excess crystal violet was carefully removed, washed with sterile water three times, and air-dried for 3 h. Then, 1 ml of 95% ethanol was added for 10 min, and the OD_{570} of each well was measured.

Statistical Analysis

Statistical analysis was performed using one-way ANOVA or Student's *t*-test. A $p < 0.05$ was considered statistically significant.

RESULTS

Isolation and Identification of Bacteriophage Against *Enterococcus faecalis*

Phage plaques were first discovered on the top agar lawn that contained a clinical strain of *E. faecalis* EFa02. Clear plaques, approximately 2 mm in diameter, were formed on the DLA plate (**Figure 1A**). The plaque was purified three times, and the isolated phage was named EFap02. Electron microscopy showed that phage EFap02 had an icosahedron head and a long tail of approximately 210 nm (**Figure 1B**). Phage EFap02 belonged to the Siphoviridae family according to the ICTV guidelines and was named vB_EFaS_EFap02.

The Bactericidal Effect of Phage EFap02

The highest phage EFap02 amount was 2.3 × 10^{10} PFU/ml when the MOI was 0.0001. Therefore, the optimal MOI of phage EFap02 is 0.0001 (**Figure 1C**). The one-step growth curve was measured to determine the latent period of phage EFap02 (**Figure 1D**). The phage titer increased 10 min after infection and peaked at 30 min, indicating a phage lysis time of approximately 30 min. These data suggest that EFap02 is effective at infecting EFa02.

Stability of Bacteriophage

The stability of EFap02 was tested under different pH, temperature, and chloroform treatments. EFap02 was viable when the pH was 4–11 (**Figure 2A**). Moreover, EFap02 was viable at 60°C and was rapidly inactivated at 70°C (**Figure 2B**). However, the EFap02 titer dropped to approximately 5 × 10^6 PFU/mL after chloroform treatment (**Figure 2C**). These data indicate that EFap02 can tolerate a strong acid and alkali environment and high temperature but is partially sensitive to chloroform.

Genomic Characterization of Phage EFap02

The phage EFap02 genome structure is shown in **Figure 3**. The whole-genome sequencing results and phageTerm analysis results revealed that EFap02 had a linear double-stranded DNA, with a length of 39,776 bps and a (G + C) content of 35%. The complete genome sequence is available at GenBank (accession no. OL505084). The EFap02 genome sequence was searched as a query in the nucleotide database of the National Center for Biotechnology Information (NCBI). The results showed that the genome sequences of *Enterococcus* phage phiSHEF4 (GenBank accession no. NC_042022.1) and *Enterococcu*s phage phiSHEF11 (GenBank accession no. OL799257.1) shared similar query coverage above 80% and identity above 91% with the EFap02 genome. The phylogenetic tree of EFap02 was constructed based on the nucleotide sequence terminase large subunit (ORF58), as shown in **Figure 4**. The result showed that EFap02 is closely related to the *Enterococcus* phage LY0323 (GenBank accession no. MH375074) genome, but was distantly related to the other phages included in the analyses. Based on these results, phage EFap02 could be considered as a novel phage. The EFap02 genome sequence encodes 60 predicted open reading frames (ORFs), 20 were functional genes, and 40 were hypothetical proteins (**Table 3**). No homologs of bacteriophage excisionases, repressors, integrases, or transposases were predicted in the EFap02 genome, supporting phage EFap02 as a lytic phage.

The EFap02 phage genome includes DNA replication and modification, transcription regulation, phage packaging, structural protein, and host lysis protein modules. ORF53 encodes phage DNA packaging protein in the packaging module. ORF58 encodes a large terminase subunit, an enzyme that inserts a single viral genome into the viral procapsid. ORF56, a portal protein, injects DNA into the host cell through a pathway formed by portal protein. Among the structural proteins, ORF49 encodes the phage major tail protein. ORF44 encodes the phage tail fiber. ORF45 encodes the phage tail assembly protein, and ORF47 encodes the phage tail tape measure protein. These structural proteins are involved in binding to the host bacterium. For the lysis module, ORF42 is annotated as N-acetylmuramoyl-L-alanine amidase, and ORF43 is annotated as holin, which can form micron-scale holes in the inner membrane. The phage releases

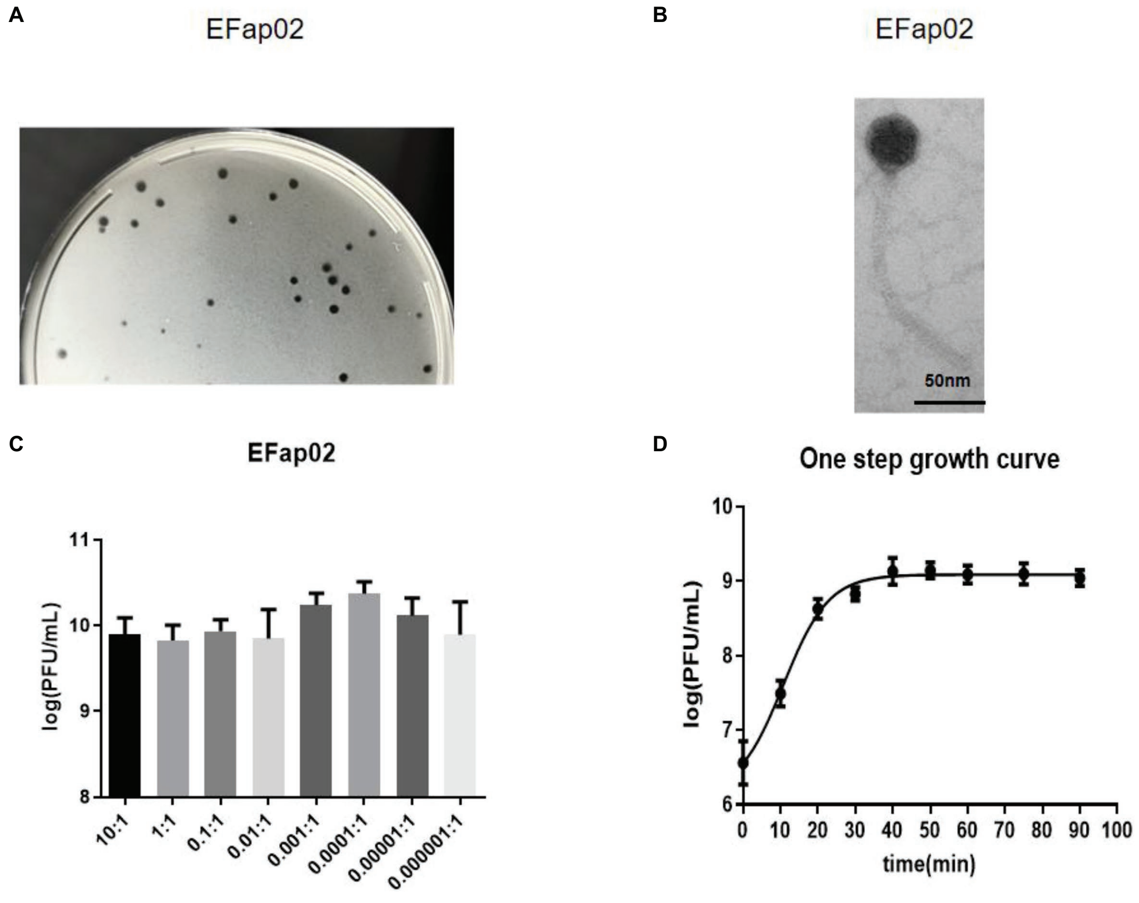


FIGURE 1 | Phage EFap02 general biological characteristics. **(A)** Phage plaques on a double-layer agar plate of EFap02. **(B)** TEM image of phage EFap02. EFap02 has an icosahedral head (length 44 nm ± 5, width 40 nm ± 5) and a long tail (210 nm ± 5). The scale bar in the right corner is 50 nm. **(C)** Different multiplicities of infection (MOI) of EFap02. When the MOI was 0.0001, the EFap02 titer was the highest. **(D)** One-step growth curve of EFap02.

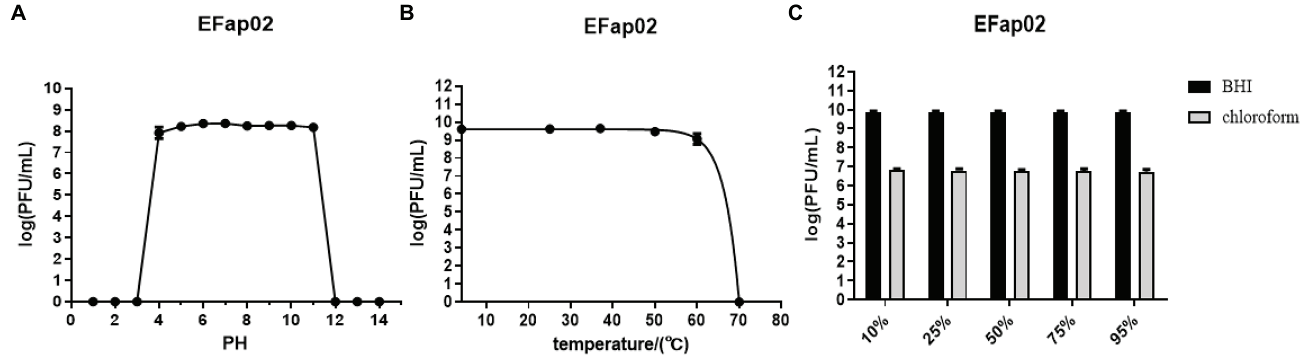


FIGURE 2 | Stability of EFap02 under different conditions. **(A)** EFap02 is stable over a broad range of pH values (pH values from 4 to 11). **(B)** EFap02 is viable over a broad range of temperatures. **(C)** EFap02 is sensitive to chloroform, and the titer drops to approximately 5 × 10^6 PFU/ml after treatment.

active endolysin (N-acetylmuramoyl-L-alanine amidase) into the periplasm to digest the peptidoglycan of *E. faecalis*. The Virulence Search Database shows that none of the ORFs encode virulence factors or antibiotic resistance genes, indicating that EFap02 can be safely used to treat *E. faecalis*-associated diseases.

Isolation and Sequencing of the Phage-Resistant Strain EFa02R

In the bacterial growth curve, the OD_{600} of the phage–bacteria coculture increased 30 h after the phage was added (**Figure 5A**). The culture medium became turbid due to the growth of

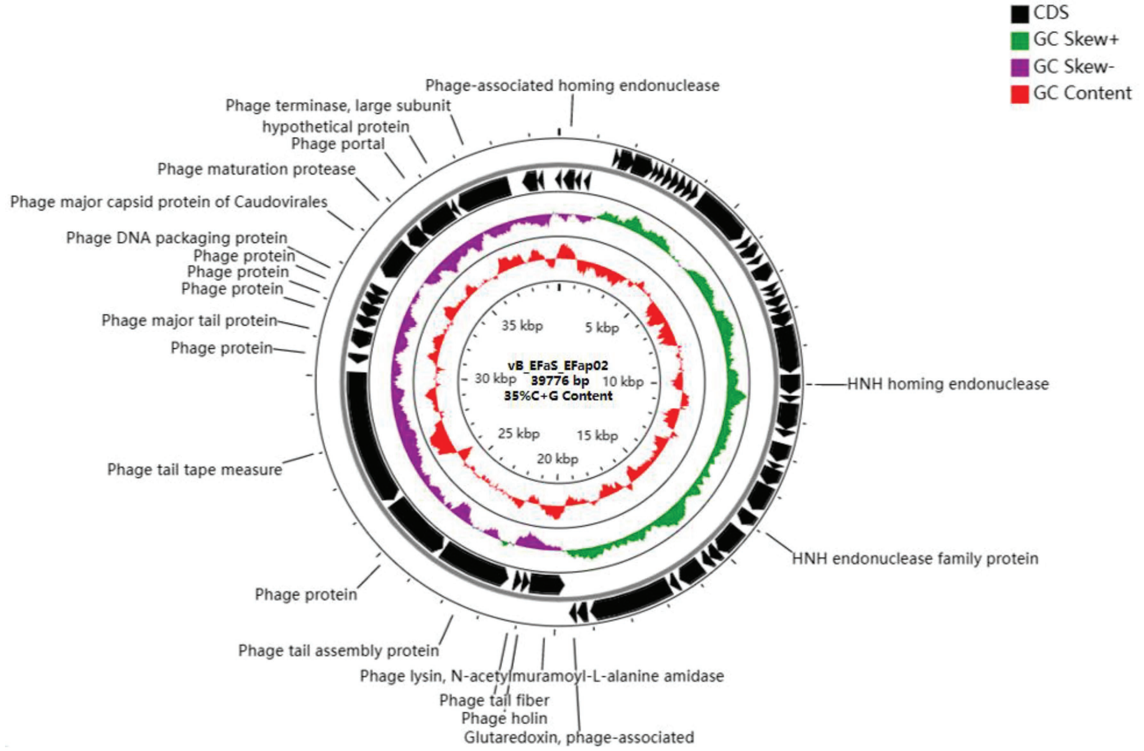


FIGURE 3 | Genome characterization of phage EFap02. Genome map of phage EFap02 and genetic characteristics. Circular genome visualization of EFap02 was performed using the CGView server database. Annotation of the specific function of ORFs was conducted using RAST and the BLASTP database. The first inner circle with the red histogram indicates the GC content, while the second inner circle with the purple and green histograms indicates the GC skew. The outer black circle indicates the predicted ORFs of phage EFap02.

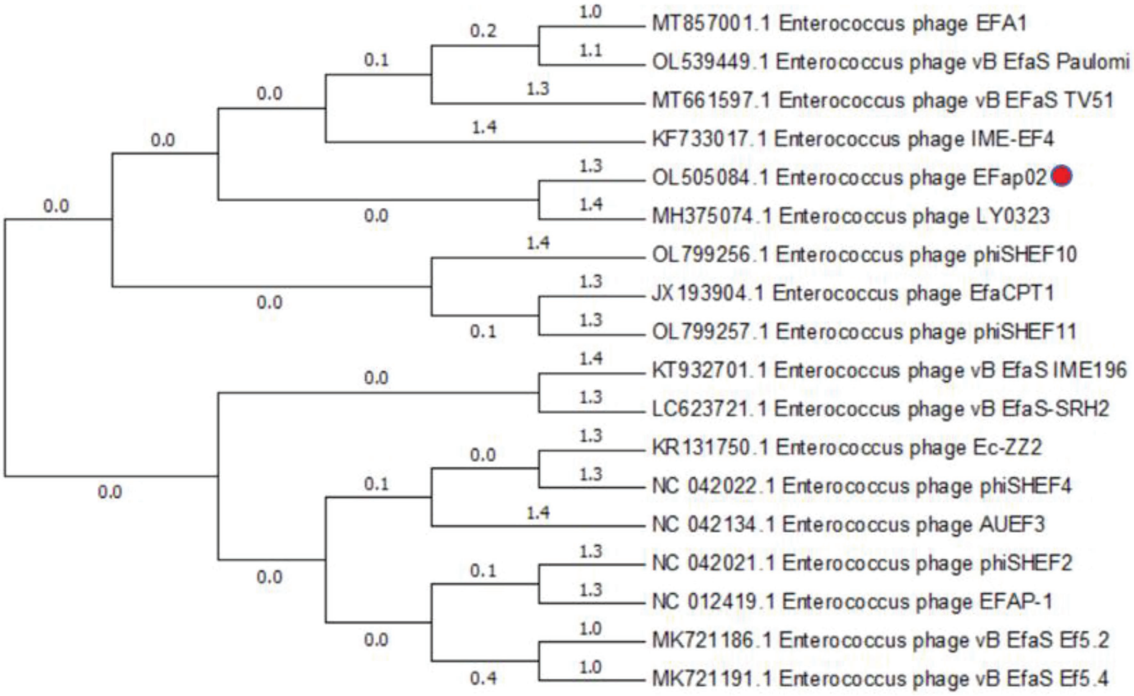

MT857001.1 Enterococcus phage EFA1
OL539449.1 Enterococcus phage vB EfaS Paulomi
MT661597.1 Enterococcus phage vB EFaS TV51
KF733017.1 Enterococcus phage IME-EF4
OL505084.1 Enterococcus phage EFap02 ●
MH375074.1 Enterococcus phage LY0323
OL799256.1 Enterococcus phage phiSHEF10
JX193904.1 Enterococcus phage EfaCPT1
OL799257.1 Enterococcus phage phiSHEF11
KT932701.1 Enterococcus phage vB EfaS IME196
LC623721.1 Enterococcus phage vB EfaS-SRH2
KR131750.1 Enterococcus phage Ec-ZZ2
NC 042022.1 Enterococcus phage phiSHEF4
NC 042134.1 Enterococcus phage AUEF3
NC 042021.1 Enterococcus phage phiSHEF2
NC 012419.1 Enterococcus phage EFAP-1
MK721186.1 Enterococcus phage vB EfaS Ef5.2
MK721191.1 Enterococcus phage vB EfaS Ef5.4
1.0
0.2
1.1
0.1
1.3
0.0
1.4
0.0
1.3
0.0
1.4
0.0
1.4
1.3
0.0
1.3
0.1
1.4
0.0
1.3
1.3
0.0
0.1
1.3
0.0
1.4
1.3
0.1
0.0
1.3
1.0
0.0
1.0
0.4

FIGURE 4 | Phylogenetic relationships of phages infecting *Enterococcus faecalis*. Terminase large subunit and the neighbor-joining method were used to construct the phylogenetic tree.

TABLE 3 | The ORFs of EFap02.

ORF	Start	Stop	Strand	Function
ORF01	481	110	–	Phage-associated homing endonuclease
ORF02	681	481	–	hypothetical protein
ORF03	967	764	–	hypothetical protein
ORF04	1,524	1,679	+	hypothetical protein
ORF05	1,676	2095	+	hypothetical protein
ORF06	2061	2,648	+	hypothetical protein
ORF07	2,673	2,792	+	hypothetical protein
ORF08	2,805	2,984	+	hypothetical protein
ORF09	2,996	3,220	+	hypothetical protein
ORF10	3,217	3,429	+	hypothetical protein
ORF11	3,426	3,644	+	hypothetical protein
ORF12	3,641	3,856	+	hypothetical protein
ORF13	3,853	4,050	+	hypothetical protein
ORF14	4,145	5,725	+	hypothetical protein
ORF15	5,817	6,011	+	hypothetical protein
ORF16	6,018	6,338	+	hypothetical protein
ORF17	6,376	6,564	+	hypothetical protein
ORF18	6,637	7,080	+	hypothetical protein
ORF19	7,213	7,398	+	hypothetical protein
ORF20	7,382	7,552	+	hypothetical protein
ORF21	7,566	7,838	+	hypothetical protein
ORF22	7,840	8,046	+	hypothetical protein
ORF23	8,030	8,404	+	hypothetical protein
ORF24	8,397	9,689	+	hypothetical protein
ORF25	9,730	10,269	+	HNH homing endonuclease
ORF26	10,306	10,500	+	hypothetical protein
ORF27	10,514	11,254	+	hypothetical protein
ORF28	11,254	11,463	+	hypothetical protein
ORF29	11,630	12,154	+	hypothetical protein
ORF30	12,211	12,366	+	hypothetical protein
ORF31	12,363	12,842	+	hypothetical protein
ORF32	12,853	13,632	+	hypothetical protein
ORF33	13,726	14,130	+	HNH endonuclease family protein
ORF34	14,277	15,095	+	hypothetical protein
ORF35	15,096	15,386	+	hypothetical protein
ORF36	15,387	15,632	+	hypothetical protein
ORF37	15,713	16,420	+	hypothetical protein
ORF38	16,490	16,714	+	hypothetical protein
ORF39	16,749	19,040	+	hypothetical protein
ORF40	19,107	19,400	+	hypothetical protein
ORF41	19,412	19,618	+	Glutaredoxin, phage-associated
ORF42	20,805	19,708	–	Phage lysin, N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28)
ORF43	21,041	20,808	–	Phage holin
ORF44	21,301	21,056	–	Phage tail fiber
ORF45	23,711	21,483	–	Phage tail assembly protein
ORF46	25,830	23,755	–	Phage protein
ORF47	30,144	25,912	–	Phage tail tape measure
ORF48	30,712	30,401	–	Phage protein
ORF49	31,451	30,888	–	Phage major tail protein
ORF50	31,895	31,530	–	Phage protein
ORF51	32,299	31,892	–	Phage protein
ORF52	32,631	32,296	–	Phage protein
ORF53	32,902	32,603	–	Phage DNA packaging protein
ORF54	34,453	33,257	–	Phage major capsid protein of Caudovirales
ORF55	35,091	34,528	–	Phage maturation protease
ORF56	36,229	35,078	–	Phage portal (connector) protein
ORF57	36,398	36,234	–	hypothetical protein
ORF58	38,189	36,468	–	Phage terminase, large subunit
ORF59	39,076	38,603	–	hypothetical protein
ORF60	39,277	39,077	–	hypothetical protein

phage-resistant bacteria. We then isolated a phage-resistant mutant named EFa02R. Since most resistant bacteria exhibit phage resistance mediated by receptor mutation (Labrie et al., 2010), an adsorption assay was conducted. The results showed that the adsorption rate of EFa02R decreased to approximately 20% (**Figure 5B**). After sequencing, it was found that resistant bacteria contain a mutation

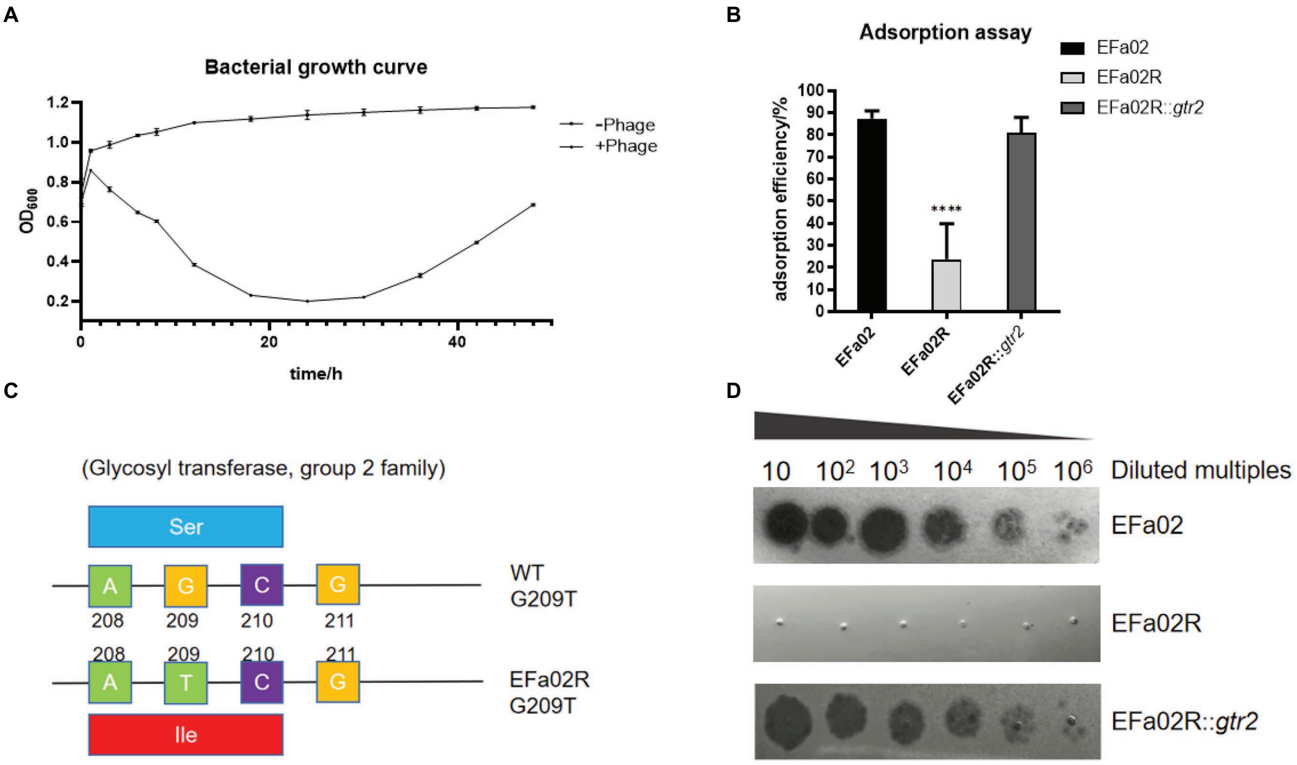


FIGURE 5 | (A) Growth curve of the phage–bacteria coculture. Thirty hours after phage addition, the OD_{600} of the phage and bacteria coculture began to increase. **(B)** Adsorption rate of phage onto three strains. This result indicates that EFap02 can adsorb EFa02 and EFa02R::*gtr2* but not EFa02R (**** $p < 0.0001$). **(C)** The mutation site of the glycosyltransferase Group 2 gene in EFa02R; the mutant site is G209T, which results in the amino acid change from serine to isoleucine. **(D)** EOP assay indicates that EFap02 can infect EFa02 and EFa02R::*gtr2* but not EFa02R.

in the *gtr2* gene. The mutant site changed from AGC to ATC, and the amino acid changed from serine to isoleucine (**Figure 5C**).

Identification of the Phage EFap02 Receptor

The *gtr2* gene is related to the synthesis of capsular polysaccharide on the surface of *Acinetobacter baumannii* (Kenyon and Hall, 2013; Singh et al., 2018). We also performed capsule staining of EFa02, EFa02R and EFa02R::*gtr2*, which showed that EFa02 and EFa02R::*gtr2* have the capsule, but EFa02R lost the capsule (**Figure 6A**). Furthermore, we performed a phage adsorption experiment to test whether capsular polysaccharide is the receptor for EFap02. We found that the adsorption capacity of EFa02R dropped dramatically to 23.9%, and the adsorption capacity recovered to 89% after complementing the *gtr2* gene (**Figure 5B**). Additionally, EOP testing showed that EFa02R::*gtr2* was resensitive to EFap02 (**Figure 5D**). These experiments suggest that the mutation of *gtr2* leads to the loss of capsular polysaccharides and prevents EFap02 from adsorbing to EFa02R.

EFa02R Is Impaired in Biofilm Formation

We measured the biofilm formation of these strains on the polystyrene surface after 24, 48, and 72 h of incubation. At each time point, the biofilm of the resistant mutant EFa02R was much less than that of the wild-type strain EFa02 and the complemented strain EFa02R::*gtr2* (**Figure 6B**), indicating that the phage-resistant mutant had impaired biofilm formation.

DISCUSSION

Phage therapy has been used to treat *E. faecalis* infectious diseases (Letkiewicz et al., 2009; Bolocan et al., 2019), as well as many other diseases. For example, phages can treat alcoholic hepatitis by killing the *E. faecalis* that produces cytolysin which leads to hepatocyte lysis (Duan et al., 2019). Thus, the characterization of *E. faecalis* phages is important for treating *E. faecalis-associated* diseases.

Identification of the phage receptors is essential for the rational selection of phages for therapy. The *E. faecalis* phage receptors had been descried previously. Duerkop et al. found that phage infection of *E. faecalis* requires a predicted integral membrane protein named PIP_{EF} (phage infection protein of *E. faecalis*). PIP_{EF} is an integral membrane protein that spans the membrane six times (Duerkop et al., 2016). In 2018, Ho et al. showed that mutation of the glycosyltransferase *epaR* leads to the resistance to phage NPV1 by preventing phage adsorption (Ho et al., 2018). And these phage receptors are cell wall-associated structures that are important for cell viability under external stresses (Canfield and Duerkop, 2020).

In this study, the isolated EFap02 phage is stable under high temperatures and other conditions, making it convenient to store

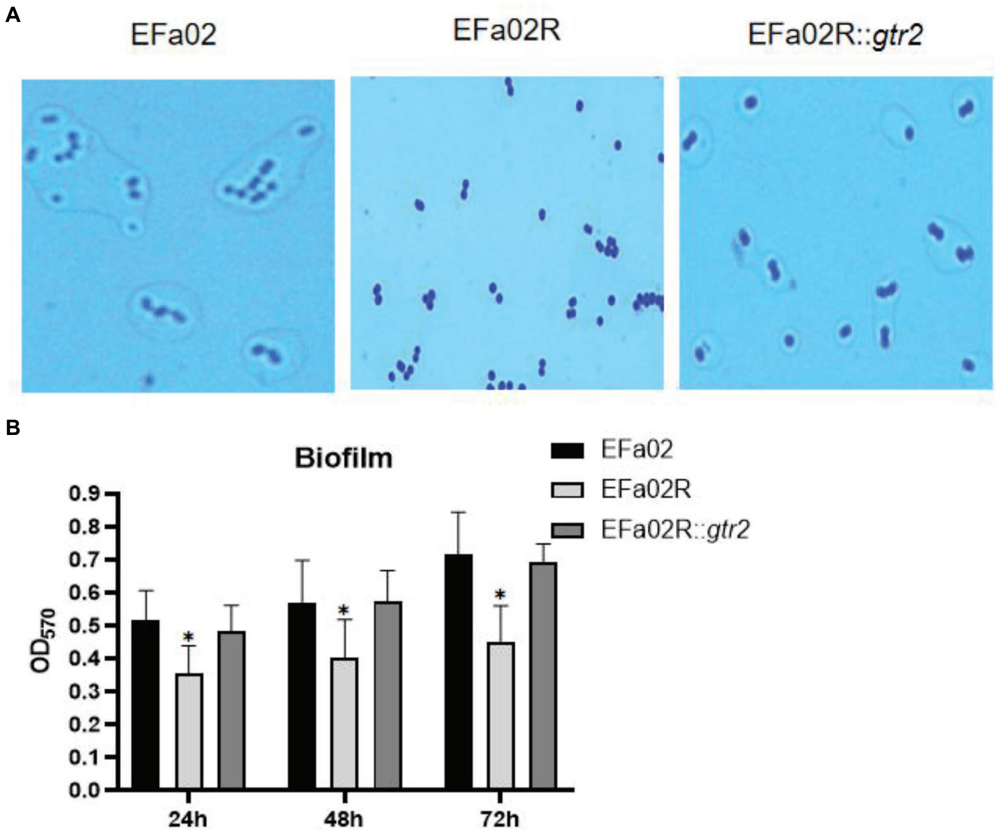


FIGURE 6 | (A) Capsule staining of EFa02, EFa02R, and EFa02R::*gtr2*. Capsule staining showed that EFa02 and EFa02R::*gtr2* have the capsule, but EFa02R lost the capsule. **(B)** The biofilm mass of EFa02R was less than those of EFa02 and EFa02R::*gtr2* (*$p < 0.05$) at the given time points.

and transport for future clinical use. Although EFap02 has strong lytic activity, we observed the rapid emergence of phage-resistant mutants. Resistance is an issue in phage therapy, and its mechanisms are diverse. The receptor mutation is the most likely to occur, and it prevents phage adsorption (Labrie et al., 2010). In this study, the phage-resistant strain EFa02R had loss-of-function mutations in the glycosyltransferase gene *Group 2 family* responsible for the biosynthesis of capsular polysaccharides (Kenyon and Hall, 2013; Singh et al., 2018). We confirmed that the *gtr2* mutation in EFa02 causes phage resistance and that the capsular polysaccharide is the phage receptor. Identifying phage receptors is essential for the rational development of phage cocktails (Gordillo Altamirano and Barr, 2021).

The loss of receptors prevents phage adsorption; however, there is always a trade-off for phage resistance. We observed that the capsular polysaccharide-loss phage-resistant mutant reduced the formation of biofilms. Previous studies found that the formation of biofilms increased the resistance to antibiotics (Ch'ng et al., 2019). Capsular polysaccharide loss may make bacteria resensitive to some antibiotics, which may allow for clearance of the bacteria by the body's immune system or antibiotics (Gordillo Altamirano and Barr, 2021).

In summary, this study revealed the biological and genomic characteristics of a phage EFap02 and identified its receptor as the capsular polysaccharide. This study suggests EFap02 as a potential phage therapy agent.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/genbank/, OL505084.

AUTHOR CONTRIBUTIONS

JL, QZ, KX, and JW conceived and designed the experiments. JL performed the experiments. JL, YZ, YLi, and YLu analyzed the data. JL, KX, QZ, and JW wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the National Natural Science Foundation of China (NSFC, 32000120 to QZ) and Military Youth Training Program (2019XQN14).

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Aziz, R. K., Bartels, D., Best, A. A., Dejongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bao, J., Wu, N., Zeng, Y., Chen, L., Li, L., Yang, L., et al. (2020). Non-active antibiotic and bacteriophage synergism to successfully treat recurrent urinary tract infection caused by extensively drug-resistant. *Emerg. Microbes Infect.* 9, 771–774. doi: 10.1080/22221751.2020.1747950

Barrick, J. E., Colburn, G., Deatherage, D. E., Traverse, C. C., Strand, M. D., Borges, J. J., et al. (2014). Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics* 15:1039. doi: 10.1186/1471-2164-15-1039

Bolocan, A. S., Upadrasta, A., Bettio, P. H. A., Clooney, A. G., Draper, L. A., Ross, R. P., et al. (2019). Evaluation of phage therapy in the context of Enterococcus faecalis and its associated diseases. *Viruses* 11:366. doi: 10.3390/v11040366

Canfield, G. S., and Duerkop, B. A. (2020). Molecular mechanisms of enterococcal-bacteriophage interactions and implications for human health. *Curr. Opin. Microbiol.* 56, 38–44. doi: 10.1016/j.mib.2020.06.003

Ch'ng, J. H., Chong, K. K. L., Lam, L. N., Wong, J. J., and Kline, K. A. (2019). Biofilm-associated infection by enterococci. *Nat. Rev. Microbiol.* 17, 82–94. doi: 10.1038/s41579-018-0107-z

Ding, T., Sun, H., Pan, Q., Zhao, F., Zhang, Z., and Ren, H. (2020). Isolation and characterization of *Vibrio parahaemolyticus* bacteriophage vB_VpaS_PG07. *Virus Res.* 286:198080. doi: 10.1016/j.virusres.2020.198080

Duan, Y., Llorente, C., Lang, S., Brandl, K., Chu, H., Jiang, L., et al. (2019). Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. *Nature* 575, 505–511. doi: 10.1038/s41586-019-1742-x

Duerkop, B. A., Huo, W., Bhardwaj, P., Palmer, K. L., and Hooper, L. V. (2016). Molecular basis for lytic bacteriophage resistance in enterococci. *MBio* 7:e01304-16. doi: 10.1128/mBio.01304-16

Fiore, E., Van Tyne, D., and Gilmore, M. S. (2019). Pathogenicity of enterococci. *Microbiol. Spectr.* 7, 4–7. doi: 10.1128/microbiolspec.GPP3-0053-2018

Forti, F., Roach, D. R., Cafora, M., Pasini, M. E., Horner, D. S., Fiscarelli, E. V., et al. (2018). Design of a broad-range bacteriophage cocktail That reduces *Pseudomonas aeruginosa* biofilms and treats acute infections in two animal models. *Antimicrob. Agents Chemother.* 62:e02573-17. doi: 10.1128/AAC.02573-17

García-Solache, M., and Rice, L. B. (2019). The Enterococcus: a model of adaptability to its environment. *Clin. Microbiol. Rev.* 32:00058-18. doi: 10.1128/CMR.00058-18

Garneau, J. R., Depardieu, F., Fortier, L.-C., Bikard, D., and Monot, M. (2017). PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Sci. Rep.* 7:8292. doi: 10.1038/s41598-017-07910-5

Gordillo Altamirano, F. L., and Barr, J. J. (2021). Unlocking the next generation of phage therapy: the key is in the receptors. *Curr. Opin. Biotechnol.* 68, 115–123. doi: 10.1016/j.copbio.2020.10.002

Ho, K., Huo, W., Pas, S., Dao, R., and Palmer, K. L. (2018). Loss-of-function mutations in confer resistance to ϕNPV1 infection in *Enterococcus faecalis* OG1RF. *Antimicrob. Agents Chemotherapy* 62:e00758-18. doi: 10.1128/AAC.00758-18

Jault, P., Leclerc, T., Jennes, S., Pirnay, J. P., Que, Y.-A., Resch, G., et al. (2019). Efficacy and tolerability of a cocktail of bacteriophages to treat burn wounds infected by *Pseudomonas aeruginosa* (PhagoBurn): a randomised, controlled, double-blind phase 1/2 trial. *Lancet Infect. Dis.* 19, 35–45. doi: 10.1016/S1473-3099(18)30482-1

Kenyon, J. J., and Hall, R. M. (2013). Variation in the complex carbohydrate biosynthesis loci of *Acinetobacter baumannii* genomes. *PLoS One* 8:e62160. doi: 10.1371/journal.pone.0062160

Kęsik-Szeloch, A., Drulis-Kawa, Z., Weber-Dąbrowska, B., Kassner, J., Majkowska-Skrobek, G., Augustyniak, D., et al. (2013). Characterising the biology of novel lytic bacteriophages infecting multidrug resistant *Klebsiella pneumoniae*. *Virol. J.* 10:100. doi: 10.1186/1743-422X-10-100

Khalifa, L., Shlezinger, M., Beyth, S., Houri-Haddad, Y., Coppenhagen-Glazer, S., Beyth, N., et al. (2016). Phage therapy against *Enterococcus faecalis* in dental root canals. *J. Oral Microbiol.* 8:32157. doi: 10.3402/jom.v8.32157

Labrie, S. J., Samson, J. E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* 8, 317–327. doi: 10.1038/nrmicro2315

Lee, D., Im, J., Na, H., Ryu, S., Yun, C. H., and Han, S. H. (2019). The novel Enterococcus phage vB_EfaS_HEf13 has broad lytic activity Against clinical isolates of *Enterococcus faecalis*. *Front. Microbiol.* 10:2877. doi: 10.3389/fmicb.2019.02877

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., and Smith, D. B. (2018). Virus taxonomy: the database of the international committee on taxonomy of viruses (ICTV). *Nucleic Acids Res.* 46, D708–D717. doi: 10.1093/nar/gkx932

Letkiewicz, S., Miedzybrodzki, R., Fortuna, W., Weber-Dabrowska, B., and Górski, A. (2009). Eradication of *Enterococcus faecalis* by phage therapy in chronic bacterial prostatitis — case report. *Folia Microbiol.* 54, 457–461. doi: 10.1007/s12223-009-0064-z

Lu, S., Le, S., Tan, Y., Zhu, J., Li, M., Rao, X., et al. (2013). Genomic and proteomic analyses of the terminally redundant genome of the *Pseudomonas aeruginosa* phage PaP1: establishment of genus PaP1-like phages. *PLoS One* 8:e62933. doi: 10.1371/journal.pone.0062933

Miller, W. R., Murray, B. E., Rice, L. B., and Arias, C. A. (2016). Vancomycin-resistant enterococci: therapeutic challenges in the 21st century. *Infect. Dis. Clin. N. Am.* 30, 415–439. doi: 10.1016/j.idc.2016.02.006

Oduor, J. M. O., Kadija, E., Nyachieo, A., Mureithi, M. W., and Skurnik, M. (2020). Bioprospecting Staphylococcus phages with therapeutic and bio-control potential. *Viruses* 12:133. doi: 10.3390/v12020133

Shen, M., Zhang, H., Shen, W., Zou, Z., Lu, S., Li, G., et al. (2018). *Pseudomonas aeruginosa* MutL promotes large chromosomal deletions through non-homologous end joining to prevent bacteriophage predation. *Nucleic Acids Res.* 46, 4505–4514. doi: 10.1093/nar/gky160

Singh, J. K., Adams, F. G., and Brown, M. H. (2018). Diversity and function of capsular polysaccharide in. *Front. Microbiol.* 9:3301. doi: 10.3389/fmicb.2018.03301

Stothard, P., and Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. *Bioinformatics* 21, 537–539. doi: 10.1093/bioinformatics/bti054

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Underwood, A. P., Mulder, A., Gharbia, S., and Green, J. (2005). Virulence searcher: a tool for searching raw genome sequences from bacterial genomes for putative virulence factors. *Clin. Microbiol. Infect.* 11, 770–772. doi: 10.1111/j.1469-0691.2005.01210.x

Yang, Y., Shen, W., Zhong, Q., Chen, Q., He, X., Baker, J. L., et al. (2020). Development of a bacteriophage cocktail to constrain the emergence of phage-resistant. *Front. Microbiol.* 11:327. doi: 10.3389/fmicb.2020.00327

Yuan, Y., Xi, H., Dai, J., Zhong, Y., Lu, S., Wang, T., et al. (2020). The characteristics and genome analysis of the novel *Y. pestis* phage JC221. *Virus Res.* 283:197982. doi: 10.1016/j.virusres.2020.197982

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

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

ORIGINAL RESEARCH
published: 01 July 2022
doi: 10.3389/fmicb.2022.907958

Characterization of *Pseudomonas aeruginosa* Bacteriophage L5 Which Requires Type IV Pili for Infection

Lan Yang[1†], Tingting Zhang[1,2†], Linlin Li[1], Chao Zheng[3], Demeng Tan[1], Nannan Wu[1,4], Mingyang Wang[3] and Tongyu Zhu[1,5]**

[1]Shanghai Institute of Phage, Shanghai Public Health Clinical Center, Fudan University, Shanghai, China, [2]Key Laboratory of Infectious Immune and Antibody Engineering of Guizhou Province, School of Biology and Engineering, Guizhou Medical University, Guiyang, China, [3]Department of Critical Care Medicine, Jiangbei District People's Hospital, Chongqing, China, [4]CreatiPhage Biotechnology Co., Ltd, Shanghai, China, [5]Shanghai Medical College, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Mao Ye,
Institute of Soil Science (CAS), China

Reviewed by:
Karishma Bisht,
Princeton University, United States
Fei Li,
Beijing University of Chemical Technology, China
Xiangyu Fan,
University of Jinan, China
Hui Zhang,
Jiangsu Academy of Agricultural Sciences (JAAS), China

***Correspondence:**
Mingyang Wang
oolaioo123@outlook.com
Tongyu Zhu
zhutongyu@shphc.org.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Virology,
a section of the journal Frontiers in Microbiology

Received: 30 March 2022
Accepted: 13 June 2022
Published: 01 July 2022

Citation:
Yang L, Zhang T, Li L, Zheng C, Tan D, Wu N, Wang M and Zhu T (2022) Characterization of Pseudomonas aeruginosa Bacteriophage L5 Which Requires Type IV Pili for Infection. Front. Microbiol. 13:907958. doi: 10.3389/fmicb.2022.907958

Pseudomonas aeruginosa is a common opportunistic human pathogen. With the emergence of multidrug-resistant (MDR) clinical infection of *P. aeruginosa*, phage therapy has received renewed attention in treating *P. aeruginosa* infections. Moreover, a detailed understanding of the host receptor of lytic phage is crucial for selecting proper phages for therapy. Here, we describe the characterization of the *P. aeruginosa* bacteriophage L5 with a double-stranded DNA genome of 42,925 bp. The genomic characteristics indicate that L5 is a lytic bacteriophage belonging to the subfamily *Autographivirinae.* In addition, the phage receptors for L5 were also identified as type IV pili, because the mutation of *pilZ*, which is involved in pili synthesis, resists phage infection, while the complementation of *pilZ* restored its phage sensitivity. This research reveals that L5 is a potential phage therapy candidate for the treatment of *P. aeruginosa* infection.

Keywords: *Pseudomonas aeruginosa*, phage (bacteriophage), *Autographivirinae*, phage-host interaction, phage receptors

INTRODUCTION

Pseudomonas aeruginosa is an important opportunistic human pathogen causing various acute and chronic infections, especially in cystic fibrosis (*CF*) patients, cancer patients, and immunocompromised individuals (Moradali et al., 2017). It is also a common pathogen that causes hospital-acquired infections, such as ventilator-associated pneumonia, urinary catheter-related infection, and surgical or transplantation infections (Moradali et al., 2017; Ibrahim et al., 2020).

In recent years, with the emergence of multidrug-resistant (MDR) *P. aeruginosa*, bacteriophages have been suggested as an alternative way to treat *P. aeruginosa* infections (Chegini et al., 2020; Yang et al., 2020). Many studies use phages to treat *P. aeruginosa* infections in animals (Forti et al., 2018; Cafora et al., 2019; Raz et al., 2019; Lin et al., 2021). In addition, phages were shown to effectively treat *P. aeruginosa* infection in patients with otitis, prosthetic knee, lung, acute kidney injury and burn wounds (Wright et al., 2009; Jennes et al., 2017; Aslam et al., 2019; Jault et al., 2019; Law et al., 2019; Ferry et al., 2021). However, one of the problems during bacteriophage therapy is phage resistance, which is often due to the mutation of receptors on the surface of the host bacterial. Therefore, it is critical to identify the receptor of new and diverse *P. aeruginosa* phages for their application to phage therapy in future.

Type IV pili (TFP) is a common bacterial surface structure in Gram-positive and Gram-negative bacteria. It confers many functions to bacteria, such as motility, adherence, DNA uptake, bacterial aggregation, and pathogenesis (Craig et al., 2004; Bahar et al., 2009; Denis et al., 2019). In addition, TFP plays a significant role in bacterial virulence and is a promising therapeutic target (Bahar et al., 2009; Dumenil, 2019). In *P. aeruginosa*, the biogenesis and function of type IV pili are controlled by over 40 genes. Some gene products are involved in biogenesis and mechanical function, whereas others play roles in transcriptional regulation and chemosensory pathways. Furthermore, the genes are expressed from unlinked gene clusters spread throughout the *P. aeruginosa* genome (Mattick, 2002; Burrows, 2012). Thus, identifying bacterial receptors for phages as type IV pili and characterizing the phage-host interaction mechanism is important in phage therapy and bacterial virulence.

To our knowledge, the majority of the *P. aeruginosa* phage receptors are the lipopolysaccharide (LPS; Jarrell and Kropinski, 1976, 1981; Temple et al., 1986; Yokota et al., 1994; Lam et al., 2011). While *P. aeruginosa* phage that targets type IV pili as receptors are not common (Heo et al., 2007; Bae and Cho, 2013). Therefore, to isolate phages that do not use the lipopolysaccharide (LPS) as a receptor, we use the *P. aeruginosa* strain PAO1r, an LPS O-antigen deficient mutant strain, as a host to isolate new phages.

In this study, we isolated a lytic bacteriophage, designated *P. aeruginosa* phage L5, from hospital sewage using the O-antigen deficient *P. aeruginosa* strain PAO1r as a host bacteria. *P. aeruginosa* phage L5 belongs to the subfamily *Autographivirinae*. In addition, we identified that phage L5 uses type IV pili as its receptor.

MATERIALS AND METHODS

Bacterial Strains, Plasmids, Media, and Growth Conditions

Bacterial strains, bacteriophages, and plasmids used in this study are listed in **Table 1**. *Escherichia coli* and *P. aeruginosa* strains were routinely grown overnight at 37°C on Luria–Bertani (LB) solid medium or in LB broth with shaking at 220 rpm. The medium was supplemented with antibiotics at the following final concentrations: gentamicin (Gm), 10 μg/ml for *E. coli,* and 35 μg/ml for *P. aeruginosa*.

Isolation and Purification of *Pseudomonas aeruginosa* Phage L5

P. aeruginosa phage L5 was isolated from hospital sewage using strain PAO1r as a host, based on a traditional double-layer agar method as described previously (Chen et al., 2019). Briefly, the sewage samples from Shanghai Public Health Clinical Center were centrifuged at 12,000 × g for 10 min and then filtered through a 0.22-μm pore size filter (Millex-GP USA). After that, 5 ml of the filtrate were mixed with 250 μl of log-phase PAO1r cells ($OD_{600} = 0.6$). After 6–8 h incubated at 37°C with shaking at 220 rpm, the mixture was centrifuged at 12,000 × g for 5 min. Then, the 300 μl supernatant was mixed with 500 μl host bacteria incubated at RT for 5 min, added to 5 ml of molten soft LB (0.5% agar), the mixtures were poured into LB plates and incubated at 37°C overnight. A single plaque was picked up and resuspended in SM buffer [5.8 g of NaCl, 2.0 g of $MgSO_4 \cdot 7H_2O$, 50 ml of Tris–HCl (PH = 7.4), 5.0 ml of 2% gelatin]. The phage-containing SM buffer was filtered through a 0.22-μm pore size filter (Millex-GP USA) and subjected to serial 10-fold dilutions in sterile SM buffer to purify the phage. Phage purification was performed by double-layer agar plate method and repeated at least three times, and the purification phage was stored at 4°C in sterile SM buffer.

TABLE 1 | Bacterial strain, phage and plasmids used in this study.

Strain, phage and plasmid	Relevant traits[a]	Source of reference
Escherichia coli		
SM10λpir	Conjugative donor strain	ZOMANBIO[b]
P. aeruginosa strain		
PAO1r	O-antigen deficient *P. aeruginosa*	(Shen et al., 2018)
PAO1rRL5	a phage L5-resistant mutant of PAO1r	This study
PAO1rΔ*pilZ*	PAO1r mutant with a deletion in the *pilZ* gene	This study
PAO1rΔ*pilZ*:: *pilZ*	PAO1rΔ*pilZ* complemented with *pilZ*	This study
Phage		
L5	*P. aeruginosa* lytic phage	This study
Plasmids		
pEXG2	Allelic exchange vector, Gm^{R}	(Rietsch et al., 2005)
pEXG2-*pilZ*	Deletion vector for deletion of *pilZ*	This study
pHB20TGm	Complementation vector, Gm^{R}	a gift [c]
pHB20TGm-*pilZ*	Complementation vector for complementation of *pilZ*	This study

[a] *Gm^{R}, gentamicin resistance.*
[b] *ZOMANBIO, Beijing Zoman Biotechnology Co., Ltd.*
[c] *a gift from Shuai Le (Army Medical University, Chongqing).*

Transmission Electron Microscopy

The morphology of the purified phage particles was observed using transmission electron microscopy (TEM). The phage particles were loaded on a carbon-coated copper grid to absorb for 15 min and negatively stained with 2% phosphotungstic acid (PTA, pH 7.0) for 2 min. Phage particles were observed using TEM (80 Kv, JEOL JEM-1200EXII, Japan Electronics and Optics Laboratory, Tokyo, Japan).

The MOI of *Pseudomonas aeruginosa* Phage L5

The host strain PAO1r was cultured to log-phase ($OD_{600} = 0.6$) and adjusted to about 1×10^8 CFU/ml. Phages were added according to the Multiplicity of Infection (MOI) of 1, 0.1, 0.01, 0.001, and 0.0001, respectively. The mixture was incubated for 3.5 h at 37°C with shaking at 220 rpm, centrifuged at 10,000 × g for 5 min at 4°C, and filtered through a 0.22-μm pore size filter. The titers were detected by the double-layer agar plate method. The experiment was repeated three times.

The One-Step Growth Curve of *Pseudomonas aeruginosa* Phage L5

The host strain PAO1r was cultured in 5 ml LB broth until log-phase ($OD_{600} = 0.6$; ~1×10^8 CFU/ml). The phage was added at a MOI = 0.01, and incubated at 37°C for 1 min without shaking. The mixture was then centrifuged at 10,000 × g for 3 min to remove free phage. Next, the precipitate was resuspended with 5 ml of fresh 37°C LB broth and cultivated at 37°C with shaking at 220 rpm. Samples were collected for 1 min, 5 min, 10 min, 20 min, 30 min, 60 min, 90 min, 120 min, and 150 min to determine the phage titer using the double-layer agar plate method. The above experiments were repeated three times.

Host Range of *Pseudomonas aeruginosa* Phage L5

Spot test was used to determine the host range of *P. aeruginosa* phage L5 on 41 clinical *P. aeruginosa* strains (**Supplementary Table S1**). All the tested strains were cultured at log-phase; 300 μl of each tested strain was mixed with 5 ml molten soft 0.7% LB agar containing 300 μl of each test bacterial culture was overlaid on 1.5% LB agar plates. Subsequently, 10 μl (~10^9 PFU/mL) of phage L5 was spotted on the soft agar. The result was observed after overnight incubation at 37°C.

DNA Extraction and Bioinformatics Analysis of *Pseudomonas aeruginosa* Phage L5

The phage genomic DNA was extracted using the phenol-chloroform protocol as described previously (Chen et al., 2019). The genome of L5 was sequenced at the Beijing Novogene using IlluminaHiSeq 2,500 platform with 200 bp read length. Protein-encoding putative open reading frames (ORFs) were predicted using RAST, tRNAs were predicted using tRNAscan-SE 2.0. The virulence genes and antibiotic resistance genes were analyzed in the virulence factors database (VFDB; VFDB: database search mgc.ac.cn) and comprehensive antibiotic resistance database (CRAD),[1] respectively. The phylogenetic tree of the phage large terminase subunit sequences was constructed using MEGA6 with 1,000 Bootstrap replications.

Screening of Phage L5-Resistant Mutants

The process of isolating phage-resistant mutants has been described previously (Johnson and Lory, 1987). Briefly, the strain PAO1r culture was infected with phage L5 particles and plated on LB agar plates for 24 h at 37°C. Single-colony was isolated and purified at least three times. The double-layer agar plate method was used to verify the resistance of isolated mutants to phage L5.

Twitching Motility Assay

Twitching motility assays were used as an indirect measurement of type IV pili function and described previously (McCutcheon et al., 2018). Briefly, a single bacterial colony was suspended in 100 μl LB broth and inoculated with a toothpick through a 3 mm thick LB agar layer (1% agar) to the bottom of the petri dish and incubated with 37°C for 72 h. Twitching motility zones between the agar and petri dish interface were visualized by gently removing the agar and staining each plate with 1% (w/v) crystal violet for 30 min followed by rinsing excess stain away with water. Stained twitching zone areas were measured using ImageJ software. Each strain was tested in biological triplicate, and the average twitching area was calculated from the three twitching zones.

Bacteriophage *Pseudomonas aeruginosa* Phage L5 Adsorption Assay

Bacteriophage adsorption assay was performed as previously described with some modifications (Chibeu et al., 2009). Briefly, bacteria that grew overnight on the LB agar plate were resuspended with fresh LB broth and adjusted to $OD_{600} = 0.6$ (~1 × 10^8 CFU/ml). Phage L5 was added to the bacterial suspension at an MOI of 0.01 and aliquoted in an equal volume to three microtubes. During the phage adsorption for 15 min at 37°C, each tube was centrifuged at13,000 × g for 5 min at the indicated time points, and the supernatants were immediately filtered through a 0.22-μm pore size filter (Millex-GP USA). The titer was immediately determined by the double-layer agar plate methods. The percentage adsorption was calculated as [(inter titer-the titer after adsorption)/ inter titer] × 100. The final adsorption rate was obtained from three independent experiments.

Construction of ΔpilZ *Pseudomonas aeruginosa* Strain PAO1r

The information of *pilZ* gene and the primers used in this study are listed in **Supplementary Table S2**; **Table 2**. In-frame deletion mutagenesis was used to construct *pilZ*-defective strain. Briefly, pilZ-up-F and pilZ-up-R were used to amplify the upstream of *pilZ*, pilZ-down-F and pilZ-down-R were used to amplify the downstream of *pilZ*. Secondly, pilZ-up-F and pilZ-down-R were used to produce the deletion fusion fragment by PCR, which has cutting sites of XhoI and XbaI. Thirdly, the XhoI/XbaI digested fragment was cloned into the XhoI/XbaI digested vector pEXG2, and transformed into SM10 λpir strain; after shaking at 37°C for 2 h, the bacterial was poured onto LB agar plate containing 15 μg/ml Gentamicin for 24 h at 37°C. Then, pilZ-up-F and pilZ-down-R were used to identify the vector pEXG2-*pilZ*. Parental strain PAO1r and the SM10 λpir strain with the vector pEXG2-*pilZ* were cultured at $OD_{600} = 1.0$, then the two strains were mixed and dropped on an LB agar plate for conjugation overnight. On the second day, scrape off the bacteria in fresh LB broth, shakes at 37°C for 1 ~ 2 h, and pour onto an LB agar plate containing 30 μg/ml gentamicin and 100 μg/ml arsenic trichloride. Next, purify clones on LB agar plates supplemented with 30 μg/ml gentamicin and 5% sucrose. Lastly, PCR and spot tests were used to identify the *pilZ*-deletion strain.

Complementation of *Pilz* Mutants

To complement *pilZ* in PAO1rΔ*pilZ*, the *pilZ* gene was amplified using primer pilZ-C-F and pilZ-C-R, and the PCR product was digested by HindIII/XbaI, and cloned into the HindIII/XbaI digested vector pHB20TGm. The product pHB20TGm-*pilZ* was electroporated into PAO1rΔ*pilZ*. Phage sensitivity of the transformants was analyzed by the efficiency of plaquing (EOP).

[1] https://omictools.com/card-tool

TABLE 2 | Primers used in this study.

Primer	Sequence (5′-3′)	Function
pilZ-up-F	TTT*CTCGAG*CAAGGTCGTGTTGCTCGAAC (XhoI)	Amplification of *pilZ* upstream
pilZ-up-R	GCATGATCCTGTCTAGCGGCAGGTTCCTGCCAGTCGAATATCAGC	Amplification of *pilZ* upstream
pilZ-down-F	GACTGGCAGGAACCTGCCGCTAGACAGGATCATGCTGGTCGATTC	Amplification of *pilZ* downstream
pilZ-down-R	TTT*TCTAGAC*GACATCGCGCACGTATTCC (XbaI)	Amplification of *pilZ* downstream
pilZ-C-F	TTT*TCTAGT*ATGAGTTTGCCACCCAATC (XbaI)	Amplification of gene *pilZ*
PilZ-C-R	TTT*AAGCTT*TTACATCGTGTGGGTCG (HindIII)	Amplification of gene *pilZ*

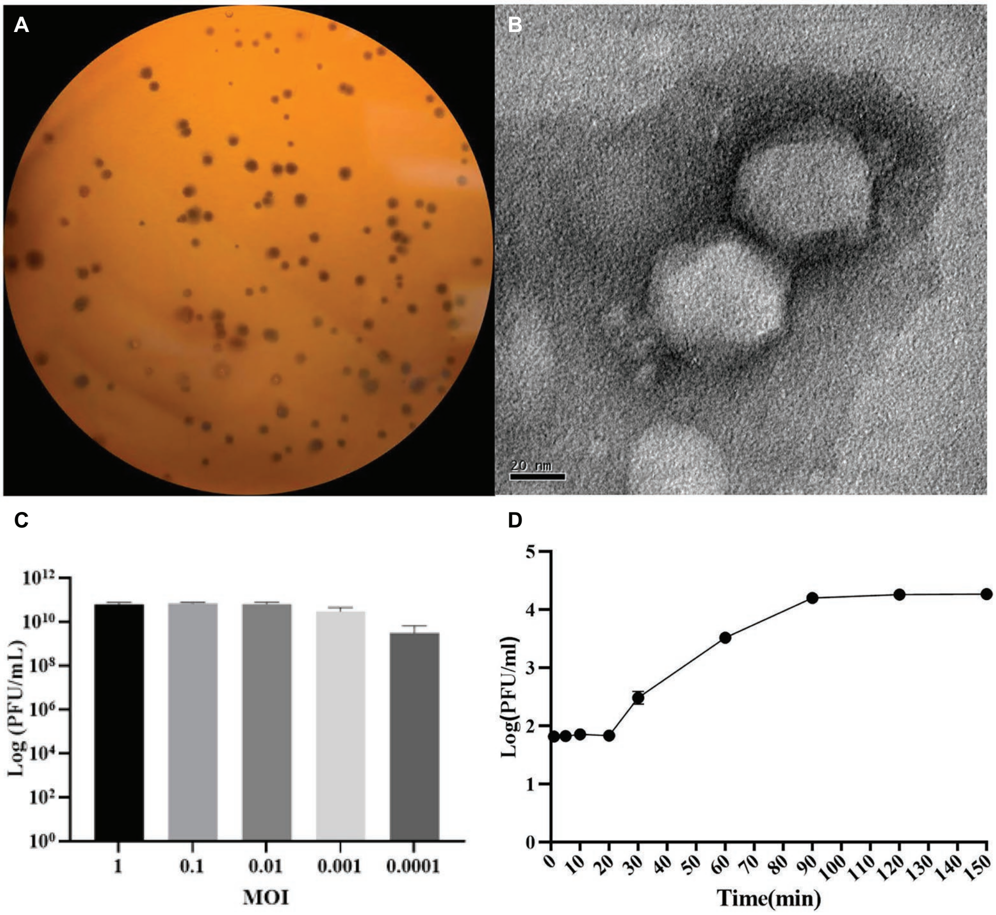


FIGURE 1 | The morphological characteristics of *Pseudomonas aeruginosa* phage L5. **(A)** Plaque morphology of L5 on *P. aeruginosa* strain PAO1r; **(B)** Transmission electron micrograph image of L5. The scale bar represents 20 nm; **(C)** The multiplicity of infection (MOI) of L5; **(D)** The one-step growth curve of L5 on *P. aeruginosa* strain PAO1r. Error bars indicate standard deviation.

Phage Plaquing Assay

The plaquing assay was determined by spotting phage on bacterial soft agar overlays. Briefly, 100 µl of overnight culture was mixed with 3 ml of soft 0.7% LB agar, overlaid onto 1.5% LB agar plates with or without antibiotics, and allowed to dry at room temperature for 30 min. Phage stocks were about 10^{11} PFU/mL on *p. aeruginosa* strain PAO1r and 10-fold serially diluted in SM to 10^{1} PFU/mL. 5 µl of each dilution was spotted onto the prepared plates and incubated for 18 h at 37°C. Each experiment was repeated in biological and technical triplicate.

RESULTS

The Biological Characteristics of *Pseudomonas aeruginosa* Phage L5

P. aeruginosa strain PAO1r was used as a host to isolate a lytic phage named *P. aeruginosa* phage L5. The results showed that the L5 could form clear plaques on LB double-layer agar plate (**Figure 1A**). The TEM showed that the L5 had an icosahedral head with a diameter of 55 nm and a very short noncontractile tail (**Figure 1B**); the morphology suggested that the L5 is a member of the *Podoviridae* family, order *Caudovirales*.

The optimum MOI of the L5 was 0.01 (**Figure 1C**). The one-step growth curve of the L5 showed that its latent and burst period was approximately 20 min and 70 min, respectively (**Figure 1D**). In addition, the host range of the L5 was analyzed against 41 clinical isolated *P. aeruginosa* strains (**Supplementary Table S1**). The result showed that 13 *P. aeruginosa* strains could be lysed, including lost LPS O-antigen strain PAO1r.

Genomic Analysis of *Pseudomonas aeruginosa* Phage L5

The complete genome sequence of L5 and annotation information is deposited in GenBank (accession number OL754589). The genomic analysis revealed that the L5 has a linear double-stranded DNA comprising 42,925 bp with a G+C content of 48.1%. A total of 55 protein-coding genes were predicted in the L5 genome. In addition, blastP analysis revealed that 32 proteins had homologs to other proteins with known functions. Meanwhile, 23 proteins were assigned as hypothetical proteins, which is common among phage populations (Meira et al., 2016; Cai et al., 2021). The detailed annotation and organization of the L5 genome are illustrated in **Figure 2**.

L5 genome can be divided into six modules, including five functionally identified modules and one functionally unknown module. In five functionally identified modules, 11 ORFs encode DNA replication and modification proteins, 3 ORFs encode packaging and assembling proteins, 13 ORFs encode structure proteins, 1 ORF encodes metabolism protein, and 3 ORFs encode host lysis proteins (**Figure 2**). In addition, no tRNA genes, antibiotic genes, toxin genes, and lysogeny genes were predicted in the L5 genome. These results revealed that the L5 satisfies several recommended criteria for selecting a therapeutic phage (Philipson et al., 2018).

Phylogenetic Analysis of *Pseudomonas aeruginosa* Phage L5

To further explore the evolutionary position of the L5, the phylogenetic analysis of L5 and other related phages was analyzed using the neighbor-joining method. Phage terminase large subunits

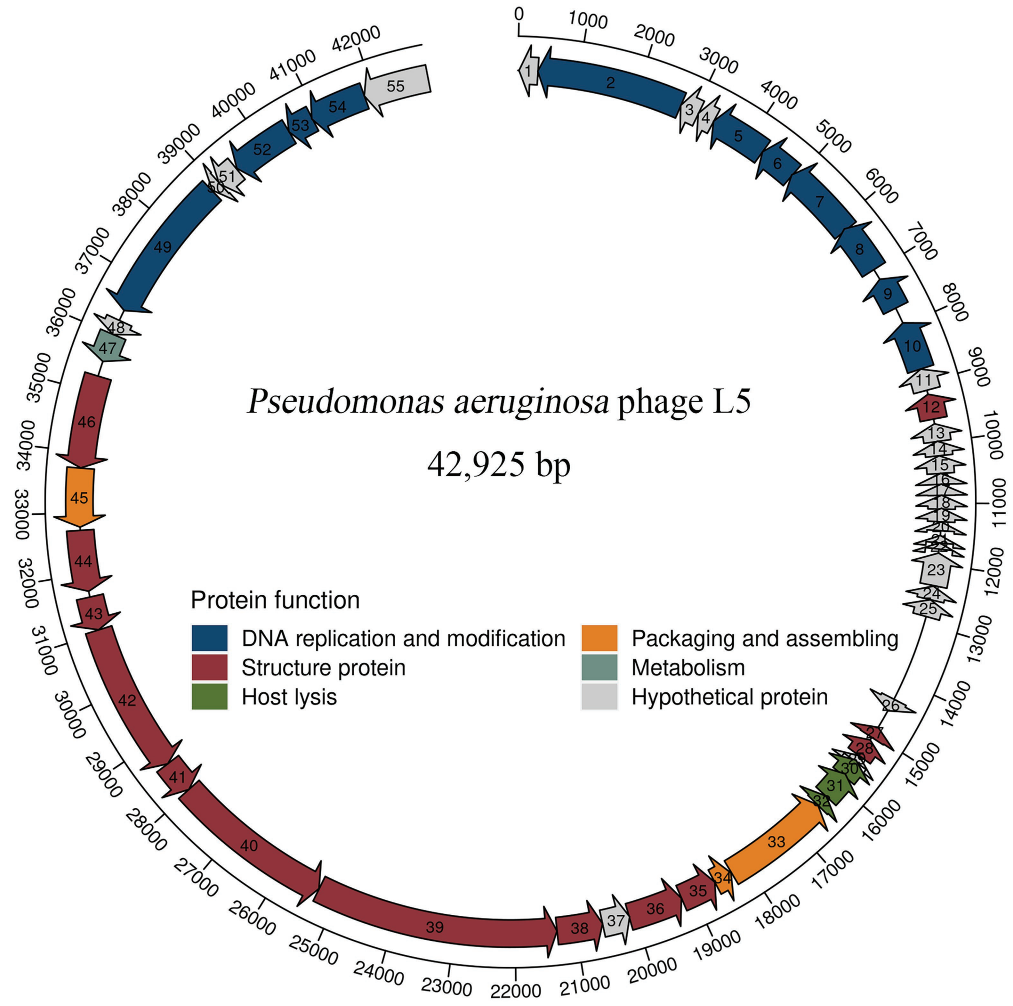


FIGURE 2 | Genome organization of *P. aeruginosa* phage L5. Arrows represent predicted CDSs, with the direction of the arrow indicating the transcription direction. The different colors indicate different functional modules of gene products. The L5 genome can be divided into six functional modules.

(TerL) is a relatively conserved protein and is mainly used as a phylogenetic marker in the comparative analysis of phage genomes (Feiss and Rao, 2012). Hence, the phylogenetic tree was constructed based on the TerL of these phages. In the phylogenetic tree of the TerL, L5 clustered together with the group including Pseudomonas phage MPK7 (YP_008431356.1), Pseudomonas phage PAXYB1 (ARB06231.1), Pseudomonas phage vB_PaeP_PAO1_Ab05 (YP_009125746.1), Pseudomonas phage RLP (YP_009820488.1) and Pseudomonas phage vB_PaeP_PPA-ABTNL (YP_009151849.1; **Figure 3A**). This result showed that L5 was the most closely related to the *Autographivirinae* subfamily. In addition, the sequences of L5 and PAXYB1 share 98.51% identity (**Figure 3B**). Therefore, combining the evolutionary analyses and some identical properties, including the morphology and linear genome, concluded that L5 is a new member of the subfamily *Autographivirinae*, the family *Podoviridae*, the order *Caudovirales*.

Phage-Resistant Mutant Isolation and Sequencing

LPS is the most common receptor for bacteriophages. However, in this study, the L5 can infect strain PAO1r, an O-antigen deficient mutant strain. In this case, it revealed the receptor of L5 is not LPS. Therefore, to investigate the receptor of strain PAO1r to L5, a phage L5-resistant mutant strain PAO1rRL5 derived from the wild-type phage-sensitive strain PAO1r, was isolated and sequencing as described in Materials and methods. By comparative genomic analysis, we found that a type IV pili *pilZ* gene has a mutant compared to the genome of strain PAO1r. The detail information of *pilZ* gene is shown in **Supplementary Table S2**. The type IV pili is a well-characterized virulence factor in *P. aeruginosa*, involved in surface motility, biofilm formation, and adherence to mammalian cells and surfaces (Burrows, 2012). Furthermore, in *P. aeruginosa*, a pilZ domain-containing protein (PA2960) involved in twitching motility (Merighi et al., 2007). Thus, in order to determine whether the type IV pili *pilZ* gene can affect twitching motility, we tested the twitching motility assays using the wild-type sensitive strain PAO1r, the phage L5-resistant strain PAO1rRL5, the *pil*Z gene knockout strain PAO1rΔ*pilZ* and the *pil*Z gene complementation strain PAO1rΔ*pilZ*::*pilZ*. The result revealed that the phage L5-resistant strain PAO1rRL5 and the *pil*Z gene knockout strain PAO1rΔ*pilZ* displayed the twitching motility is reduced by 54 and 75% relative to the phage L5-resistant strain PAO1rRL5, compared to the *pil*Z gene complementation

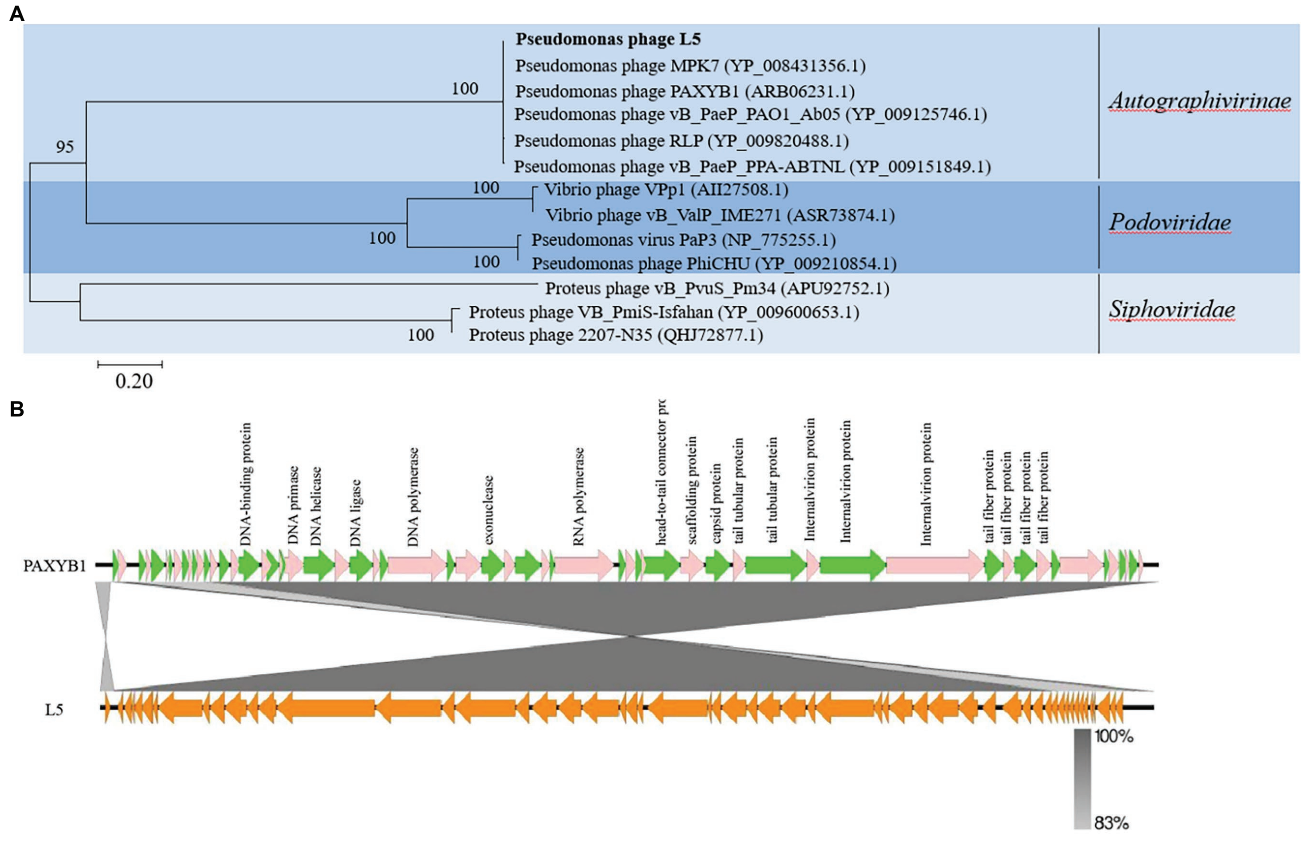

A
Pseudomonas phage L5
Pseudomonas phage MPK7 (YP_008431356.1)
Pseudomonas phage PAXYB1 (ARB06231.1)
Pseudomonas phage vB_PaeP_PAO1_Ab05 (YP_009125746.1)
Pseudomonas phage RLP (YP_009820488.1)
Pseudomonas phage vB_PaeP_PPA-ABTNL (YP_009151849.1)
Vibrio phage VPp1 (AII27508.1)
Vibrio phage vB_ValP_IME271 (ASR73874.1)
Pseudomonas virus PaP3 (NP_775255.1)
Pseudomonas phage PhiCHU (YP_009210854.1)
Proteus phage vB_PvuS_Pm34 (APU92752.1)
Proteus phage VB_PmiS-Isfahan (YP_009600653.1)
Proteus phage 2207-N35 (QHJ72877.1)
100
95
100
100
100
100
100
Autographivirinae
Podoviridae
Siphoviridae
0.20
B
DNA-binding protein
DNA primase
DNA helicase
DNA ligase
DNA polymerase
exonuclease
RNA polymerase
head-to-tail connector pro
scaffolding protein
capsid protein
tail tubular protein
tail tubular protein
Internalvirion protein
Internalvirion protein
Internalvirion protein
tail fiber protein
tail fiber protein
tail fiber protein
tail fiber protein
PAXYB1
L5
100%
83%

FIGURE 3 | (A) Phylogenetic analysis of *P. aeruginosa* phage L5 and related phages. The terminase large subunits of selected phages were compared using the ClustalW program, and the phylogenetic tree was generated using the neighbor-joining method with 1,000 bootstrap replicates. **(B)** Comparison of the genome sequence of phage PAXYB1(top) with L5 (bottom). Predicted ORFs and the direction of transcription are indicated by block arrows. Conserved regions are shaded in gray. The color intensity corresponds to the sequence identity level (83 to 100%). Genomic comparisons were performed using BLASTn. Similarities with E values lower than 0.00001 are plotted. The figure was produced using Easyfig 2.2.5 (Sullivan et al., 2011).

strain PAO1rΔ*pilZ*::*pilZ* restoring twitching motility to 83% of the phage L5-resistant strain PAO1rRL5. In addition, to perform the phage adsorption assay, about 71.43% of the L5 virions had adsorbed onto the strain PAO1r control cells. However, adsorption on the phage L5-resistant mutant strain PAO1rRL5 decreased to 1% (**Figure 4**). These results indicated that phage L5-resistant mutants prevent phage adsorption.

Pilz is Responsible for *Pseudomonas aeruginosa* Phage L5 Infection

To further confirm the type IV pili *pilZ* gene is necessary for the infection of L5 to the host bacteria, we performed the phage plaque assay using the wild-type sensitive strain PAO1r, the phage L5-resistant strain PAO1rRL5, the *pil*Z gene knockout strain PAO1rΔ*pilZ* and the *pil*Z gene complementation strain PAO1rΔ*pilZ*::*pilZ*. The result showed that L5 loses sensitivity to the strain PAO1rRL5 and the strain PAO1rΔ*pilZ*. However, L5 restored the sensitivity to the complementation strain PAO1rΔ*pilZ*::*pilZ* (**Figure 5**). These results supported the type IV pili *pil*Z gene is required for L5 infection to the host strain PAO1r.

DISCUSSION

Pseudomonas aeruginosa is one of the most important bacterial pathogens with high mortality rates in patients diagnosed with cystic fibrosis, cancer, severe burns, and immunocompromised patients (Moradali et al., 2017). This bacterium can survive on water, different surfaces, and medical devices using influential binding factors such as flagella, pili, and biofilms (Moradali et al., 2017). The emergence of MDR *P. aeruginosa* strains led to phage therapy against *P. aeruginosa*

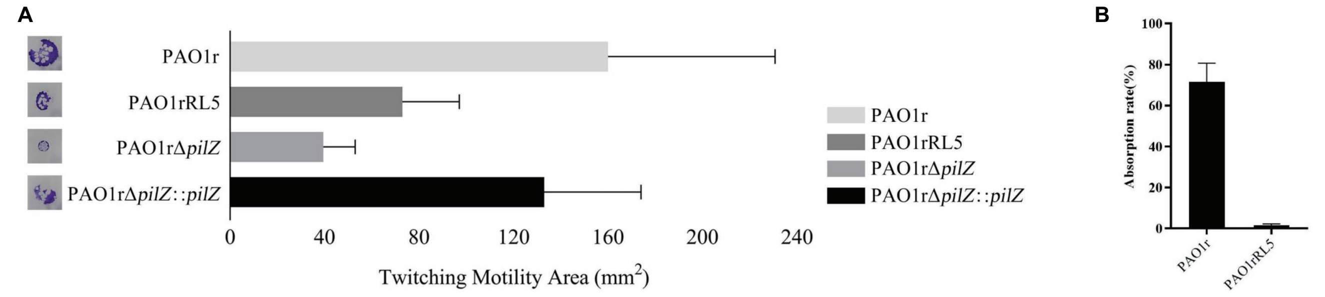


FIGURE 4 | (A) Twitching motility zones by the wild-type sensitive strain PAO1r, the L5-resistant mutant strain PAO1rRL5, the knockout strain PAO1rΔ*pilZ*, and the complementation strain PAO1rΔ*pilZ*::*pilZ*. **(B)** Phage Adsorption Assay of *P. aeruginosa* phage L5 for the strain PAO1r and the strain PAO1rRL5. The phage adsorption percent was calculated as described in the MATERIALS and METHODS.

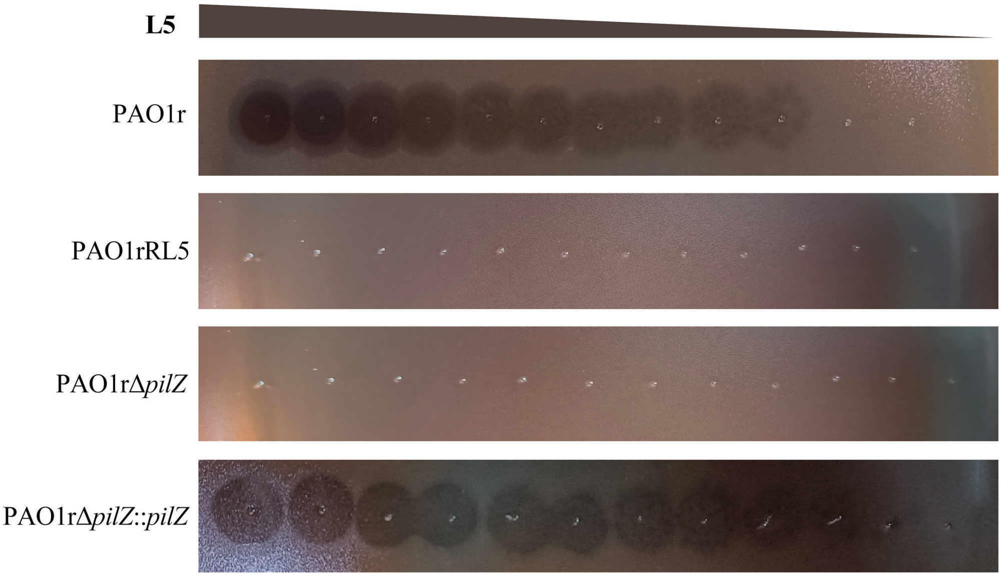


FIGURE 5 | Phage Plaquing Assay of *P. aeruginosa* phage L5. The strain PAO1r and the strain PAO1rΔ*pilZ*::*pilZ* are susceptible to L5, while the strain PAO1rRL5 and strain PAO1rΔ*pilZ* are resistant to infection.

has renewed interest (Chegini et al., 2020). Bacteriophage adsorption initiates the infection process (Rakhuba et al., 2010; Bertozzi Silva et al., 2016). Therefore, it is crucial in the infection process and plays an essential role in phage therapy against bacteria (Gordillo Altamirano and Barr, 2021). Thus, identifying phage receptors is vital and could expand phage therapy applications.

Type IV pili (T4P) are thin and flexible filaments found on the surface of a wide range of Gram-negative bacteria. T4P is involved in a broad range of functions, including twitching motility, adhesion, cell orientation, biofilm formation, pathogenicity, natural transformations, and bacteriophage infection (Craig et al., 2004; Persat et al., 2015; Tala et al., 2019). The type IV pili is a receptor for some *P. aeruginosa* phages has been identified before (Pemberton, 1973; Bae and Cho, 2013; McCutcheon et al., 2018); and in this study, we also identified the type IV pili *pil*Z gene is required for L5 infection. In addition, previous studies revealed that cyclic diguanylate (c-di-GMP) is a ubiquitous bacterial second messenger responsible for regulating cellular processes, including motility and biofilm formation. In *P. aeruginosa* genome encodes seven PilZ domain-containing proteins that have been shown to bind to c-di-GMP and an eighth PilZ domain protein that lacks c-di-GMP binding.

In this study, we found that the *pilZ* gene mutant led to a lower twitching motility zone. This means that if the *pil*Z gene has not mutant, leading to cell lysis when L5 infects the cells of the *P. aeruginosa* strain PAO1r. Meanwhile, the *pil*Z gene mutant leads to reduced motility. Therefore, the application of L5 may prove to be an effective therapy for *P. aeruginosa* infection.

In conclusion, we isolated a virulent bacteriophage L5. Genomic sequencing and analysis showed L5 belongs to the subfamily *Autographivirinae*, the family *Podoviridae,* and the order *Caudovirales*. Further, we identified the type IV pili as a receptor for phage L5 using genetic approaches.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/genbank/, OL754589.

AUTHOR CONTRIBUTIONS

ToZ and MW conceived and designed the experiments. LY and TiZ carried out the experiments and wrote the manuscript. NW, CZ, DT, and LL analyzed the data. MW and ToZ revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Natural Science Foundation of Chongqing (Grant No. cstc2019jcyj-msxmX0794), the National Natural Science Foundation of China (Grant No. 82070772) and the Shanghai Public Health Clinical Center (Grant No. KY-GW-2021-27).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.907958/full#supplementary-material

REFERENCES

Aslam, S., Courtwright, A. M., Koval, C., Lehman, S. M., Morales, S., Furr, C. L., et al. (2019). Early clinical experience of bacteriophage therapy in 3 lung transplant recipients. *Am. J. Transplant.* 19, 2631–2639. doi: 10.1111/ajt.15503

Bae, H. W., and Cho, Y. H. (2013). Complete genome sequence of *Pseudomonas aeruginosa* Podophage MPK7, which requires type IV Pili for infection. *Genome Announc.* 1:e00744-13. doi: 10.1128/genomeA.00744-13

Bahar, O., Goffer, T., and Burdman, S. (2009). Type IV Pili are required for virulence, twitching motility, and biofilm formation of *acidovorax avenae* subsp. Citrulli. *Mol. Plant Microbe. Interact* 22, 909–920. doi: 10.1094/MPMI-22-8-0909

Bertozzi Silva, J., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* 363:fnw002. doi: 10.1093/femsle/fnw002

Burrows, L. L. (2012). *Pseudomonas aeruginosa* twitching motility: type IV pili in action. *Annu. Rev. Microbiol.* 66, 493–520. doi: 10.1146/annurev-micro-092611-150055

Cafora, M., Deflorian, G., Forti, F., Ferrari, L., Binelli, G., Briani, F., et al. (2019). Phage therapy against *Pseudomonas aeruginosa* infections in a cystic fibrosis zebrafish model. *Sci. Rep.* 9:1527. doi: 10.1038/s41598-018-37636-x

Cai, X., Tian, F., Teng, L., Liu, H., Tong, Y., Le, S., et al. (2021). Cultivation of a lytic double-stranded RNA bacteriophage infecting *Microvirgula aerodenitrificans* reveals a mutualistic parasitic lifestyle. *J. Virol.* 95:e0039921. doi: 10.1128/JVI.00399-21

Chegini, Z., Khoshbayan, A., Taati Moghadam, M., Farahani, I., Jazireian, P., and Shariati, A. (2020). Bacteriophage therapy against *Pseudomonas aeruginosa* biofilms: a review. *Ann. Clin. Microbiol. Antimicrob.* 19:45. doi: 10.1186/s12941-020-00389-5

Chen, Y., Yang, L., Sun, E., Song, J., and Wu, B. (2019). Characterisation of a newly detected bacteriophage infecting *Bordetella bronchiseptica* in swine. *Arch. Virol.* 164, 33–40. doi: 10.1007/s00705-018-4034-0

Chibeu, A., Ceyssens, P. J., Hertveldt, K., Volckaert, G., Cornelis, P., Matthijs, S., et al. (2009). The adsorption of *Pseudomonas aeruginosa* bacteriophage phiKMV is dependent on expression regulation of type IV pili genes. *FEMS Microbiol. Lett.* 296, 210–218. doi: 10.1111/j.1574-6968.2009.01640.x

Craig, L., Pique, M. E., and Tainer, J. A. (2004). Type IV pilus structure and bacterial pathogenicity. *Nat. Rev. Microbiol.* 2, 363–378. doi: 10.1038/nrmicro885

Denis, K., Le Bris, M., Le Guennec, L., Barnier, J. P., Faure, C., Gouge, A., et al. (2019). Targeting type IV pili as an antivirulence strategy against invasive meningococcal disease. *Nat. Microbiol.* 4, 972–984. doi: 10.1038/s41564-019-0395-8

Dumenil, G. (2019). Type IV Pili as a therapeutic target. *Trends Microbiol.* 27, 658–661. doi: 10.1016/j.tim.2019.05.005

Feiss, M., and Rao, V. B. (2012). The bacteriophage DNA packaging machine. *Adv. Exp. Med. Biol.* 726, 489–509. doi: 10.1007/978-1-4614-0980-9_22

Ferry, T., Kolenda, C., Batailler, C., Gaillard, R., Gustave, C. A., Lustig, S., et al. (2021). Case report: arthroscopic "debridement antibiotics and implant retention" With local injection of personalized phage therapy to salvage a relapsing *Pseudomonas Aeruginosa* prosthetic knee infection. *Front. Med.* 8:569159. doi: 10.3389/fmed.2021.569159

Forti, F., Roach, D. R., Cafora, M., Pasini, M. E., Horner, D. S., Fiscarelli, E. V., et al. (2018). Design of a Broad-Range Bacteriophage Cocktail That Reduces *Pseudomonas aeruginosa* biofilms and treats acute infections in two animal models. *Antimicrob. Agents Chemother.* 62:e02573-17. doi: 10.1128/AAC.02573-17

Gordillo Altamirano, F. L., and Barr, J. J. (2021). Unlocking the next generation of phage therapy: the key is in the receptors. *Curr. Opin. Biotechnol.* 68, 115–123. doi: 10.1016/j.copbio.2020.10.002

Heo, Y. J., Chung, I. Y., Choi, K. B., Lau, G. W., and Cho, Y. H. (2007). Genome sequence comparison and superinfection between two related *Pseudomonas aeruginosa* phages, D3112 and MP22. *Microbiology* 153, 2885–2895. doi: 10.1099/mic.0.2007/007260-0

Ibrahim, D., Jabbour, J. F., and Kanj, S. S. (2020). Current choices of antibiotic treatment for *Pseudomonas aeruginosa* infections. *Curr. Opin. Infect. Dis.* 33, 464–473. doi: 10.1097/QCO.0000000000000677

Jarrell, K., and Kropinski, A. M. (1976). The isolation and characterization of a lipopolysaccharide-specific *Pseudomonas aeruginosa* bacteriophage. *J. Gen. Virol.* 33, 99–106. doi: 10.1099/0022-1317-33-1-99

Jarrell, K. F., and Kropinski, A. M. (1981). *Pseudomonas aeruginosa* bacteriophage phi PLS27-lipopolysaccharide interactions. *J. Virol.* 40, 411–420. doi: 10.1128/JVI.40.2.411-420.1981

Jault, P., Leclerc, T., Jennes, S., Pirnay, J. P., Que, Y. A., Resch, G., et al. (2019). Efficacy and tolerability of a cocktail of bacteriophages to treat burn wounds infected by *Pseudomonas aeruginosa* (PhagoBurn): a randomised, controlled, double-blind phase 1/2 trial. *Lancet Infect. Dis.* 19, 35–45. doi: 10.1016/S1473-3099(18)30482-1

Jennes, S., Merabishvili, M., Soentjens, P., Pang, K. W., Rose, T., Keersebilck, E., et al. (2017). Use of bacteriophages in the treatment of colistin-only-sensitive *Pseudomonas aeruginosa* septicaemia in a patient with acute kidney injury-a case report. *Crit. Care* 21:129. doi: 10.1186/s13054-017-1709-y

Johnson, K., and Lory, S. (1987). Characterization of *Pseudomonas aeruginosa* mutants with altered piliation. *J. Bacteriol.* 169, 5663–5667. doi: 10.1128/jb.169.12.5663-5667.1987

Lam, J. S., Taylor, V. L., Islam, S. T., Hao, Y., and Kocincova, D. (2011). Genetic and functional diversity of *Pseudomonas aeruginosa* lipopolysaccharide. *Front. Microbiol.* 2:118. doi: 10.3389/fmicb.2011.00118

Law, N., Logan, C., Yung, G., Furr, C. L., Lehman, S. M., Morales, S., et al. (2019). Successful adjunctive use of bacteriophage therapy for treatment of multidrug-resistant *Pseudomonas aeruginosa* infection in a cystic fibrosis patient. *Infection* 47, 665–668. doi: 10.1007/s15010-019-01319-0

Lin, Y., Quan, D., Chang, R. Y. K., Chow, M. Y. T., Wang, Y., Li, M., et al. (2021). Synergistic activity of phage PEV20-ciprofloxacin combination powder formulation-A proof-of-principle study in a *P. aeruginosa* lung infection model. *Eur. J. Pharm. Biopharm.* 158, 166–171. doi: 10.1016/j.ejpb.2020.11.019

Mattick, J. S. (2002). Type IV pili and twitching motility. *Annu. Rev. Microbiol.* 56, 289–314. doi: 10.1146/annurev.micro.56.012302.160938

McCutcheon, J. G., Peters, D. L., and Dennis, J. J. (2018). Identification and characterization of type IV Pili as the cellular receptor of broad host range *Stenotrophomonas maltophilia* bacteriophages DLP1 and DLP2. *Viruses* 10:338. doi: 10.3390/v10060338

Meira, G. L., Campos, F. S., Albuquerque, J. P., Cabral, M. C., Fracalanzza, S. E., Campos, R. M., et al. (2016). Genome sequence of KP-Rio/2015, a novel *Klebsiella pneumoniae* (*Podoviridae*) phage. *Genome Announc.* 4:e01298-16. doi: 10.1128/genomeA.01298-16

Merighi, M., Lee, V. T., Hyodo, M., Hayakawa, Y., and Lory, S. (2007). The second messenger bis-(3′-5′)-cyclic-GMP and its PilZ domain-containing receptor Alg44 are required for alginate biosynthesis in *Pseudomonas aeruginosa*. *Mol. Microbiol.* 65, 876–895. doi: 10.1111/j.1365-2958.2007.05817.x

Moradali, M. F., Ghods, S., and Rehm, B. H. (2017). *Pseudomonas aeruginosa* lifestyle: A paradigm for adaptation, survival, and persistence. *Front. Cell. Infect. Microbiol.* 7:39. doi: 10.3389/fcimb.2017.00039

Pemberton, J. M. (1973). F116: a DNA bacteriophage specific for the pili of *Pseudomonas aeruginosa* strain PAO. *Virology* 55, 558–560. doi: 10.1016/0042-6822(73)90203-1

Persat, A., Inclan, Y. F., Engel, J. N., Stone, H. A., and Gitai, Z. (2015). Type IV pili mechanochemically regulate virulence factors in *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* 112, 7563–7568. doi: 10.1073/pnas.1502025112

Philipson, C. W., Voegtly, L. J., Lueder, M. R., Long, K. A., Rice, G. K., Frey, K. G., et al. (2018). Characterizing phage genomes for therapeutic applications. *Viruses* 10:188. doi: 10.3390/v10040188

Rakhuba, D. V., Kolomiets, E. I., Dey, E. S., and Novik, G. I. (2010). Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. *Pol. J. Microbiol.* 59, 145–155. doi: 10.33073/PJM-2010-023

Raz, A., Serrano, A., Hernandez, A., Euler, C. W., and Fischetti, V. A. (2019). Isolation of phage Lysins That effectively kill *Pseudomonas aeruginosa* in mouse models of lung and skin infection. *Antimicrob. Agents Chemother.* 63:e00024-19. doi: 10.1128/AAC.00024-19

Rietsch, A., Vallet-Gely, I., Dove, S. L., and Mekalanos, J. J. (2005). ExsE, a secreted regulator of type III secretion genes in *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. USA* 102, 8006–8011. doi: 10.1073/pnas.0503005102

Shen, M., Zhang, H., Shen, W., Zou, Z., Lu, S., Li, G., et al. (2018). *Pseudomonas aeruginosa* MutL promotes large chromosomal deletions through non-homologous end joining to prevent bacteriophage predation. *Nucleic Acids Res.* 46, 4505–4514. doi: 10.1093/nar/gky160

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Tala, L., Fineberg, A., Kukura, P., and Persat, A. (2019). *Pseudomonas aeruginosa* orchestrates twitching motility by sequential control of type IV pili movements. *Nat. Microbiol.* 4, 774–780. doi: 10.1038/s41564-019-0378-9

Temple, G. S., Ayling, P. D., and Wilkinson, S. G. (1986). Isolation and characterization of a lipopolysaccharide-specific bacteriophage of *Pseudomonas aeruginosa*. *Microbios* 45, 81–91.

Wright, A., Hawkins, C. H., Anggard, E. E., and Harper, D. R. (2009). A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant *Pseudomonas aeruginosa*; a preliminary report of efficacy. *Clin. Otolaryngol.* 34, 349–357. doi: 10.1111/j.1749-4486.2009.01973.x

Yang, Y., Shen, W., Zhong, Q., Chen, Q., He, X., Baker, J. L., et al. (2020). Development of a bacteriophage cocktail to constrain the emergence of phage-resistant *Pseudomonas aeruginosa*. *Front. Microbiol.* 11:327. doi: 10.3389/fmicb.2020.00327

Yokota, S., Hayashi, T., and Matsumoto, H. (1994). Identification of the lipopolysaccharide core region as the receptor site for a cytotoxin-converting phage, phi CTX, of *Pseudomonas aeruginosa*. *J. Bacteriol.* 176, 5262–5269. doi: 10.1128/jb.176.17.5262-5269.1994

Conflict of Interest: NW was employed by the company CreatiPhage Biotechnology Co., Ltd.

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 07 July 2022
doi: 10.3389/fmicb.2022.863620

Auxiliary Metabolic Gene Functions in Pelagic and Benthic Viruses of the Baltic Sea

Benedikt Heyerhoff, Bert Engelen* and Carina Bunse

Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany

Marine microbial communities are facing various ecosystem fluctuations (e.g., temperature, organic matter concentration, salinity, or redox regimes) and thus have to be highly adaptive. This might be supported by the acquisition of auxiliary metabolic genes (AMGs) originating from virus infections. Marine bacteriophages frequently contain AMGs, which allow them to augment their host's metabolism or enhance virus fitness. These genes encode proteins for the same metabolic functions as their highly similar host homologs. In the present study, we analyzed the diversity, distribution, and composition of marine viruses, focusing on AMGs to identify their putative ecologic role. We analyzed viruses and assemblies of 212 publicly available metagenomes obtained from sediment and water samples across the Baltic Sea. In general, the virus composition in both compartments differed compositionally. While the predominant viral lifestyle was found to be lytic, lysogeny was more prevalent in sediments than in the pelagic samples. The highest proportion of AMGs was identified in the genomes of *Myoviridae*. Overall, the most abundantly occurring AMGs are encoded for functions that protect viruses from degradation by their hosts, such as methylases. Additionally, some detected AMGs are known to be involved in photosynthesis, 7-cyano-7-deazaguanine synthesis, and cobalamin biosynthesis among other functions. Several AMGs that were identified in this study were previously detected in a large-scale analysis including metagenomes from various origins, i.e., different marine sites, wastewater, and the human gut. This supports the theory of globally conserved core AMGs that are spread over virus genomes, regardless of host or environment.

Keywords: bacteriophage, AMGs, salinity, marine, sediment

OPEN ACCESS

Edited by:
Minh-Thu Nguyen,
University Hospital Münster, Germany

Reviewed by:
Yosuke Nishimura,
The University of Tokyo, Japan
Xupeng Cao,
Dalian Institute of Chemical Physics (CAS), China

****Correspondence:***
Bert Engelen
engelen@icbm.de

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *27 January 2022*
Accepted: *02 June 2022*
Published: *07 July 2022*

Citation:
Heyerhoff B, Engelen B and Bunse C (2022) Auxiliary Metabolic Gene Functions in Pelagic and Benthic Viruses of the Baltic Sea. Front. Microbiol. 13:863620. doi: 10.3389/fmicb.2022.863620

INTRODUCTION

Viruses are the most abundant biotic entities on Earth and are ubiquitous in the marine environment. Bacteriophages, viruses that infect bacteria, occur in concentrations of up to 10^7 viruses per ml marine surface waters, often outnumbering their hosts by 10-fold (Wommack and Colwell, 2000). Abundances of viruses in marine sediments are even higher with 10^7-10^{10} viral particles per g of dry sediment (Danovaro et al., 2008a). With an estimated number of 10^{30} viruses in the world's oceans (Breitbart, 2012), viruses play an important role in controlling marine bacterial populations through virus-induced mortality and represent a substantial reservoir of genetic diversity (Suttle, 2007). The exact numbers of virus-induced mortality are environment-dependent but increase with water depth and are as high as 90% at depths below 1,000 m

(Danovaro et al., 2008b; Breitbart et al., 2018). Virus-induced mortality has major implications on global carbon and nutrient cycling, as it leads to a conversion of biomass to dissolved organic matter (DOM), enabling a reuptake by prokaryotes as well as preventing the transfer of DOM into higher trophic levels (Fuhrman, 1999; Wilhelm and Suttle, 1999; Suttle, 2005).

The Baltic Sea is one of the largest brackish water bodies on Earth, characterized by high rates of sedimentation (Ilus et al., 2001), high nutrient and DOC concentrations, and seasonal temperature variations of > 15°C (Bunse et al., 2019), as well as substantial riverine influx of freshwater that establishes the north-southerly salinity gradient. These mechanisms result in a stratification of the Baltic Sea and a constant halocline at a water depth of 40–80 m (Vali et al., 2013). However, the connection to the North Sea allows inflow events of saline and oxygenated water to occur irregularly (Meier et al., 2006; Reissmann et al., 2009). Prolonged stagnation further divides, e.g., deep basins of the Baltic Sea into an oxygenated layer and underlying anoxic waters, separated by the pelagic redoxcline (Labrenz et al., 2007). This distinct zonation also divides bacterial mortality factors, such as grazing and viral lysis (Pernthaler, 2005). The majority of grazing occurs in oxygenated waters, while viral lysis becomes the predominant mortality factor in anoxic layers (Weinbauer et al., 2003; Kostner et al., 2017). The functional importance of viruses in the Baltic Sea becomes apparent at the phosphorous (P)-limited Ore Estuary in the northern Baltic Sea. Here, viral lysis is supplying the dissolved DNA pool with up to 25% of its total volume. The uptake of dissolved DNA covers up to 70% of the bacterioplankton's P-demand and thus supports their growth (Riemann et al., 2009). Stratification continues through Baltic Sea sediments, which are, like other marine sediments, vertically stratified and follow a redox gradient exhibiting decreasing free energy yield (Sørensen et al., 1979). High sedimentation rates and high organic matter concentrations hence harbor highly active bacterial communities and associated phages, even in deep subsurface sediments (Jørgensen et al., 2020). The most studied viruses of the Baltic Sea today are the bacteriophages of the Bacteroidetes phylum (Šulčius and Holmfeldt, 2016). Studies investigating other phyla such as Proteobacteria and Cyanobacteria remain scarce (Zeigler Allen et al., 2017; Nilsson et al., 2022, 2019).

A typical trait of marine bacteriophages is the ability to augment their host's metabolism through the promotion of auxiliary metabolic genes (AMGs) (Breitbart et al., 2007; Williamson et al., 2008). Viral AMGs are genes of high similarity to host homologs. They are introduced during viral infection and encode for the same metabolic functions as those proteins of the hosts the originate from Thompson et al. (2011). AMGs were first discovered in marine heterotrophs and Cyanobacteria in the early 2000s (Rohwer et al., 2000; Mann et al., 2003; Lindell et al., 2004b). AMGs of cyanophages are associated with a variety of functions, such as energy conservation as part of Photosystem II (Mann et al., 2003; Sharon et al., 2011). Some cyanophage genomes contain over 20 AMGs that can alter the electron transport chain or enhance the carbon metabolism of their hosts (Hellweger, 2009; Sullivan et al., 2010; Thompson et al., 2011; Crummett et al., 2016). Other functions of AMGs include, e.g., the acceleration of nucleotide biosynthesis in roseophages or sulfur oxidation genes in deep-sea viruses (Anantharaman et al., 2014; Zheng et al., 2021). Through contextual distribution and maintenance of particular AMGs in the environment, viruses increase their own fitness. While most AMGs seem to affect functions of global biogeochemical cycles, genes that increase host virulence occur as well. Here, the most famous example is the filamentous CTX bacteriophage, which carries the toxin that causes the virulence of *Vibrio cholerae* (Waldor and Mekalanos, 1996). In the past, AMG identification was performed through manual inspection and functional annotation. The advance toward scalable approaches in AMG identification through new bioinformatic tools has recently allowed for large-scale assessments across whole ecosystems and has further emphasized the ecological importance of viruses (Kieft et al., 2020, 2021). The aim of the present study was to examine the diversity and composition of benthic and pelagic viral assemblies across the Baltic Sea. We focused on how salinity as an environmental driver influenced their latitudinal distribution. Therefore, we downloaded and analyzed 212 publicly available Baltic Sea metagenomes from the National Center for Biotechnology Information (NCBI) sequence read archive (SRA). We separately analyzed the viral composition and distribution in Baltic Sea sediments and the water column. In the current study, we hypothesize that virus diversity differs in both compartments due to characteristic environmental factors. We further identified AMGs within the metagenomes and analyzed their composition and distribution along the north-southerly salinity gradient of the Baltic Sea. We hypothesize that the identified AMGs enhance the fitness of the viruses and putatively support their host.

MATERIALS AND METHODS

Metagenomic Data From the Baltic Sea

We retrieved a total of 212 publicly available metagenomes from sequence-based metagenomic studies of Baltic Sea sediments and water samples (**Figure 1**). The included data result from the project IDs PRJEB22997 (Alneberg et al., 2018), PRJEB34883 (Alneberg et al., 2020), PRJEB6616 (Thureborn et al., 2016), PRJEB8682 (Kopf et al., 2015), PRJNA308531 (Andren et al., 2015), PRJNA273799 (Hugerth et al., 2015), PRJNA297401, PRJNA322246 (Asplund-Samuelsson et al., 2016), PRJNA367442, PRJNA337783, PRJNA433242 (Zinke et al., 2017), and PRJNA337783 (Espínola et al., 2018), which were obtained from the NCBI SRA (accessed May–June 2020). The data originate from different sample sets that comprise different filter fractions, community members, and environments and likely also differ in sampling method as well as DNA extraction. The location of all metagenomic samples analyzed within this study were plotted using the R package oceanmaps (Bauer, 2020). All metadata and published environmental data available from these projects are summarized in **Supplementary Table 1**.

Sequence Quality Analysis and Assembly

Sequence quality control analysis was performed using FastQC (Andrews, 2010). Metagenomic read files were then trimmed of adapters using BBDuk (Bushnell, 2018). Quality trimming was also performed using BBDuk with the quality threshold

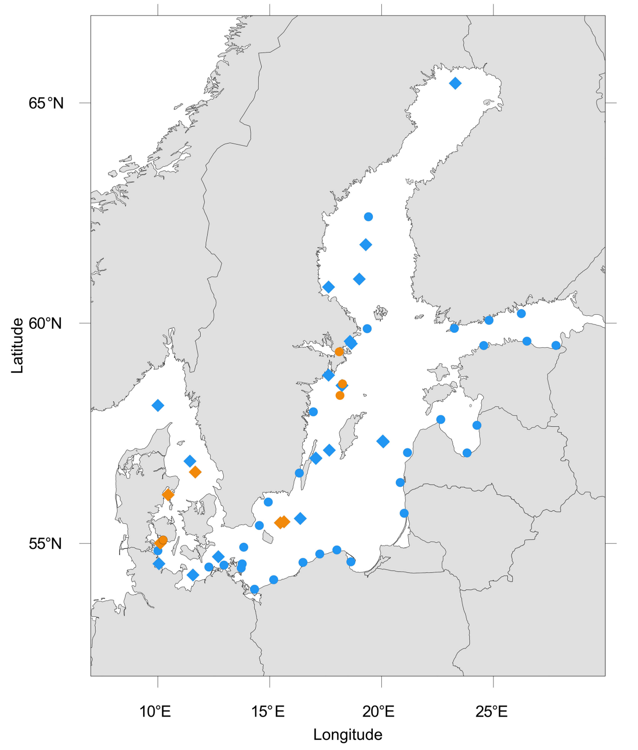


FIGURE 1 | Location of analyzed Baltic Sea metagenomes downloaded from NCBI. The blue color represents metagenomes from the water column and orange represents those from sediments. Dots represent single metagenomes, diamond shapes represent a depth profile of sediment or water metagenomes, or time-series samples. Some metagenomes were sampled in close proximity to another such that overlaying symbols may entail more than one metagenome.

set to Q30. High-quality metagenomes were assembled using MEGAHIT (Li et al., 2015) with the meta-large flag, as suggested for complex metagenomes. Statistics about the quality of assembled metagenomes were analyzed using MetaQUAST v5.0.2 (Mikheenko et al., 2016), **Supplementary Table 2**.

Identification of Viral Contigs

The *in silico* prediction of phage scaffolds and viral AMGs was done on all 212 metagenomes using VIBRANT v1.2.1 (Kieft et al., 2020) with default settings. Vibrant can accurately recover viruses and AMGs by applying machine learning and a protein similarity approach. The quality of contigs identified by VIBRANT was further assessed with CheckV (Nayfach et al., 2021) retaining contigs > 3 kb and filtering any non-viral contigs. Abundance profiles of AMGs identified by VIBRANT were generated by mapping quality-controlled metagenome reads to the AMGs using Bowtie2 (v1.2.2) (Langmead and Salzberg, 2012). The sequence mapping files were handled and converted using SAMtools (v1.9-58) (Li et al., 2009). Fragments Per Kilobase per Million (FPKM) mapped reads were calculated as the number of mapped reads times 10^9 divided by the total number of mapped reads per sample multiplied by the gene length. Viral taxonomy of AMGs located on filtered contigs was assigned using DIAMOND BLASTp v0.9.30 (E-value of < 0.0001, bit score ≥ 50) and the "—very-sensitive" preset (Buchfink et al., 2015). Viral hosts of identified contigs were assigned with VirHostMatcher-Net using default settings (Wang et al., 2020).

To gain broader context over the identified viral contigs, they were compared to viral contigs in the following public databases: (1) Global oceans virome (GOV) 2.0 Seawater (Gregory et al., 2019) and (2) Stordalen thawing permafrost (Emerson et al., 2018). Open reading frames for each viral contig were called using Prodigal V2.6.3 (Hyatt et al., 2010). Predicted protein sequences were used as input from vConTACT2 (Bin Jang et al., 2019). Viral Refseq (211) was used as a reference database (O'Leary et al., 2016). Diamond BLASTp was used for the protein-protein similarity method. All other parameters were set as default. The gene network was visualized using Cytoscape v3.9.1 (Shannon et al., 2003).

Inferring viral taxonomy through clusters identified by vConTACT2 resulted in small numbers of contigs that could be taxonomically assigned. Thus, to gain an overview of viral families present in Baltic Sea metagenomes, we used Kraken 2 to assign viral taxonomy from filtered high-quality unassembled reads applying default settings and using viral sequences from the NCBI non-redundant nucleotide database (release 211) as reference (O'Leary et al., 2016; Wood et al., 2019). Kraken 2 infers taxonomic classification by using exact k-mer matching and assigning query sequences to the lowest common ancestor. Accurate species abundance re-estimation was calculated using Bayesian Reestimation of Abundance with KrakEN (Bracken) with default settings for all metagenomes (Lu et al., 2017).

Statistical Analysis

Data processing and visualization were carried out with R version 4.0.5 (R Core Team, 2021) and the tidyverse package (Wickham et al., 2019). Bray–Curtis distances of relative viral abundances at each station were visualized by non-metric multidimensional scaling (NMDS) ($k = 2$; 999 permutations) using the vegan (v2.5-7) package (Oksanen et al., 2013). The top nine most abundant virus families were fitted to the ordination using the vegan envfit-function with 999 permutations and removal of unavailable data enabled. Salinity isobars were added using the ordiplot function. Alpha diversity was calculated with the Shannon diversity index and centered log-ratio normalized counts using the phyloseq package in R (McMurdie and Holmes, 2013; Gloor et al., 2017). A Wilcoxon test was conducted to test the significant difference in alpha diversity between the viral composition of water and sediment metagenomes using the Wilcox test R function. Beta diversity was analyzed by using the Aitchinson distance by applying principal component analysis (PCA) to the centered log-ratio transformed counts. Zero counts were avoided by adding a pseudo count to avoid errors during clr transformation. Differential abundance testing was done using DESeq2 normalized counts (Love et al., 2014). A permutational multivariate analysis of variance (PERMANOVA; function Adonis, method = "Euclidean," Permutations = 999) was done to test if beta diversity was significantly different in water or sediment metagenomes. The 20 most differentially abundant taxa with the smallest p.adj values ($p.adj < 0.001$) were plotted in a heatmap using the ComplexHeatmap package, using the dendextend R package for hierarchical clustering analysis (Galili, 2015; Gu et al., 2016).

RESULTS

The Viral Composition Differs Between Baltic Sea Sediments and Water Column Samples

In this study, we assembled and analyzed 212 publicly available metagenomes from Baltic Sea sediments and water column samples. The metagenomes were constructed from samples collected between 53°N and 65°N latitudes in the years 2008–2015. In total, 37 of the analyzed metagenomes originated from sediments and 175 from the water column samples. We identified 102,892 viral contigs > 3 kb after quality filtering with CheckV of which 7,540 contained at least one AMG (**Supplementary Table 3**).

We investigated the relationship of Baltic Sea viral contigs, with other publicly available viral sequences from different ecosystems (**Figure 2A**). Baltic Sea sediment and water column viral contigs as well as viral contigs from permafrost, GOV seawater, and RefSeq were grouped into 2,638 clusters (**Supplementary Table 4**). Viral contigs originating from the Baltic Sea water column overlapped relatively well with GOV seawater viruses; however, some small outliers occurred with the largest one displayed on the right side of the network graph (**Figure 2A**). The separate cluster of water column viruses was exclusively lytic but appeared throughout the Baltic Sea from 53° N to 65° N from surface water to 241.7 m water depth in the Skagerrak and seemed not to be impacted by salinity or temperature. None of the clusters contained vOTUs of all analyzed ecosystems. Rather, the Baltic Sea water column shared 35 clusters with GOV seawater and 11 clusters with Baltic Sea sediments but only 5 clusters contained vOTUs from Baltic Sea sediments, water column, and GOV seawater. Baltic Sea sediment shared 4 clusters with Stordalen permafrost and only very few vOTUs (0.4%) clustered with taxonomically known genomes from viral RefSeq. The limited number of clusters between viral genomes of analyzed ecosystems may reflect the high habitat specificity of viruses. The limited number of taxonomically identifiable viral genomes led us to use another method of taxonomic identification of viruses *via* Kraken 2 and allowed for a more detailed look at present viral families.

The Baltic Sea viral composition was versatile and locally differentiated. The most evenly distributed viruses in analyzed water and sediment samples were *Myoviridae*, showing abundances of around 20–40% (**Figure 2B**). One outlier was observed at 58°N, where they comprised 83.98% of the total viral composition. In sediments, the lowest abundance of *Myoviridae* occurred at around 55°N. *Siphoviridae* represented the second most abundant viral family and occurred in a less evenly distributed manner than *Myoviridae*. They dominated the sediment viral communities between 55°N and 56°N and between 57° and 58°N, where they made up to 55% of the viral assemblies. In the water column, a larger fraction of unknown and *Phycodnaviridae* viral families were found compared to sediments, of which the latter accounted for more than 80% of some communities in the water column. Overall, we did not observe a major trend along the latitudinal gradient.

Members of the *Phycodnaviridae* family were the most differentially abundant viruses and distinguished pelagic from benthic viral assemblies. *Phycodnaviridae* infect *Bathycoccus, Micromonas,* and *Ostreococcus* genera, which belong to the green algae (**Figure 3**). Sediment stations from the Bornholm Basin and the Bay of Aarhus (SRR3081534, SRR3085416, SRR3085435, SRR3085585, SRR3089827, SRR3091743, SRR3095933, SRR3095939, and SRR7067081) sampled at depths of 0.75–3 m below sea floor (mbsf) displayed higher counts of the differentially abundant Mycobacterium phage Sparkdehlily. Deep subsurface stations (SRR12059190, SRR12059191, and SRR12059199) sampled at 24.1, 24.1, and 67.5 mbsf, close to the island of Anholt and the Little Belt, were defined by the differentially abundant Ralstonia Phage RSS30.

Viruses in Sediments and Water Column are Similarly Diverse

Beta diversity of viral communities revealed a distinct pattern in virus composition between sediment and water viral assemblies, explaining 22.6% of total variation (Adonis, $p = 0.001$). The two plotted components separated viral taxa from sediments and water samples distinctively, with a small remaining overlap (**Figure 4A**). The median viral alpha diversities (Shannon index) of the sediment stations were 4.25, and 4.2 for water stations, indicating no statistically significant difference calculated by the Wilcoxon test (p-value = 0.7861) (**Figure 4B**). Due to the compositional nature of the metagenomic data used in this study, we assessed a possible batch effect *via* Bray–Curtis distance and visualized the results in an NMDS ordination (**Supplementary Figure 3**). We additionally conducted a PERMANOVA to test whether the different sequencing projects caused a batch effect within our NMDS ordinations. While some project-specific clustering could be observed within the ordination, the PERMANOVA showed these effects to be less important ($R^2 = 0.30723$, $P < 0.001$) (**Supplementary Figure 3**). The batch effect among just water and sediment stations was also not relevant ($R^2 = 0.23696$, $p < 0.001$ and $R^2 = 0.28178$, $p < 0.001$).

Viral community dissimilarity was investigated using non-metric multidimensional scaling (NMDS) analysis (**Figure 4C**). Samples close to the center of the ordination represented samples with similar viral compositions. Sediment samples spread more throughout the ordination and appeared more toward higher salinity, while water stations clustered closer together around the center of the ordination and around the 10 PSU isobar. The top nine most abundant viral families were plotted into the ordination. Among these, *Autographiviridae* and *Myoviridae* aligned more with the y-axis, whereas the other viral families aligned more with the cluster of samples that aligned with the x-axis in the lower left quadrant. The *Siphoviridae* and

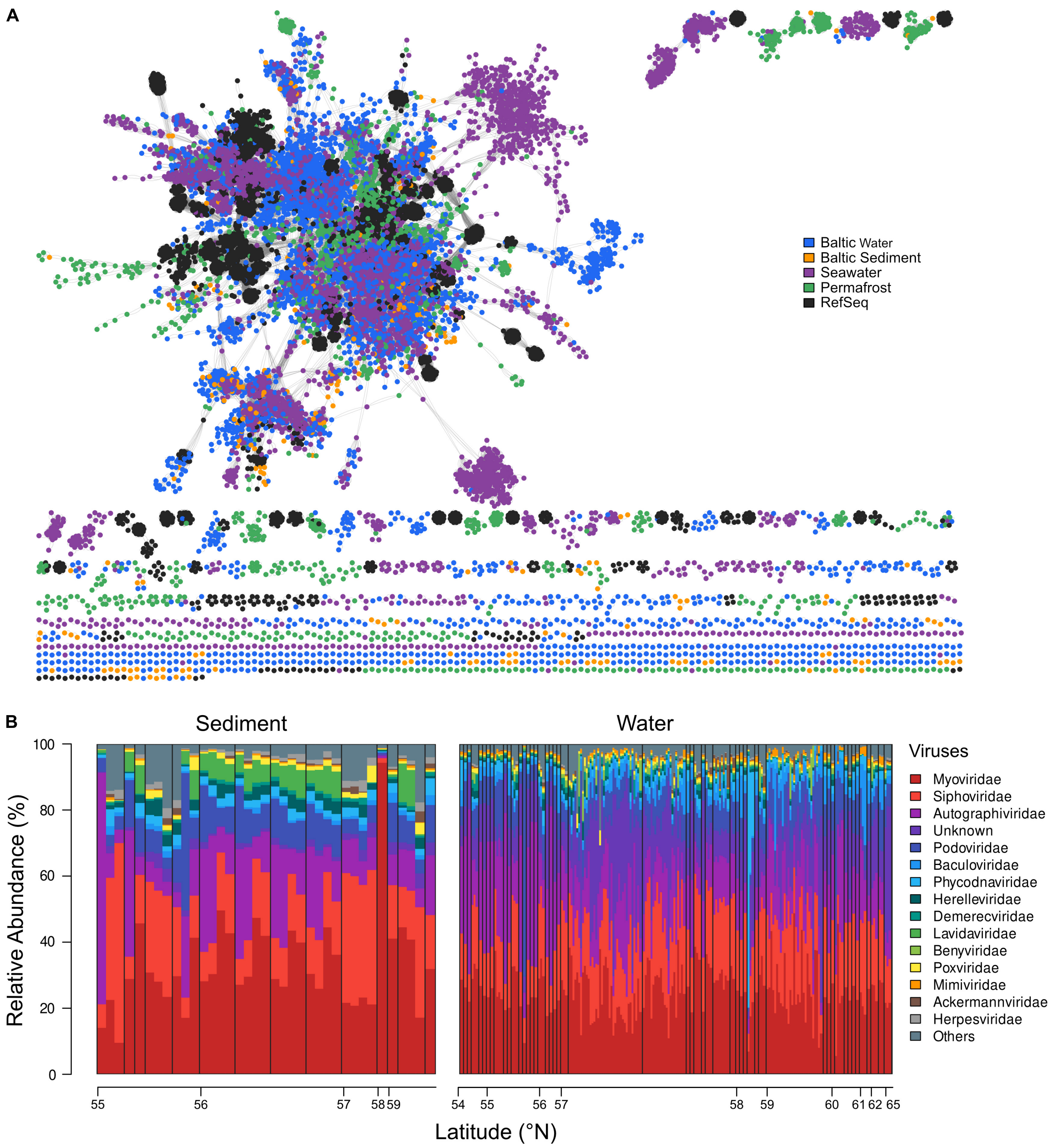


FIGURE 2 | (A) Gene-sharing network of viral sequence space based upon assembled viral genomes from Baltic Sea sediment and water column, GOV seawater, permafrost, and viral RefSeq genomes. **(B)** Relative abundance (%) and distribution of the top 15 most abundant virus families, ordered in a south-northerly arrangement by latitude (°N). Viral families are plotted in an increasing order across all water and sediment stations. The remaining families are summarized in "Others." Black bars delimit depth profiles and are ordered by increasing sediment or water depth, respectively.

Phycodnaviridae vectors were located somewhat separately from other viral families. While the lytic lifestyle aligned more with water stations, the lysogenic lifestyle appeared more in the sediments. However, the lytic lifestyle was found to be the overall dominant viral lifestyle in the Baltic Sea as shown in **Supplementary Figure 1**.

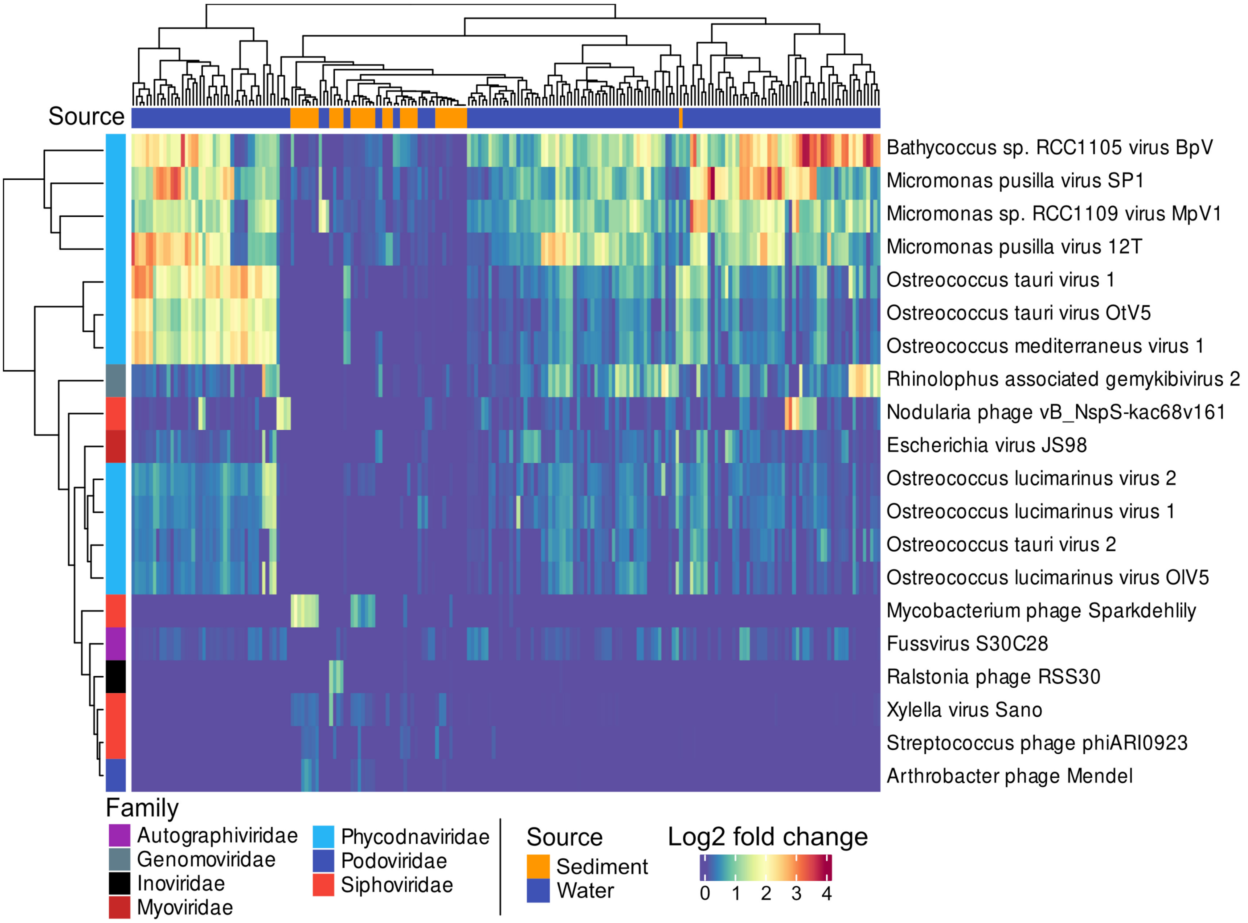


FIGURE 3 | A heatmap showing the 20 most significant, differentially abundant viral taxa between sediments and water samples. The Matrix was DESeq2 normalized showing the 20 most differentially abundant taxa with the smallest *p*.adj values (< *0.0001*). Sources of samples are indicated by blue or orange in the top color bar, and viral families by the color bar on the left side.

Viruses From Water and Sediment Carry Auxiliary Metabolic Genes Specific to the Environment

In both, the sediment and water metagenome AMGs could be assigned to 322 unique KEGG orthologs (**Supplementary Table 5**). The water column was the more diverse habitat with 173 unique KEGG orthologs assigned in total. While 36 unique KEGG orthologs could be identified in sediments, 113 identified AMGs were found in both habitats (**Figure 5A**). In the water column, AMGs of unknown function accounted for 13.1% of all mappable FPKM and 15.2% in sediments.

The most abundant AMGs are visualized by the percent occurrence of stations over their log sum of FPKM (**Figure 5B** and **Table 1**). Six outliers stand out among the AMGs identified in the water column: *dcm, cobS, gale, P4HA, gmd,* and *psbA*. The two most abundantly occurring AMGs in this study were encoding for *dcm/DNMT1* and *cobS* occurring in 84 and 75% of stations, while *gale, P4HA, gmd,* and *psbA* occurred in 62, 62, 60, and 54% of water respectively. In contrast, the most abundant outliers of sediment AMGs, *cysH, dcm,* and *queC* occurred in 65, 61, and 57% of stations, respectively. These genes encode for the phosphoadenosine phosphosulfate reductase, DNA (cytosine-5)-methyltransferase 1, and the 7-cyano-7-deazaguanine synthase.

Auxiliary Metabolic Gene Distribution Along the Salinity Gradient of the Baltic Sea

While the most predominant salinity gradient stretches from the south to the north, differences in salinity also occur by depth due to the higher density of saline water. "Amino acid" and "carbohydrate metabolism" AMGs as well as "Metabolism of cofactors and vitamins" appeared most consistently along the salinity gradient. The highest number of fragments were assigned to the "Energy metabolism" at salinities of 6.7, 6.71, and 12.3. However, while "Metabolism of cofactors and vitamins" did not occur in high numbers at certain salinities, 23.2% of all assigned FPKM were assigned to this pathway as it appears most consistently along the salinity gradient. It is closely followed by "Amino acid metabolism" with 18.1% and

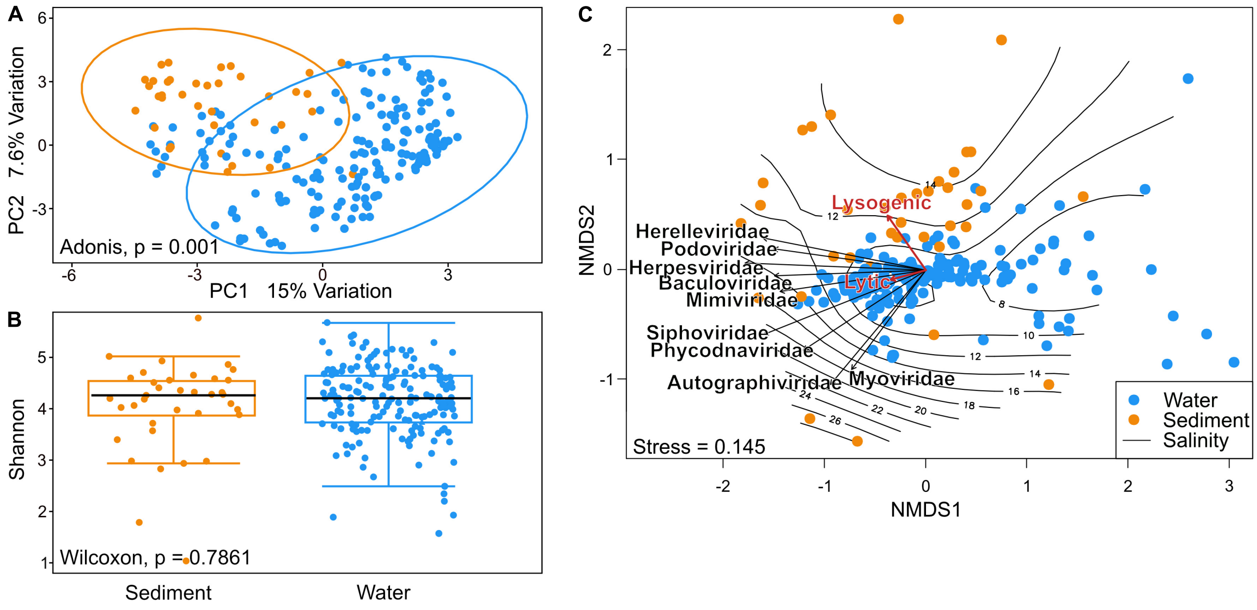


FIGURE 4 | (A) Principal component analysis of beta diversity using Aitchinson distance shows distinct patterns in community composition of sediment and water communities. *(B)* Shannon alpha diversity of Baltic Sea viral communities in Wilcoxon *p*-value of 0.7861 indicates no relevant difference in per-sample diversity when comparing species richness of sediment and water communities. **(C)** Non-metric multidimensional scaling (NMDS) analysis using Bray–Curtis dissimilarity. Vectors of the top nine most abundant families and of environmental factors were fitted to the ordination using the envfit function. A separation of sediment and water stations can be observed along the salinity gradient plotted using the ordisurf function. Envfit vectors indicate a tendency of viruses toward the lysogenic lifestyle in sediment metagenomes and lytic lifestyle in the water columns.

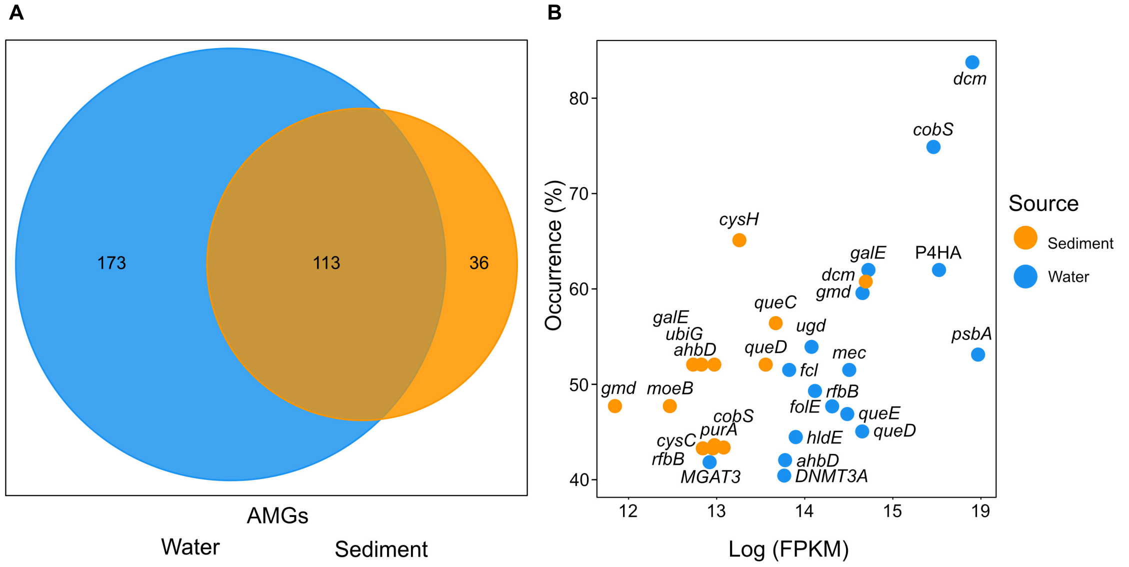


FIGURE 5 | (A) Number of unique KEGG orthologs specific to either water or sediment viral composition. The overlap displays the number of unique KEGG orthologs found in both sediments and the water column. **(B)** The distribution of the most abundant KEGG orthologs plotted by station occurrence over Fragments Per Kilobase per Million reads (FPKM) assigned. The x-axis displays the log FPKM count of KEGG orthologs. The *y*-axis indicates the occurrence in stations (%) of samples in which the KEGG orthologs were found calculated per source. The cutoff was chosen at 40% of stations. Colors indicate the gene origin.

"Energy metabolism" and unknown pathways with 14.9 and 14.1%, respectively (**Figure 6A** and **Supplementary Table 6**). PCA of Hellinger transformed AMG counts clustered most of the analyzed metagenomic samples evenly throughout the ordination (**Figure 6B**). While most pathway vectors cluster relatively closely together, "Amino acid metabolism", "Metabolism of cofactors and vitamins", and "Carbohydrate metabolism" are separated from the other vectors.

TABLE 1 | Most abundant KEGG orthologs with assigned metabolic pathway, the sum of assigned FPKM, and occurrence (% samples) in which the KEGG orthologs were found.

Gene name	KEGG ortholog	Log (FPKM)	Station occurrence (%)
Sediment			
cysH	K00390	12.85	65.2
DNMT1, dcm	K00558	14.29	60.9
queC	K06920	13.27	56.5
ubiG	K00568	12.42	52.2
queD	K01737	13.15	52.2
galE	K01784	12.33	52.2
ahbD	K22227	12.57	52.2
Water			
DNMT1, *dcm*	K00558	15.50	83.9
cobS	K09882	15.05	75.0
P4HA	K00472	15.12	62.1
galE	K01784	14.32	62.1
gmd	K01711	14.25	59.7
ugd	K00012	13.67	54.0
psbA	K02703	15.56	53.2
fcl	K02377	13.42	51.6
mec	K21140	14.10	51.6

Cyanobacteria and Proteobacteria Are the Most Abundant Prokaryotic Hosts

Prokaryotic hosts of AMGs were assigned with the VirHostMatcher-Net tool, allowing the identification of hosts affected most by viral metabolic interference. In the water column, *Cyanobacteria* were found to be the hosts with the highest number of assigned FPKM with 43.9% of all FPKM assigned to them. *Proteobacteria* and *Bacteroidetes* followed closely with 34 and 20.2% of total assigned FPKM respectively. In sediments, most FPKM were assigned to *Proteobacteria* (53.5%), Cyanobacteria (24.1%), and Bacteroidetes (21.9%) (**Figures 7A,B** and **Supplementary Table 7**).

Most FPKM Assigned to Auxiliary Metabolic Genes of Myo- and Siphoviridae

In the water column, 48.8% of FPKM were assigned to AMGs identified in the *Myoviridae* family and 21.7% in the *Siphoviridae* family. While some FPKM were also assigned to AMGs of other viral families (i.e., *Herelleviridae* and *Ackermannviridae*), they only accounted for 1.9 and 0.6%, respectively (**Figure 7A**). In the sediment, 40.5% of FPKM could be assigned to *Siphoviridae* and 23% to *Myoviridae*, while *Podoviridae* and *Herelleviridae* only accounted for 1.88 and 0.31%, respectively (**Supplementary Table 7**).

Myoviridae procured the most diverse pathways. AMGs assigned to this family were found in nine out of twelve KEGG pathways that are detected in all metagenomes. The second most versatile family of viruses was *Siphoviridae*, which carried AMGs belonging to 8 out of 12 of all KEGG pathways, while *Siphoviridae* carried AMGs from 7 out of 12 KEGG hierarchies. Notably, more AMGs of the “Metabolism of cofactors and vitamins” pathways were present in *Myoviridae* in sediments, whereas AMGs of “Biosynthesis of other secondary metabolites” and “Energy metabolism” in the water columns were more pronounced compared to sediments. AMGs assigned to the “Amino acid metabolism” and “Metabolism of cofactors and vitamins” were the most abundantly occurring type of AMGs that were present in the top four most abundant viral families, which made up 98% of all assigned AMGs.

DISCUSSION

Diverse Viral Composition in Sediments and the Water Column

In general, *Myoviridae* and *Siphoviridae* were detected as the most abundant viral families within the whole metagenomic data set. However, while some overlaps occur, the viral composition of sediment and water stations clustered separately from each other. Even though the viral assembly in sediments and the water columns were similar in species richness, they differed in beta diversity. Minor differences in viral diversity of the samples can be expected due to varying sampling methods and filter fractions used in the individual projects, that are summarized in this meta-analysis. Specifically, differences in the viral composition of sediments and the water column were defined through the differentially abundant *Phycodnaviridae*, which contributed up to 80% of total relative abundance at the surface. While in compositional datasets, relative abundances cannot be used to infer absolute viral absolute viral particles, cell counts, or gene abundances, we noted decreasing relative abundances of *Phycodnaviridae* with decreasing water depth. The high relative abundances of *Phycodnaviridae* in the subsurface sediments at 10.1 mbsf are likely the result of high sedimentation rates in the Baltic Sea and a lack of benthic animals due to limited oxygen supply, allowing the undisturbed formation of sediment layers. While most sediment stations in our study were similar to each other, small increases of Ralstonia phage RSS30 were observed in subsurface stations sampled from Aarhus Bay sediments, Denmark. Cyanobacteria in our study were abundant hosts, yet the identification of *Prochloraceae* as hosts likely resulted from a database bias, as previous studies have not detected them in the Baltic Sea (Bertos-Fortis et al., 2016; Celepli et al., 2017). This interpretation is supported by Celepli et al. (2017), who have generated hits of *Prochloraceae* in their metagenomic data set, while their 16S rRNA analysis also revealed the absence of *Prochloraceae*. Other unicellular Cyanobacteria (*Synechococcus* and *Cyanobium*) are frequently found in datasets of the Baltic Sea and likely contributed as cyanophage hosts also in these metagenomes (Haverkamp et al., 2009; Larsson et al., 2014; Broman et al., 2021).

The Lytic Lifestyle Prevails in the Baltic Sea

In the past, viruses have mostly been studied under laboratory conditions with a focus on three life cycles: chronic, lytic, and lysogenic. During the chronic lifestyle, phages enter

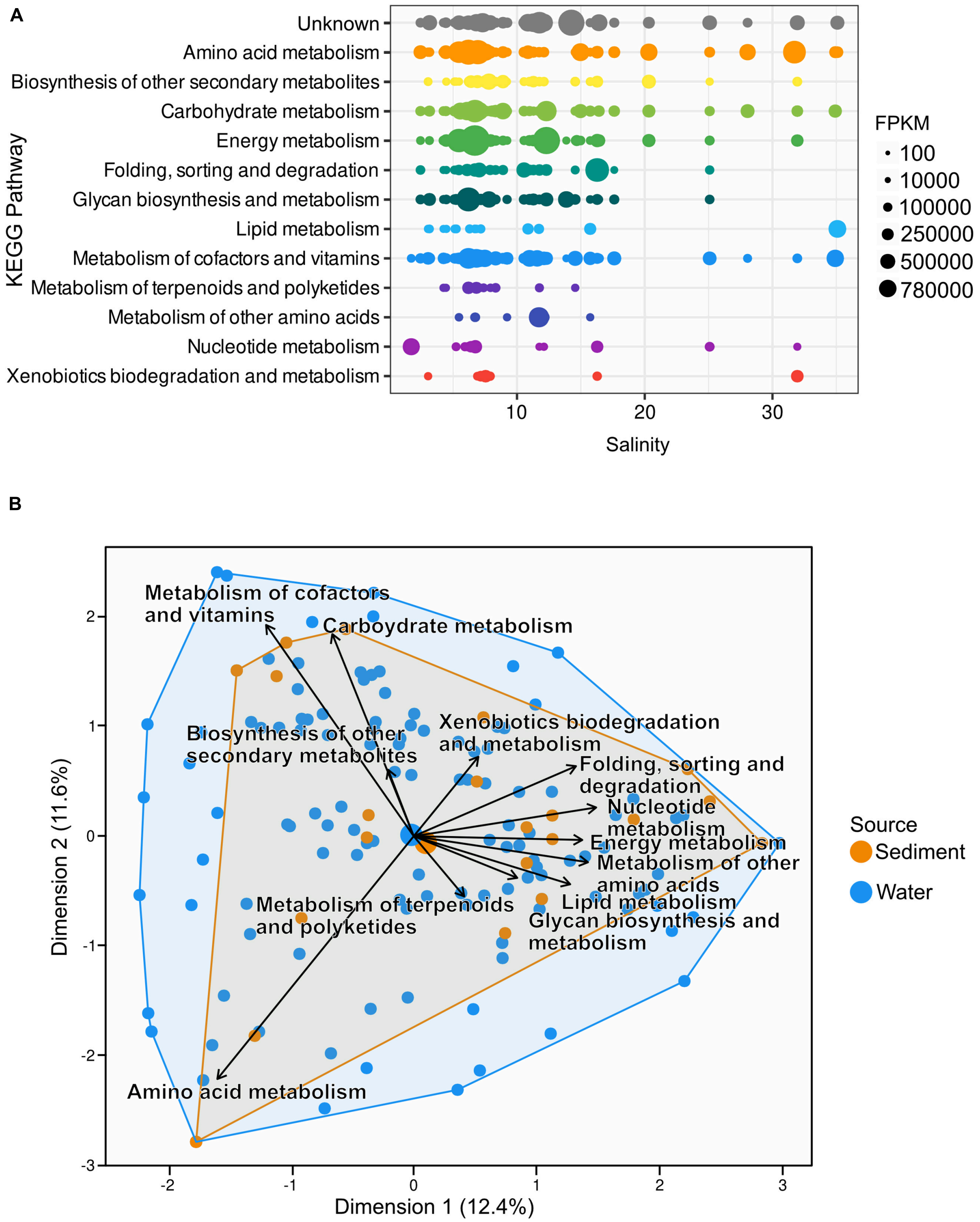


FIGURE 6 | (A) Distribution of AMGs along the salinity gradient of the Baltic Sea; both water and sediment samples are depicted together. The number of AMGs was normalized to AMG per million reads. **(B)** Principal component analysis of Hellinger transformed AMG counts shows most samples cluster close to the center of the ordination. However, vectors of metabolic pathways indicate no uniform presence of metabolic pathways at all stations.

a productive replication cycle, releasing virions without lysing their host. Exhibiting the lytic lifestyle, viruses lyse their hosts upon infection, releasing viral progeny into the environment (Correa et al., 2021). In coastal waters, Wilcox and Fuhrman (1994) reported the lytic lifestyle to be the most abundant. During the lysogenic cycle, a non-productive infection occurs by integrating the virus into the host's genome and replicating it along with the host. A virus may exit the lysogenic cycle and become lytic through specific factors or by spontaneous switching of lifestyles. Previous laboratory studies have looked at abiotic factors, where life lifestyle switching was influenced by phosphate availability or salinity

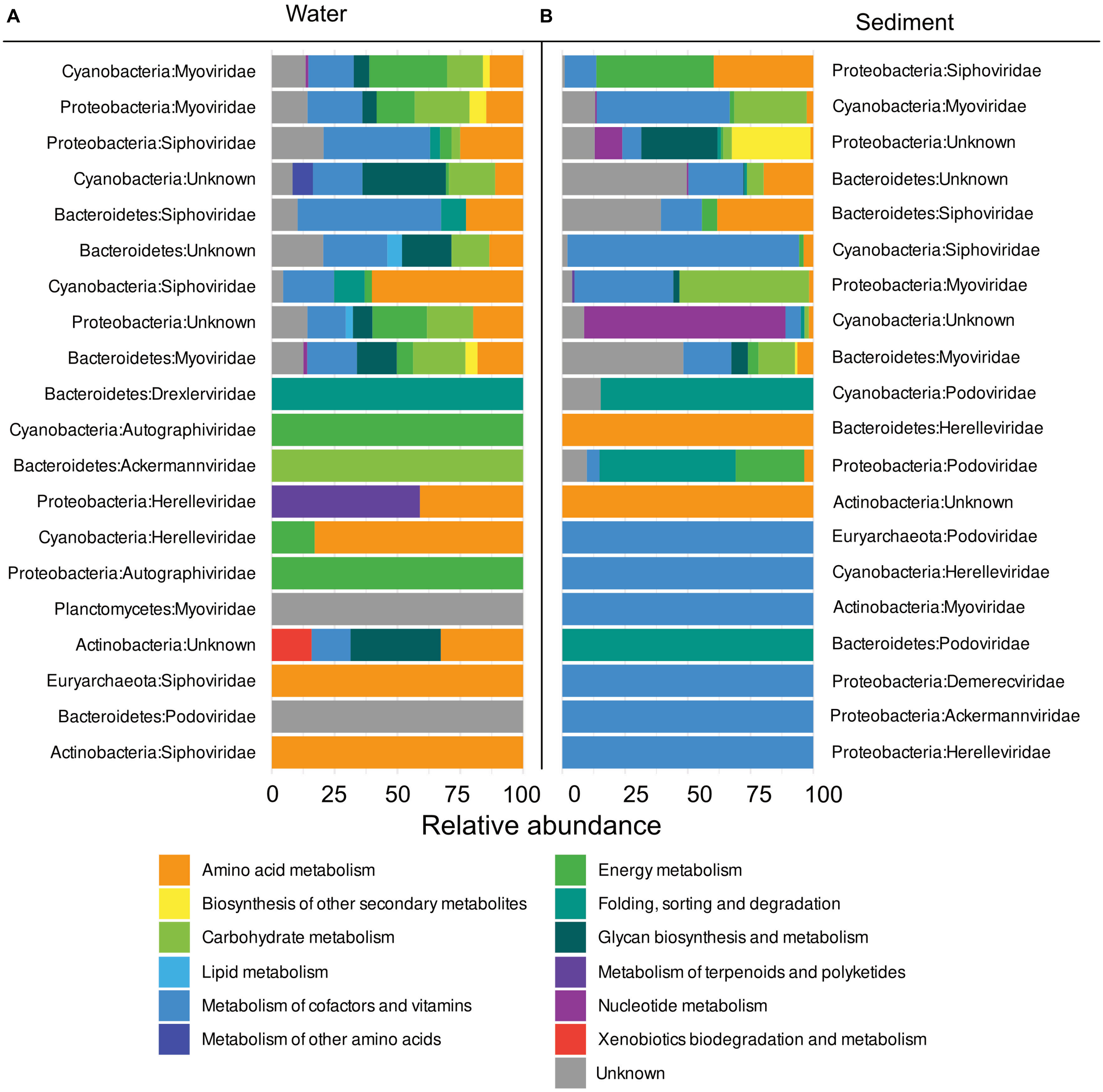


FIGURE 7 | Relative abundance of metabolic pathways per host phylum and viral family in descending order by the sum of total FPKM assigned to metabolic pathways in **(A)** the water columns and **(B)** sediments.

(Wilson et al., 1996; McDaniel et al., 2002; Bettarel et al., 2011).

However, while lytic, lysogenic, and chronic lifestyles are reflective of viral behavior in the laboratory, they are not entirely representative of natural behavior. Instead, viral lifestyles are controlled by complex interactions and represent a continuum rather than infection categories (Weitz et al., 2019; Correa et al., 2021). External mechanisms such as diel and seasonal changes may influence viral lifestyles (Ballaud et al., 2016; Brum et al., 2016; Puxty et al., 2018). However, recent studies postulate that switching between lysogeny and lysis is especially influenced by host density (Erez et al., 2017). In the environment, lysogeny has been found to be a low-density refugium occurring at low host abundance. The refugium theory assumes exponentially growing communities to be rich in intracellular energy which favors lysis, whereas communities of low abundances are depleted of intracellular energy sources, which favors lysogeny (Erez et al., 2017). Yet, lysogeny has been found to be a survival strategy in low and high host-density conditions (Mizuno et al., 2016; Kim and Bae, 2018; Coutinho et al., 2019; Luo et al., 2020).

Lysogeny, as a result of high host density, has been described as the "piggyback-the-winner" model (Knowles et al., 2016). Additionally, the "killing the winner" theory predicts that hosts of the highest density are lysed (Thingstad, 2000). Low density and energy availability favor lysogeny, but increasing host density facilitates induction and lysis as denser communities administer more internal energy. Lysogeny continues to decrease until host densities of $\sim 10^6$ cells ml^{-1} are reached, which are typically observed in the open oceans (Luque and Silveira, 2020). Higher densities increase the chances of coinfections with other viruses. Thus, lysogeny becomes a favorable lifestyle (Luque and Silveira, 2020). This switching might be communicated among phages as observed by Erez et al. (2017). Here, the *Bacillus* infecting phages of the SPbeta group released small peptides into the medium, signaling switching to the lysogenic lifestyle at higher concentrations of the respective compounds. This mechanism has been identified in different phages, each of which utilized different versions of the communication peptide (Erez et al., 2017).

Considering the abovementioned complexity of viral lifestyles, the observed dominance of identified lytic viral contigs in the Baltic Sea (**Supplementary Figure 1**) provides just a snapshot derived from genomic sequences of complex and dynamic systems. For instance, the metagenomic samples which are the basis of this study were taken at different time points and locations, thus allowing only remarks about the moment of sampling. Furthermore, the methodological limitation of the VIBRANT tool used to classify contigs as lytic or lysogenic has to be considered. VIBRANT assigns the lysogenic lifestyle by using surrounding host genome elements or integrases as evidence, limiting the identification of lysogenic contigs. The detected viral contigs might be lysogenic but the absence of the aforementioned properties in partial genomes could lead VIBRANT to falsely categorize them as lytic rather than lysogenic.

Auxiliary Metabolic Genes Catalyze Virus-Host Interactions

Viruses utilize AMGs to alter their host's rate-limited cellular processes during infection (Sullivan et al., 2006). The roles of such AMGs are not random but critical for the successful proliferation of the viruses. Here, the most abundantly distributed AMG was the *dcm* gene, encoding for a methyltransferase. These enzymes are ubiquitously found in prokaryotes and are often associated with cognate restriction endonucleases, forming a restriction-modification system that protects bacterial cells from foreign DNA invasion. In bacteriophages, the so-called orphan methyltransferases appear without these endonucleases and are involved in regulatory activities to protect the phage DNA from being digested (Schlagman et al., 1986; Boye and Løbner-Olesen, 1990; Palmer and Marinus, 1994; Kossykh et al., 1995). While methylation is a well-known way of escaping host restriction in viruses, they also procured other means of nucleotide modifications. The genes *folE, queD, queE,* and *queC*, are necessary for 7-cyano-7-deazaguanine (preQ$_0$) synthesis. These genes were among the most abundantly occurring AMGs in our dataset with the genes *queD* and *queC* both occurring in 52% of sediment and *queE* in 46% of water samples. This indicates that preQ$_0$ is important in viral replication. Queuosine is a hypermodified guanosine found in tRNAs specific for four amino acids (Asp, Asn, His, Tyr) and increases translation efficiency (Sabri et al., 2011; El Yacoubi et al., 2012). The presence of preQ$_0$ synthesis genes in viruses has been reported previously and protects the virus from host restriction enzymes (Hutinet et al., 2019).

Photosynthesis genes, such as the *cobS* gene among the *psbA* gene, are considered core AMGs in cyanophages (Ignacio-Espinoza and Sullivan, 2012). In our study, these occurred especially abundantly in water samples. The *cobS* gene encodes for a protein that catalyzes the final step in bacterial cobalamin (vitamin B12) biosynthesis (Magnusdottir et al., 2015). Speculations about the involvement of viruses in the cobalamin biosynthesis in the pelagic ecosystem are tempting, yet more targeted analyses and experimental evidence would be needed for a conclusive answer. The *psbA* gene is among the best-studied AMGs. It encodes for the photosystem II protein D1. Together with photosystem II protein D2, it forms a heterodimer and binds P680, which is a specific chlorophyll a and the primary electron donor of photosystem II. Marine picocyanobacteria, such as those of the genus *Synechococcus,* are among the most abundant photosynthetic organisms on Earth and are responsible for the fixation of approximately 25% of the carbon in the marine environment (Scanlan et al., 2009; Flombaum et al., 2013). Viral production in *Cyanobacteria* is limited by the availability of energy for protein synthesis during late infection. Cyanophage production correlates with irradiation intensity and is inhibited by darkness (Puxty et al., 2016; Thompson et al., 2016). To circumvent energy limitations, cyanophages augment their hosts' metabolism by introducing genes for the photosynthetic light reactions. In the early stage of infection, CO_2 fixation can be actively inhibited by the phages, diverting the hosts' metabolism toward the pentose phosphate pathway, thus increasing NADPH and ribose-5-phosphate production, facilitating viral protein and DNA synthesis rather than increasing photosynthetic activity (Thompson et al., 2011). The regulatory ability of cyanophages on global carbon cycling and primary production through lysis and active augmentation of carbon fixation rates implies the importance of these phages. The AdoMet-dependent heme synthase *ahbD* is involved in protoheme biosynthesis by catalyzing the conversion of Fe-coproporphyrin III into heme. This has been studied in sulfate-reducing bacteria of the *Desulfivibrio* genus and in methanogenic Archaea (Buchenau et al., 2006). Heme is an essential prosthetic group and, among other biological processes such as respiration, is very important in photosynthesis (Layer et al., 2010). Procuring such genes could increase the energy metabolism and speed up virus production by reducing the latent period (Mann et al., 2003; Lindell et al., 2004a; Millard et al., 2004). However, while we found 261 instances of the *ahbD* AMG, only three contigs contained both, the radical SAM domain PF13186 and radical SAM domain PF04055, and were not classified as Archaea nor of *Desulfovibrio.* Hence, inferences about the function of this AMG are rather speculative. While the *ahbD* gene requires further investigation, the genes *psbA* and *cobS* highlight the importance and distribution of

AMGs involved in photosynthesis in pelagic phages, emphasized also by high cyanobacteria abundances in the Baltic Sea.

The *ubiG, galE,* and *P4HA* genes cannot easily be assigned to greater metabolic functions such as photosynthesis but appear to be of similar importance. The *ubiG* gene encoding for the last step in the pathway of ubiquinone biosynthesis likely provides the phages with the ability to affect the electron transport chain. The *galE* gene encoding the UDP-glucose 4-epimerase placed third among the most abundant AMGs in the water column. It mediates the conversion of UDP-galactose and UDP-glucose in galactose metabolism (Thoden and Holden, 1998). Thus, the introduction of *galE* likely allows the virus to participate in the carbohydrate metabolism to generate energy. The similarly abundant *P4HA* gene, encoding for the prolyl 4-hydroxylase, catalyzes the hydroxylation of proline residues in peptide linkages in collagens, forming 4-hydroxyproline (Myllyharju, 2003). In viruses, collagen can be part of the tail fibers and was first detected in Paramecium bursaria Chlorella virus-1 (Eriksson et al., 1999; Rasmussen et al., 2003). In which way viruses utilize prolyl 4-hydroxylase remains cryptic. However, biological consequences of prolyl hydroxylation include altering protein conformation and protein–protein interactions but also contributing to collagen-helix stability in general (Rao and Adams, 1979).

Analogous to water samples in our study, nucleotide modification through the *dcm* and *queC* gene are the most important functions provided by viruses. As photosynthesis is less relevant in sediments, especially those of greater depths, other functions prevail. Here, the most abundantly occurring AMG not involved in DNA modification is the *cysH* gene, encoding for the phosphoadenosine 5′-phosphosulfate reductase. It was found in 65% of all sediment stations but also in 39,5% of the water stations. The presence of the *cysH* gene suggests the viral involvement in sulfur cycling, especially in viruses found in Baltic Sea sediments. The enzyme is involved in the synthesis of sulfite from phosphoadenosine 5’-phosphosulfate (PAPS) and thus part of the sulfate reduction pathway (Bick et al., 2000). The *cysH* gene has been identified in phages infecting members of the SAR11 clade, which lack the phosphoadenosine 5’-phosphosulfate reductase and other genes required in assimilatory sulfate reduction but has recently also been found to be generally widespread in marine phages (Du et al., 2021; Kieft et al., 2021).

Conserved Core Auxiliary Metabolic Genes

Recently, Kieft et al. (2020) identified a set of AMGs found in highly different viral assemblies from various origins, i.e., human gut, marine sediment, deep subsurface, and others. Their set of globally conserved AMGs consisted of the *dcm, cysH, folE, phnP, ubiG, ubiE, waaF, moeB, ahbD, cobS, mec, queE, queD, queC* genes and occurred in at least 10 out of 12 of their studied samples. These genes were also identified in the Baltic Sea, though not all of them are among the most abundant. However, the genes *dcm, cysH, folE, cobS, mec, queE, queD, queC* are concurrent with the findings of Kieft et al. (2020) and suggest the existence of a globally distributed set of conserved core AMGs, which are present regardless of host and environment. Locally, the core AMGs might then be extended by genes specific to an environment of interest such as genes for photosynthesis. A definition of core AMGs is difficult, as setting a threshold for their occurrence at a given station or environment would be arbitrary. However, the similarity in composition of most abundantly occurring AMGs in the Baltic Sea and other environments is striking.

CONCLUSION

The metagenomic analysis revealed a predominantly lytic viral life mode in the Baltic Sea, possibly aided by high nutrient availabilities and increasing lysogeny traits in sediments. We did not find major virus community differences along the north-southerly salinity gradient of the Baltic Sea. Yet, the composition of pelagic and benthic virus assemblies differed, especially in the relative abundances of *Phycodnaviridae*. Also, the functional virus AMGs differed between the pelagic and benthic samples. While viruses from the water columns procured AMGs specific for photosynthesis, viruses in sediments acquired AMGs that are part of the nutrient cycling pathways such as sulfur cycling. Other AMGs that exclusively occurred in sediments or the water column were found in low abundances and are likely linked to functions that specifically increase virus fitness in the respective ecosystem. Viruses use AMGs to evade host restriction mechanisms, i.e., by modifying their DNA through methylation or utilization of preQ$_0$. These DNA modification AMGs were highly abundant in the Baltic Sea and have also been observed to be globally conserved. Our findings, therefore, strengthen the hypothesis on the existence of global core AMGs that are central to viral replication, regardless of environment and host.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

BH performed the conceptualization, datamining, bioinformatics, and analyses. BE secured the funding. BH and BE wrote the manuscript. CB helped conceptualize the study and helped with data evaluation. All authors read and approved the manuscript.

FUNDING

This research was supported by a grant from the German Research Foundation (DFG) as part of the Collaborative Research Center TRR51 Roseobacter. Funding to CB was generously provided by the Ministry for Science and Culture of Lower Saxony Vorab grant "Ecology of Molecules, EcoMol."

ACKNOWLEDGMENTS

We thank the personnel and shipboard scientists for sampling and data provision of the public metagenomes included in our study: OSD-CONSORTIUM, Kungliga Tekniska Hogskolan Science for Life Laboratory, UNIMIB, Aarhus University, Stockholm University, JGI, University of Southern California, IODP Leg 347. We would like to express our gratitude to the scientist at the mentioned institutes who sampled and sequenced the metagenomic data that is the basis of this study. We thank the reviewers for their constructive feedback, which helped improve our manuscript. Furthermore, we also thank Leon Dlugosch for his help with scripting and discussing general ideas.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.863620/full#supplementary-material

REFERENCES

Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K., et al. (2020). Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Commun. Biol.* 3:119. doi: 10.1038/s42003-020-0856-x

Alneberg, J., Sundh, J., Bennke, C., Beier, S., Lundin, D., Hugerth, L. W., et al. (2018). Barm and balticmicrobedb, a reference metagenome and interface to meta-omic data for the baltic sea. *Sci. Data* 5:180146. doi: 10.1038/sdata.2018.146

Anantharaman, K., Duhaime, M. B., Breier, J. A., Wendt, K. A., Toner, B. M., and Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science* 344, 757–760. doi: 10.1126/science.1252229

Andren, T., Barker Jørgensen, B., Cotterill, C., Green, S., and The IODP expedition 347 scientific party (2015). IODP expedition 347: Baltic sea' basin paleoenvironment and biosphere. *Sci. Dril.* 20, 1–12.

Andrews, S. (2010). *Babraham Bioinformatics-fastqc a Quality Control Tool for High Throughput Sequence Data*. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed April, 2020).

Asplund-Samuelsson, J., Sundh, J., Dupont, C. L., Allen, A. E., McCrow, J. P., Celepli, N. A., et al. (2016). Diversity and expression of bacterial metacaspases in an aquatic ecosystem. *Front. Microbiol.* 7:1043. doi: 10.3389/fmicb.2016.01043

Ballaud, F., Dufresne, A., Francez, A.-J., Colombet, J., Sime-Ngando, T., and Quaiser, A. (2016). Dynamics of viral abundance and diversity in a sphagnum-dominated peatland: temporal fluctuations prevail over habitat. *Front. Microbiol.* 6:1494. doi: 10.3389/fmicb.2015.01494

Bauer, R. K. (2020). *"Oceanmap: a Plotting Toolbox for 2D Oceanographic Data." R Package Version.*

Bertos-Fortis, M., Farnelid, H. M., Lindh, M. V., Casini, M., Andersson, A., Pinhassi, J., et al. (2016). Unscrambling cyanobacteria community dynamics related to environmental factors. *Front. Microbiol.* 7:625. doi: 10.3389/fmicb.2016.00625

Bettarel, Y., Bouvier, T., Bouvier, C., Carre, C., Desnues, A., Domaizon, I., et al. (2011). Ecological traits' of planktonic viruses and prokaryotes along a full-salinity gradient. *FEMS Microbiol. Ecol.* 76, 360–372.

Bick, J. A., Dennis, J. J., Zylstra, G. J., Nowack, J., and Leustek, T. (2000). Identification of a new class of 5'-adenylylsulfate (aps) reductases from sulfate-assimilating bacteria. *J. Bacteriol.* 182, 135–142. doi: 10.1128/JB.182.1.135-142.2000

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* 37, 632–639. doi: 10.1038/s41587-019-0100-8

Boye, E., and Løbner-Olesen, A. (1990). The role of dam methyltransferase in the control of dna replication in e. coli. *Cell* 62, 981–989. doi: 10.1016/0092-8674(90)90272-g

Breitbart, M. (2012). Marine viruses: truth or dare. *Annu. Rev. Mar. Sci.* 4, 425–448. doi: 10.1146/annurev-marine-120709-142805

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* 3, 754–766. doi: 10.1038/s41564-018-0166-y

Breitbart, M., Thompson, L. R., Suttle, C. A., and Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. *Oceanography* 20, 135–139.

Broman, E., Holmfeldt, K., Bonaglia, S., Hall, P. O., and Nascimento, F. J. (2021). Cyanophage diversity and community structure in dead zone sediments. *mSphere* 6:e00208-21. doi: 10.1128/mSphere.00208-21

Brum, J. R., Hurwitz, B. L., Schofield, O., Ducklow, H. W., and Sullivan, M. B. (2016). Seasonal time bombs: dominant temperate viruses affect southern ocean microbial dynamics. *ISME J.* 10, 437–449.

Buchenau, B., Kahnt, J., Heinemann, I. U., Jahn, D., and Thauer, R. K. (2006). Heme biosynthesis in *Methanosarcina barkeri* via a pathway involving two methylation reactions. *J. Bacteriol.* 188, 8666–8668. doi: 10.1128/JB.01349-06

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Bunse, C., Israelsson, S., Baltar, F., Bertos-Fortis, M., Fridolfsson, E., Legrand, C., et al. (2019). High frequency multi-year variability in baltic sea microbial plankton stocks and activities. *Front. Microbiol.* 9:3296. doi: 10.3389/fmicb.2018.03296

Bushnell, B. (2018). *Bbtools: A Suite of Fast, Multithreaded Bioinformatics Tools Designed for Analysis of DNA and RNA Sequence Data*. Available online at: https://sourceforge.net/projects/bbmap/ (accessed April, 2020).

Celepli, N., Sundh, J., Ekman, M., Dupont, C. L., Yooseph, S., Bergman, B., et al. (2017). Meta-omic analyses of baltic sea cyanobacteria: diversity, community structure and salt acclimation. *Environ. Microbiol.* 19, 673–686. doi: 10.1111/1462-2920.13592

Correa, A. M., Howard-Varona, C., Coy, S. R., Buchan, A., Sullivan, M. B., and Weitz, J. S. (2021). Revisiting the rules of life for viruses of microorganisms. *Nat. Rev. Microbiol.* 19, 501–513. doi: 10.1038/s41579-021-00530-x

Coutinho, F. H., Rosselli, R., and Rodríguez-Valera, F. (2019). Trends of microdiversity reveal depth-dependent evolutionary strategies of viruses in the Mediterranean. *mSystems* 4:e00554-19. doi: 10.1128/mSystems.00554-19

Crummett, L. T., Puxty, R. J., Weihe, C., Marston, M. F., and Martiny, J. B. (2016). The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology* 499, 219–229. doi: 10.1016/j.virol.2016.09.016

Danovaro, R., Corinaldesi, C., Filippini, M., Fischer, U. R., Gessner, M. O., Jacquet, S., et al. (2008a). Viriobenthos in freshwater and marine sediments: a review. *Freshw. Biol.* 53, 1186–1213.

Danovaro, R., Dell'Anno, A., Corinaldesi, C., Magagnini, M., Noble, R., Tamburini, C., et al. (2008b). Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* 454, 1084–1087. doi: 10.1038/nature07268

Du, S., Qin, F., Zhang, Z., Tian, Z., Yang, M., Liu, X., et al. (2021). Genomic diversity, life strategies and ecology of marine htvc010p-type pelagiphages. *Microb. Genom.* 7:000596. doi: 10.1099/mgen.0.000596

El Yacoubi, B., Bailly, M., and de Crecy-Lagard, V. (2012). Biosynthesis and function of posttranscriptional' modifications of transfer rnas. *Annu. Rev. Genet.* 46, 69–95.

Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. B., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* 3, 870–880. doi: 10.1038/s41564-018-0190-y

Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., et al. (2017). Communication between viruses guides lysis–lysogeny decisions. *Nature* 541, 488–493.

Eriksson, M., Myllyharju, J., Tu, H., Hellman, M., and Kivirikko, K. I. (1999). Evidence for 4hydroxyproline in viral proteins: characterization of a viral prolyl 4-hydroxylase and its peptide substrates. *J. Biol. Chem.* 274, 22131–22134. doi: 10.1074/jbc.274.32.22131

Espínola, F., Dionisi, H. M., Borglin, S., Brislawn, C. J., Jansson, J. K., Mac Cormack, W. P., et al. (2018). Metagenomic analysis of subtidal sediments from polar and subpolar coastal environments highlights the relevance of anaerobic hydrocarbon degradation processes. *Microb. Ecol.* 75, 123–139. doi: 10.1007/s00248-017-1028-5

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci. U.S.A.* 110, 9824–9829. doi: 10.1073/pnas.1307701110

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31, 3718–3720. doi: 10.1093/bioinformatics/btv428

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., et al. (2019). Marine DNA viral macro-and microdiversity from pole to pole. *Cell* 177, 1109–1123.e14. doi: 10.1016/j.cell.2019.03.040

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849. doi: 10.1093/bioinformatics/btw313

Haverkamp, T. H., Schouten, D., Doeleman, M., Wollenzien, J., Ute, denm Huisman, et al. (2009). Colorful microdiversity of synechococcus strains (picocyanobacteria) isolated from the baltic sea. *ISME J.* 3, 397–408. doi: 10.1038/ismej.2008.118

Hellweger, F. L. (2009). Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. *Environ. Microbiol.* 11, 1386–1394. doi: 10.1111/j.1462-2920.2009.01866.x

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., et al. (2015). Metagenomeassembled genomes uncover a global brackish microbiome. *Genome Biol.* 16:279. doi: 10.1186/s13059-015-0834-7

Hutinet, G., Kot, W., Cui, L., Hillebrand, R., Balamkundu, S., Gnanakalai, S., et al. (2019). 7-Deazaguanine modifications protect phage DNA from host restriction systems. *Nat. Commun.* 10:5442. doi: 10.1038/s41467-019-13384-y

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Ignacio-Espinoza, J. C., and Sullivan, M. B. (2012). Phylogenomics of t4 cyanophages: lateral gene transfer in the 'core' and origins of host genes. *Environ. Microbiol.* 14, 2113–2126. doi: 10.1111/j.1462-2920.2012.02704.x

Ilus, E., Mattila, J., Klemola, S., Ikaheimonen, T., and Niemisto, L. (2001). *Sedimentation Rate in the Baltic Sea. Tech. Rep. NKS–8*. Available online at: https://www.osti.gov/etdeweb/servlets/purl/20226330 (accessed October, 2021).

Jørgensen, B. B., Andren, T., and Marshall, I. P. (2020). Sub-seafloor biogeochemical processes and' microbial life in the Baltic Sea. *Environ. Microbiol.* 22, 1688–1706.

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). Vibrant: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8:90. doi: 10.1186/s40168-020-00867-0

Kieft, K., Zhou, Z., Anderson, R. E., Buchan, A., Campbell, B. J., Hallam, S. J., et al. (2021). Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat. Commun.* 12:3503. doi: 10.1038/s41467-021-23698-5

Kim, M. S., and Bae, J. W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J.* 12, 1127–1141. doi: 10.1038/s41396-018-0061-9

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobián-Güemes, A. G., et al. (2016). Lytic to temperate switching of viral communities. *Nature* 531, 466–470.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., et al. (2015). The ocean sampling day consortium. *Gigascience* 4:27. doi: 10.1186/s13742-015-0066-5

Kossykh, V. G., Schlagman, S. L., and Hattman, S. (1995). Phage t4 dna [n]-adenine6methyltransferase. overexpression, purification, and characterization. *J. Biol. Chem.* 270, 14389–14393.

Kostner, N., Scharnreitner, L., Jürgens, K., Labrenz, M., Herndl, G. J., and Winter, C. (2017). High viral abundance as a consequence of low viral decay in the Baltic sea Redoxcline. *PLoS One* 12:e0178467. doi: 10.1371/journal.pone.0178467

Labrenz, M., Jost, G., and Jürgens, K. (2007). Distribution of abundant prokaryotic organisms in the water column of the central Baltic Sea with an oxic-anoxic interface. *Aquat. Microb. Ecol.* 46, 177–190.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Larsson, J., Celepli, N., Ininbergs, K., Dupont, C. L., Yooseph, S., Bergman, B., et al. (2014). Picocyanobacteria containing a novel pigment gene cluster dominate the brackish water Baltic sea. *ISME J.* 8, 1892–1903. doi: 10.1038/ismej.2014.35

Layer, G., Reichelt, J., Jahn, D., and Heinz, D. W. (2010). Structure and function of enzymes in heme biosynthesis. *Protein Sci.* 19, 1137–1161. doi: 10.1002/pro.405

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., and Chisholm, S. W. (2004a). Photosynthesis genes in *Prochlorococcus* cyanophage. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11013–11018.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., and Chisholm, S. W. (2004b). Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11013–11018. doi: 10.1073/pnas.0401526101

Love, M., Anders, S., and Huber, W. (2014). Differential analysis of count data–the deseq2 package. *Genome Biol.* 15, 10–1186.

Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* 3:e104.

Luo, E., Eppley, J. M., Romano, A. E., Mende, D. R., and DeLong, E. F. (2020). Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME J.* 14, 1304–1315. doi: 10.1038/s41396-020-0604-8

Luque, A., and Silveira, C. B. (2020). Quantification of lysogeny caused by phage coinfections in microbial communities from biophysical principles. *mSystems* 5:e00353-20. doi: 10.1128/mSystems.00353-20

Magnusdottir, S., Ravcheev, D., de Crecy-Lagard, V., and Thiele, I. (2015). Systematic genome assessment' of b-vitamin biosynthesis suggests co-operation among gut microbes. *Front. Genet.* 6:148. doi: 10.3389/fgene.2015.00148

Mann, N. H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature* 424:741. doi: 10.1038/424741a

McDaniel, L., Houchin, L., Williamson, S., and Paul, J. (2002). Lysogeny in marine synechococcus. *Nature* 415, 496–496. doi: 10.1038/415496a

McMurdie, P. J., and Holmes, S. (2013). phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217

Meier, H., Feistel, R., Piechura, J., Arneborg, L., Burchard, H., Fiekas, V., et al. (2006). Ventilation of the baltic sea deep water: a brief review of present knowledge from observations and models. *Oceanologia* 48, 133–164.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). Metaquast: evaluation of metagenome assemblies. *Bioinformatics* 32, 1088–1090. doi: 10.1093/bioinformatics/btv697

Millard, A., Clokie, M. R., Shub, D. A., and Mann, N. H. (2004). Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11007–11012. doi: 10.1073/pnas.0401478101

Mizuno, C. M., Ghai, R., Saghaï, A., López-García, P., and Rodriguez-Valera, F. (2016). Genomes of abundant and widespread viruses from the deep ocean. *mBio* 7:e00805-16. doi: 10.1128/mBio.00805-16

Myllyharju, J. (2003). Prolyl 4-hydroxylases, the key enzymes of collagen biosynthesis. *Matrix Biol.* 22, 15–24. doi: 10.1016/s0945-053x(03)00006-4

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* 39, 578–585. doi: 10.1038/s41587-020-00774-7

Nilsson, E., Li, K., Fridlund, J., Sulčius, S., Bunse, C., Karlsson, C. M., et al. (2019). Genomic and seasonal variations among aquatic phages infecting the baltic sea Gammaproteobacterium *Rheinheimera* sp. strain BAL341. *Appl. Environ. Microbiol.* 85:e01003-19. doi: 10.1128/AEM.01003-19

Nilsson, E., Li, K., Hoetzinger, M., and Holmfeldt, K. (2022). Nutrient driven transcriptional changes during phage infection in an aquatic gammaproteobacterium. *Environ. Microbiol.* 24, 2270–2281. doi: 10.1111/1462-2920.15904

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R., et al. (2013). *Package 'vegan'. Community Ecology Package, version* 2.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Palmer, B. R., and Marinus, M. G. (1994). The dam and DCM strains of *Escherichia coli*—a review. *Gene* 143, 1–12. doi: 10.1016/0378-1119(94)90597-5

Pernthaler, J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nat. Rev. Microbiol.* 3, 537–546. doi: 10.1038/nrmicro1180

Puxty, R. J., Evans, D. J., Millard, A. D., and Scanlan, D. J. (2018). Energy limitation of cyanophage development: implications for marine carbon cycling. *ISME J.* 12, 1273–1286. doi: 10.1038/s41396-017-0043-3

Puxty, R. J., Millard, A. D., Evans, D. J., and Scanlan, D. J. (2016). Viruses inhibit CO2 fixation in the most abundant phototrophs on Earth. *Curr. Biol.* 26, 1585–1589. doi: 10.1016/j.cub.2016.04.036

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rao, N. V., and Adams, E. (1979). Collagen helix stabilization by hydroxyproline in (ala-hyp-gly)n. *Biochem. Biophys. Res. Commun.* 86, 654–660. doi: 10.1016/0006-291x(79)91763-7

Rasmussen, M., Jacobsson, M., and Bjorck, L. (2003). Genome-based identification and analysis of¨ collagen-related structural motifs in bacterial and viral proteins. *J. Biol. Chem.* 278, 32313–32316.

Reissmann, J. H., Burchard, H., Feistel, R., Hagen, E., Lass, H. U., Mohrholz, V., et al. (2009). Vertical mixing in the Baltic Sea and consequences for eutrophication - A review. *Prog. Oceanogr.* 82, 47–80.

Riemann, L., Holmfeldt, K., and Titelman, J. (2009). Importance of viral lysis and dissolved DNA for bacterioplankton activity in a P-limited estuary, Northern Baltic sea. *Microb. Ecol.* 57, 286–294. doi: 10.1007/s00248-008-9429-0

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., et al. (2000). The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* 45, 408–418.

Sabri, M., Hauser, R., Ouellette, M., Liu, J., Dehbi, M., Moeck, G., et al. (2011). Genome annotation and¨ intraviral interactome for the streptococcus pneumoniae virulent phage dp-1. *J. Bacteriol.* 193, 551–562.

Scanlan, D. J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W. R., et al. (2009). Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* 73, 249–299. doi: 10.1128/MMBR.00035-08

Schlagman, S. L., Hattman, S., and Marinus, M. G. (1986). Direct role of the *Escherichia coli* dam DNA methyltransferase in methylation-directed mismatch repair. *J. Bacteriol.* 165, 896–900.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharon, I., Battchikova, N., Aro, E.-M., Giglione, C., Meinnel, T., Glaser, F., et al. (2011). Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J.* 5, 1178–1190. doi: 10.1038/ismej.2011.2

Sørensen, J., Jørgensen, B. B., and Revsbech, N. P. (1979). A comparison of oxygen, nitrate, and sulfate respiration in coastal marine sediments. *Microb. Ecol.* 5, 105–115. doi: 10.1007/BF02010501

Šulčius, S., and Holmfeldt, K. (2016). Viruses of microorganisms in the baltic sea: current state of researchˆ and perspectives. *Mar. Biol. Res.* 12, 115–124.

Sullivan, M. B., Huang, K. H., Ignacio-Espinoza, J. C., Berlin, A. M., Kelly, L., Weigele, P. R., et al. (2010). Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* 12, 3035–3056. doi: 10.1111/j.1462-2920.2010.02280.x

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., and Chisholm, S. W. (2006). Prevalence and evolution of core photosystem ii genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* 4:e234. doi: 10.1371/journal.pbio.0040234

Suttle, C. (2005). Crystal ball. The viriosphere: the greatest biological diversity on Earth and driver of global processes. *Environ. Microbiol.* 7, 481–482. doi: 10.1111/j.1462-2920.2005.803_11.x

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812. doi: 10.1038/nrmicro1750

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* 45, 1320–1328.

Thoden, J. B., and Holden, H. M. (1998). Dramatic differences in the binding of udp-galactose and udp-glucose to udp-galactose 4-epimerase from *Escherichia coli*. *Biochemistry* 37, 11469–11477. doi: 10.1021/bi9808969

Thompson, L. R., Zeng, Q., and Chisholm, S. W. (2016). Gene expression patterns during light and dark infection of Prochlorococcus by cyanophage. *PLoS One* 11:e0165375. doi: 10.1371/journal.pone.0165375

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J. A., et al. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 108, E757–E764. doi: 10.1073/pnas.1102164108

Thureborn, P., Franzetti, A., Lundin, D., and Sjoling, S. (2016). Reconstructing ecosystem functions of the¨ active microbial community of the baltic sea oxygen depleted sediments. *PeerJ* 4:e1593.

Vali, G., Meier, M., and Elken, J. (2013). Simulated halocline variability in the baltic sea and its impact on hypoxia during 1961-2007. *J. Geophys. Res. Oceans* 118, 6982–7000.

Waldor, M. K., and Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272, 1910–1913. doi: 10.1126/science.272.5270.1910

Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J. C., Fuhrman, J. A., et al. (2020). A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom. Bioinform.* 2:lqaa044. doi: 10.1093/nargab/lqaa044

Weinbauer, M. G., Brettar, I., and Höfle, M. G. (2003). Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters. *Limnol. Oceanogr.* 48, 1457–1465.

Weitz, J. S., Li, G., Gulbudak, H., Cortez, M. H., and Whitaker, R. J. (2019). Viral invasion fitness across a continuum from lysis to latency. *Virus Evol.* 5:vez006. doi: 10.1093/ve/vez006

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Softw.* 4:1686.

Wilcox, R. M., and Fuhrman, J. A. (1994). Bacterial viruses in coastal seawater: lytic rather than lysogenic production. *Mar. Ecol. Prog. Ser.* 114, 35–45.

Wilhelm, S. W., and Suttle, C. A. (1999). Viruses and nutrient cycles in the sea. *Bioscience* 49, 781–788.

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., et al. (2008). The sorcerer II global ocean sampling expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* 3:1456. doi: 10.1371/journal.pone.0001456

Wilson, W. H., Carr, N. G., and Mann, N. H. (1996). The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium synechococcus sp. wh7803 1. *J. Phycol.* 32, 506–516.

Wommack, K. E., and Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* 64, 69–114. doi: 10.1128/MMBR.64.1.69-114.2000

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biol.* 20:257. doi: 10.1186/s13059-019-1891-0

Zeigler Allen, L., McCrow, J. P., Ininbergs, K., Dupont, C. L., Badger, J. H., Hoffman, J. M., et al. (2017). The baltic sea virome: diversity and transcriptional activity of DNA and RNA viruses. *mSystems* 2:e00125-16. doi: 10.1128/mSystems.00125-16

Zheng, X., Liu, W., Dai, X., Zhu, Y., Wang, J., Zhu, Y., et al. (2021). Extraordinary diversity of viruses in deep-sea sediments as revealed by metagenomics without prior virion separation. *Environ. Microbiol.* 23, 728–743. doi: 10.1111/1462-2920.15154

Zinke, L. A., Mullis, M. M., Bird, J. T., Marshall, I. P., Jørgensen, B. B., Lloyd, K. G., et al. (2017). Thriving or surviving? evaluating active microbial guilds in baltic sea sediment. *Environ. Microbiol. Rep.* 9, 528–536. doi: 10.1111/1758-2229.12578

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

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

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 04 August 2022
DOI 10.3389/fmicb.2022.874319

Expanding the environmental virome: Infection profile in a native rainforest tree species

Anderson Carvalho Vieira[1†], Ícaro Santos Lopes[2†], Paula Luize Camargos Fonseca[1,3], Roenick Proveti Olmo[4], Flora Bittencourt[1], Letícia Maróstica de Vasconcelos[1], Carlos Priminho Pirovani[1], Fernanda Amato Gaiotto[1*†] and Eric Roberto Guimarães Rocha Aguiar[1*†]

[1]Department of Biological Science, Center of Biotechnology and Genetics, Universidade Estadual de Santa Cruz, Ilhéus, Brazil, [2]Department of Biochemistry and Immunology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [3]Department of Genetics, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [4]Université de Strasbourg, CNRS UPR9022, Inserm, Strasbourg, France

OPEN ACCESS

EDITED BY
Minh-Thu Nguyen,
University Hospital Münster, Germany

REVIEWED BY
Pedro Luis Ramos-González,
Biological Institute of São Paulo, Brazil
Laura Miozzi,
Institute for Sustainable Plant Protection (CNR), Italy

*CORRESPONDENCE
Fernanda Amato Gaiotto
gaiotto@uesc.br
Eric Roberto Guimarães Rocha Aguiar
ericgdp@gmail.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

RECEIVED 12 February 2022
ACCEPTED 29 June 2022
PUBLISHED 04 August 2022

CITATION
Vieira AC, Lopes ÍS, Fonseca PLC, Olmo RP, Bittencourt F, de Vasconcelos LM, Pirovani CP, Gaiotto FA and Aguiar ERGR (2022) Expanding the environmental virome: Infection profile in a native rainforest tree species.
Front. Microbiol. 13:874319.
doi: 10.3389/fmicb.2022.874319

Agroforestry systems (AFS) for cocoa production combine traditional land-use practices with local biodiversity conservation, resulting in both ecological and agricultural benefits. The cacao-cabruca AFS model is widely implemented in regions of the Brazilian Atlantic Forest. *Carpotroche brasiliensis* (Raddi) A. Gray (Achariaceae) is a tree found in cabruca landscapes that is often used for reforestation and biotechnological applications. Despite its importance, we still lack information about viruses circulating in *C. brasiliensis*, particularly considering the possibility of spillover that could affect cocoa production. In our study, we analyzed the *Carpotroche brasiliensis* virome from Atlantic Forest and cacao-cabruca AFS regions using metatranscriptomics from several vegetative and reproductive organs. Our results revealed a diverse virome detecting near-complete or partial coding sequences of single- and double-stranded DNA and RNA viruses classified into at least six families (*Botourmiaviridae*, *Bromoviridae*, *Caulimoviridae*, *Genomoviridae*, *Mitoviridae*, and *Rhabdoviridae*) plus unclassified elements. We described with high confidence the near-complete and the partial genomes of two tentative novel viruses: Carpotroche-associated ilarvirus and Carpotroche-associated genomovirus, respectively. Interestingly, we also described sequences likely derived from a rhabdovirus, which could represent a novel member of the genus *Gammanucleorhabdovirus*. We observed higher viral diversity in cacao-cabruca AFS and reproductive organs of *C. brasiliensis* with preferential tropism to fruits, which could directly affect production. Altogether, our results provide data to better understand the virome in this unexplored agroecological interface, such as cacao-cabruca AFS and forest ecosystem, providing information on the aspects of virus–plant interactions.

KEYWORDS

virome, agroforestry, metatranscriptomics, rainforest tree, Carpotroche brasiliensis

Introduction

The agroforestry-based management systems (agroforestry systems, AFS) aggregate types of traditional land-use practices involving the deliberate combination of crops, animals, and tree vegetation for agricultural commodities production, and are often used by small and large agricultural producers around the world (Nair, 1993, 2012). AFS usually provide the greatest agricultural gain in ecosystem services that are beneficial to productivity, and, in turn, biodiversity is maintained in the environment (Bhagwat et al., 2008; Teixeira et al., 2021). Nevertheless, plantation management must occur in areas close to the conserved forests for increased benefits (Faria et al., 2007). In Brazil, pastureland, cropland, monoculture tree plantations, and mosaics of AFS have been used in regions of younger native forests or remnants of old-growth forests to maintain the ecosystem services (Joly et al., 2014; Rosa et al., 2021).

One example of AFS use in Brazil is the cacao-AFS. Cocoa fruits are produced by the *Theobroma cacao* L. (*Malvaceae*) and are cultivated in over 606,794 hectares (ha) of Brazilian territory, responsible for producing 5% of the world's cocoa (Gama-Rodrigues et al., 2021). Of this area, 430,051 ha are part of the Atlantic forest in the States of Bahia and Espírito Santo (Gama-Rodrigues et al., 2021). The cacao-cabruca AFS model is practiced in about 250,000 ha of Atlantic Forest in southern Bahia state, an area that holds most of the forest in northeastern Brazil (Martini et al., 2007; Sambuichi et al., 2012; Gama-Rodrigues et al., 2021). This AFS model contributes to 44% of the 259,425 tons of national cocoa produced in 2019 (Gama-Rodrigues et al., 2021). The *cabruca* system consists of random intercropping of cocoa in native forest strata, exploiting native or exotic trees to provide shade. This ecosystem environment is characterized by a high diversity of trees that may vary according to age, environmental conditions, and management of the areas (Sambuichi et al., 2009, 2012; Gama-Rodrigues et al., 2021).

Carpotroche brasiliensis (Raddi) A. Gray (Achariaceae) is a plant species native to the Atlantic Forest that is found in *cabruca* landscapes (Sambuichi and Haridasan, 2007). The species, whose vernacular name is *sapucainha*, is widely used in reforestation programs, environmental restoration, and agroforestry systems due to shade tolerance (Brito-Rocha et al., 2017; Cerqueira et al., 2018) and its capacity to provide resources for wildlife (Marangon et al., 2010; Zucaratto et al., 2010). The fruits of *C. brasiliensis* are consumable by wild animals, mainly rodents, and marketed to the cosmetics industry that makes use of its oil in esthetics products (Lima et al., 2020). The Chaulmoogra oil can be extracted from *C. brasiliensis* seeds, and bioactive fatty acids like chaulmoogric and hydnocarpic acids have anti-inflammatory, analgesic, and antiparasitic pharmacological properties (Sharma and Hall, 1991; Parascandola, 2003; Lima et al., 2005; dos Santos et al., 2008; Krist, 2020). Therefore, the valorization of *C. brasiliensis* is fundamental to adding economic and ecological value to the cacao-AFS, and the evaluation of aspects of the microbial biodiversity, such as the viral diversity, results in a greater understanding of these unexplored ecosystems of tree species, to avoid the reduction of native biodiversity in agroforestry systems (Alexander et al., 2014; Piasentin et al., 2014).

Advances in plant virology after the emergence of high-throughput sequencing (HTS) have accelerated the identification of novel virus species from crops and fruit trees in agricultural ecosystems, expanding the knowledge of viral epidemiology in intensive and diverse production systems such as AFS models (Maclot et al., 2020). In Brazil, due to recurrent virus outbreaks in crops of economic importance, more than 200 virus species infecting plants were cataloged and officially recognized by the International Committee on Taxonomy of Viruses (ICTV) until 2018 (Kitajima, 2020). However, little is known about the ecology of viruses infecting wild hosts in native ecosystems adjacent to the crop fields, and the latter is known as an agroecological interface (Alexander et al., 2014; Roossinck and García-Arenal, 2015; Rodríguez-Nevado et al., 2020). Moreover, there is a real risk of spillover of virus infection between plants, which is common among plant viruses and emergent species, such as cacao swollen shoot virus (CSSV) (Muller, 2016). This virus seems to have originated from indigenous forest trees that work as alternative hosts for pathogens (Posnette et al., 1950; Dzahini-Obiatey et al., 2010; Topolovec-Pintaric, 2020). Interestingly, other groups of viruses are undergoing evolutionary radiation, adapting to infect new plant species in different environments (Pagán, 2018; Moury and Desbiez, 2020).

Plant viruses may act in different ecological contexts along the evolutionary relationships established with the host. These viral interactions may lead to asymptomatic infections in latent or persistent viral cycles, through the integration of fragments derived from the viral genome into the host DNA as endogenous viral elements (EVEs) (Lefeuvre et al., 2019). EVEs may have a beneficial relationship for adaptation in environments or promote competition in plant communities by affecting virulence and host tolerance/susceptibility, and subsequently affect the epidemiological profile of viral pathology in the ecosystem (Engering et al., 2013; Lefeuvre et al., 2019; Takahashi et al., 2019). Understanding how evolutionary forces act in distinct environments, together with the anthropic influence on the emergence of diseases, can provide knowledge to forecast and avert alterations in natural ecosystems that cause crop damage (Lefeuvre et al., 2019). The advances in plant virology are needed mainly for forest tree species that have a current few viral diversity data in their respective forest ecosystems (Rumbou et al., 2021).

In our study, we analyzed the virome of *C. brasiliensis* from Atlantic Forest regions and private properties of southern Bahia using cacao-cabruca AFS through HTS of the RNA samples obtained from vegetative and reproductive organs. Our results showed that the *C. brasiliensis* virome is composed of members spanning almost all members of the Baltimore classification of viruses [+ssRNA, double-stranded (ds)RNA, –ssRNA, reverse-transcribing (RT)-DNA, –ssDNA, dsDNA] that could be classified into at least six distinct families (*Botourmiaviridae*, *Bromoviridae*, *Caulimoviridae*, *Genomoviridae*, *Mitoviridae*, and *Rhabdoviridae*) together with three different unclassified elements. Of note, we successfully reconstituted with high confidence the complete coding sequence of two novel viruses (Carpotroche-associated ilarvirus and Carpotroche-associated genomovirus) and partial sequences from a virus likely representing a novel member of the genus *Gammanucleorhabdovirus*. Proteomics assay from Carpotroche seeds detected peptides from many viral sequences, further confirming the virus presence. Finally, we show the restricted occurrence of viruses from distinct families in exclusive ecosystems and specific organs of *C. brasiliensis*, as well as discrepant abundance among samples. Therefore, our results provide background for a better understanding of the viral diversity in the context of the agroecological interface, such as cacao-cabruca AFS and forest ecosystem.

Materials and methods

Sampling design

Samples from six different organs of *C. brasiliensis* (leaf, flower, flower bud, root, fruit, and seed) were obtained from each individual in June 2014 in Camamu-Maraú Country, State of Bahia, Brazil. All samples were extracted from asymptomatic individuals with no discoloration of any kind, wilting, or necrotic lesions. Sixteen adult trees were randomly selected to compose two sampling groups: eight from agroforestry systems and eight from natural Atlantic Forest ecosystems (**Figure 1A**; **Supplementary Table 1**). All samples were immediately frozen in liquid nitrogen and stored at −80°C for RNA extraction at the Center of Biotechnology and Genetics, Laboratory of Molecular Markers, Universidade Estadual de Santa Cruz (UESC), Brazil.

Plant material, RNA extraction, cDNA library preparation, and high-throughput sequencing

Total RNA was extracted from each sample using RNAqueous® Total RNA Isolation Kit (AM1912, Thermo Fisher Scientific), following the manufacturer's recommendations. The integrity and quantity of RNA were confirmed using the TapeStation Agilent 2200 instrument (Agilent Technologies Co., Santa Clara, CA, United States), considering an RNA Integrity Number (RIN) value above 5. Only RNA samples of vegetative and reproductive organs with integrity and quantity acceptable were used to generate the 20 cDNA libraries. The cDNA libraries were constructed using 10 ng of total RNA and using the NEBNext Ultra RNA Library Prep Kit for Illumina (E7530S, New England Biolabs, Inc., Ipswich, MA, United States) and the NEBNext Multiplex (E7335, New England Biolabs, Inc., Ipswich, MA, United States) Oligos, following the manufacturer's protocols (Illumina, San Diego, CA, United States). The libraries were quantified with KAPA Library Quantification Kit for Illumina Platforms (KAPA Biosystems, Wilmington, MA, United States) and Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, DE, United States). A total of 20 cDNA libraries (11 from samples of agroforestry systems and 9 samples of Atlantic Forest ecosystems) were sequenced on Illumina MiSeq 2 × 250 bp pair-end sequencing (**Supplementary Table 1**) to provide biological replicates. Raw sequencing data were deposited at NCBI Sequence Read Archive (SRA)[1] under project number PRJNA858666.

Pre-processing, quality control, and transcriptome *de novo* assembly

Trimmomatic software v.0.32 was used to filter out sequences not fitting in following criteria: minimal average Phred score: 33, Leading: 30, Trailing: 30, Sliding window 4: 30, and Minlen: 50 (Bolger et al., 2014). Pre-processed libraries obtained from different plant tissues were used as input for the transcriptome assembly in the SPAdes tool (Bankevich et al., 2012) in two steps: initially, an assembly of the 20 libraries was made using the *-rna* parameter, with all other parameters kept as default; subsequently, contigs that showed sequence similarity to viral sequences in the previous step were used as an anchor for the new assembly, setting the *–trusted-contigs* parameter. An overview of the methods is shown in **Figure 1B**. We also performed a metagenomic analysis to determine the presence of possible contaminations using the Kaiju platform (Menzel et al., 2016).

Virome analysis

Assembled contigs were used as queries in the Diamond tool (Buchfink et al., 2015) against amino acid sequence databases to identify sequences of possible viral origin. In this analysis, we considered only the best hit for each contig and with an e-value < 1e-3. The result was further filtered using regular expression and manually inspected to select only hits of viral origin. The contigs with putative viral origin were submitted to CAP3 (Huang and Madan, 1999) and CD-HIT (Fu et al., 2012) tools to remove redundancy and extend

[1] https://www.ncbi.nlm.nih.gov/sra

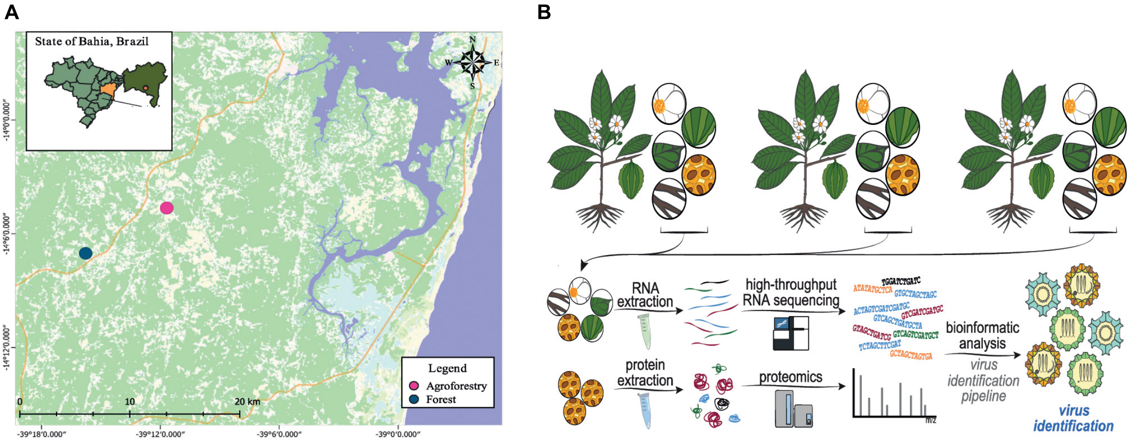


FIGURE 1
Experimental design of the study. **(A)** Geographic location in the Brazilian Atlantic forest (blue spots) and agroforestry systems (pink spots) where the *Carpotroche brasiliensis* samples were collected. **(B)** Simplified scheme showing the strategy applied in our work, including tissue collection, RNA or protein extraction, RNA deep sequencing, mass spectrometry, and bioinformatics analysis.

assembled contigs. For CD-HIT clustering, sequences with 90% coverage and 90% identity were merged, and a representative sequence was defined. The putative viral contigs were analyzed by the NCBI BLAST online tool (Johnson et al., 2008) to guarantee the last updated version of nucleic acid databases. The sequences were submitted to BLASTN against the *nt* database and to BLASTX against the *nr* database to obtain the best hit, to confirm the results. The contigs that showed characteristics of viral sequences (contig length and hit with viral sequences) were subsequently analyzed in the ORFfinder tool[2] to perform its structural annotation. Finally, conserved domains were annotated in each contig using the PHMMER tool (Potter et al., 2018) under the following parameters: sequence e-value 0.01 and hit e-value 0.03. For the quantification of viral sequences, we built the transcriptome index using the putative viral contigs that were compared to each library using Salmon (Patro et al., 2017) with standard parameters. Counts were used to estimate the relative abundance at the transcript level through transcripts per million (TPM) and plotted as heatmap using the ComplexHeatmap package in R (Gu et al., 2016). Reconstituted high-quality viral sequences used in the phylogeny were deposited at the NCBI nucleotide database (GenBank) under accessions OL964097–OL964101. All viral sequences assembled in our work are also available in **Supplementary Data Sheet 1**.

Phylogenetic analyses

The assembled contigs were translated into amino acid sequences and aligned with the closest viral sequences available in the NCBI protein database (Geer et al., 2010) using the MAFFT program (Katoh and Standley, 2013). The best-fit model was selected for each alignment using the ProtTest 3.2 program considering the Akaike Information Criterion (AIC) (Akaike, 1974; Abascal et al., 2005). Maximum likelihood (ML) trees were constructed in MEGA X (Kumar et al., 2018), considering 1,000 bootstrap replicates. The trees generated were mid-term rooted and edited in Geneious Prime 2021.[3] The sequences from the viruses common bean curly stunt virus (unclassified, tentative member of the family *Geminiviridae*), peanut clump virus (*Pecluvirus* genus) (NP_620047), and ampivirus A1 (YP_00913521) were used as outgroup for the phylogenies produced.

Protein analysis and mass spectrometry

Sample preparation

Total proteins were extracted from separate pools of whole seeds derived from staminated individuals using a protocol developed by Pirovani et al. (2008). Briefly, seeds were macerated in the presence of liquid nitrogen and 7% polyvinylpolypyrrolidone. In total, 0.1 g of seeds were used for three replications for stages S1 and S2. The samples were resuspended in 500 μL of 1-butanol: chloroform (1:9), and the mixture was vortexed and centrifuged for 5 min at 13,400 *xg* at 4°C. This procedure was repeated two times. Subsequently, the precipitate generated by centrifugation was washed two times with 500 μL of 100% acetone and centrifuged as in the

[2] https://www.ncbi.nlm.nih.gov/orffinder/

[3] https://www.geneious.com

previous step. Finally, the precipitate was washed with 500 μL of petroleum ether and centrifuged for 5 min at 13,400 x*g* at 4°C. After completion of delipidation, the samples were subjected to precipitation and protein extraction (Pirovani et al., 2008), with the modification being the incubation of the samples overnight at −20°C in the washing step with 10% trichloroacetic acid (TCA) in water. The final precipitate was resuspended in 400 μL of urea at 8 M. At the end, the proteins from the samples were stored in a freezer at −20°C until further use. The protein extracts obtained were quantified with the 2D Quant Kit (GE Healthcare Life Sciences, Chalfont, United Kingdom) following the manufacturer's recommendations. A standard curve of bovine serum albumin (BSA) was constructed, which served as a basis for the quantification of samples of *C. brasiliensis* seeds.

Mass spectrometry

The solution containing the digested peptides was desalted using tips with C18 resin (10 μL; Millipore® Ziptips C18). The peptides were eluted in 50 μL of a solution containing 75% acetonitrile, 25% water, and 0.1% formic acid. The peptides were analyzed in a liquid chromatography system (Agilent 1290 Infinity II HPLC) coupled to a quadrupole/time-of-flight mass spectrometer (Agilent 6545 LC/QTOF) (Agilent Technologies, Santa Clara, CA, United States). Samples were separated using a reversed-phase column (C18; AdvanceBio Peptide Mapping 2.1 × 250 mm; Agilent), maintaining a temperature of 55°C. A 20-min gradient was applied with mobile phases A (water and 0.1% formic acid) and B (acetonitrile and 0.1% formic acid). The percentages of phase B along the grid were 5–35% (1–10 min.), 35–70% (11–14 min.), 70–100% (1618 min.), and 100% (16–20 min.). In addition, a final period of 5 min was programmed for column stabilization. Then, the samples were injected into three technical replicates. The samples were injected into the QTOF through an electrospray source, using the Auto MS/MS acquisition mode, with a maximum selection of 10 precursors per cycle. The parameters for the selection of precursors were as follows: threshold of 1,000, 10,000 counts/spectrum, the stringency of 100% purity, a cut-off of 30% purity, peptide isotopic model, charge preference of 2, 3, >3, and unknown. The instrument parameters were set as follows: gas temperature of 325°C, the gas flow of 13 L/min, a capillary voltage of 4,000 V, and skimmer voltage of 56 V. Nitrogen gas was used for the induced dissociation collision. Instrument control (HPLC and QTOF) and parameter configuration were performed using the Agilent MassHunter Acquisition software.

Identification of virus-derived peptides

The data integration of transcriptomic and proteomic analyses was implemented through mass spectrometry (MS) analysis to discover proteomics from *C. brasiliensis*. The identification of peptides derived from viral proteins was performed from spectral extraction and merging through analysis with Spectrum Mill Proteomics Workbench (Agilent Technologies, Santa Clara, CA, United States) against proteomic data from *C. brasiliensis* seeds. The static modification to carbamidomethylation was set to default, with a mass (MH +) range of 200–6,000 mass–charge ratio (m/z). Retention time tolerance was ± 60 s, m/z tolerance was ± 1.4 m/z, MS noise threshold was set to 10 counts, and charge general was set to default. The data searched were filtered for validation by score threshold with a false discovery rate (FDR) > 1%. The spectral intensity of identified proteins was searched against the viral ORFs analyzed. The MS/MS spectral comparison included four miss-cleavage sites and fixed modifications: carbamidomethylation on cysteine residues (C), differential modifications for oxidized methionine (M), pyroglutamic acid (N-termQ), deamidated (N), phosphorylated S (S), phosphorylated T (T), and phosphorylated Y (Y). To determine the combined score of minimal intensity, 10% with a mass tolerance of 20 ppm was validated and filtered by false discovery rate (FDR) > 1%. All MS files produced in this study were deposited at MassIVE[4] with the identifier MSV000089145 and can be accessed at ftp://massive.ucsd.edu/MSV000089145/.

Results

Metavirome analysis

We deep sequenced 20 cDNA libraries derived from the tissues of *C. brasiliensis* from forest and agroforestry ecosystems (cacao-cabruca AFS), totalizing 28,177,426 raw reads (**Figure 1**). After pre-processing steps, including quality and length filters, 24,689,370 (87.62%) reads were kept (**Supplementary Table 1**). Transcriptome assembly produced 281,643 transcripts with N50 of 481 bp and a total number of bases in transcripts of 160,909,318 (**Supplementary Table 2**). From the total, 184 sequences showed sequence similarity to viral sequences stored in NCBI public databases, of which 136 (75%) showed hits related to elements with non-retroviral origin. Retroviral sequences were discarded from the analysis since they were of low reliability and often are misidentified with transposable elements. As quality control, we performed metagenomic analyses that did not detect sequences from animals, indicating a low chance of contamination from external sources (**Supplementary Figure 1** and **Supplementary Table 3**). After removal of redundancy and manual curation, we selected 30 sequences larger than 400 nt for further characterization. The length of putative viral contigs ranged from 503 nt to 8,695 nt.

[4] http://massive.ucsd.edu

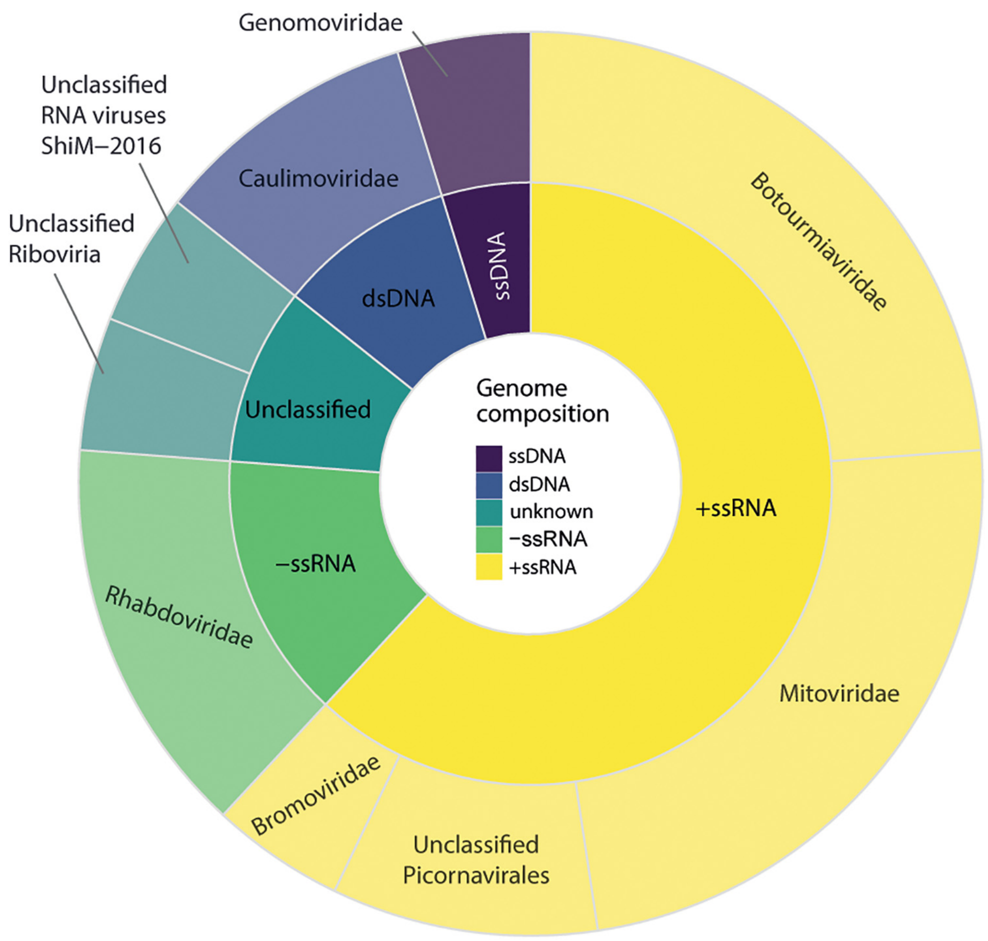


FIGURE 2
Diversity of viral sequences identified in *Carpotroche brasiliensis* samples. Virus classification was performed based on the information of the closest related sequence present in NCBI databases classified by genome structure and viral family. The diversity analysis only included considered sequences derived from RNA-dependent RNA polymerases, replicase, or polyproteins.

Diversity and distribution of viral sequences

We observed that 30 sequences of possible viral origin are related to viruses from at least six different viral families. Of these, 17 sequences showed similarity to either RNA-dependent RNA polymerases, replicases, or polyproteins, referred hereafter as key sequences, and were considered in the diversity analysis (**Figure 2**). From the total of non-retroviral viral sequences identified, 16 sequences could be identified by nucleotide similarity, suggesting they are possibly strains of known viruses, and 14 were classified only at the amino acid level (**Supplementary Table 4**). Overall, sequences showed similarity to closely related viruses from at least six different families or unclassified. Eight transcripts showed similarity to Maize fine streak virus genus *Gammanucleorhabdovirus* (*Rhabdoviridae*, −ssRNA); four transcripts to viruses of the genus *Ourmiavirus*, three to an unclassified virus of the family *Botourmiaviridae* (+ssRNA), five transcripts to viruses from genus *Mitovirus* (*Mitoviridae*, +ssRNA), three transcripts to viruses of the genus *Ilarvirus* (*Bromoviridae*, +ssRNA), and two unclassified sequences showing similarity with picornavirus (+ssRNA); one transcript to viruses of the genus *Caulimovirus*, and one to viruses of the genus *Solendovirus* (*Caulimoviridae*, +dsDNA); one transcript to viruses of the genus *Genomovirus* (*Genomoviridae*, −ssDNA) and other two unclassified transcripts (**Supplementary Table 4**).

These sequences were detected in the sum of all 20 *C. brasiliensis* libraries analyzed. Co-occurrence analyses indicated that all three segments of a tentative new virus from the genus *Ilarvirus* were detected in 16 samples, while at least two sequences appeared together in 18 different samples. The second most abundant family was *Rhabdoviridae*, genus *Gammanucleorhabdovirus*. In total, the sequences showing similarity to the viruses of the genus *Gammanucleorhabdovirus* were detected in nine libraries of *C. brasiliensis* with a maximum number of seven different contigs in a single library. Then, in decreasing order of abundance sequences presenting similarity with viruses from the families, *Mitoviridae* was detected in 12 libraries, *Botourmiaviridae* was found in 4 libraries, *Caulimoviridae* was found in 3 libraries, and elements of *Picornavirales*, *Genomoviridae*, and unclassified *Riboviria* were detected in two different libraries. Only one contig showing similarity with unclassified RNA viruses was restricted to a single library.

Characterization of near-complete viral genomes

To increase certainty for high-quality viral sequences, we performed an extra assembly step with Cap3. Four out of the 30 transcripts had their lengths extended, of which three sequences showed similarity to ilaviruses and one was likely derived from an unclassified picornavirus. Through manual curation *via* BLAST (**Supplementary Table 4**), it was possible to confirm the similarity of these elements to previously identified viruses available in the NCBI sequence databases.

The largest transcript containing 8,636 nt was closely related to Skokie picorna-like virus (SPV), with a size 8,401 bp, 94% nucleotide sequence identity, and >97% of amino acid sequence identity, which is an unclassified picornavirus that infects the mite Dermatophagoides pteronyssinus. Remarkably, we observed very similar patterns of ORF distribution and domain organization when comparing this assembled transcript with SPV (**Figure 3A**; **Supplementary Figure 2**), suggesting a close evolutionary relationship. Indeed, they share conserved domains encoded by RNA-dependent RNA polymerase (PFAM PF00680) and capsid protein (PFAM PF00073) of viruses from the order Picornavirales (**Figure 3A**). For sequences presenting similarity with the elements of the family Bromoviridae, the Contig1 (2,442 nt) was closely related to the phytopathogenic virus Lilac ring mottle virus (LRMV), with 2,287 nt and 73% identity, which belongs to the genus Ilarvirus. Both sequences also share the same pattern of ORFs, although the main ORF in Contig1 is 11 amino acids longer than the main ORF of LMRV (**Figure 3B**; **Supplementary Figure 3**). Domain analysis indicates that this major ORF encodes for a Bromovirus movement protein (PFAM PF01573), increasingly the possibility of functionality (**Figure 3B**; **Supplementary Figure 3**). Contig2 (2,954 nt), on the other hand, was most similar to another phytopathogenic ilarvirus Citrus leaf rugose virus, with 2,990 nt and 72% identity. Besides having little difference in length, they also had similar ORF patterns (**Figure 3B**; **Supplementary Figure 3**). The major ORF differs from one other by only 36 bp, and both code for the RNA-dependent RNA polymerase enzyme (PFAM PF00978) (**Figure 3B**; **Supplementary Figure 3**). Contig3 (3,467 nt) was closely related to a phytopathogenic ilarvirus, Tomato necrotic streak virus (with 3,378 nt and 68% identity with segment RNA 1). Structural annotation shows that the longest ORF in both sequences differs only by 4 nt, which was only identified at the amino acid level as the replicase protein. Domain analysis also shows that both longest ORFs share methyltransferase (PFAM PF01660) and helicase (PFAM PF01443) conserved domains (**Figure 3B**; **Supplementary Figure 3**).

We also selected one transcript showing similarity to that of the family *Genomoviridae* to undergo further structural and functional annotations. This transcript was chosen because the sequence recovered corresponds to an intact replicase with a length consistent with genomoviruses. This transcript (943 nt) showed the highest identity to chicken genomovirus mg4_1247 (complete genome: 2,142 nt; replicase: 1,008 nt, 62% identity) (**Figure 3C**; **Supplementary Figure 4**). Domain analysis revealed that both sequences present a conserved Gemini_AL1 domain (PFAM00799) (**Figure 3C**; **Supplementary Figure 4**).

Phylogenetic analysis of Carpotroche-associated viruses

We selected putative viral sequences with near-complete or complete coding sequences according to the closest relative virus in public databases to perform further phylogenetic characterization. Therefore, here we assessed the phylogenetic relationship of the three viral transcripts with their closely related viruses from the genera *Picornavirus*, *Ilarvirus*, and *Genomovirus*, respectively (**Figure 4**). The phylogeny of the sequence corresponding to a virus of the family *Picornaviridae* confirmed the similarity observed by local alignment with the Skokie picorna-like virus with 100% of bootstrap. This sequence was also related to the Bat posalivirus, suggesting that the assembled transcript belongs to the genus *Picornavirus* (family *Picornaviridae*) (**Figure 4A**). This virus was named Skokie picorna-like virus Carpotroche isolate. The transcript showing similarity to viruses of the genus *Ilarvirus* was close to the clade of plant ilarviruses containing viruses of the species *Citrus leaf rugose virus*, *Tulare apple mosaic virus*, *Tomato necrotic streak virus*, *Spinach latent virus*, *Asparagus virus 2*, *Citrus variegation virus*, and *Elm mottle virus* (**Figure 4B**), thus we can infer that this sequence is from the genus *Ilarvirus* (*Bromoviridae*) and the putative virus was named Carpotroche-associated ilarvirus. In the phylogeny of the element previously associated to *Genomovirus*, we observed that the sequence grouped with the virus chicken genomovirus mg4, and Gemycircularvirus sp. with a bootstrap value of 97.4%, indicating that this species belongs to the *Genomoviridae* (**Figure 4C**). This putative new virus was named Carpotroche-associated genomovirus.

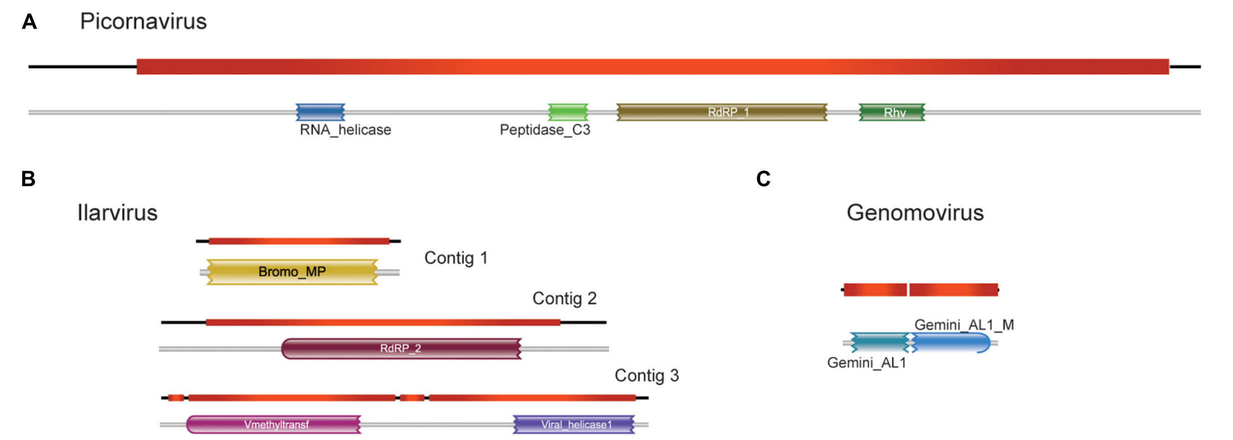


FIGURE 3
Structural and functional annotation of high-confident viral sequences. For sequences showing similarity to viruses of the families **(A)** *Picornavirus*, **(B)** *Ilarvirus*, and **(C)** *Genomovirus*, the detected open reading frames (ORF) and the presence of conserved domain are shown.

Characterization of a tentative new virus from the genus *Gammanucleorhabdovirus*

Metavirome analysis revealed a considerable number of viral transcripts likely derived from a plant-infecting rhabdovirus related to the viruses of the *Gammanucleorhabidovirus* genus. Indeed, 8 out of 30 viral transcripts presented similarity identity values at the protein level, ranging from 32.54 to 71.88%, with Maize fine streak virus (MFSV), the unique known member of the genus *Gammanucleorhabdovirus* genus accepted by ICTV (**Figure 5A**). These transcripts ranged from 550 to 1,321 nt totalizing 6,561 nt with an average size of ∼820 nt and were distributed along the MFSV genome. Six out of the eight contigs presented conserved domains commonly identified in rhabdoviruses, including MFSV (**Figure 5A**). Of note, we identified one transcript of 983 nt showing similarity to the RdRp gene, containing a Mononeg_RNA_pol (PF00946) domain. We took advantage of this transcript likely derived from viral polymerase to assess the phylogenetic relationship of the tentative virus with MFSV and other rhabdoviruses. According to our maximum likelihood tree, we observed the presence of six main clades, which refer to different genera of the subfamily *Betarhabdovirinae—Rhabdoviridae* (*Cytorhabdovirus*, *Varicosavirus*, *Alphanucleorhabdovirus*, *Betanucleorhabdovirus*, *Dichorhabdovirus*, and *Gammanucleorhabdovirus*) (**Figure 5B**). The transcript assembled clustered with Maize fine streak virus in a clade-specific manner with a bootstrap of 100, suggesting that the sequence probably belongs to a species from this genus (**Figure 5B**). However, since we were not able to reconstitute the complete genome of the virus, we named the viral species as a Carpotroche-associated gammanucleorhabdovirus-like virus.

Ecosystem distribution and organ tropism

Since we deep sequenced the RNA of several tissues from vegetative and reproductive organs of *C. brasiliensis* originated from forestal ecosystems and agroforestry, we decided to explore the tropism of the identified viral sequences among the individuals assessed in this study. We also analyzed the relative abundance of viral elements detected in organ samples for each ecosystem. We noticed that the seed was the tissue with the greatest diversity of viral sequences, with the presence of 20 (11 key sequences) out of the 30 viral sequences identified in our study (**Figures 6A,B**). The libraries of floral bud, root, and small fruit also showed high percentages of viral diversity, with 18 (eight key sequences), 10 (eight key sequences), and 8 (three key sequences) viral sequences, respectively (**Figures 6A,B**). However, we acknowledge that these data have to be interpreted carefully, since we analyzed only one library of small fruit while we assessed nine libraries for floral buds. Besides, in one of the libraries from *C. brasiliensis* leaves, we were not able to detect any sequence of viral origin (**Figure 6A**). Of note, over half of the key sequences (10) were organ-specific, while approximately 22% were found in at least three different organs (**Figure 6B**). Surprisingly, only one viral sequence of Carpotroche-associated ilarvirus was identified in all tissues assessed (**Figure 6B**).

Since tissue tropism analysis indicated seeds as containing the major number of viral sequences, we decided to perform a proteomics assay for the sample derived from whole seeds collected from staminated individuals to provide an additional level of confidence. From the 17 putative viral family elements identified in seeds, we were able to detect peptides for nine of them. We also detected peptides from NODE_15442 (*Rhabdoviridae*), NODE_18845 (*Rhabdoviridae*), NODE_68365 (*Botourmiaviridae*), and NODE_74445

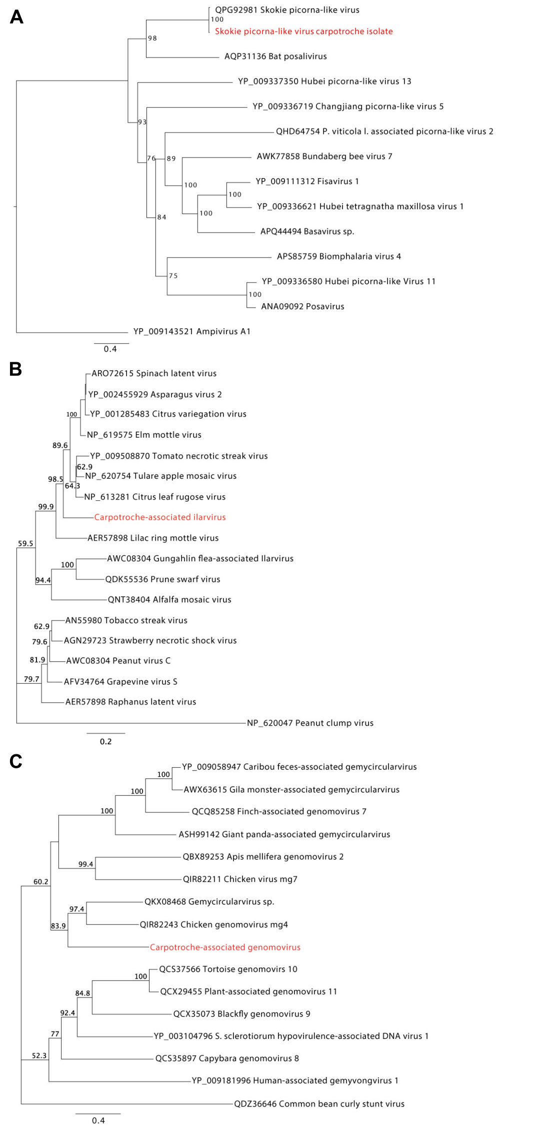

A
QPG92981 Skokie picorna-like virus
100
Skokie picorna-like virus carpotroche isolate
98
AQP31136 Bat posalivirus
YP_009337350 Hubei picorna-like virus 13
YP_009336719 Changjiang picorna-like virus 5
93
QHD64754 P. viticola l. associated picorna-like virus 2
76
89
AWK77858 Bundaberg bee virus 7
100
YP_009111312 Fisavirus 1
100
YP_009336621 Hubei tetragnatha maxillosa virus 1
100
84
APQ44494 Basavirus sp.
APS85759 Biomphalaria virus 4
75
YP_009336580 Hubei picorna-like Virus 11
100
ANA09092 Posavirus
YP_009143521 Ampivirus A1
0.4
B
ARO72615 Spinach latent virus
YP_002455929 Asparagus virus 2
100
YP_001285483 Citrus variegation virus
NP_619575 Elm mottle virus
89.6
YP_009508870 Tomato necrotic streak virus
62.9
NP_620754 Tulare apple mosaic virus
98.5
64.3
NP_613281 Citrus leaf rugose virus
99.9
Carpotroche-associated ilarvirus
AER57898 Lilac ring mottle virus
59.5
100
AWC08304 Gungahlin flea-associated Ilarvirus
QDK55536 Prune swarf virus
94.4
QNT38404 Alfalfa mosaic virus
AN55980 Tobacco streak virus
62.9
79.6
AGN29723 Strawberry necrotic shock virus
81.9
AWC08304 Peanut virus C
79.7
AFV34764 Grapevine virus S
AER57898 Raphanus latent virus
NP_620047 Peanut clump virus
0.2
C
YP_009058947 Caribou feces-associated gemycircularvirus
100
AWX63615 Gila monster-associated gemycircularvirus
100
100
QCQ85258 Finch-associated genomovirus 7
ASH99142 Giant panda-associated gemycircularvirus
99.4
QBX89253 Apis mellifera genomovirus 2
QIR82211 Chicken virus mg7
60.2
97.4
QKX08468 Gemycircularvirus sp.
83.9
QIR82243 Chicken genomovirus mg4
Carpotroche-associated genomovirus
QCS37566 Tortoise genomovirs 10
100
84.8
QCX29455 Plant-associated genomovirus 11
92.4
QCX35073 Blackfly genomovirus 9
77
YP_003104796 S. sclerotiorum hypovirulence-associated DNA virus 1
52.3
QCS35897 Capybara genomovirus 8
YP_009181996 Human-associated gemyvongvirus 1
QDZ36646 Common bean curly stunt virus
0.4

FIGURE 4
Phylogenetic analyses of high-confident viral sequences identified in *Carpotroche brasiliensis*. Phylogenetic trees inferred by maximum likelihood for transcripts associated with *C. brasiliensis* and showing similarity to **(A)** picornaviruses, **(B)** ilarviruses, and **(C)** genomoviruses species. RNA-dependent RNA polymerase or polyprotein aminoacid sequences were used in the analysis. Branch support was calculated using the bootstrap method, requiring 1,000 pseudoreplicates. Highlighted regions in red correspond to the assembled sequences obtained in our work.

(*Caulimoviridae*) sequences previously identified in other organs (**Figure 6A—horizontal** bars; **Supplementary Table 5**).

Relative to ecosystems, the abundance of viral elements in equivalent tissue samples is slightly higher in specimens originating from the cacao-cabruca AFS (41.2% of total sequences) when compared to the forest ecosystem. Viral sequences detected exclusively in the cacao-cabruca AFS were Skokie picorna-like virus isolate Carpotroche, Carpotroche-associated genomovirus, botourmiavirus (3), non-classified sequence (1), and mitovirus (1). In contrast, one sequence related to element from botourmiavirus and one related to unclassified picornavirus were identified exclusively in samples obtained from the forestal ecosystem (**Figures 6A–C**).

Regarding the reproductive organs, in both ecosystems, samples obtained from stamen tissues show a lower abundance and diversity in viral sequences compared to those obtained from pistil tissues (**Figure 6D**). While we noticed species that were organ-specific, none of the viral sequences were exclusively found in samples derived from staminate plants (**Figure 6D**).

Discussion

In this study, we identified viral sequences representing the *C. brasiliensis*-associated virome in native forest and cacao-cabruca AFS. Through HTS technologies associated with proteomics assay, it was possible to perform an in-depth analysis of viruses putatively associated with this plant, expanding our knowledge about environmental virology and describing the profile of viral infection among samples and ecosystems. Also, we enlightened the abundance and distribution of virus infection per vegetative and reproductive organs of *C. brasiliensis* and the composition of viral communities per sampled landscape.

The patterns of diversity, abundance, and virus tropism between vegetative and reproductive organs observed in this study can be related to the host preference of a virus, as well as to the immunological capacity in different cell types and tissues of the host (Navarro et al., 2019; Keesing and Ostfeld, 2021). Therefore, this can impact the amplification effects arising from increased transmissibility among hosts or dilution effects of infections in agricultural environments or in natural ecosystems (Keesing et al., 2006; Keesing and Ostfeld, 2021; Susi and Laine, 2021). Interestingly, we identified *C. brasiliensis* putative viruses related to two different families that have species described to infect different plant hosts of forest trees or crops, i.e., *Ilarvirus* and *Mitovirus* (Pallas et al., 2013; Rumbou et al., 2021). Of note, the highest diversity of these virus sequences was observed in reproductive organs. It is likely that viral exchange exists at the agroecological interface through the transmissibility of viruses detected in the reproductive structures of *C. brasiliensis* during pollination processes and dispersal over longer distances due to animal intermediation (Marangon et al., 2010; Zucaratto et al., 2010). Viruses with these characteristics can also mutate acquiring to the possibility to infect unrelated alternative hosts in the same ecological community, such as detected in cacao-cabruca AFS or forest ecosystem (Elena et al., 2009).

We observed widespread infection in *C. brasiliensis* by a tentative new virus of the genus *Ilarvirus*. Viruses of this genus can infect woody plants, such as *C. brasiliensis*, as well as fruits, vegetables, and forages of economic interest, such as peach, apricot, tomato, apple, vine, cucurbitaceae, banana, and alfalfa (Pallas et al., 2013). Ilarvirus infections range from no observable symptoms to the occurrence of leaf mosaics, structural malformations, abortive flowers, and atrophy in fruits and seeds, and in some cases, may lead to host death (Girgis et al., 2009). Contigs belonging to detected ilavirus (Carpotroche-associated ilarvirus 1, 2, 3) presented homology with three different viruses: citrus leaf rugose virus (CLRV), lilac ring mottle virus (LiRMoV), and tomato necrotic streak virus (TomNSV). CLRV can infect a wide range of citrus hosts with induction of milder symptoms, while LiRMoV can induce leaf deformation and reduce leaf size, ring spots, and line patterns (Sharma-Poudyal et al., 2016; Zhou et al., 2020). The infection with LiRMoV was observed in herbaceous plants in the experimental conditions, although it is possible that cryptic infection may exist in crop plants for many years (Sharma-Poudyal et al., 2016). On the other hand, TomNSV infects the *Solanaceae* and *Chenopodiaceae* families, producing serious damage to the leaves (Badillo-Vargas et al., 2016). However, we did not find any symptoms in *C. brasiliensis* individuals sampled.

Virus diversity observed in organs suggests an established viral community in *C. brasiliensis in situ*, with possible transmission between ecosystems *via* arthropod vectors, such as for the virus alfalfa mosaic virus, where thrips and aphid contamination is mediated by feeding on pollen grains or *via* direct contact with infected pollen, seeds, or fruits (Pallas et al., 2013; Silvestre et al., 2020). The arthropod vector-mediated contamination may also be related to the transmission of virus of the genus *Gammanucleorhabdovirus* (*Rhabdoviridae*), with different sequences detected in the samples of leaves, fruits, seed, and flower buds obtained from *C. brasiliensis*, showing similarity to the segments of MFSV, such as nucleocapsid protein (N), glycoprotein (G), and polymerase (L) (Dietzgen et al., 2020). Plant viruses of the family *Rhabdoviridae* have global circulation with damage to diverse commodities, including maize and wheat crops in South American countries, such as Argentina and Peru (Willie and Stewart, 2017; Maurino et al., 2018; Dietzgen et al., 2020). Infection symptoms by these viruses include yellowing, chlorosis, and streak formation on leaves, as well as dwarfism and leaf deformation. Although the specific vector of MFSV, the leafhopper *Graminella nigrifrons* (*Cicadellidae*) has been reported only in the United States and the Caribbean. These insects belong to the families *Cicadellidae* and *Delphacidae*, which have the potential to vector rhabdoviruses that are commonly found infecting grasses in Brazil. Although only partial sequence assembled, after further validation, this find could represent the first report of a virus from the genus *Gammanucleorhabdovirus* in South America (Todd et al., 2010; Oliveira et al., 2013; Dietzgen et al., 2020).

Viral sequences found in the root of *C. brasiliensis* may be associated with variation in soil biodiversity between forest and agricultural ecosystems (Pacchioni et al., 2014). Soil particularities between ecosystems (chemical composition, temperature, oxygenation, and organic matter) may influence the diversity and abundance of the microbiome, with a potential to be vectors or virus hosts in the soil, making the root susceptible to infections from different sources (Rodelo-Urrego et al., 2013; Souza et al., 2016; Tripathi et al., 2016). We identified sequences from viruses of the families *Mitoviridae* and *Botourmiaviridae* in seeds, floral buds, and roots. Mitoviruses have genomes containing a single ORF that encodes to an RdRp that presents genetic code specific for mitochondrial genomes, and possibiy the virus uses exclusively the mitochondrial machinery of the host in their replication cycle (Hillman and Cai, 2013; Nibert et al., 2018). Mitoviruses can cause latent infection in hosts and are thought to be transmitted by cell division or through the dispersal of spores (Nibert et al., 2018). In phytopathogenic fungi, the infection can result in fungal hypovirulence, revealing the potential of mitoviruses for use as biocontrol agents (Wu et al., 2010; Xu et al., 2015). Evidence shows that mitoviruses possibly adapted their genome for a cross-kingdom transmission due to the co-evolutive relationship between fungi and other organisms, such as flowering plants (Nibert et al., 2018; Fonseca et al., 2021; Wang et al., 2022). Viruses of the family *Botourmiaviridae* can be grouped into six genera. The first one presents three genomic segments and infects the cytoplasm of plants, whereas some other members infect filamentous fungi (*Ourmiavirus*). The second genus comprises viruses with a single segment and able to infect fungi and plants (*Scleroulivirus*), and the viruses of the four other genera (*Botoulivirus*, *Magoulivirus*, *Penoulivirus*, and *Rhizoulivirus*) infect mainly fungal hosts (Ayllón et al., 2020). Recently, representatives of the families *Mitoviridae* and *Botourmiaviridae* have also been detected in mycorrhizae, which are responsible for forming inter- or intracellular structures in roots of more than 90% of plant species, which assist in nutrient cycling, water uptake, and disease resistance (Bonfante and Genre, 2010; Sutela et al., 2020). The results found in our study indicate that the presence of viral genomes from this family can be ubiquitous in plant samples and may indicate a possible interaction among virus, plant, and fungal cells.

We also observed the presence of picornaviruses in the vegetative organs of *C. brasiliensis*, which had the closest hit with a polyprotein of Biomphalaria virus 2 (coverage of 95% and identity of 54.3%). In the phylogenetic analysis, the studied virus was grouped with Pittsburgh sewage-associated virus 1, a virus that was isolated from an urban sewage sample with an undetermined host (Cantalupo and Pipas, 2018). Biomphalaria virus 2 was originally identified in the microbiota of health snails *Biomphalaria glabrata* and *B. pfeifferi*, vectors of protozoa from the genus *Schistosoma* (Palasio et al., 2021). Previous studies using RdRp indicated the phylogenetic relationship of Biomphalaria virus 2 with viruses of the family *Secoviridae*, the only one from the order *Picornavirales* that has been

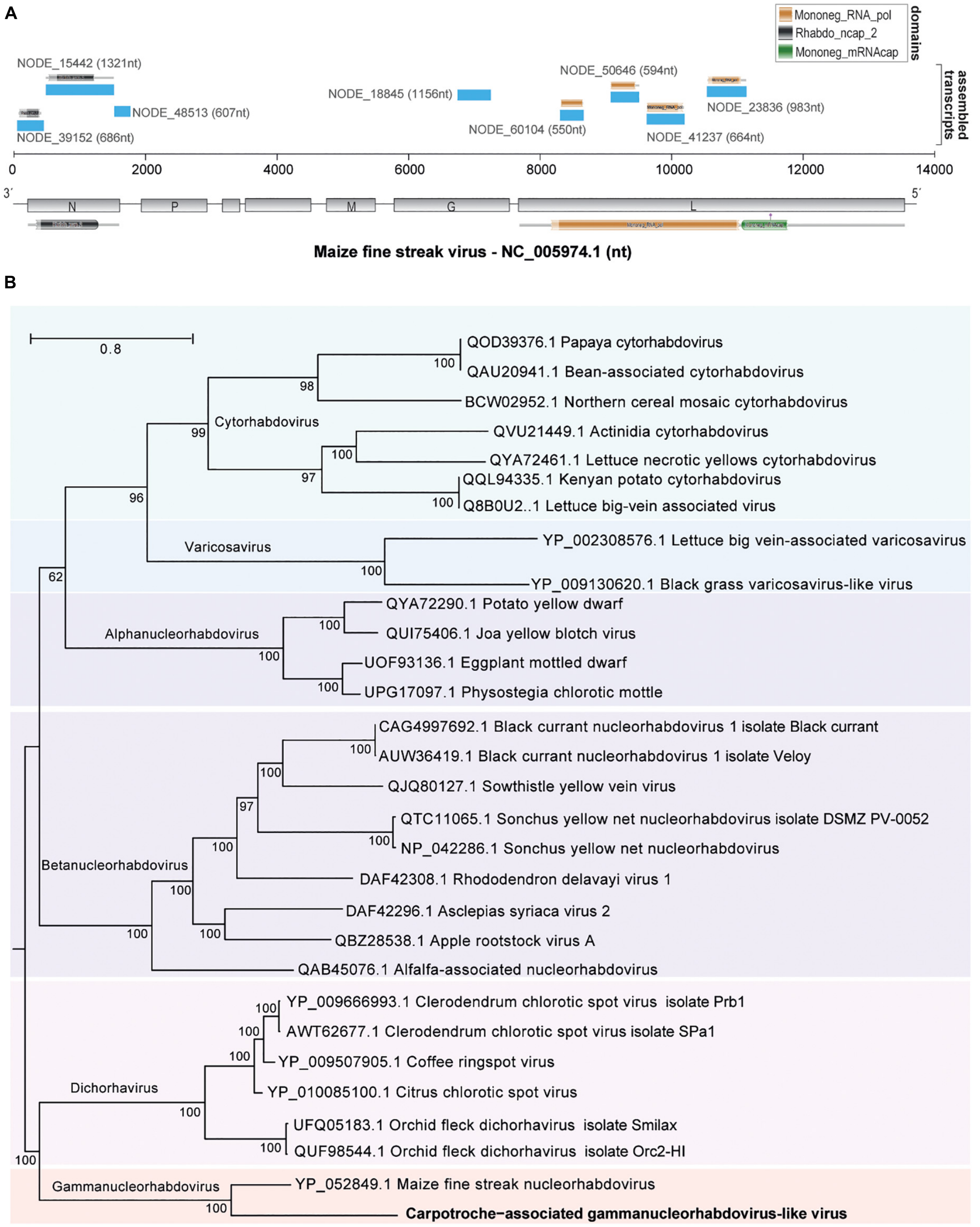


FIGURE 5
Characterization of a putative gammanucleorhabdovirus. **(A)** Distribution of assembled transcripts along the genome of the closest relative virus, Maize fine streak virus (MFSV). Conserved domains are indicated above each assembled transcript or below the MFSV genome. **(B)** Maximum likelihood tree containing sequences from viruses of the subfamily *Betarhabdovirinae* (*Rhabdoviridae*) based on the largest fragment, 983 nt, derived from L (polymerase) gene.

identified to be infecting plants to date. Therefore, the presence of a picorna-like sequence only in the *C. brasiliensis* root sample obtained in the forest ecosystem could be driven by the presence of snails in this region. A second virus showing similarity to picornavirus (Skokie picorna-like virus carpotroche isolate), specifically to Skokie picorna-like virus, was found in our analysis. Another virus close to the studied picornavirus in the phylogenetic tree is Bat posalivirus, originally isolated from bat feces in Cameroon (Yinda et al., 2017). Coupled with the fact that the isolate of Skokie picorna-like virus is

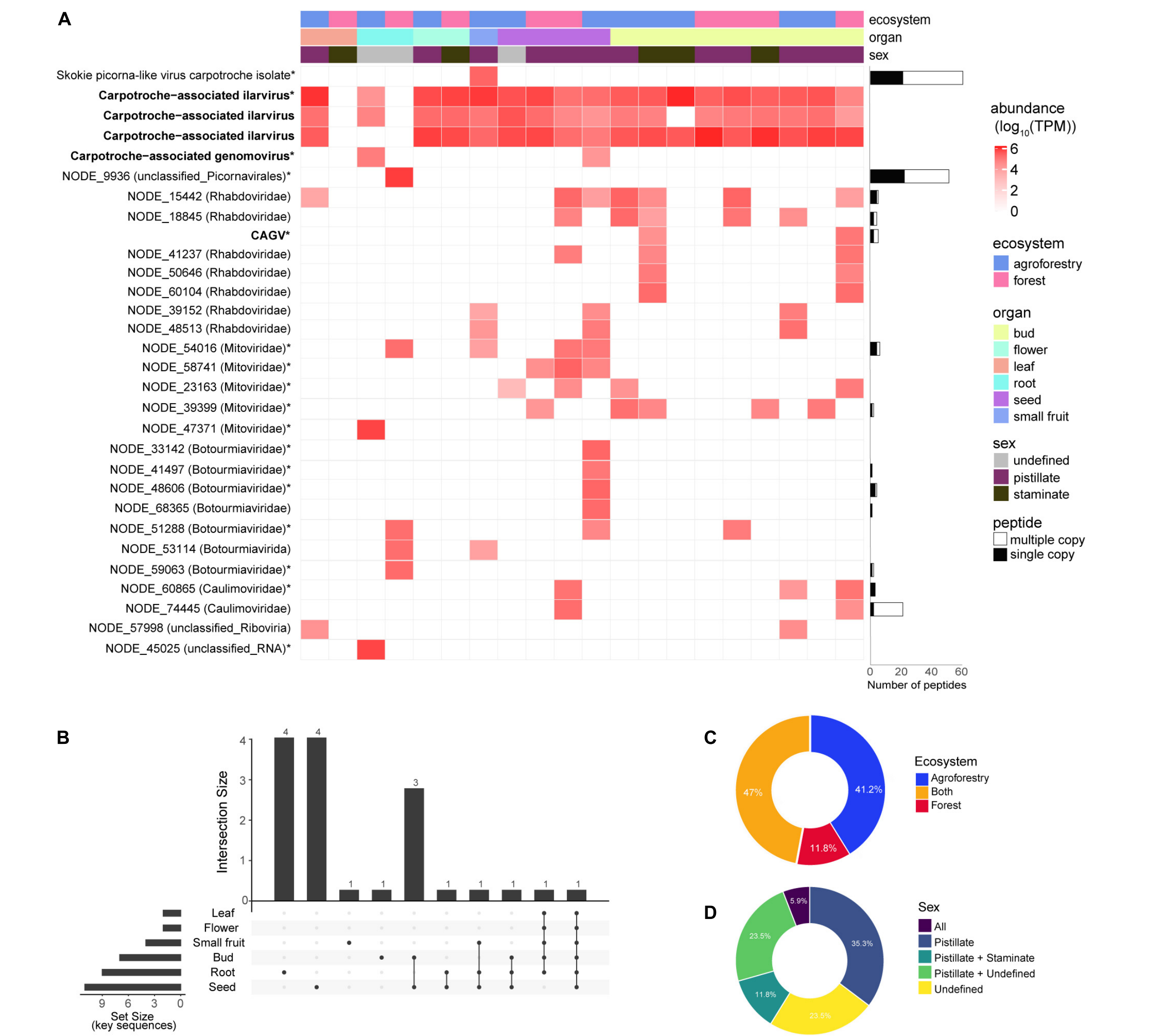


FIGURE 6

Abundance and distribution of Carpotroche-associated viruses. **(A)** Heatmap was constructed based on the relative abundance of each viral transcript identified classified by ecosystem, tissue, and sex. Quantification was computed using transcripts per million (TPM). Horizontal bars indicate the abundance and uniqueness of peptides derived for each of the assembled transcript key sequences and are indicated with *. CAGV, Carpotroche-associated gammanucleorhabdovirus-like virus. The presence of key viral sequences by plant tissue **(B)**, ecosystem **(C)**, or sex **(D)**.

present only in fruit samples, this information indicates the possibility of this virus infecting *C. brasiliensis* to be transmitted by a vector. Considering the pathogenicity of picornaviruses in several species and the economic impact of severe vector-mediated plant diseases, the identification of this viral sequence is of central importance in the investigation of the *C. brasiliensis* virome (Jia et al., 2018).

Among the DNA viruses detected, a putative caulimovirus was detected in seed and floral buds. Viruses of the family *Caulimoviridae* have a circular dsDNA genome, with no envelope and a reverse transcription (RT) step in their life cycle, and are transmitted *via* arthropod vectors with several representatives considered important pathogens for a wide diversity of monocot and dicot plants, including apple, citrus, banana, cocoa, grape, cassava, rice, potato, corn, papaya, soybean, tomato, and others (Bhat et al., 2016; Teycheney et al., 2020). Some family members also possess the ability to integrate part of their genomes in the form of minichromosomes into the host genome during their replication cycle (Diop et al., 2018). Particularly for *T. cacao*, the studies mostly focused on symptomatic infections that affect the economic potential of the production with the genus *Badnavirus*, and correlate to the outbreak of Cacao Swollen Shoot Virus Disease (CSSVD) that started in 1922 in West Africa and is still devastating cacao production in Eastern African regions (Mondego et al., 2016). Thus, 10 species of viruses were described in West Africa or in some Caribbean islands, and most recently, one isolate of CaMMV-BR-like virus was identified in the Bahia state, Brazil, and also one species (*Cacao bacilliform Sri Lanka virus*) from Sri Lanka (Muller et al., 2021; Ramos-Sobrinho et al., 2021). It is likely that the caulimovirus found in *C. brasiliensis* is an EVE, given the sequence detected has a more likely link to the Aristotelia chilensis virus 1, an endogenous virus corresponding to the genus *Petuvirus* (Villacreses et al., 2015). DNA viruses of the family *Genomoviridae* were also detected in *C. brasiliensis.* This group of ssDNA viruses is considered a sister group of the family *Geminiviridae* and has the potential to cause fungal hypovirulence, identified in different environmental samples associated with plants and animals, including in Brazil infecting common beans and citrus (Lamas et al., 2016; Chabi-Jesus et al., 2020; Fontenele et al., 2020; Varsani and Krupovic, 2021). Phylogenetic analysis of the putative viral sequence suggests its association with viruses from the family *Genomoviridae*, indicating a close relationship to chicken genomovirus mg4 and other viruses of the genus *Gemycircularvirus*. While the first is pathogenic to birds, the latter can infect the phytopathogenic fungus *Sclerotinia sclerotiorum* and decrease its virulence (Varsani and Krupovic, 2017).

To corroborate the viral sequences present in our samples, we performed a proteomics assay to explore the presence of viral proteins. Mass spectrometry-based proteomics for viral infection analyses employed in this study is recognized for both structural virology and virus–host interaction studies (Dülfer et al., 2019). The detection of the charge state spectrum at high resolution allows the investigation of size, mass, stability, and shape of viral protein complexes (Uetrecht and Heck, 2011; Dülfer et al., 2019). Thus, the strategy to compare peptide sequences in different MS spectra readouts is ideal to detect the direct presence of viral proteins. In our analysis, we detected peptides derived from 13 viral sequences, eight of them representing key sequences from the variants of known viruses or viruses likely belonging to novel species. In seeds, the same organ for which we performed proteomics assay, we were able to detect viral peptides for 9 out of 17 sequences identified in RNA sequencing analysis.

Data availability statement

The datasets presented in this study can be found in online repositories. Specific repositories and accession codes for each dataset are specified at “Materials and methods” section.

Author contributions

EA and FG: conceptualization, methodology, and supervision. ÍL, PF, RO, LV, and FB: formal analysis. EA, FG, and CP: resources. AV, ÍL, PF, FG, and EA: writing of the original draft preparation. AV, ÍL, FG, and EA: reviewing and editing. FG: funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Finance Code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq # 306160/2017-0 and # 403670/2020-9), Fundação de Apoio á Pesquisa do Estado da Bahia (Fapesb # TSC0017/2014), and Financiadora de Estudos e Projetos (FINEP # 01.18.0087.00). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We would like to thank the ideas and suggestions raised during the scientific discussions at Conservation Genetics Group and Virus Bioinformatics Laboratory. We also acknowledge the editor and reviewers for the careful reading and suggestions in the manuscript, which substantially improved the final version of this manuscript.

Conflict of interest

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

Publisher’s note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.874319/full#supplementary-material

References

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105. doi: 10.1093/bioinformatics/bti263

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Alexander, H. M., Mauck, K. E., Whitfield, A. E., Garrett, K. A., and Malmstrom, C. M. (2014). Plant-virus interactions and the agro-ecological interface. *Eur. J. Plant Pathol.* 138, 529–547. doi: 10.1007/s10658-013-0317-1

Ayllón, M. A., Turina, M., Xie, J., Nerva, L., Marzano, S.-Y. L., Donaire, L., et al. (2020). ICTV virus taxonomy profile: Botourmiaviridae. *J. Gen. Virol.* 101, 454–455. doi: 10.1099/jgv.0.001409

Badillo-Vargas, I. E., Baker, C. A., Turechek, W. W., Frantz, G., Mellinger, H. C., Funderburk, J. E., et al. (2016). Genomic and biological characterization of tomato necrotic streak virus, a novel subgroup 2 *Ilarvirus* infecting tomato in Florida. *Plant Dis.* 100, 1046–1053. doi: 10.1094/PDIS-12-15-1437-RE

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Computat. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bhagwat, S. A., Willis, K. J., Birks, H. J. B., and Whittaker, R. J. (2008). Agroforestry: a refuge for tropical biodiversity? *Trends Ecol. Evol.* 23, 261–267. doi: 10.1016/j.tree.2008.01.005

Bhat, A. I., Hohn, T., and Selvarajan, R. (2016). *Badnaviruses*: the current global scenario. *Viruses* 8:177. doi: 10.3390/v8060177

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bonfante, P., and Genre, A. (2010). Mechanisms underlying beneficial plant–fungus interactions in mycorrhizal symbiosis. *Nat. Commun.* 1:48. doi: 10.1038/ncomms1046

Brito-Rocha, E., dos Anjos, L., Schilling, A. C., Dalmolin, ÂC., and Mielke, M. S. (2017). Individual leaf area estimations of a dioecious tropical tree species *Carpotroche brasiliensis* (Raddi) A. Gray, Achariaceae. *Agroforest. Syst.* 91, 9–15. doi: 10.1007/s10457-016-9927-x

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Cantalupo, P. G., and Pipas, J. M. (2018). Complete genome sequence of Pittsburgh sewage-associated virus 1. *Genome Announc.* 6:e01460–17. doi: 10.1128/genomeA.01460-17

Cerqueira, A. F., Dalmolin, ÂC., dos Anjos, L., da Silva Ledo, C. A., da Costa Silva, D., and Mielke, M. S. (2018). Photosynthetic plasticity of young plants of *Carpotroche brasiliensis* (Raddi) A. Gray, Achariaceae. *Trees* 32, 191–202. doi: 10.1007/s00468-017-1623-6

Chabi-Jesus, C., Najar, A., Fontenele, R. S., Kumari, S. G., Ramos-González, P. L., Freitas-Astúa, J., et al. (2020). Viruses representing two new genomovirus species identified in citrus from Tunisia. *Arch. Virol.* 165, 1225–1229. doi: 10.1007/s00705-020-04569-8

Dietzgen, R. G., Bejerman, N. E., Goodin, M. M., Higgins, C. M., Huot, O. B., Kondo, H., et al. (2020). Diversity and epidemiology of plant rhabdoviruses. *Virus Res.* 281:197942. doi: 10.1016/j.virusres.2020.197942

Diop, S. I., Geering, A. D. W., Alfama-Depauw, F., Loaec, M., Teycheney, P.-Y., and Maumus, F. (2018). Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. *Sci. Rep.* 8:572. doi: 10.1038/s41598-017-16399-x

dos Santos, F. S. D., de Souza, L. P. A., and Siani, A. C. (2008). [Chaulmoogra oil as scientific knowledge: the construction of a treatment for leprosy]. *Hist Cienc Saude Manguinhos* 15, 29–47. doi: 10.1590/s0104-59702008000100003

Dülfer, J., Kadek, A., Kopicki, J.-D., Krichel, B., and Uetrecht, C. (2019). "Chapter seven – Structural mass spectrometry goes viral," in *Advances in Virus Research Complementary Strategies to Understand Virus Structure and Function*, ed. F. A. Rey (Cambridge, MA: Academic Press), 189–238. doi: 10.1016/bs.aivir.2019.07.003

Dzahini-Obiatey, H., Domfeh, O., and Amoah, F. M. (2010). Over seventy years of a viral disease of cocoa in Ghana: from researchers perspective. *AJAR* 5, 476–485. doi: 10.5897/AJAR09.625

Elena, S. F., Agudelo-Romero, P., and Lalić, J. (2009). The evolution of viruses in multi-host fitness landscapes. *Open Virol. J.* 3, 1–6. doi: 10.2174/1874357900903010001

Engering, A., Hogerwerf, L., and Slingenbergh, J. (2013). Pathogen–host–environment interplay and disease emergence. *Emerg. Microbes Infect.* 2, 1–7. doi: 10.1038/emi.2013.5

Faria, D., Paciencia, M. L. B., Dixo, M., Laps, R. R., and Baumgarten, J. (2007). Ferns, frogs, lizards, birds and bats in forest fragments and shade cacao plantations in two contrasting landscapes in the Atlantic forest, Brazil. *Biodivers. Conserv.* 16, 2335–2357. doi: 10.1007/s10531-007-9189-z

Fonseca, P., Ferreira, F., da Silva, F., Oliveira, L. S., Marques, J. T., Goes-Neto, A., et al. (2021). Characterization of a novel *Mitovirus* of the sand fly *Lutzomyia longipalpis* using genomic and virus–host interaction signatures. *Viruses* 13:9. doi: 10.3390/v13010009

Fontenele, R. S., Roumagnac, P., Richet, C., Kraberger, S., Stainton, D., Aleamotu'a, M., et al. (2020). Diverse genomoviruses representing twenty-nine species identified associated with plants. *Arch. Virol.* 165, 2891–2901. doi: 10.1007/s00705-020-04801-5

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gama-Rodrigues, A. C., Müller, M. W., Gama-Rodrigues, E. F., and Mendes, F. A. T. (2021). Cacao-based agroforestry systems in the Atlantic forest and Amazon biomes: an ecoregional analysis of land use. *Agric. Syst.* 194:103270. doi: 10.1016/j.agsy.2021.103270

Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., et al. (2010). The NCBI biosystems database. *Nucleic Acids Res.* 38, D492–D496. doi: 10.1093/nar/gkp858

Girgis, S. M., Bem, F. P., Dovas, C. I., Sclavounos, A., Avgelis, A. D., Tsagris, M., et al. (2009). Characterisation of a novel *Ilarvirus* causing grapevine angular mosaic disease. *Eur. J. Plant Pathol.* 125, 203–211. doi: 10.1007/s10658-009-9472-9

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849. doi: 10.1093/bioinformatics/btw313

Hillman, B. I., and Cai, G. (2013). "Chapter six – The family Narnaviridae: simplest of RNA viruses," in *Advances in Virus Research Mycoviruses*, ed. S. A. Ghabrial (Cambridge, MA: Academic Press), 149–176. doi: 10.1016/B978-0-12-394315-6.00006-4

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Jia, D., Chen, Q., Mao, Q., Zhang, X., Wu, W., Chen, H., et al. (2018). Vector mediated transmission of persistently transmitted plant viruses. *Curr. Opin. Virol.* 28, 127–132. doi: 10.1016/j.coviro.2017.12.004

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9. doi: 10.1093/nar/gkn201

Joly, C. A., Metzger, J. P., and Tabarelli, M. (2014). Experiences from the Brazilian Atlantic forest: ecological findings and conservation initiatives. *N. Phytol.* 204, 459–473. doi: 10.1111/nph.12989

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Keesing, F., and Ostfeld, R. S. (2021). Dilution effects in disease ecology. *Ecol. Lett.* 24, 2490–2505. doi: 10.1111/ele.13875

Keesing, F., Holt, R. D., and Ostfeld, R. S. (2006). Effects of species diversity on disease risk. *Ecol. Lett.* 9, 485–498. doi: 10.1111/j.1461-0248.2006.00885.x

Kitajima, E. W. (2020). An annotated list of plant viruses and viroids described in Brazil (1926-2018). *Biota Neotrop.* 20, 1–101. doi: 10.1590/1676-0611-BN-2019-0932

Krist, S. (2020). "Chaulmoogra oil," in *Vegetable Fats and Oils*, ed. S. Krist (Cham: Springer International Publishing), 223–226. doi: 10.1007/978-3-030-30314-3_34

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lamas, N. S., Fontenele, R. S., Melo, F. L., Costa, A. F., Varsani, A., and Ribeiro, S. G. (2016). Complete genome sequence of a genomovirus associated with common bean plant leaves in Brazil. *Genome Announc.* 4:e01247–16. doi: 10.1128/genomeA.01247-16

Lefeuvre, P., Martin, D. P., Elena, S. F., Shepherd, D. N., Roumagnac, P., and Varsani, A. (2019). Evolution and ecology of plant viruses. *Nat. Rev. Microbiol.* 17, 632–644. doi: 10.1038/s41579-019-0232-3

Lima, J. A., Oliveira, A. S., de Miranda, A. L. P., Rezende, C. M., and Pinto, A. C. (2005). Anti-inflammatory and antinociceptive activities of an acid fraction of the seeds of *Carpotroche brasiliensis* (Raddi) (Flacourtiaceae). *Braz. J. Med. Biol. Res.* 38, 1095–1103. doi: 10.1590/S0100-879X2005000700013

Lima, T. M., Amaral, E. S., Gaiotto, F. A., dos Anjos, L., Dalmolin, ÂC., Santos, A. S., et al. (2020). Fruit and seed biometry of *Carpotroche brasiliensis* (RB) A. Gray (Achariaceae), a tropical tree with great potential to provide natural forest products. *Austral. J. Crop Sci.* 14, 1826–1833. doi: 10.21475/ajcs.20.14.11.p2596

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C. M., Roumagnac, P., van der Vlugt, R., et al. (2020). Illuminating an ecological blackbox: using high throughput sequencing to characterize the plant virome across scales. *Front. Microbiol.* 11:578064. doi: 10.3389/fmicb.2020.578064

Marangon, G. P., Cruz, A. F., Barbosa, W. B., Loureiro, G. H., and de Holanda, A. C. (2010). Dispersão de sementes de uma comunidade arbórea em um remanescente de mata atlântica, município de Bonito, PE. *Revista Verde Agroecologia Desenvolvimento Sustentável* 5:19.

Martini, A. M. Z., Fiaschi, P., Amorim, A. M., and da Paixão, J. L. (2007). A hot-point within a hot-spot: a high diversity site in Brazil's Atlantic forest. *Biodivers. Conserv.* 16, 3111–3128. doi: 10.1007/s10531-007-9166-6

Maurino, F., Dumón, A. D., Llauger, G., Alemandri, V., de Haro, L. A., Mattio, M. F., et al. (2018). Complete genome sequence of maize yellow striate virus, a new *Cytorhabdovirus* infecting maize and wheat crops in Argentina. *Arch. Virol.* 163, 291–295. doi: 10.1007/s00705-017-3579-7

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257. doi: 10.1038/ncomms11257

Mondego, J., Thomazella, D., Teixeira, P. J., and Pereira, G. (2016). "Genomics, transcriptomics, and beyond: the fifteen years of Cacao's Witches' broom disease genome project," in *Cacao Diseases*, eds B. Bailey and L. Meinhardt (Cham: Springer). doi: 10.1007/978-3-319-24789-2_6

Moury, B., and Desbiez, C. (2020). Host range evolution of potyviruses: a global phylogenetic analysis. *Viruses* 12:111. doi: 10.3390/v12010111

Muller, E. (2016). "Cacao swollen shoot virus (CSSV): history, biology, and genome," in *Cacao Diseases: A History of Old Enemies and New Encounters*, eds B. A. Bailey and L. W. Meinhardt (Cham: Springer International Publishing), 337–358. doi: 10.1007/978-3-319-24789-2_10

Muller, E., Ullah, I., Dunwell, J. M., Daymond, A. J., Richardson, M., Allainguillaume, J., et al. (2021). Identification and distribution of novel badnaviral sequences integrated in the genome of cacao (*Theobroma cacao*). *Sci. Rep.* 11:8270. doi: 10.1038/s41598-021-87690-1

Nair, P. K. R. (1993). *An Introduction to Agroforestry*. Berlin: Springer Science & Business Media.

Nair, P. K. R. (2012). "Climate change mitigation: a low-hanging fruit of agroforestry," in *Agroforestry – The Future of Global Land Use Advances in Agroforestry*, eds P. K. R. Nair and D. Garrity (Dordrecht: Springer), 31–67. doi: 10.1007/978-94-007-4676-3_7

Navarro, J. A., Sanchez-Navarro, J. A., and Pallas, V. (2019). Key checkpoints in the movement of plant viruses through the host. *Adv. Virus. Res.* 104, 1–64. doi: 10.1016/bs.aivir.2019.05.001

Nibert, M. L., Vong, M., Fugate, K. K., and Debat, H. J. (2018). Evidence for contemporary plant mitoviruses. *Virology* 518, 14–24. doi: 10.1016/j.virol.2018.02.005

Oliveira, C. M. D., Oliveira, E. D., Souza, I. R. P. D., Alves, E., Dolezal, W., Paradell, S., et al. (2013). Abundance and species richness of leafhoppers and planthoppers (Hemiptera: Cicadellidae and Delphacidae) in Brazilian maize crops. *Florida Entomol.* 96, 1470–1481. doi: 10.1653/024.096.0427

Pacchioni, R. G., Carvalho, F. M., Thompson, C. E., Faustino, A. L. F., Nicolini, F., Pereira, T. S., et al. (2014). Taxonomic and functional profiles of soil samples from Atlantic forest and Caatinga biomes in northeastern Brazil. *Microbiol. Open* 3, 299–315. doi: 10.1002/mbo3.169

Pagán, I. (2018). The diversity, evolution and epidemiology of plant viruses: a phylogenetic view. *Infect. Genet. Evol.* 65, 187–199. doi: 10.1016/j.meegid.2018.07.033

Palasio, R. G. S., de Azevedo, T. S., Tuan, R., and Chiaravalloti-Neto, F. (2021). Modelling the present and future distribution of *Biomphalaria* species along the watershed of the Middle Paranapanema region, São Paulo, Brazil. *Acta Trop.* 214:105764. doi: 10.1016/j.actatropica.2020.105764

Pallas, V., Aparicio, F., Herranz, M. C., Sanchez-Navarro, J. A., and Scott, S. W. (2013). "Chapter five – The molecular biology of *Ilarviruses*," in *Advances in Virus Research*, eds K. Maramorosch and F. A. Murphy (Cambridge, MA: Academic Press), 139–181. doi: 10.1016/B978-0-12-407698-3.00005-3

Parascandola, J. (2003). Chaulmoogra oil and the treatment of leprosy. *Pharm. Hist.* 45, 47–57.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Piasentin, F. B., Saito, C. H., and Sambuichi, R. H. R. (2014). Preferências locais quanto às árvores do sistema cacau-cabruca no sudeste da Bahia1. *Ambiente Sociedade* 17, 55–78.

Pirovani, C. P., Carvalho, H. A. S., Machado, R. C. R., Gomes, D. S., Alvim, F. C., Pomella, A. W. V., et al. (2008). Protein extraction for proteome analysis from cacao leaves and meristems, organs infected by *Moniliophthora perniciosa*, the causal agent of the witches' broom disease. *Electrophoresis* 29, 2391–2401. doi: 10.1002/elps.200700743

Posnette, A. F., Robertson, N. F., and McA Todd, J. (1950). Virus diseases of cacao in West Africa. *Ann. Appl. Biol.* 37, 229–240. doi: 10.1111/j.1744-7348.1950.tb01041.x

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448

Ramos-Sobrinho, R., Ferro, M. M. M., Nagata, T., Puig, A. S., Keith, C. V., Britto, D. S., et al. (2021). Complete genome sequences of three newly discovered cacao mild mosaic virus isolates from *Theobroma cacao* L. in Brazil and Puerto Rico and evidence for recombination. *Arch. Virol.* 166, 2027–2031. doi: 10.1007/s00705-021-05063-5

Rodelo-Urrego, M., Pagán, I., González-Jara, P., Betancourt, M., Moreno-Letelier, A., Ayllón, M. A., et al. (2013). Landscape heterogeneity shapes host-parasite interactions and results in apparent plant–virus codivergence. *Mol. Ecol.* 22, 2325–2340. doi: 10.1111/mec.12232

Rodríguez-Nevado, C., Gavilán, R., and Pagán, I. (2020). Host abundance and identity determine the epidemiology and evolution of a generalist plant virus in a wild ecosystem. *Phytopathology* 110, 94–105. doi: 10.1094/PHYTO-07-19-0271-FI

Roossinck, M. J., and García-Arenal, F. (2015). Ecosystem simplification, biodiversity loss and plant virus emergence. *Curr. Opin. Virol.* 10, 56–62. doi: 10.1016/j.coviro.2015.01.005

Rosa, M. R., Brancalion, P. H. S., Crouzeilles, R., Tambosi, L. R., Piffer, P. R., Lenti, F. E. B., et al. (2021). Hidden destruction of older forests threatens Brazil's Atlantic forest and challenges restoration programs. *Sci. Adv.* 7:eabc4547. doi: 10.1126/sciadv.abc4547

Rumbou, A., Vainio, E. J., and Büttner, C. (2021). Towards the forest virome: high-throughput sequencing drastically expands our understanding on virosphere in temperate forest ecosystems. *Microorganisms* 9:1730. doi: 10.3390/microorganisms9081730

Sambuichi, R. H. R., Mielke, M. S., and Pereira, C. E. (2009). Nossas árvores: conservação, uso e manejo de árvores nativas no sul da Bahia. Ilhéus: Editus.

Sambuichi, R. H. R., Vidal, D. B., Piasentin, F. B., Jardim, J. G., Viana, T. G., Menezes, A. A., et al. (2012). Cabruca agroforests in southern Bahia, Brazil: tree component, management practices and tree species conservation. *Biodivers. Conserv.* 21, 1055–1077. doi: 10.1007/s10531-012-0240-3

Sambuichi, R., and Haridasan, M. (2007). Recovery of species richness and conservation of native Atlantic forest trees in the cacao plantations of southern Bahia in Brazil. *Biodivers. Conserv.* 16, 3681–3701. doi: 10.1007/s10531-006-9017-x

Sharma, D. K., and Hall, I. H. (1991). Hypolipidemic, anti-inflammatory, and antineoplastic activity and cytotoxicity of flavonolignans isolated from *Hydnocarpus wightiana* seeds. *J. Nat. Prod.* 54, 1298–1302. doi: 10.1021/np50077a010

Sharma-Poudyal, D., Osterbauer, N. K., Putnam, M. L., and Scott, S. W. (2016). First report of lilac ring mottle virus infecting lilac in the United States. *Plant Health Prog.* 17, 158–159. doi: 10.1094/PHP-BR-15-0055

Silvestre, R., Fuentes, S., Risco, R., Berrocal, A., Adams, I., Fox, A., et al. (2020). Characterization of distinct strains of an aphid-transmitted *Ilarvirus* (Fam. Bromoviridae) infecting different hosts from South America. *Virus Res.* 282:197944. doi: 10.1016/j.virusres.2020.197944

Souza, R. C., Mendes, I. C., Reis-Junior, F. B., Carvalho, F. M., Nogueira, M. A., Vasconcelos, A. T. R., et al. (2016). Shifts in taxonomic and functional microbial diversity with agriculture: how fragile is the Brazilian Cerrado? *BMC Microbiol.* 16:42. doi: 10.1186/s12866-016-0657-z

Susi, H., and Laine, A.-L. (2021). Agricultural land use disrupts biodiversity mediation of virus infections in wild plant populations. *New Phytol.* 230, 2447–2458. doi: 10.1111/nph.17156

Sutela, S., Forgia, M., Vainio, E. J., Chiapello, M., Daghino, S., Vallino, M., et al. (2020). The virome from a collection of endomycorrhizal fungi reveals new viral taxa with unprecedented genome organization. *Virus Evol.* 6:veaa076. doi: 10.1093/ve/veaa076

Takahashi, H., Fukuhara, T., Kitazawa, H., and Kormelink, R. (2019). Virus latency and the impact on plants. *Front. Microbiol.* 10:2764. doi: 10.3389/fmicb.2019.02764

Teixeira, H. M., Bianchi, F. J. J. A., Cardoso, I. M., Tittonell, P., and Peña-Claros, M. (2021). Impact of agroecological management on plant diversity and soil-based ecosystem services in pasture and coffee systems in the Atlantic forest of Brazil. *Agric. Ecosyst. Environ.* 305:107171. doi: 10.1016/j.agee.2020.107171

Teycheney, P.-Y., Geering, A. D. W., Dasgupta, I., Hull, R., Kreuze, J. F., Lockhart, B., et al. (2020). ICTV virus taxonomy profile: Caulimoviridae. *J. Gen. Virol.* 101, 1025–1026. doi: 10.1099/jgv.0.001497

Todd, J. C., Ammar, E.-D., Redinbaugh, M. G., Hoy, C., and Hogenhout, S. A. (2010). Plant host range and leafhopper transmission of maize fine streak virus. *Phytopathology* 100, 1138–1145. doi: 10.1094/PHYTO-05-10-0144

Topolovec-Pintaric, S. (2020). *Plant Diseases: Current Threats and Management Trends*. Norderstedt: BoD – Books on Demand.

Tripathi, B. M., Song, W., Slik, J. W. F., Sukri, R. S., Jaafar, S., Dong, K., et al. (2016). Distinctive tropical forest variants have unique soil microbial communities, but not always low microbial diversity. *Front. Microbiol.* 7:376. doi: 10.3389/fmicb.2016.00376

Uetrecht, C., and Heck, A. J. R. (2011). Modern biomolecular mass spectrometry and its role in studying virus structure, dynamics, and assembly. *Angewandte Chemie International Edition* 50, 8248–8262. doi: 10.1002/anie.201008120

Varsani, A., and Krupovic, M. (2017). Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. *Virus Evol.* 3:vew037. doi: 10.1093/ve/vew037

Varsani, A., and Krupovic, M. (2021). Family Genomoviridae: 2021 taxonomy update. *Arch. Virol.* 166, 2911–2926. doi: 10.1007/s00705-021-05183-y

Villacreses, J., Rojas-Herrera, M., Sánchez, C., Hewstone, N., Undurraga, S. F., Alzate, J. F., et al. (2015). Deep sequencing reveals the complete genome and evidence for transcriptional activity of the first virus-like sequences identified in *Aristotelia chilensis* (Maqui Berry). *Viruses* 7, 1685–1699. doi: 10.3390/v7041685

Wang, Q., Zou, Q., Dai, Z., Hong, N., Wang, G., and Wang, L. (2022). Four novel mycoviruses from the hypovirulent *Botrytis cinerea* SZ-2-3y isolate from *Paris polyphylla*: molecular characterisation and mitoviral sequence transboundary entry into plants. *Viruses* 14:151. doi: 10.3390/v14010151

Willie, K., and Stewart, L. R. (2017). Complete genome sequence of a new maize-associated *Cytorhabdovirus*. *Genome Announc.* 5:e00591–17. doi: 10.1128/genomeA.00591-17

Wu, M., Zhang, L., Li, G., Jiang, D., and Ghabrial, S. A. (2010). Genome characterization of a debilitation-associated mitovirus infecting the phytopathogenic fungus *Botrytis cinerea*. *Virology* 406, 117–126. doi: 10.1016/j.virol.2010.07.010

Xu, Z., Wu, S., Liu, L., Cheng, J., Fu, Y., Jiang, D., et al. (2015). A mitovirus related to plant mitochondrial gene confers hypovirulence on the phytopathogenic fungus *Sclerotinia sclerotiorum*. *Virus Res.* 197, 127–136. doi: 10.1016/j.virusres.2014.12.023

Yinda, C. K., Zell, R., Deboutte, W., Zeller, M., Conceição-Neto, N., Heylen, E., et al. (2017). Highly diverse population of Picornaviridae and other members of the Picornavirales, in Cameroonian fruit bats. *BMC Genomics* 18:249. doi: 10.1186/s12864-017-3632-7

Zhou, C., da Graça, J. V., Freitas-Astúa, J., Vidalakis, G., Duran-Vila, N., and Lavagi, I. (2020). "Chapter 19 – Citrus viruses and viroids," in *The Genus Citrus*, eds M. Talon, M. Caruso, and F. G. Gmitter (Sawston: Woodhead Publishing), 391–410. doi: 10.1016/B978-0-12-812163-4.00019-X

Zucaratto, R., Carrara, R., and Franco, B. K. S. (2010). Dieta da paca (*Cuniculus paca*) usando métodos indiretos numa área de cultura agrícola na floresta Atlântica Brasileira. *Biotemas* 23, 235–239. doi: 10.5007/2175-7925.2010v23n1p235

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership